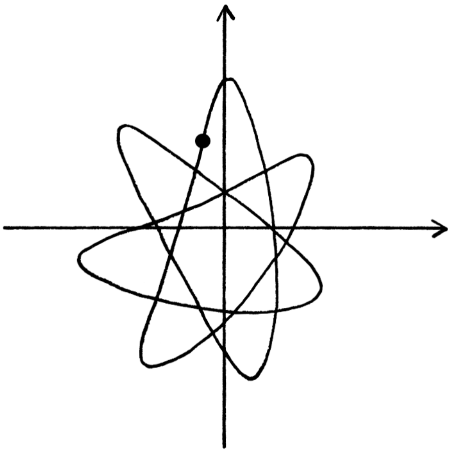
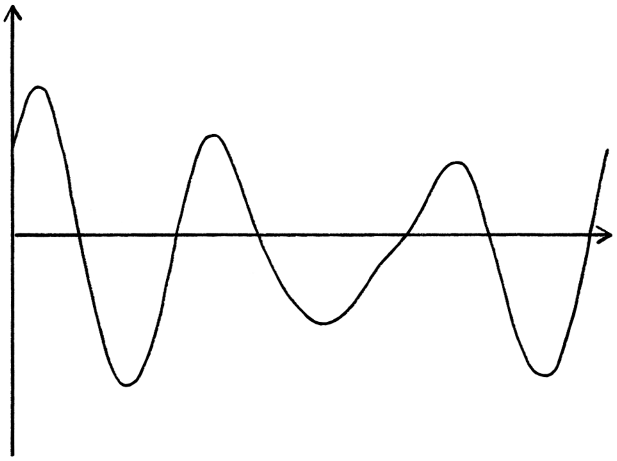
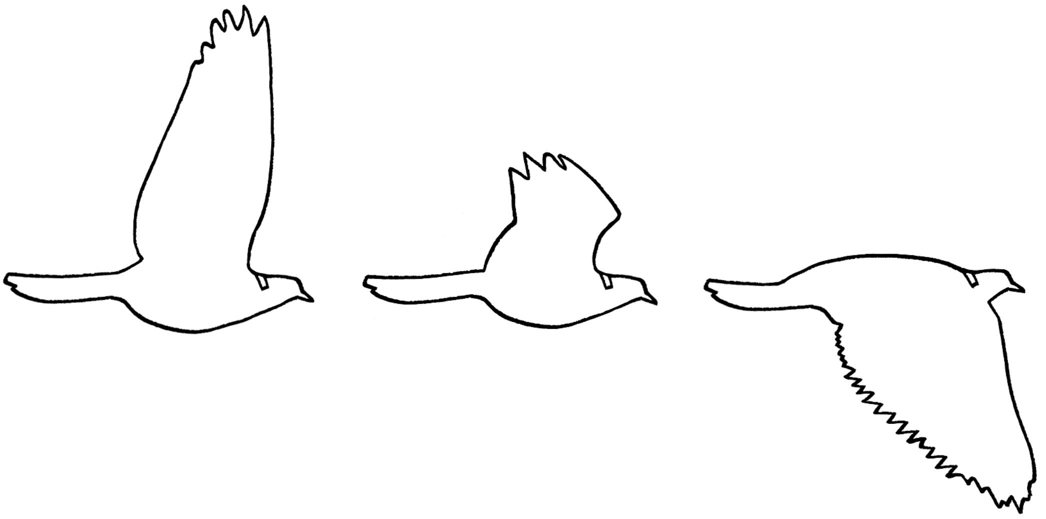




A Book About Waves
Part One

By Tim Warriner



7th December 2022

A Book About Waves
Part 1

by Tim Warriner

www.timwarriner.com
September 2021
[Last edited on 2022-12-07]

To see if this is the latest version of this book, visit www.timwarriner.com

Contents

Introduction

This book is a simple introduction to waves and signals for people who are new to the subject, find the subject confusing, or are not particularly skilled at maths. This book starts at the very beginning with explanations of Sine and Cosine.

I started writing this book because I could not find any good explanations of the basics of waves. Most digital signal processing books are aimed at university students and graduates, and start with the assumption that the reader is an expert in maths and already knows everything about the subject. Maths books intended for schoolchildren generally give limited and simplified information with the idea that children just need to remember it blindly for the purposes of passing exams. It is very difficult to find a book in the middle that is both easy to understand and actually explains things. Coupled with that, the teaching of mathematical subjects is often terrible. Maths is the only subject where it is possible to read about something you already understand, and find the explanation incomprehensible.

You do not need to know much maths to learn from this book, but it is probably useful to have a technical mind or an interest in mathematical puzzles. If you enjoy puzzles, you will find it easier to remember how things work.

This is not intended to be a school textbook. I explain many things that are not relevant to an academic curriculum, and I omit many things that would be relevant. To remove ambiguity, I sometimes use different terms to those commonly used in academia. However, after having read both parts of this book, you will find it much easier to understand certain aspects of academic maths and physics.

This book should be read sequentially from the start to the finish. If you think you already know the subject being discussed in one of the chapters, you should read it anyway because there will be points I make that will be important later on. Every chapter follows on from the previous ones, so if you skip chapters, it will be harder to understand later chapters.

You do not need anything to learn from this book, but you will progress more quickly if you can experiment with waves on a graphing calculator. Generally, graphing calculator computer programs and phone apps are better and cheaper than real-world calculators. At the time of writing, the best calculator I have seen by a very long way is “Graphing Calculator” app for Android by MathLab Apps LLC.

To see if this is the latest version of this book, visit www.timwarriner.com

Notes on the content of this book

For much of the first part of this book, I use degrees instead of radians. This is because everything is much more intuitive with degrees, and I want this book to be easy to understand.

I use the asterisk "*" as the symbol for multiplication. I use "÷" or "/" as the symbol for division, depending on which is easiest to read in the context in which they are being used. The trouble with using "/" to mean division is that it requires context to prevent it being confused with "/" to mean "or". The phrase "amplitude / frequency" could mean "amplitude divided by frequency" in a mathematical calculation, but could also mean "amplitude or frequency" in an English sentence. The symbol "÷" is unambiguous in an English-language text.

When I have numbers that are not integers, I generally round them up to the first four decimal places or the first four decimal places after the first non-zero decimal place. In other words, the square root of 2 is 1.4142. The number 1 divided by 3000 is 0.0003333.

I sometimes use the nouns "Sine" and "Cosine" as verbs. I tend to put formulas in inverted commas to make them easier to read. I frequently break the standard rules of punctuation to make explanations and calculations clearer. My maths formulas often use programming-style variable names instead of mathematical symbols as that can be a better way of explaining things. For example, I might say "amplitude^2" instead of "x^2". I often include unnecessary ones and zeroes in formulas and Complex numbers to make their meaning clearer. I capitalise some terms to make them more distinct or unambiguous.

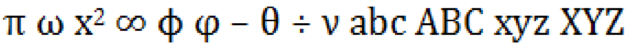
Being from the UK, I say "anticlockwise" instead of "counterclockwise". I say "right-angled triangle" instead of "right triangle". I say "maths" instead of "math". I say "gradient" instead of "slope".

When I describe symbols as being "Greek" letters, I mean that they are from ancient and modern Greek. Those languages have the same alphabet. When I describe a symbol as being a "Latin" letter, I mean that it appears in modern alphabets that originally derived from the Latin alphabet, such as English, French, German, and so on.

To tell if this PDF is being displayed correctly, this line of text:

π ω x^2 ∞ ϕ φ – θ ÷ ν abc ABC xyz XYZ

... should appear roughly the same as in this picture:



π ω x^2 ∞ ϕ φ – θ ÷ ν abc ABC xyz XYZ

Thoughts on learning

- Different people have different ways of thinking. If you do not understand something, it does not mean that you are incapable of understanding it. It just means that you have not found a way of understanding it *yet*.

- The more you see something that you do not understand from different viewpoints, the more you will begin to understand it.

- The more you search for explanations for things you do not understand, the higher your chance of finding something that explains it for your way of thinking.

- Thinking about the subjects in this book when you have nothing better to do is a good way of figuring out how everything works. Do not just read everything – think about it too.

- If you immerse yourself in a subject, you will learn it much more quickly. If you struggle to understand it, immersing yourself will help you move closer to understanding it. Therefore, read everything you can find on a subject even if it makes no sense.

- With most books on waves, signals, digital signal processing, and so on, if you do not understand the maths, you will still learn a lot by skipping the maths parts.

- If you want to learn well, you have to become involved – draw some triangles, draw a sine wave, try calculating π, play with some Complex numbers and so on.

- The most important factor in learning is having the desire to learn. If you *really* want to learn about something, you will enjoy seeking out the information and trying to understand it.

- There is a difference between being told a rule, and being able to visualise why that rule is correct. If you understand why a rule works, you will not need to make an effort to remember it.

- Wikipedia is excellent for many topics, but it is often overly complicated when explaining mathematical subjects. It is easy to become discouraged when trying to understand mathematical ideas on Wikipedia. For complicated subjects, the best explanations are often those available free on the websites of university lecturers.

Acknowledgements

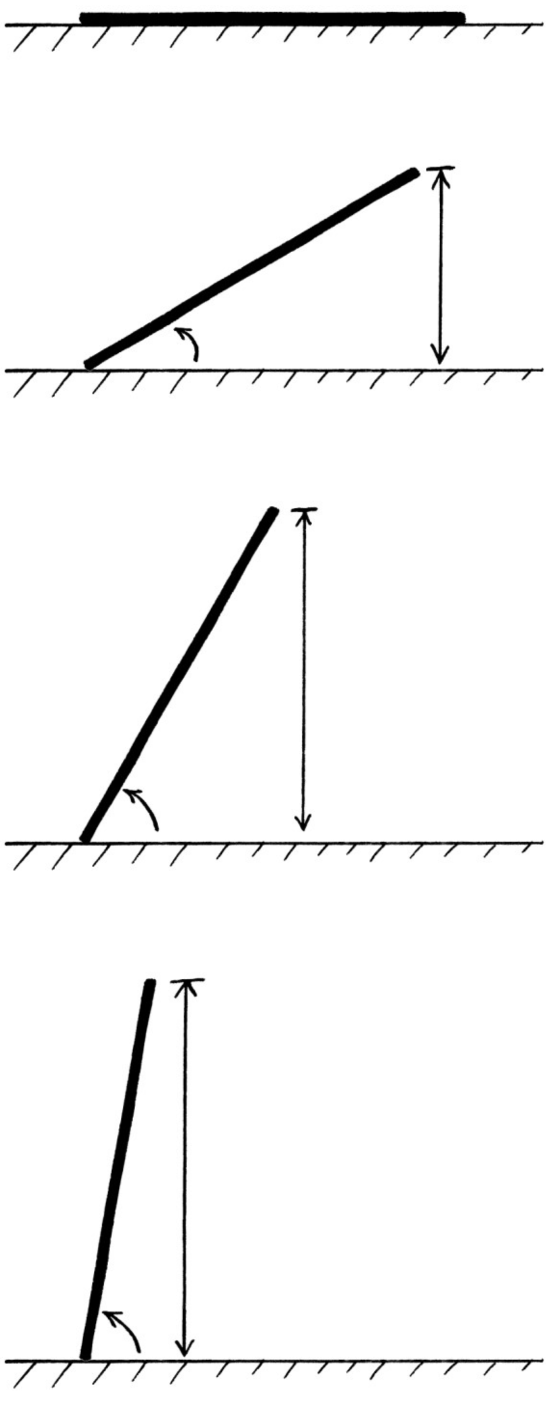
The following people and organisations have unknowingly assisted me with writing this book, and without them, this book would never have existed:

- M
- F+
- L
- ME
- BRM
- Michael Ossmann (HackRf)
- Mike Walters (Inspectrum)
- Alexandru Csete (GQRX)
- Airspy.com (Airspy Hf+)
- GNU Radio
- Paul Dawkins: tutorial.math.lamar.edu
- The anonymous author of “The Radio Spectrum – UK Allocations” list from 2012: ukspec.tripod.com/spectrum.html
- Intel

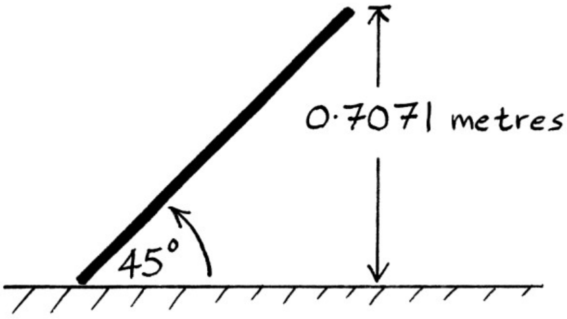
Chapter 1: Triangles and circles

Sticks

If you lay a metre-long stick on the ground, and gradually raise one end while keeping the other end still, you will notice a consistent relationship between the angle of the stick and the height of the end.



The height of the end of the stick when the stick is held at a particular angle is *always* the same. For example, if the stick is at 45 degrees, then the height of the end of the stick will *always* be 0.7071 metres above the ground. It will never be, and can never be, anything else.



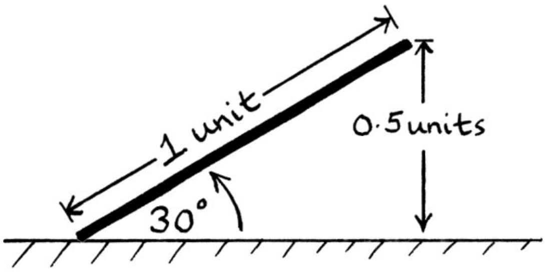
If the stick is at 10 degrees, the end of the stick will *always* be 0.1736 metres above the ground. If the stick is at 70 degrees, the end of the stick will *always* be 0.9397 metres above the ground. If the stick is at 0 degrees (in other words, it is lying flat on the ground), then the end of the stick will be 0 metres above the ground. If the stick is at 90 degrees (in other words, it is pointing directly upwards), the end will be 1 metre above the ground (because the stick is 1 metre long).

There is a direct relationship between the angle of the stick and the height of the end. This relationship is not a simple one – for example it is not the case that if we double the angle, the height doubles, and similarly, it is not the case that if we add 10 degrees to the angle, the height will increase by 10 centimetres. Instead, the height increases rapidly as the end of the stick moves from the ground, and then the rate of increase slows down as the stick approaches 90 degrees. The relationship cannot be portrayed using a concise, simple mathematical formula, and that is one of the reasons it is considered an interesting one.

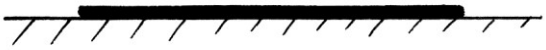
The name given to the relationship between the height of the end of the stick and its angle is called the "Sine function". The "Sine of the angle" gives the height of the end of the stick. If we wanted to find out the height of the end of a 1-metre long stick using a calculator, we would enter the stick's angle and then press the Sine button.

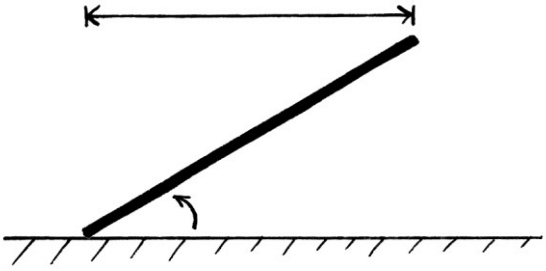
The particular units and the size of the stick are irrelevant to the Sine function. We could use metres, feet, miles, chains, leagues or anything. If the stick is one unit long of any unit, then the Sine function gives its height in terms of that unit. For example, if the stick is one kilometre long and at 30 degrees, the height of its end will be 0.5 kilometres above the ground. If the stick is one millimetre long and at 30 degrees, its height will be 0.5 millimetres. If the stick is 13.405 feet long and at

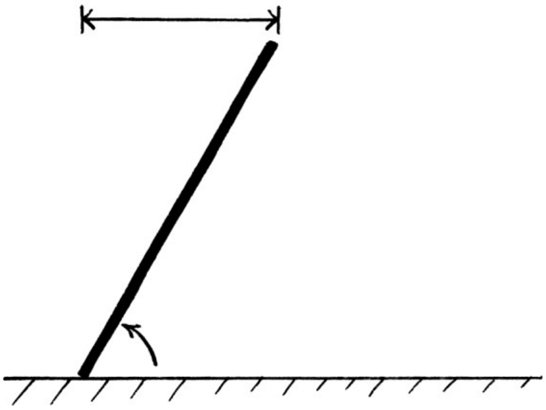
30 degrees, we can treat 13.405 feet as one unit, and the height will be 0.5 of those units.

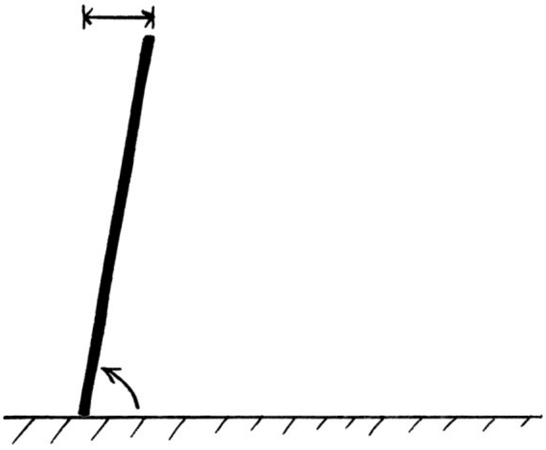


In the same way that the angle of the stick dictates at what height the end of the stick will be, it also dictates how far away the top end of the stick will be *horizontally* from the bottom end.

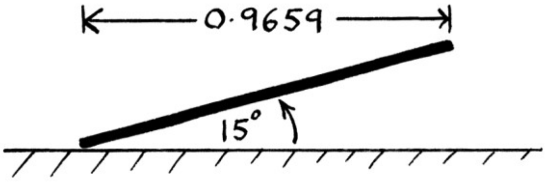




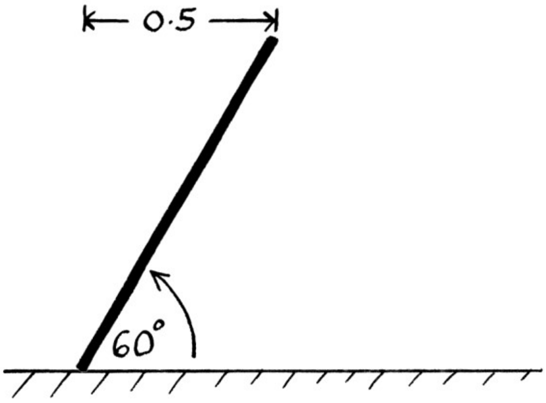




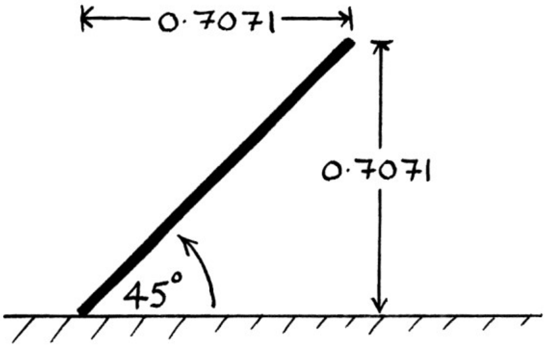
The horizontal distance from one end of the stick to the other is *always* the same for a particular angle. For example, if the metre-long stick is at 15 degrees, the horizontal distance will *always* be 0.9659 metres.



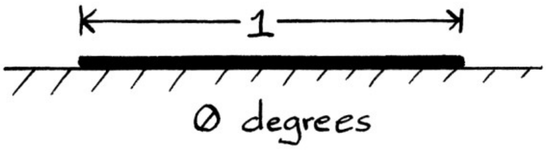
If the stick is at 60 degrees, the horizontal distance will *always* be 0.5 metres.



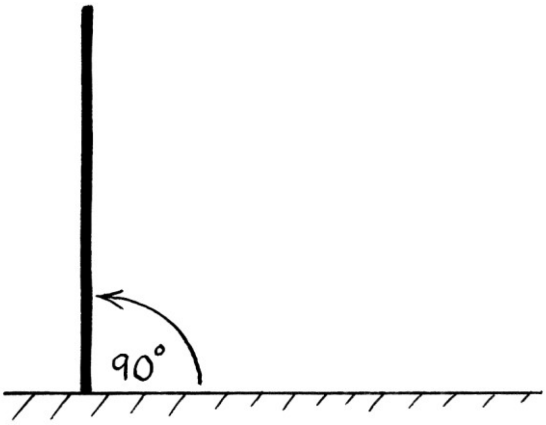
If the stick is at 45 degrees, then the horizontal distance will *always* be 0.7071 metres. This happens to be the same as the height of that end when the stick is at that angle.



If the stick is at 0 degrees – in other words if it is lying on the ground – the horizontal distance will be the length of the stick: 1 metre.



If the stick is at 90 degrees – in other words, if it is pointing directly upwards – the horizontal distance will be 0 metres. One end of the stick is directly above the other.



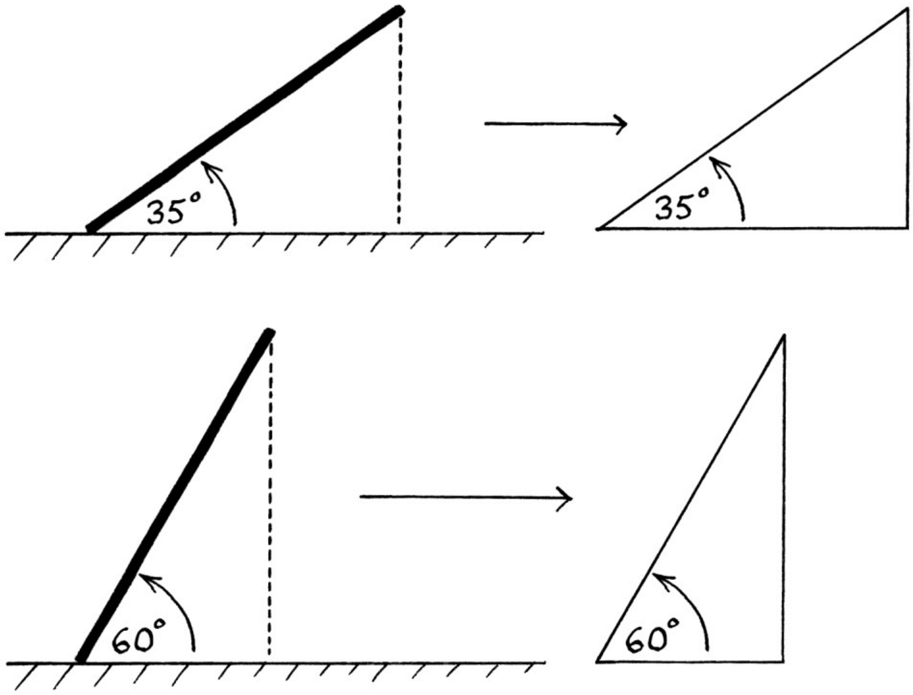
The relationship between the angle of the stick and the horizontal distance from one end to the other is called the "Cosine function". The Cosine of the angle gives the horizontal distance between the ends of the stick. If we wanted to find out the horizontal distance between the two ends of a one-metre long stick with a calculator, we would enter the stick's angle and then press the Cosine button.

As with Sine, the particular units and size of the stick are irrelevant to the Cosine function. If the stick is one unit long of any unit, then the Cosine function gives the horizontal distance between the ends in terms of that unit. For example, if the stick is one kilometre long and at 30 degrees, the distance will be 0.8660 kilometres. If the stick is one millimetre long and at 30 degrees, the distance will be 0.8660 millimetres.

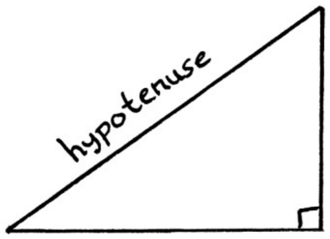
Sine and Cosine are completely linked. We can see this by raising or lowering the end of the stick. As the height of the end increases, that is to say, as the Sine of the angle increases, the horizontal distance between the ends, that is to say the Cosine of the angle, decreases. Conversely, as the horizontal distance between the ends (the Cosine of the angle) increases, the vertical distance between the ends (the Sine of the angle) decreases.

Triangles

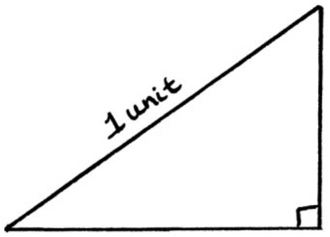
The nature of the stick as it is moved from lying flat on the ground to pointing directly upwards can be portrayed using right-angled triangles: [In the United States, these are called "right triangles".]



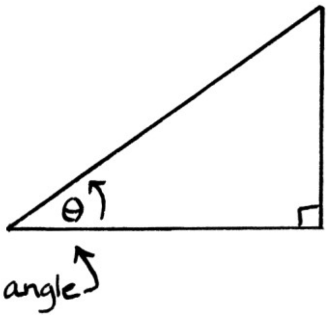
So we can use right-angled triangles to describe the behaviour of the stick, the stick itself will be portrayed by the longer side of the right-angled triangle, which is generally called the “hypotenuse”.



We will ignore that the stick is one metre long, and just treat it as being one unspecified unit long. Its length will be one unit. Therefore, the hypotenuse of the right-angled triangle will also be one unit long.

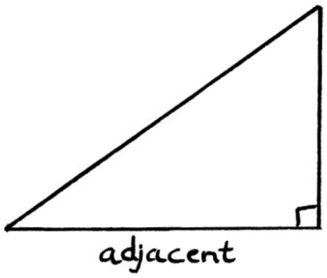


The angle of the stick will be portrayed by the angle of the hypotenuse with respect to the horizontal base of the triangle. The angle representing the angle of the stick will be referred to as “the angle” or “the angle of interest”. Although there are two other angles in the triangle (the 90-degree one and the other non-90-degree one), we will ignore them for the purposes of this explanation. By convention, we will use the lower-case Greek letter “θ”, pronounced “theta”, as the symbol for the angle of interest when either it is unknown or its exact value is irrelevant to the explanation.

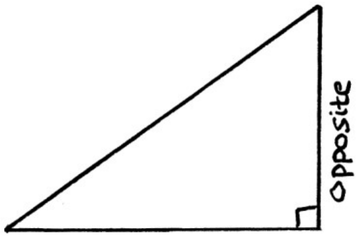


[Ancient and Modern Greek use the same alphabet, but with slightly different pronunciations. The letter "θ" is the Greek equivalent of the "th" sound in English. Do not confuse "θ" with a zero with a diagonal line drawn through it to distinguish it from the letter "O" for "Oscar".]

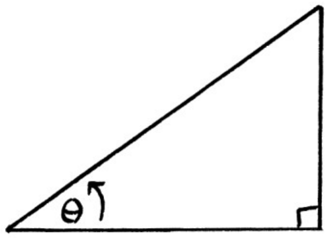
The horizontal distance between the ends of the stick is portrayed by the "adjacent side" of the triangle. This side is acting as the stretch of ground directly underneath the stick. It is called the adjacent side because it is *adjacent* to, or next to, our angle of interest.



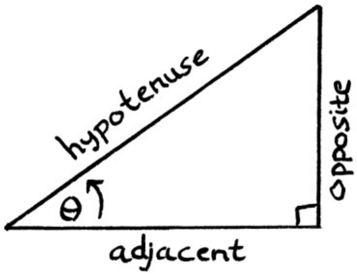
The height of the end of the stick is portrayed by the "opposite side" of the triangle. It is called the opposite side because it is *opposite* our angle of interest.



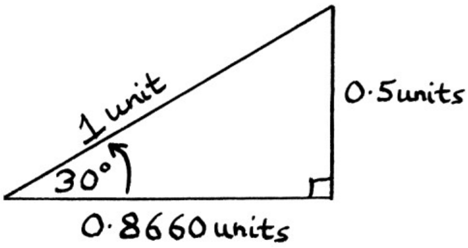
We will treat the basic example of a triangle drawn with the adjacent side lowest and with the angle of interest on the left as the "correct" way to draw it – this is completely arbitrary, but we need to do this if we are going to be consistent with generally accepted mathematical conventions. By convention, angles increase anticlockwise. This means that a triangle with a higher angle will have a steeper hypotenuse.



The fully labelled triangle looks like this:



The properties of the triangle remain true to the original properties of the stick. For example, if we draw a right-angled triangle with a hypotenuse 1 unit long at a 30-degree angle, the opposite side of that triangle will always be 0.5 units. The adjacent side will always be 0.8660 units. These are the same as if the metre-long stick were held at a 30-degree angle.



In a right-angled triangle, the angle of a unit-long hypotenuse dictates exactly how long the opposite and adjacent sides will be. As with the stick, there is a fixed, unchangeable relationship between the angle of the hypotenuse and the length of the opposite side (the Sine function), and there is a fixed, unchangeable relationship between the angle of the hypotenuse and the length of the adjacent side (the Cosine function).

Given the fixed relationship, it might be apparent that if the opposite side of a unit-long hypotenuse right-angled triangle has a particular length, then the angle of the hypotenuse can only ever be one value. Similarly, if the adjacent side of a unit-long hypotenuse right-angled triangle has a particular length, then the angle of the hypotenuse can only ever be one value.

To refer to the Sine relationship, the expression “Sine” or “sin” is used. “sin 20” means the Sine function performed on the number 20, which is another way of saying “the height of the opposite side of a right-angled triangle that has a unit-long hypotenuse at an angle of 20 degrees.”

Similarly, to refer to the Cosine relationship, the expression “Cosine” or “cos” is used. For example, “cos 20” means the Cosine of 20, which is another way of saying “the length of the adjacent side of a right-angled triangle that has a unit-long hypotenuse at an angle of 20 degrees”.

In this book, when I give the result of Sine, Cosine or any other calculation, I will generally round the number up to the fourth decimal place, or the fourth decimal place after any preceding zeroes.

Sine and Cosine examples

To develop an understanding of how the Sine and Cosine functions work, here are the results of the Sine and Cosine of various angles in degrees:

The Sine of 0 degrees = 0 units.	The Cosine of 0 degrees = 1 unit.
Sine 1 = 0.01745	Cosine 1 = 0.9998
Sine 2 = 0.03490	Cosine 2 = 0.9994
Sine 3 = 0.05234	Cosine 3 = 0.9986
Sine 4 = 0.06976	Cosine 4 = 0.9976
Sine 5 = 0.08716	Cosine 5 = 0.9962
Sine 10 = 0.1736	Cosine 10 = 0.9848
Sine 20 = 0.3420	Cosine 20 = 0.9397
Sine 30 = 0.5	Cosine 30 = 0.8660
Sine 40 = 0.6428	Cosine 40 = 0.7660
Sine 44 = 0.6947	Cosine 44 = 0.7193
Sine 45 = 0.7071	Cosine 45 = 0.7071
Sine 46 = 0.7193	Cosine 46 = 0.6947
Sine 60 = 0.8660	Cosine 60 = 0.5
Sine 70 = 0.9397	Cosine 70 = 0.3420
Sine 80 = 0.9848	Cosine 80 = 0.1736
Sine 88 = 0.9994	Cosine 88 = 0.03490
Sine 89 = 0.9998	Cosine 89 = 0.01745
Sine 90 = 1	Cosine 90 = 0

You might notice that the Sines of angles from 0 rising up to 90 degrees are identical to the Cosines of angles from 90 going down to 0 degrees. The reason for this becomes clearer if we think about the stick again. If the stick is flat on the ground, the vertical distance between the ends is 0 units and the horizontal distance is 1 unit; if the stick is pointing upwards, the vertical distance is 1 unit and the horizontal distance is 0 units. As the end of the stick moves upwards, the height between the ends increases while the horizontal distance between the ends decreases. As the end of the stick moves downwards, the horizontal distance between the ends increases while the height between the ends decreases. As the stick is raised and lowered, the values of the vertical distances between its ends will pass through the same range of values that are possible for the horizontal distance between its ends (but not at the same time).

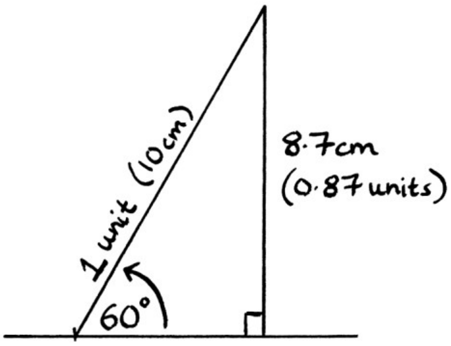
From examining the list, we can make the observation that the Sine of an angle is the same as the Cosine of that angle subtracted from 90 degrees, and the Cosine of an angle is the same as the Sine of that angle subtracted from 90 degrees.

A more mathematical way of saying this is:
$\sin \theta = \cos (90 - \theta)$
$\cos \theta = \sin (90 - \theta)$

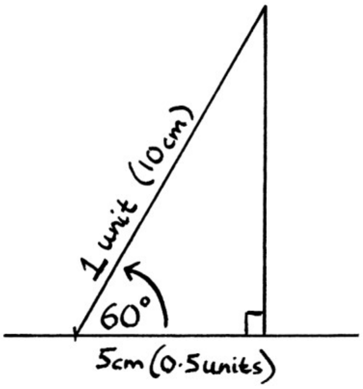
Calculating Sine and Cosine

We can calculate the Sine of an angle between just over 0 degrees and just under 90 degrees by drawing a reasonably accurate right-angled triangle with a one-unit long hypotenuse. We draw the hypotenuse at the angle for which we want to find the Sine, and then we measure the length of the opposite side of the triangle to reveal the Sine of the angle. The larger the unit system we use, the more accurate the result will be. Given that we will probably be doing this on an ordinarily sized piece of paper, it is best to create our own unit of length and say that 10 centimetres is one unit, instead of using 1 inch, 1 centimetre, or 1 foot. [Strictly speaking, 10 centimetres *is* a unit – it is 1 decimetre, but rulers do not usually count in decimetres]. With 10 centimetres as one unit, all measurements in centimetres on the triangle will be 10 times the number of units they represent.

To calculate the Sine of 60 degrees, we can draw a right-angled triangle with a hypotenuse that is 10 centimetres long and at an angle of 60 degrees. We then measure the opposite side of the triangle in centimetres and divide that by 10 to give the number of our units. Without putting in much effort to be accurate, it is possible to calculate the Sine of 60 degrees as 0.87 units. [A calculator would give the Sine of 60 degrees as 0.8660 units]. Generally, when drawing triangles in this way, we should be able to achieve a result within 0.01 units of the correct answer.



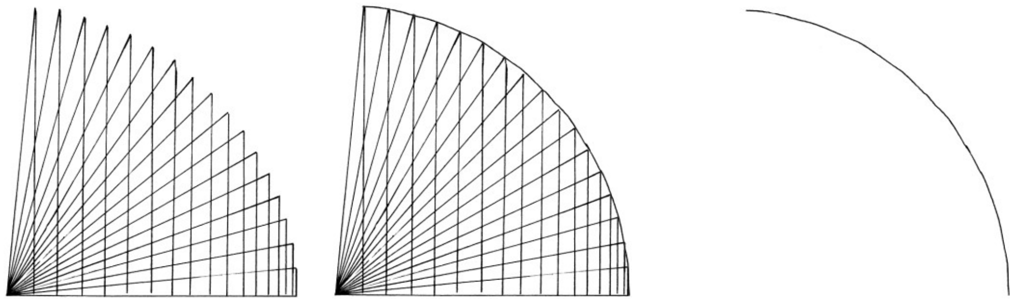
We can use the same method to calculate the Cosine of a number between just over 0 and just under 90 degrees, but in that case, we would measure the length of the *adjacent* side to reveal the Cosine of the angle. Without much effort, it is easy to calculate the Cosine of 60 degrees to be 0.5 units. [A calculator would give the Cosine of 60 degrees as 0.5 units, but the difference is that on a calculator, we know the result is exact. When we draw a triangle using 10 centimetres as one unit, we are limited to how well we can measure millimetres, so we could never be truly sure that the actual result was not something such as 0.49 or 0.51].



The larger the triangle we draw, the more accurate the result will be, but on the other hand, the harder it will be to make an accurate measurement using everyday geometry tools.

Circles

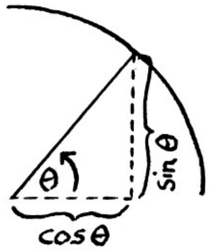
If we draw a series of right-angled triangles starting from the same point, all with a unit hypotenuse, with angles of 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80 and 85 degrees, the points at the end of the hypotenuses will draw out most of a quarter of a circle.



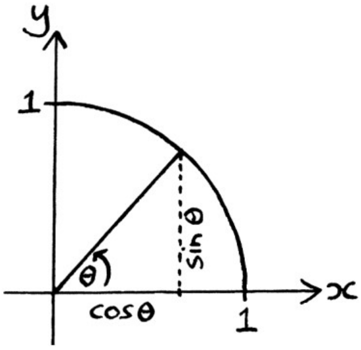
The same quarter of a circle would be portrayed by moving the end of the metre-long stick from just above horizontal to just below vertical.

That these triangles draw out a quarter circle means that we could just as easily draw a quarter of a circle with a unit radius, and we would have a look-up table for Sine and Cosine for nearly every possible right-angled triangle.

We would draw a line from the origin of the quarter circle out to the edge at the angle for which we want to find the Sine or Cosine. The Sine will be the height where that line crosses the quarter circle's edge, and the Cosine will be the horizontal distance from the centre to where that line crosses the quarter circle's edge.

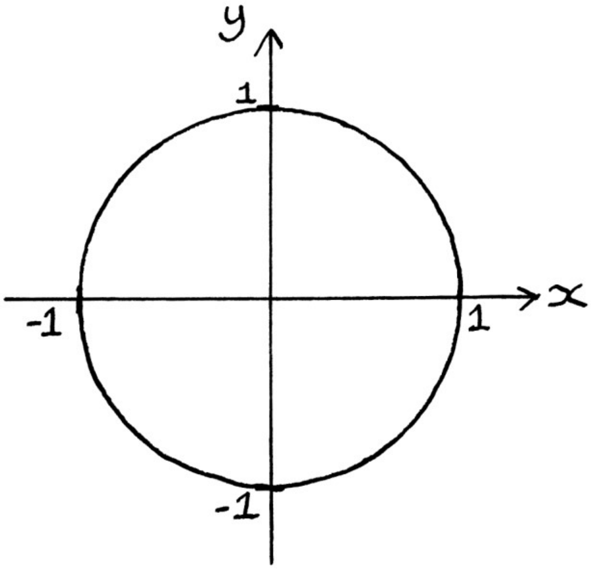


This is much easier to do if the quarter circle is drawn over x and y-axes. In which case, the Sine of an angle will be the y-axis value of the point where the line crosses the quarter circle's edge, and the Cosine of an angle will be the x-axis value of the point where the line crosses the quarter circle's edge:

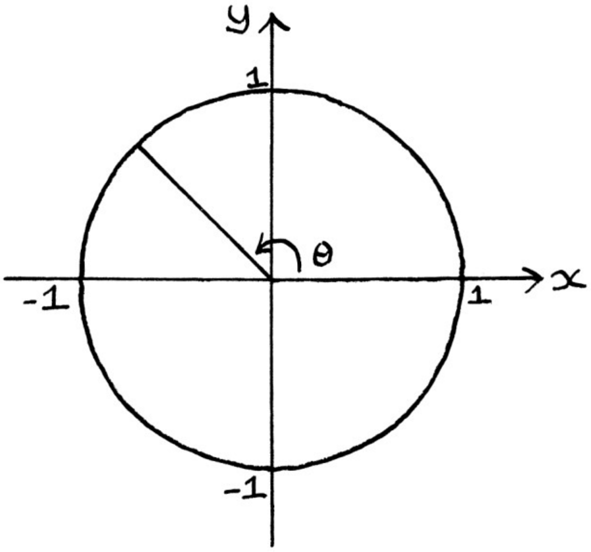


One problem here is that a right-angled triangle cannot have an angle of zero degrees, 90 degrees, or more than 90 degrees. From remembering the behaviour of the stick, we know that for an angle of zero degrees, Cosine would produce 1, and Sine would produce 0 (the stick would be lying on the ground). For an angle of 90 degrees, Cosine would produce 0, and Sine would produce 1 (the stick would be vertical). However, that does not help with angles over 90 degrees. Fortunately, we can measure the Sine and Cosine of any angle from 0 to 360 degrees by using a full circle.

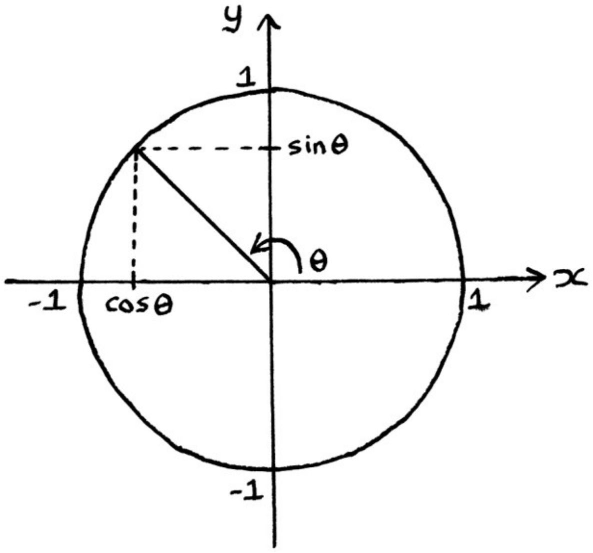
We draw a circle with a 1-unit radius, centred on the origin of x and y-axes:



Then, we draw a line from the centre of the axes out to the edge of the circle at the angle for which we want the Sine or Cosine.



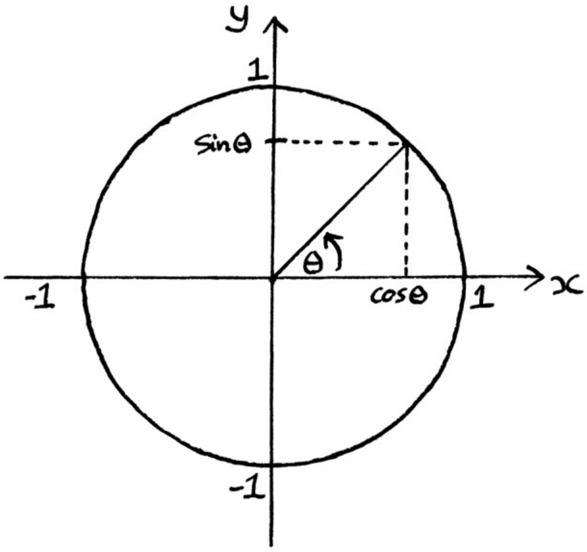
The height of where that line crosses the circle's edge is the Sine of that angle. The horizontal distance from the origin out to where that line crosses the circle's edge is the Cosine of that angle. In other words, the y-axis value of that point is the Sine of the angle; the x-axis value of that point is the Cosine of the angle.



We can think of the circle as being made up of the far ends of the hypotenuses of an infinite number of right-angled triangles that point to the right, to the left, up and down. The radius of the circle (1 unit) is the same as the hypotenuses of the triangles it is made from (also 1 unit). Strictly speaking, we cannot have a right-angled triangle with an angle of interest that is 0 or 90 degrees, so there are 4 points on the circle that are not really based on triangles – those at angles of 0 degrees, 90 degrees, 180 degrees and 270 degrees.

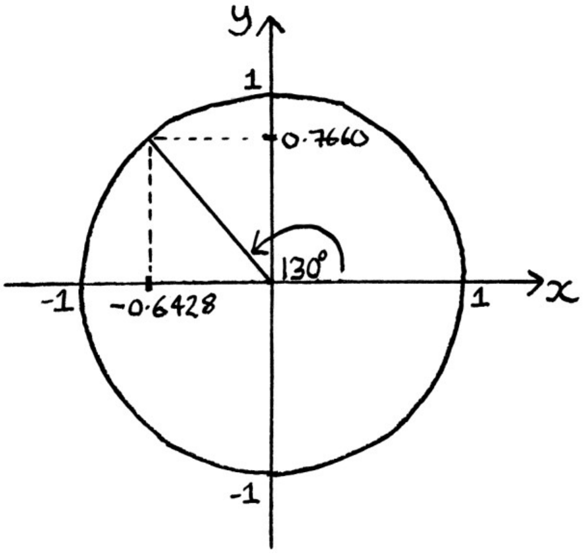
Results for Sine and Cosine on the circle

For angles from 0 to 90 degrees, the results of Sine and Cosine will be positive, as we have seen several times by now. The line drawn from the origin at the chosen angle will end up in the top right quarter of the circle.



For angles from just over 90 to 180 degrees, the *Sine* of the angle will still be positive, but the *Cosine* of the angle will be negative. The point to which we would measure is in the negative half of the x-axis. A line drawn from the origin of the axes at the chosen angle would end up in the top left quarter of the circle. Back when we were dealing with an angled stick, treating a distance as negative would not have made much sense as we were measuring between two points. However, when it comes to the circle and triangles, it is important to pay attention to the direction of the lines away from the origin of the axes. A negative x-axis value means that a line is pointing to the left; a negative y-axis value means that the line is pointing downwards.

With a 130-degree angle, the circle would look like this:



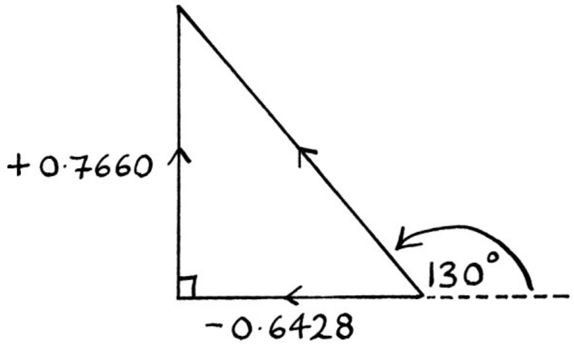
On the circle, the hypotenuse of a right-angled triangle is the line going from the centre of the circle out to the edge. We knew this already. For the opposite and adjacent sides, things are slightly more complicated when we deal with angles over 90 degrees.

The opposite side of the triangle is a vertical line going from the x-axis towards the end of the hypotenuse. When we deal with any angle around the circle, it means that sometimes that vertical line will be drawn pointing upwards, and sometimes it will be drawn pointing downwards – it all depends on whether the hypotenuse is pointing into the top half or the bottom half of the circle.

The adjacent side of the triangle is the horizontal line drawn from the y-axis out to the base of the opposite side. This means that the adjacent side might be pointing to the right or it might be pointing to the left – it all depends on whether the hypotenuse is pointing to the right or left.

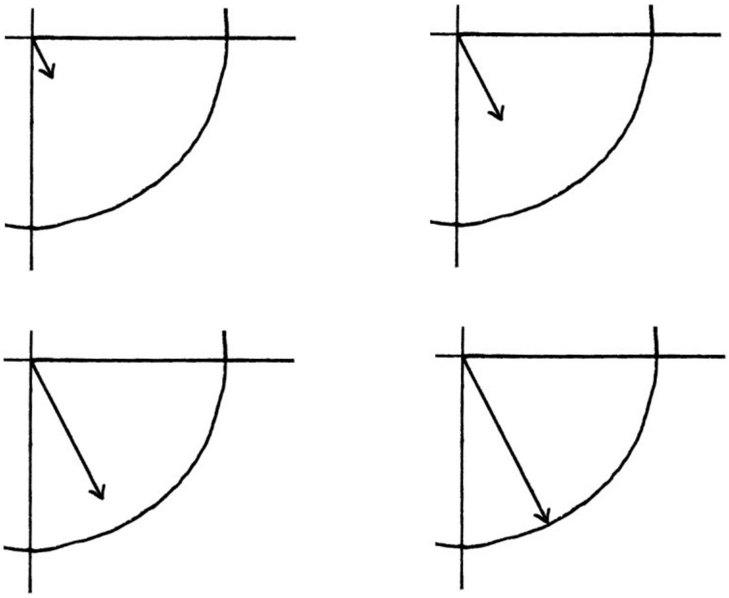
All of this means that a triangle might be the wrong way around, upside down, or both, depending on the angle of the hypotenuse.

If we were to draw a corresponding triangle for the angle of 130 degrees, and yet still be consistent with the thinking behind our original triangles, the angle would be on the outside. The angle of the hypotenuse line is still 130 degrees, but the triangle drawn around that hypotenuse is the wrong way around. The hypotenuse points upwards and to the left. The adjacent side points to the left; the opposite side points upwards. Therefore, the length of the adjacent side is *negative*, and the length of the opposite side is still positive. The triangle is the right way up, but it faces to the left:

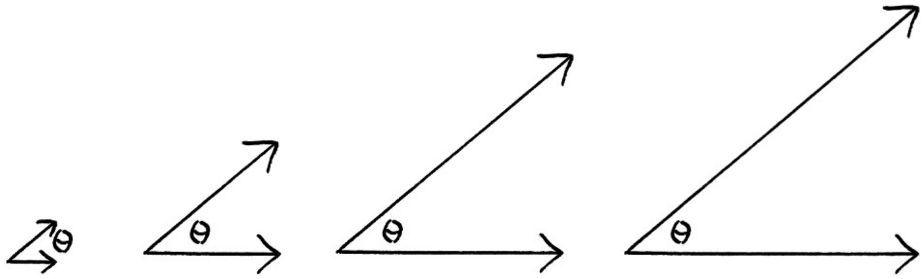


Side note: direction of lines

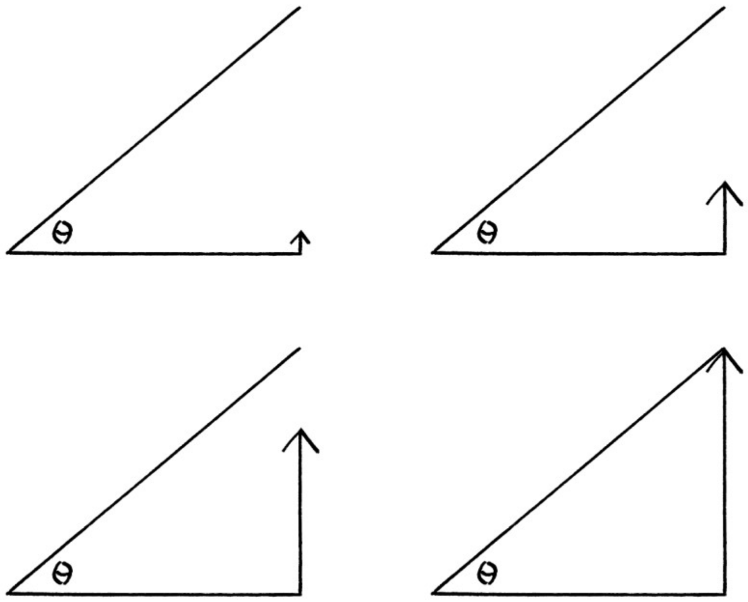
In the above picture, I have drawn arrows on the adjacent and opposite sides to indicate the direction of the sides away from the angle of interest (which is the same thing as the direction of the sides away from the origin of the x and y-axes). When drawing triangles or lines around the circle, it can help to clarify matters if we indicate the direction of the line. We can imagine that every line is drawn starting at the origin of the circle's axes and moving out to its end:



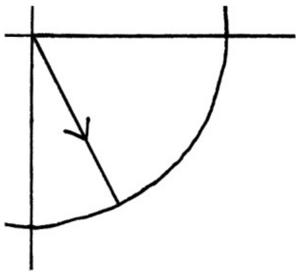
On a triangle, we can imagine that the hypotenuse and adjacent side are drawn starting at the angle of interest and moving out towards the end:

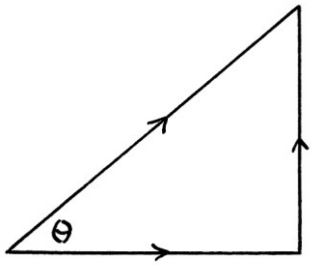


Think of the opposite side being drawn from the level of the angle of interest up to, or down to, the end of the hypotenuse:



By thinking in this way, the lines all have a "direction" – they point in a particular direction. We can mark arrows on the lines to indicate the direction to make them clearer.

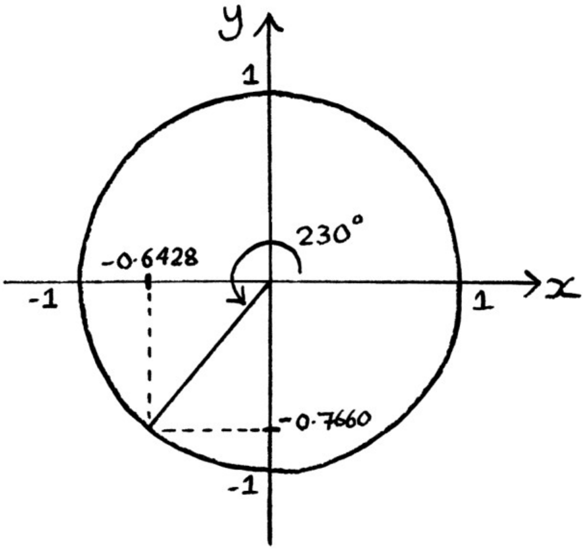




Thinking of the direction of the lines allows us to know if the length of a line, or the result of Sine or Cosine, will be positive or negative. If a horizontal line points to the right, it will have a positive length; if it points to the left, it will have a negative length. Similarly, if a vertical line points upwards, it will have a positive length; if it points downwards, it will have a negative length. A hypotenuse that points upwards and to the right will be part of a triangle with a positive length opposite side and a positive length adjacent side. A hypotenuse that points downwards and to the left will be part of a triangle with a negative length opposite side and a negative length adjacent side.

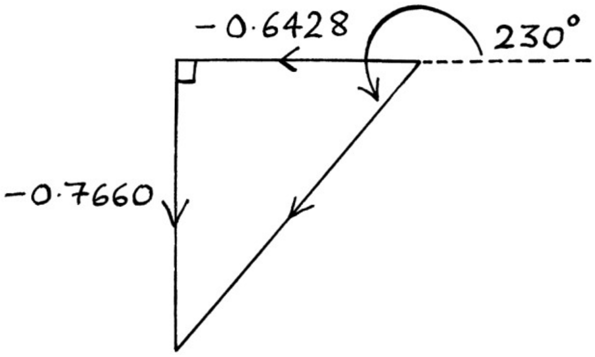
Back to circles

For angles from just over 180 to 270 degrees, the Sine of the angle and the Cosine of the angle will both be negative. The points on the x-axis and y-axis to which we would measure are negative. The line points into the lower left hand corner of the circle. For example, with the angle of 230 degrees, the Sine is −0.7660 units and the Cosine is −0.6428 units:

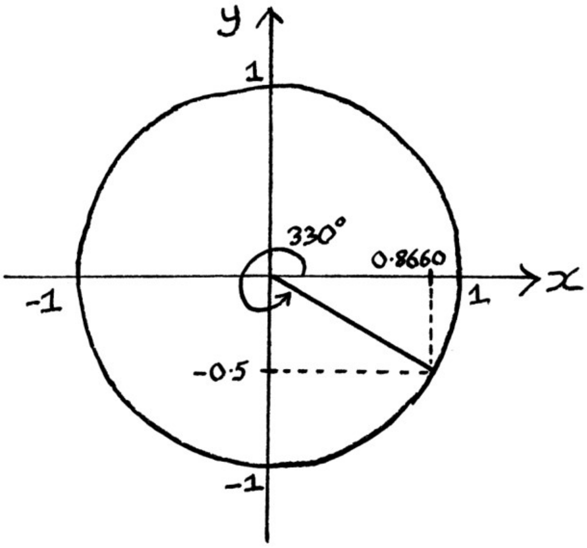


The corresponding triangle for this would be upside down and face to the left. The hypotenuse points down to the left, the opposite side points downwards, and the adjacent side points to the left.

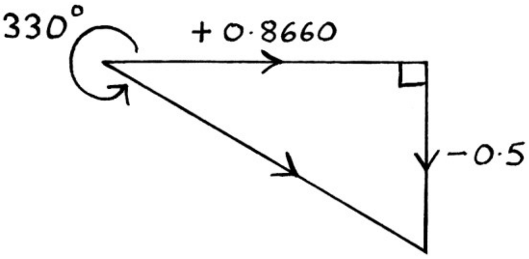
It looks like this:



For angles between 270 and 360 degrees, the Sine of the angle will be negative (as the point is in the negative half of the y-axis), but the Cosine of the angle will be positive. For example, for an angle of 330 degrees, the Sine is −0.5 units and the Cosine is 0.8660 units.

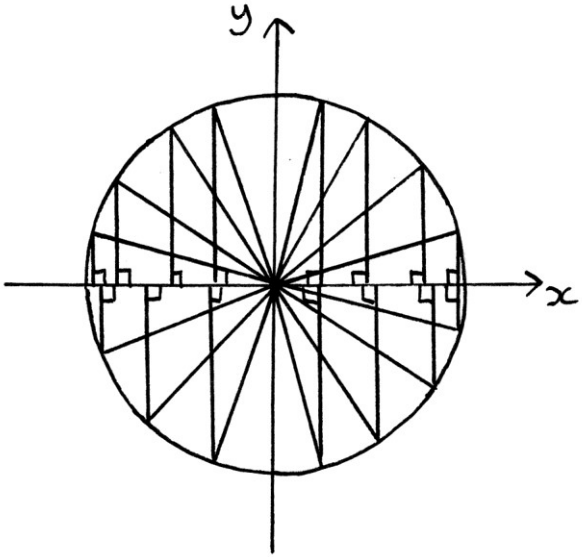


The corresponding triangle for this would be the “right way around”, but upside down. The hypotenuse points down to the right, the opposite side points downwards, and the adjacent side points to the right.



Intuitively knowing some results for Sine and Cosine

As we have seen, when using the circle to measure the Sine and Cosine of angles, we are still essentially dealing with triangles, but they are not necessarily pointing in the same direction as before. We can think of the circle's edge as being made up of the end points of an infinite number of triangle hypotenuses.



For angles that are specifically at 0, 90, 180, 270 and 360 degrees, we cannot draw corresponding triangles, as such triangles would be flat, and so, by definition, would not be triangles. The circle still works though, and it is possible to visualise the circle without drawing it, and so know what the Sines and Cosines would be at certain places:

- At 0 degrees, the Sine of the angle is 0, and the Cosine of the angle is 1. We know this because if it were possible to draw a right-angled triangle in such a case, there would be no height to its opposite side (it would be zero), and the length of the adjacent side would be the same as the length of the hypotenuse (1 unit). This situation is equivalent to the stick in the first examples lying on the ground.

- At 90 degrees, the Sine of the angle is 1, and the Cosine of the angle is 0. We know this because the opposite side of a right-angled triangle would be pointing directly upwards and so have the same length as the hypotenuse. The adjacent side would have no length, so would be zero. This situation is equivalent to the stick pointing vertically upwards.

- At 180 degrees, the Sine of the angle is 0, and the Cosine of the angle is −1. The opposite side of a right-angled triangle would have no length, and the adjacent side would be the same length as the hypotenuse but pointing in the negative direction. This situation is equivalent to the stick lying on the ground, but pointing in the other direction to before.

- At 270 degrees, the Sine of the angle is −1 and the Cosine of the angle is 0. The opposite side of a right-angled triangle would be pointing directly downwards and so be the same length as the hypotenuse (1 unit), while the adjacent side would have no length. This is equivalent to the stick being pushed into the ground, so the top is now underground and beneath what was the bottom.

- At 360 degrees (which is the same as 0 degrees), the Sine of the angle is 0 and the Cosine of the angle is 1. This is, again, equivalent to the stick lying on the ground in its original position.

Maximum lengths

Now we can see Sine and Cosine's relationship with the circle, it is important to realise that the Sine or Cosine of any number can never be more than +1 or less than −1 because:

- Sine indicates the height of a point on a unit circle (a circle with a radius of 1 unit) as measured on the y-axis, and no point on that circle reaches beyond y = 1 or below y = −1.

- Cosine indicates the horizontal distance out from a point on a unit circle as measured on the x-axis, and no point on that circle goes beyond x = 1 or under x = −1.

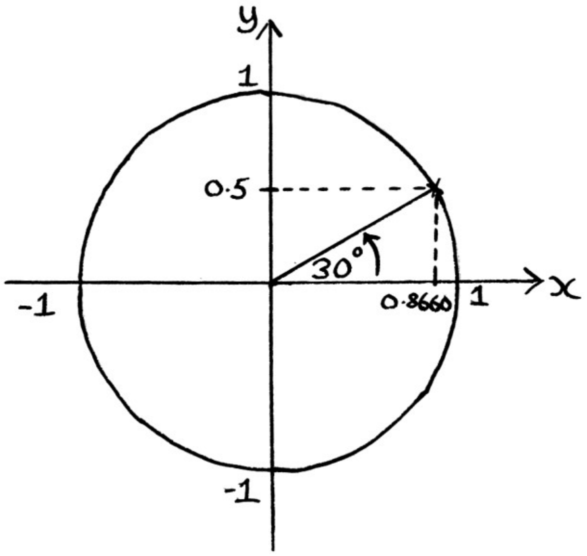
To put all this in another way, if a triangle has a hypotenuse of 1 unit, then no matter the angle of the hypotenuse, the adjacent and opposite sides can never be longer than it is.

To express this idea in yet another way, if we have a 1-metre long stick, no matter at which angle we hold it, the top end will never be more than 1 metre vertically or horizontally from the other end.

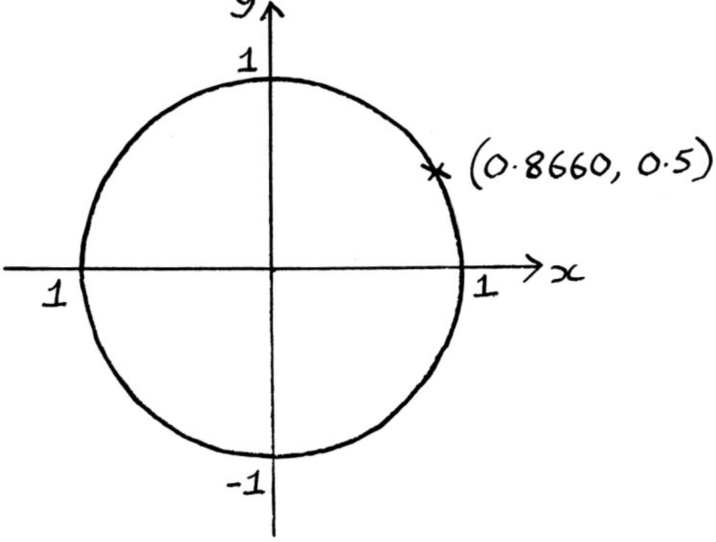
Coordinates on the circle

As we have seen, measuring the position of a point on the circle's edge gives the Sine and Cosine of that point's angle from the origin.

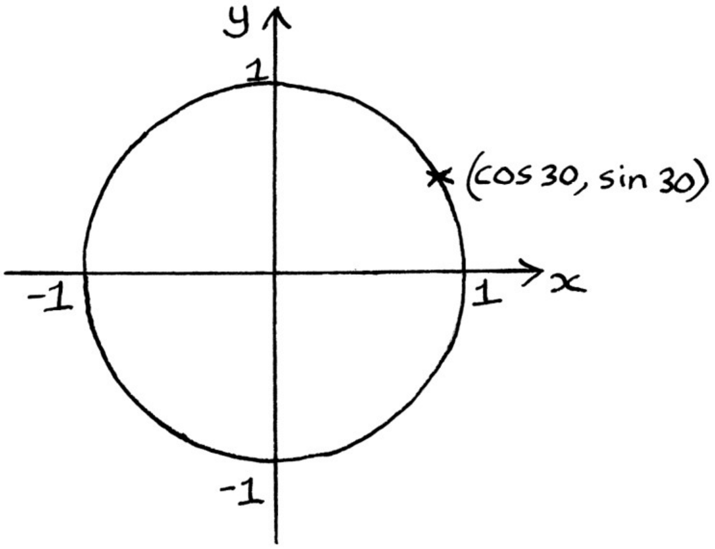
Another way of thinking about this is that the Sine and Cosine of that point's angle give the *coordinates* of that point on the circle's edge. If a point on the circle's edge is at 30 degrees from the origin, then we know that its y-axis value is Sine 30 = 0.5 and that its x-axis value is Cosine 30 = 0.8660.



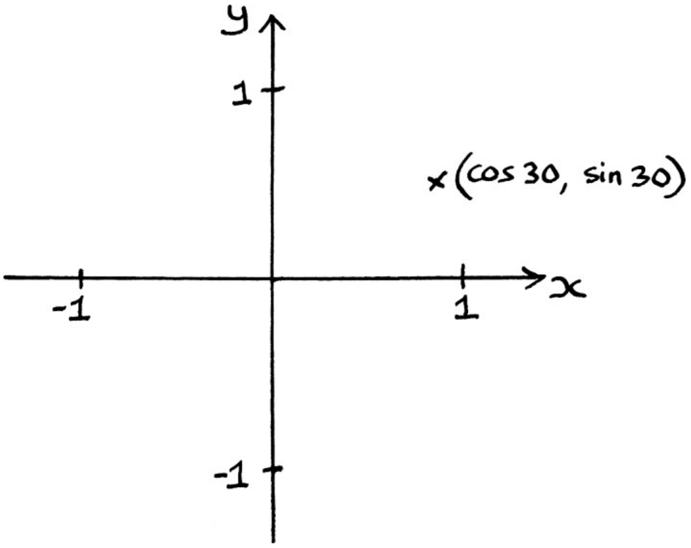
Therefore, we can write the Cosine result followed by the Sine result to give the coordinates of the point: (0.8660, 0.5). The Cosine and Sine of the angle have given us the coordinates.



We could also write the coordinates in the form (cos 30, sin 30), which means the same thing, and in some situations can be just as useful.



In fact, the coordinates of any point on the circle can be given by just calculating the Cosine and Sine of that point's angle from the origin. We do not really need to draw the circle, and we can use Sine and Cosine to identify any point 1 unit away from the origin of the axes by its angle.

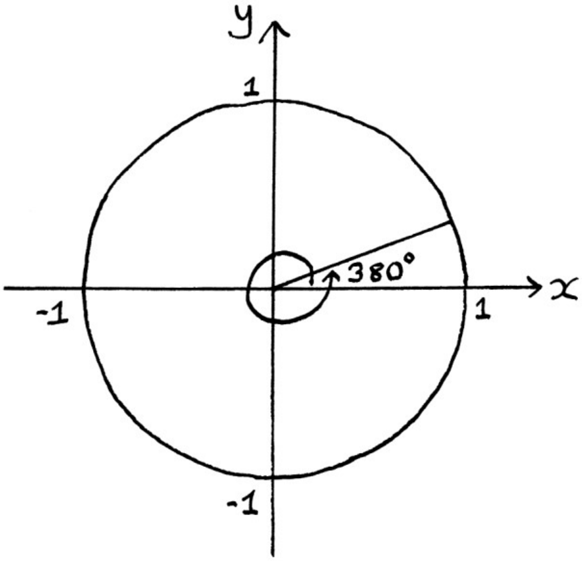


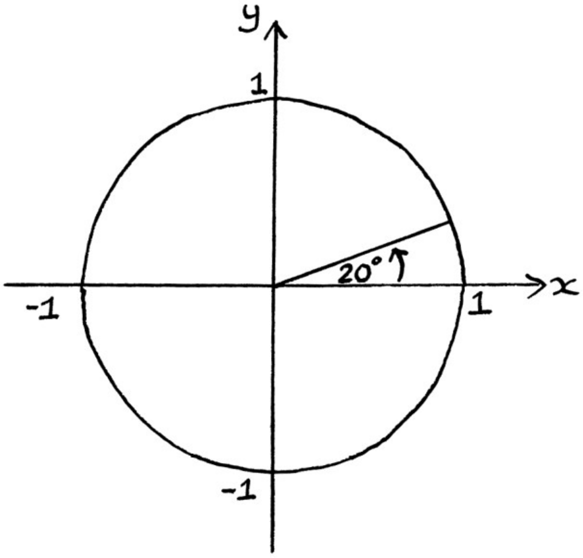
At the moment, using Sine and Cosine for coordinates is merely interesting, but it will become more useful later on.

More Sine and Cosine results

Angles over 360 degrees

If we want to calculate the Sine or Cosine of a number over 360 degrees, we can still use the circle. For example, if we measure an angle of 380 degrees, it would wrap around the circle to be in the same place as an angle of 20 degrees.





Therefore, the Sine and Cosine of 380 degrees are the same as the Sine and Cosine of 20 degrees. From the point of view of the circle, 380 degrees *is* 20 degrees.

The results of the Sine and Cosine functions repeat every 360 degrees. To calculate the Sine or Cosine of very large numbers, we can just keep subtracting 360 until we end up with a number less than 360, and then draw a line at the resulting angle.

A mathematical way of expressing the relationship between angles below 360 degrees and angles of 360 degrees or above is:

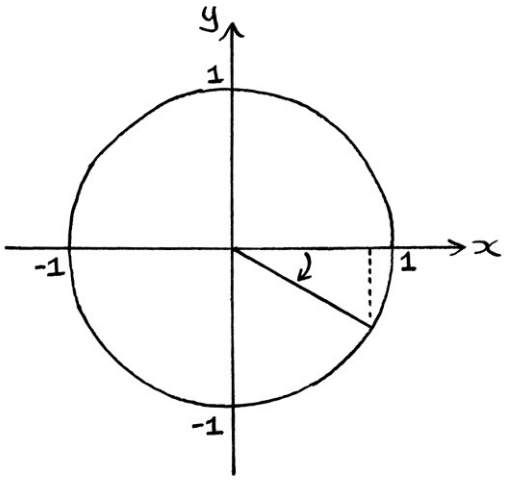
sin (θ + 360) = sin (θ)
cos (θ + 360) = cos (θ)

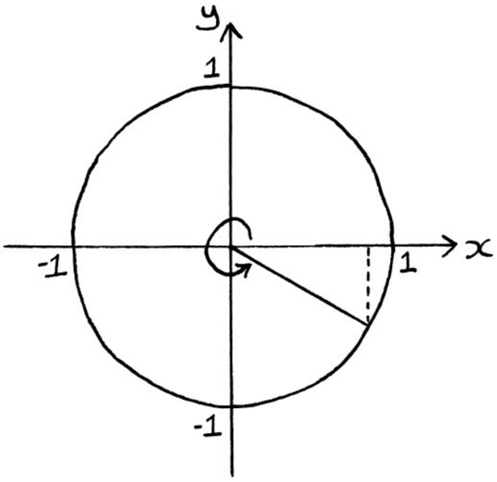
We could also give these as:

sin (θ) = sin (θ + 360)
cos (θ) = cos (θ + 360)

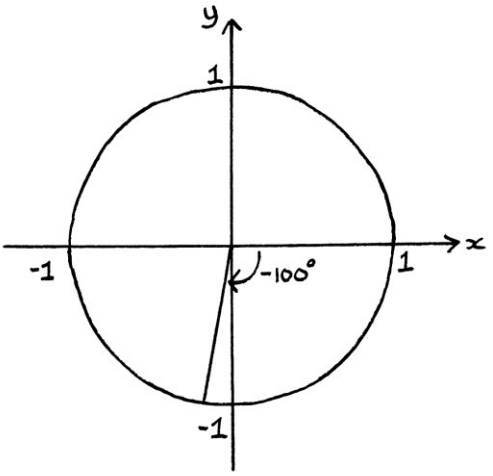
Negative angles

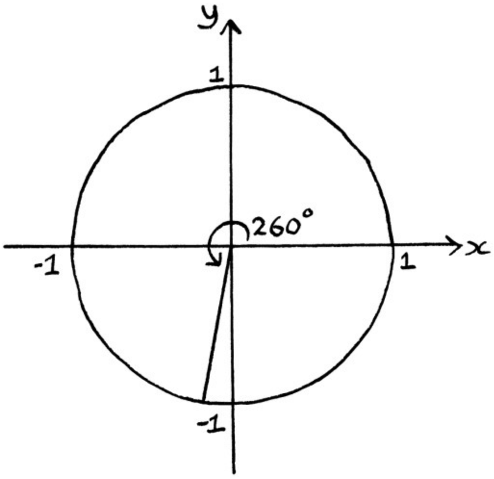
We can also use the circle to calculate the Sine and Cosine of *negative* numbers. We just have to realise that negative numbers mean going around the circle in the other direction (clockwise). We then measure the y-axis and x-axis values as before. If we draw a line at a negative angle out to the edge of a circle, we will have the same result as if we had added 360 to the negative number and drawn a line at that angle.



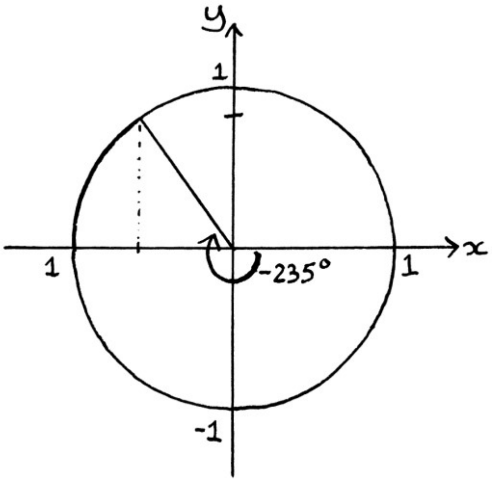


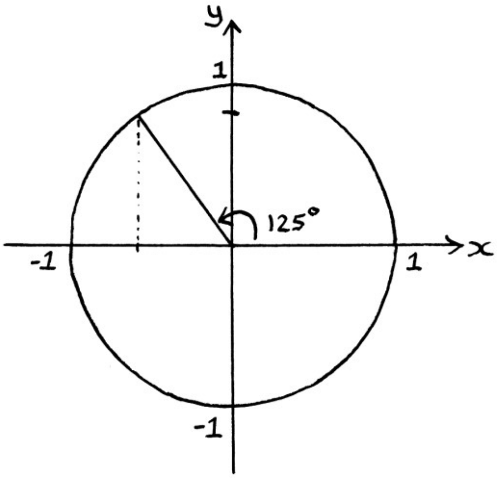
As an example, if we want the Sine or Cosine of −100, then that is just another way of saying the Sine or Cosine of 360 – 100, which is the Sine or Cosine of 260.





The Sine of −235 is the same as the Sine of 360 − 235 = the Sine of 125.





Similarly, the "Sine of −90" is really another way of saying the "Sine of 360 – 90", or "Sine 270". From the circle's point of view, −90 degrees is exactly the same thing as 270 degrees.

A mathematical way of expressing the relationship between positive and negative angles is:

sin (−θ) = sin (360 – θ)
cos (−θ) = cos (360 – θ)

We can rephrase these to be similar to the second pair of formulas for angles over 360 degrees:

$\sin(-\theta) = \sin(-\theta + 360)$
$\cos(-\theta) = \cos(-\theta + 360)$

Dealing with negative angles (which is another way of saying angles below zero) and dealing with angles over 360 degrees really involve the same process – we keep adding 360 to, or subtracting 360 from, the number until we end up with something between 0 and just under 360, and then find the Sine or Cosine of that value instead.

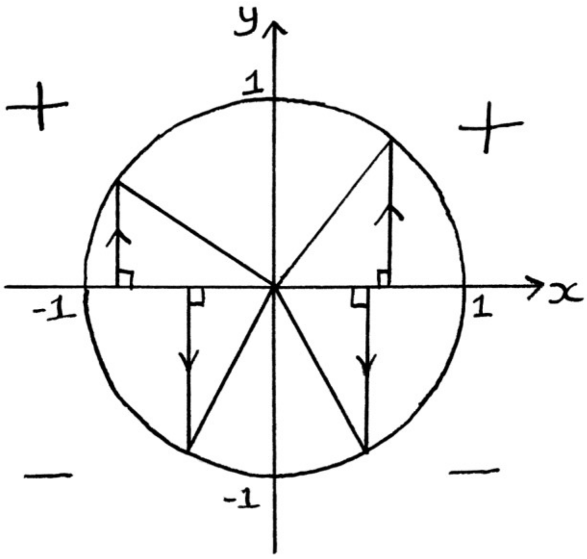
Sine and Cosine observations

Thinking about the circle helps us to know some facts about Sine and Cosine. Among other things, knowing these can be useful when checking for mistakes in calculations.

Positive and negative results

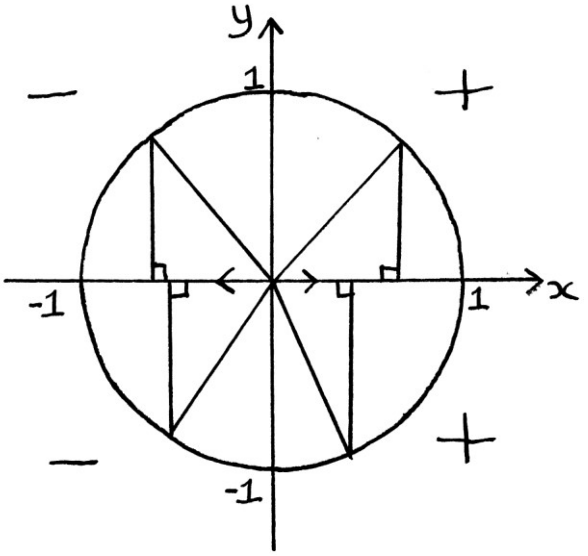
For Sine, any angle that results in a line pointing into the bottom half of the circle will have a negative result. Any angle that results in a line pointing into the top half of the circle will have a positive result. This might be obvious – the Sine of an angle gives the y-axis value of that point on the circle's edge. Therefore, if the y-axis value is positive, the Sine of that angle will be positive, and if the y-axis value is negative, the Sine of that angle will be negative.

Another way of saying all of this is that the Sine of any angle between 180 and 360 degrees will be negative, and the Sine of any angle between 0 and 180 degrees will be positive.



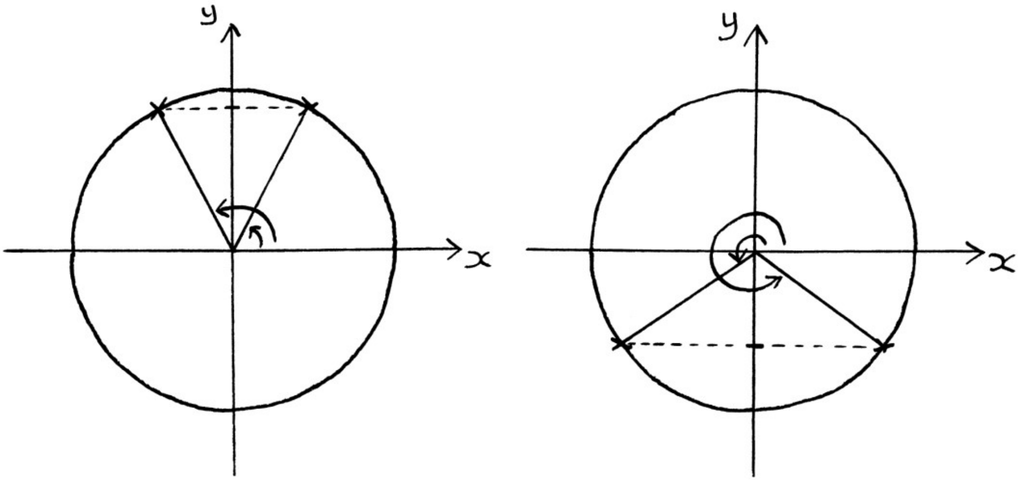
For Cosine, any angle that results in a line pointing into the left hand half of the circle will have a negative result. Any angle that results in a line pointing into the right hand half of the circle will have a positive result. Again, this might be obvious – the Cosine of an angle gives the x-axis value of a point on the circle's edge, so if the x-axis value is negative, the Cosine of the angle will be negative, and if the x-axis value is positive, the Cosine of the angle will be positive.

Another way of saying all this is that the Cosine of an angle between 90 and 270 degrees will be negative, and the Cosine of an angle between 270 and 90 degrees will be positive.



Symmetry

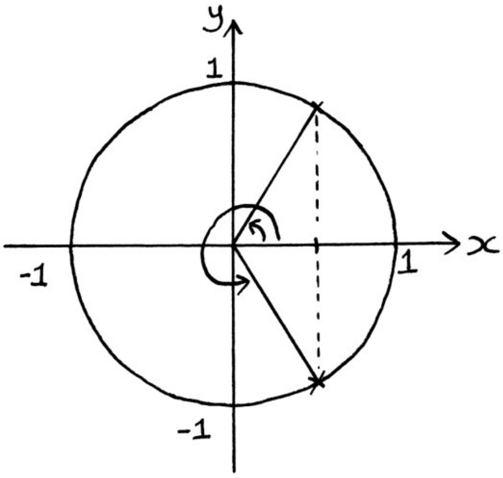
There is symmetry in the circle and the results of the Sine function – the Sine of the left half of angles is the same as the Sine of the right half of angles for points on the circle's edge at the same height. For example, Sine 10 is the same as Sine 170. Both are 0.1736 units. Sine 200 is the same as Sine 340. Both are −0.3420 units. We can tell this is true because the circle itself is symmetrical. Therefore, if a point on one side has a certain y-axis value, then the corresponding point on the other side will be at the same height, and so have the same y-axis value.

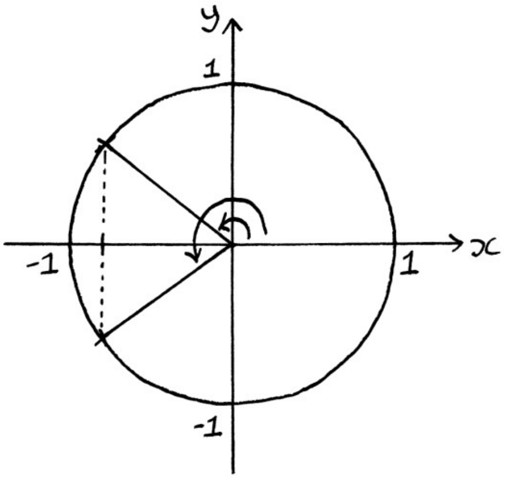


The rule stated more clearly is that the Sine of an angle so many degrees below 90 degrees is the same as the Sine of an angle that number of degrees above 90 degrees. A more mathematical way of expressing this is:

$\sin(90 - z) = \sin(90 + z)$

A similar rule exists for Cosine, where the Cosine of the top half of angles is the same as the Cosine of the bottom half of angles if the relevant lines meet the circle's edge at the same x-axis value. For example, Cosine 10 is the same as Cosine 350. Both are 0.9848 units; Cosine 170 is the same as Cosine 190. Both are −0.9848 units. Again, this might be obvious because if two angles result in lines that meet the circle's edge at the same x-axis point, then they will clearly have the same Cosine result.

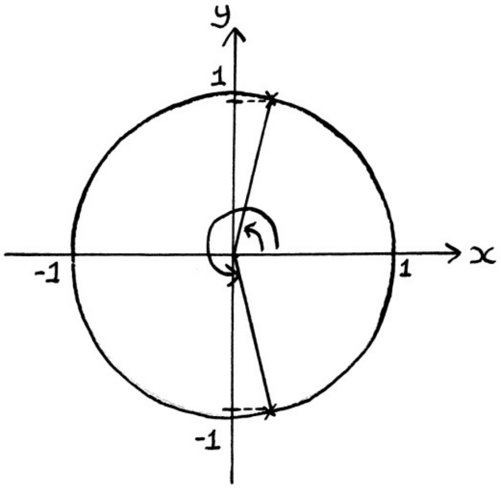


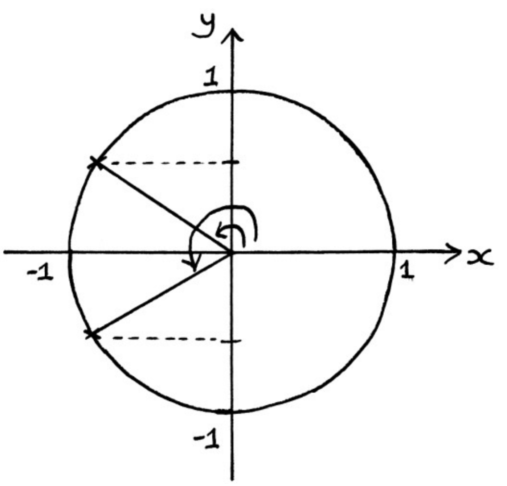


The rule for Cosine stated more clearly is that the Cosine of an angle so many degrees below 180 degrees will be the same as the Cosine of an angle that number of degrees above 180 degrees. A more mathematical way of expressing this is:
"cos (180 – z) = cos (180 + z)"

Negative Symmetry

The Sines of the angles in the bottom half of the circle are the *negative* of the corresponding ones in the top half. For example, Sine 190 is the negative of Sine 170. They are −0.1736 units and +0.1736 units. Sine 350 is the negative of Sine 10. They are −0.1736 units and +0.1736 units. We can tell that this is true because the circle is symmetrical, so the y-axis value of one point on its edge in the top half must be the negative of the corresponding point in the bottom half.

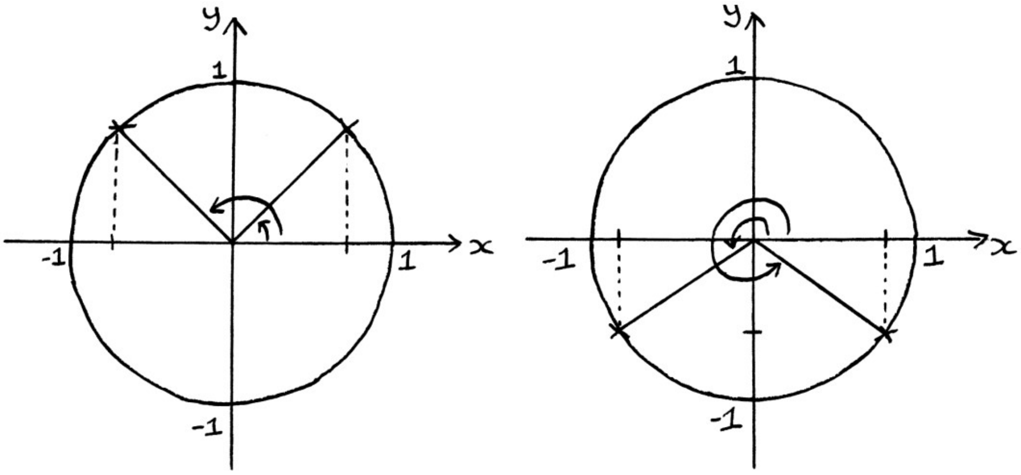




The basic rule is that the Sine of an angle so many degrees below 180 degrees will be the negative of the Sine of an angle that number of degrees above 180 degrees. Or to put it mathematically:

sin (180 − z) = −sin (180 + z)

For Cosine, the Cosines of the angles in the right half are the negative of the corresponding ones in the left half. For example, Cosine 190 is the negative of Cosine 350. They are −0.9848 units and +0.9848 units. Cosine 170 is the negative of Cosine 10. They are −0.9848 units and +0.9848 units. Again, this might be obvious from thinking about the circle.

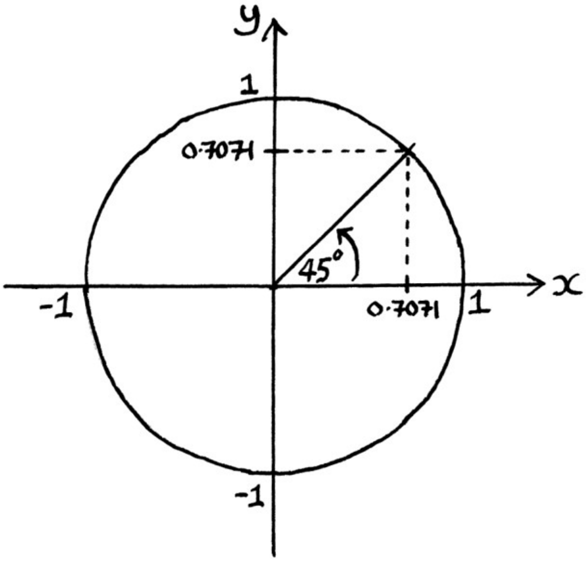


The rule for Cosine stated more clearly is that the Cosine of an angle so many degrees below 90 degrees will be the negative of the Cosine of an angle that number of degrees above 90 degrees. A more mathematical way of expressing this is:

cos (90 − z) = −cos (90 + z)

Diagonal symmetry

In a right-angled triangle and in the circle, we can see that the Sine of 45 degrees is the same as the Cosine of 45 degrees – both are 0.7071 units. Such a triangle has the same length opposite side as adjacent side.

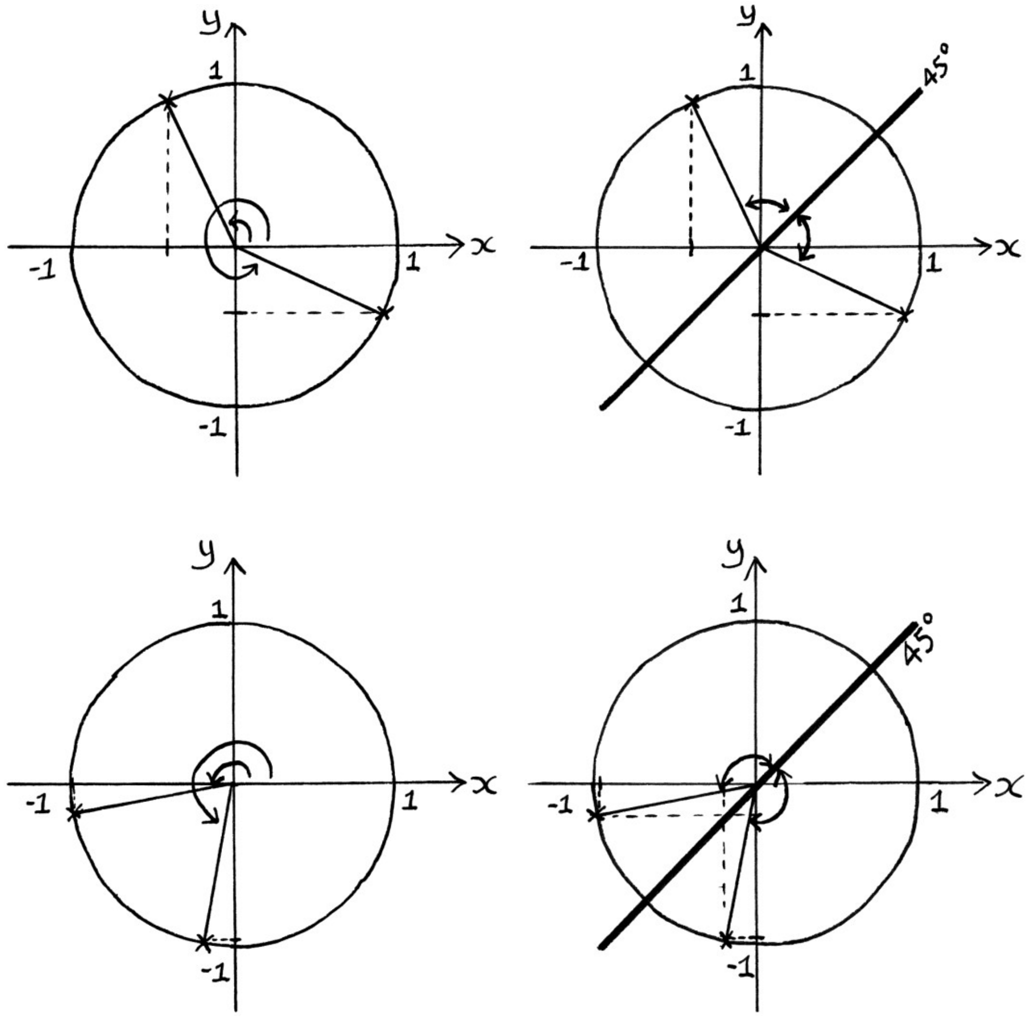


It is also true that the Sine of 46 is equal to the Cosine of 44, and the Cosine of 46 is equal to the Sine of 44. More equivalences include:
Sine 47 = Cosine 43; Cosine 47 = Sine 43
Sine 48 = Cosine 42; Cosine 48 = Sine 42
Sine 49 = Cosine 41; Cosine 49 = Sine 41
... and so on.

In fact, the Sine of any angle that is a particular number of degrees *above* 45 degrees will be the same as the Cosine of an angle that number of degrees *below* 45 degrees. This means that the Sine of any angle that is a particular number of degrees *below* 45 degrees will be the same as the Cosine of an angle that number of degrees *above* 45 degrees.

Saying exactly the same thing with different words: the Cosine of any angle a particular number of degrees above or below 45 degrees will be the same as the Sine of an angle that number of degrees below or above 45 degrees.

In other words, Sine and Cosine are equal to each other when the points on the circle's edge are reflected diagonally on the circle.



A mathematical way of saying this is:

sin (45 − z) = cos (45 + z)
cos (45 − z) = sin (45 + z)

Strictly speaking, this is exactly the same as the observation that the Sine of a value is equal to the Cosine of that value subtracted from 90 degrees, and vice versa. Subtracting an angle from 90 degrees results in an angle that is the "diagonal mirror image" of the original angle. As an example, we will think about the angle 23.5 degrees. We will do this calculation 90 − 23.5 = 66.5 degrees. The angles 23.5 and 66.5 are both exactly 21.5 degrees either side of 45 degrees. The Sine of 66.5 degrees is the same as the Cosine of 23.5 degrees; the Cosine of 66.5 degrees is the same as the Sine of 23.5 degrees.

As a second example, we will think about the very large negative angle of −15,246.12 degrees in the same way. We do this calculation: 90 − −15,246.12 = 15,336.12. Although it is hard to tell easily, −15,246.12 and 15,336.12 are both equidistant from 45 degrees. The Sine of −15,246.12 and the Cosine of 15,336.12 are the same as each other; the Cosine of −15,246.12 and the Sine of 15,336.12 are the same as each other.

Yet another Sine and Cosine equivalence

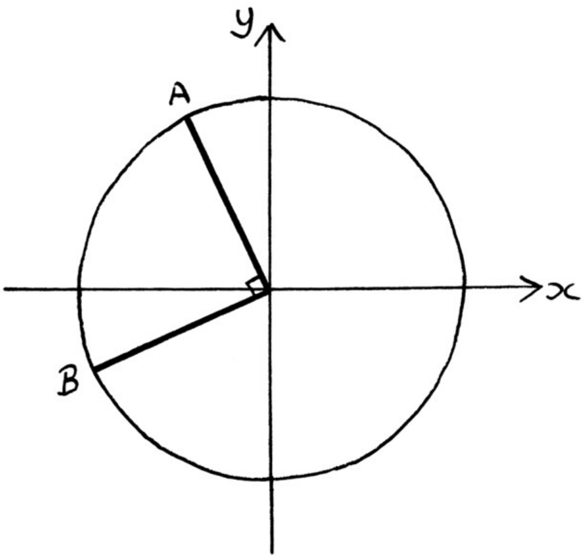
Another equivalence between Sine and Cosine is that the Sine of an angle is the same as the Cosine of that angle minus 90, and that the Cosine of an angle is the same as the Sine of that angle added to 90. Note that this is different from the idea that "$\sin\theta = \cos(90 - \theta)$" and "$\cos\theta = \sin(90 - \theta)$". In this case, the mathematical way of expressing the idea is as follows:

$\sin\theta = \cos(\theta - 90)$
$\cos\theta = \sin(\theta + 90)$

As an example, the Sine of 123.45 degrees is the same as the Cosine of 123.45 – 90 = 33.45 degrees. They are both 0.8344 units.

The Cosine of −11.28 degrees is the same as the Sine of −11.28 + 90 = 78.72 degrees. Both are 0.9807 units.

This is more obvious when looking at the circle and angles that are 90 degrees apart. For any two angles that are 90 degrees apart, the Cosine of the smaller angle will be the same as the Sine of the larger angle. Here, the Cosine of the angle for A will be equal to the Sine of the angle for B.

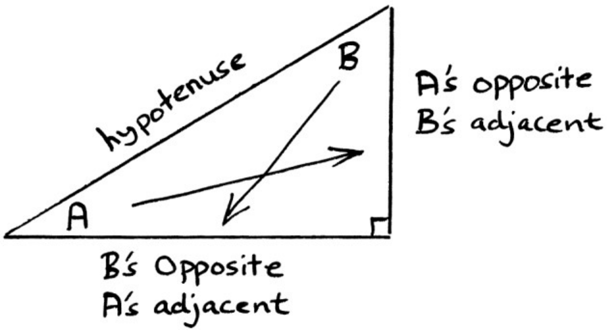


This becomes clearer still if we imagine the thick black right-angle in the picture moving around the circle.

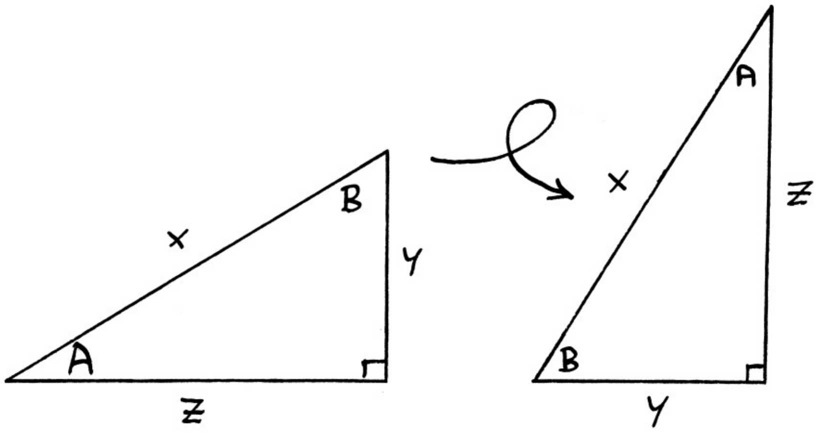
The most obvious example of the equivalence is when we look at the Cosine of 0 degrees and the Sine of 90 degrees.

Sine and Cosine together

Once we understand what Sine and Cosine are, one useful idea is that Sine and Cosine are *the same thing* seen from different perspectives. This is most obvious when we realise that the "opposite side" with reference to one angle in a right-angled triangle, is also the "adjacent side" with reference to the other angle. In other words, one angle's opposite side is the other angle's adjacent side.

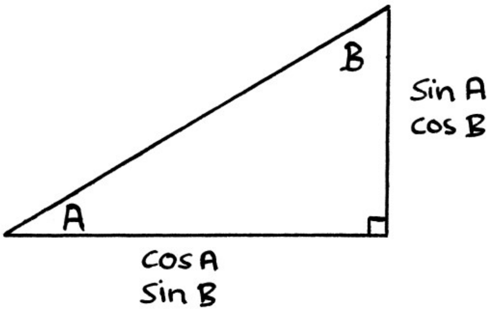


Generally, the direction of a line out from the origin is important, and the direction of the angle from the horizontal is important. However, if we were to ignore that, and concentrate only on the absolute sizes of the sides of a triangle with no reference to their being positive or negative, we could flip a triangle around as so:



In the first triangle, the length of side Y is Sine A. The length of side Z is Cosine A. In the flipped triangle, the length of side Y is Cosine B. The length of side Z is Sine B.

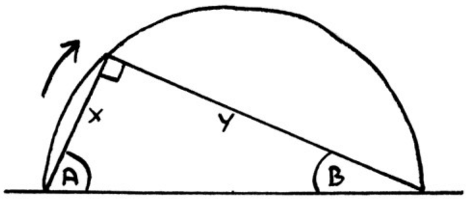
If we are ignoring direction in the sides of a triangle, the length of side Y is equal to both Sine A or Cosine B. The absolute values of Sine A and Cosine B are the same. Similarly, the length of side Z is both Cosine A and Sine B – their absolute values are the same thing.



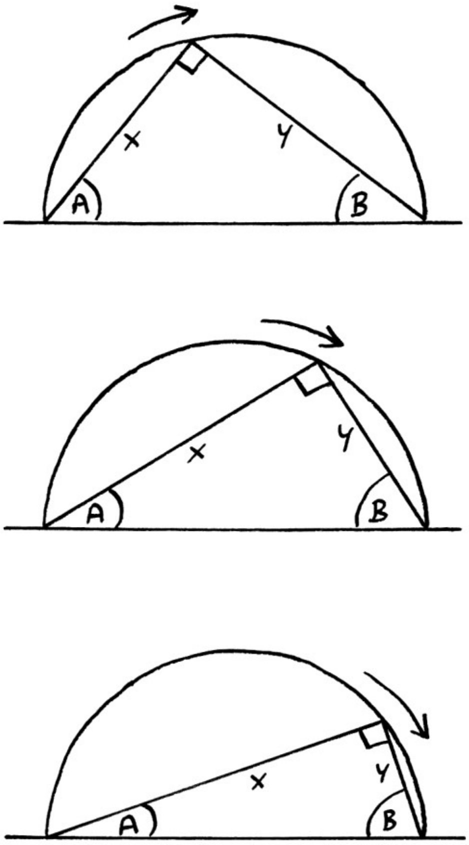
In nearly all usage of Sine and Cosine, the direction of the lines from the angle of interest, and whether the values are positive or negative, are vitally important. Therefore, we should not flip triangles around in this way, and we should pay attention to the direction of lines and whether the results of Sine and Cosine are positive or negative.

Another way of visualising Sine and Cosine's similarity

One way of clearly seeing how Sine and Cosine are the same thing but seen from different points of view is to imagine a right-angled triangle lying on its hypotenuse side, put in a half circle like this:



We then imagine the right-angled corner moving back and forth along the edge of the semicircle. As the right-angled corner moves clockwise around the semicircle, angle A decreases and angle B increases. As this happens, the length of side Y decreases, and the length of side X increases.



As one angle increases, the other angle decreases. As one side increases, the other decreases. This is the same thing as saying as the Sine of one angle increases, the Sine of the other angle decreases, or as the Cosine of one angle decreases, the Cosine of the other angle increases. There is no difference in how the angles and sides change as the corner of the triangle moves from one side to the other.

Sine and Cosine summary

The most important idea in this chapter is that the Sine function, the Cosine function, right-angled triangles and circles are completely linked. Sine gives the y-axis value of the point where a line at a particular angle meets the edge of a unit-radius circle. Cosine gives the x-axis value. The angle of the line and the y-axis and x-axis values also describe a right-angled triangle. On a calculator, the Sine and Cosine buttons are essentially "look-up tables" for the countless measurements possible on a circle, but they give results that are more accurate than those obtained by drawing a circle on a piece of paper. We enter an angle (or a value that is being treated as an angle) and the calculator will give us the result.

Conclusion

Sine and Cosine are the basis for understanding waves, so it is important to understand what they are and how they work.

The basis for everything in this chapter can be deduced by drawing triangles and circles, and observing their characteristics. If you have a protractor and a ruler, you could have figured out everything here yourself. Obviously, the generally used names for everything cannot be deduced, as they are arbitrary and vary from country to country and from language to language. Similarly, the use of an angle system based on 360 degrees, and which way around to measure angles, are social constructions. However, the way that the aspects of a triangle or a circle relate to each other is a universal fact of reality. An alien at the far end of the universe could come up with the same deductions about triangles and circles.

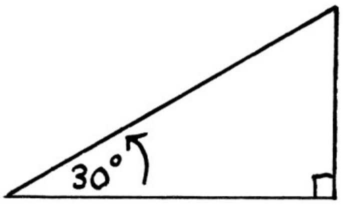
The path to understanding and remembering the information in this chapter, and in fact every chapter in this book, is to become involved with the subject. If you do not understand something in this chapter, or even if you do, find a pencil, ruler and protractor, do some drawing and measuring, and see how angles, triangles and circles relate to each other. If you want to progress, try to think about circles, triangles, and Sine and Cosine in your everyday life.

Chapter 2: More about triangles

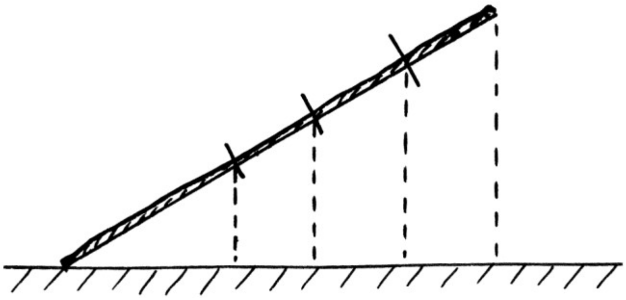
In this chapter, we will look at more concepts relating to Sine and Cosine that will be useful in the future.

The shape of triangles

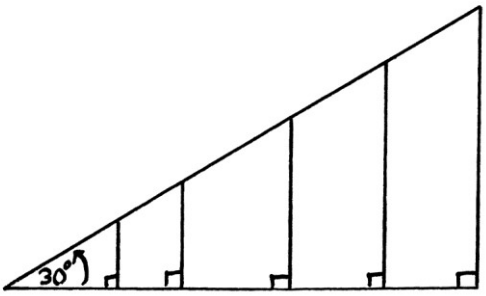
So far, we have looked at triangles with a unit long hypotenuse, and circles with a unit long radius. All right-angled triangles with hypotenuses at a particular angle, no matter what length they are, will be of an identical shape, and therefore scaled versions of each other. In other words, the length of a hypotenuse at a particular angle does not affect the shape of the triangle. For example, if we have some right-angled triangles with hypotenuses of varying lengths, but all with an angle of 30 degrees, they will all look like a scaled version of this:



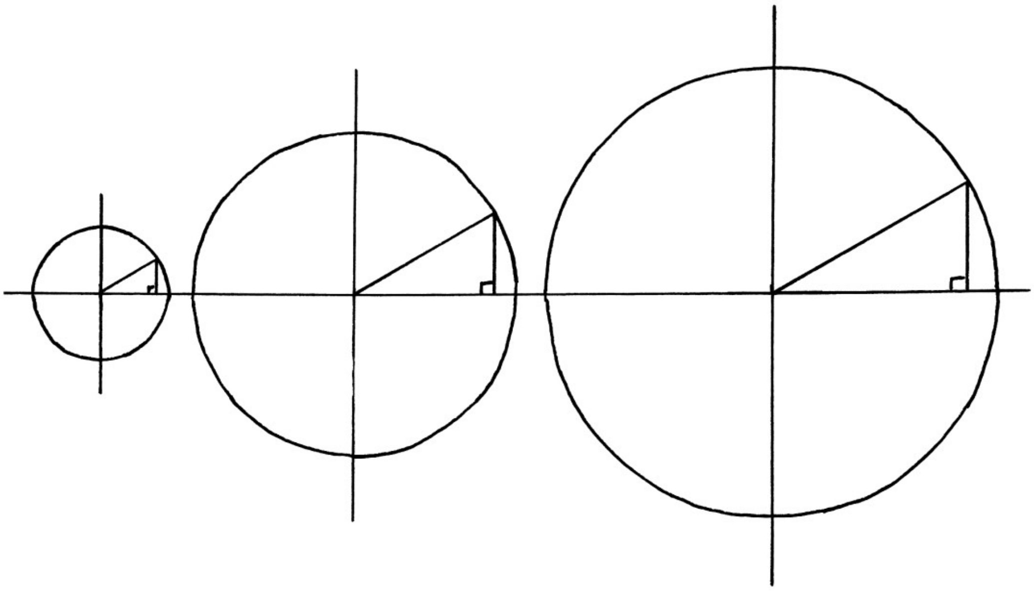
There are countless ways to know that this is true. One way is by imagining a stick placed in the ground at a particular angle. No matter how much we cut off the end of the stick, the triangle formed by its start and end will always have the same shape:



A very similar way is to overlay several triangles with different length hypotenuses, but all with the same angle:



Yet another way of understanding the common shape of equally angled triangles is by thinking of the circle they would fit in. However big the circle is, a triangle formed by drawing a line from the origin out to the edge of the circle at a particular angle will always have the same shape.



Yet another way of thinking about this is by realising that the unit for measurement is completely arbitrary. The shape of the triangle is independent of whether we say the hypotenuse is 1 metre or 1 kilometre or 1 anything. If the hypotenuse is 1.34232 metres, we could create a new measurement system where 1 of our invented units is equal to 1.34232 metres. In such a case, the hypotenuse would be equal to one of our invented units.

The main thing to understand and remember is that all right-angled triangles with the same angle have the same shape.

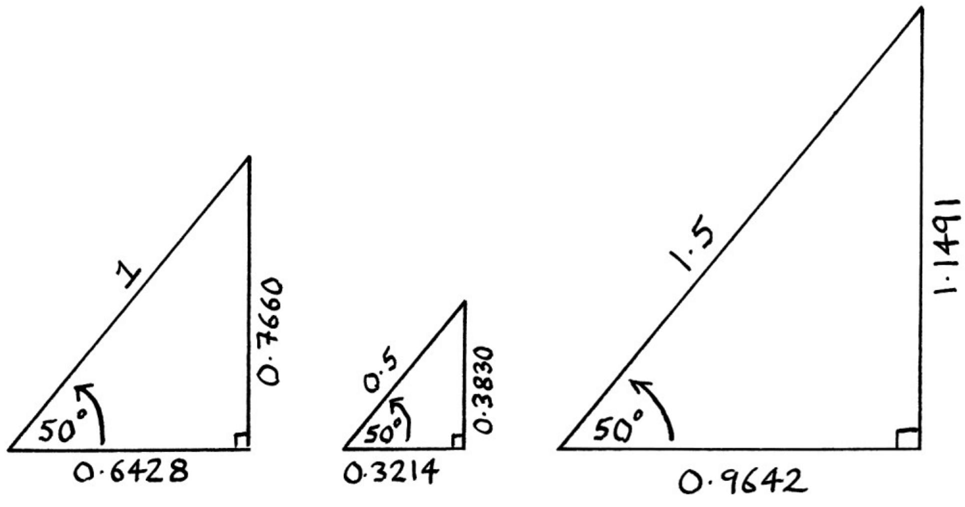
Proportions

A 30-degree right-angled triangle with a unit long hypotenuse has an opposite side of 0.5 units, and an adjacent side of 0.8660 units. The same shaped triangle (that is, one still with a 30 degree angle), but with a hypotenuse of 2 units, will have an opposite side of 1 unit and an adjacent side of 1.7321 units. All the sides of the 2-unit hypotenuse triangle are double those of the 1-unit hypotenuse triangle.

It is actually the case that if the hypotenuse of *any* right-angled triangle is doubled, the opposite and adjacent sides will be doubled too. In fact, if the hypotenuse is scaled by any value, then the opposite and adjacent sides will also be scaled by exactly the same value. Everything is scaled in proportion. If a triangle has a hypotenuse of half a unit, the opposite and adjacent sides will be half the length that they would be if the triangle had a hypotenuse of one unit.

For any set of right-angled triangles with a particular angle, they will all have the same shape, and the size of the opposite and adjacent sides will be scaled to the same amount as the length of the hypotenuse is scaled to the number 1.

The benefit of knowing this is that if we want to calculate the length of the opposite side of a right-angled triangle that has a hypotenuse other than 1, we calculate the opposite side as if the hypotenuse were 1, and then multiply it by the length of the actual hypotenuse to scale it up or down. The same goes for the adjacent side – we calculate the adjacent side as if the hypotenuse were 1, and then multiply it by the length of the actual hypotenuse to scale it up or down.



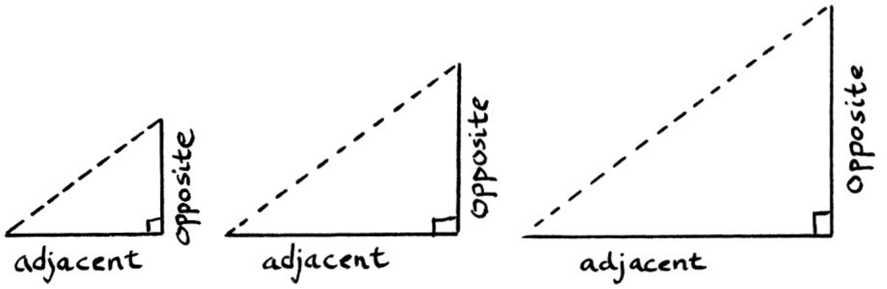
In school education, this idea is usually summarised with the standard formulas:
opposite side = hypotenuse * Sine θ
... and:
adjacent side = hypotenuse * Cosine θ

These are often rearranged to be:
Sine θ = opposite ÷ hypotenuse
... and:
Cosine θ = adjacent ÷ hypotenuse

The formulas are useful, but it is better to understand how right-angled triangles work first. Understanding right-angled triangles will make the formulas intuitive, so no effort will be needed to remember them.

Gradient

As we have seen, a right-angled triangle with the same angle will *always* have the same shape. If it has the same shape, it will *always* have the same ratio between any two of the three sides – in other words, one particular side divided by a second particular side will always result in the same number for any size of triangle of a particular shape. Of interest in this section is the ratio between the opposite and adjacent sides. The ratio between the opposite side and adjacent side of a right-angled triangle will be identical for every right-angled triangle of the same shape.



One way to know this is true is by how the sides of triangles scale together – if we draw a right-angled triangle with a 2 unit long hypotenuse, the opposite and adjacent sides will have twice the length they would have if the hypotenuse were 1 unit long. If the hypotenuse is 0.5 units long, the adjacent and opposite sides will have half the length that they would have if the hypotenuse were 1 unit long. The result of the opposite side divided by the adjacent side will be the same no matter how the sides are scaled because the scaling factor becomes cancelled out in the calculation.

For example, this:

$$\frac{2 * \text{"}length\ of\ opposite\ side\ when\ hypotenuse\ is\ 1\text{"}}{2 * \text{"}length\ of\ adjacent\ side\ when\ hypotenuse\ is\ 1\text{"}}$$

... becomes this:

$$\frac{\text{"}length\ of\ opposite\ side\ when\ hypotenuse\ is\ 1\text{"}}{\text{"}length\ of\ adjacent\ side\ when\ hypotenuse\ is\ 1\text{"}}$$

This:

$$\frac{0.5 * \text{"}length\ of\ opposite\ side\ when\ hypotenuse\ is\ 1\text{"}}{0.5 * \text{"}length\ of\ adjacent\ side\ when\ hypotenuse\ is\ 1\text{"}}$$

... becomes this as well:

$$\frac{\text{"}length\ of\ opposite\ side\ when\ hypotenuse\ is\ 1\text{"}}{\text{"}length\ of\ adjacent\ side\ when\ hypotenuse\ is\ 1\text{"}}$$

... and this:

$$\frac{568.456 * \text{"}length\ of\ opposite\ side\ when\ hypotenuse\ is\ 1\text{"}}{568.456 * \text{"}length\ of\ adjacent\ side\ when\ hypotenuse\ is\ 1\text{"}}$$

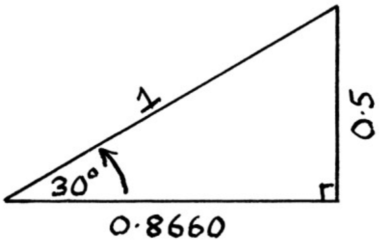
... also becomes this:

$$\frac{\text{"}length\ of\ opposite\ side\ when\ hypotenuse\ is\ 1\text{"}}{\text{"}length\ of\ adjacent\ side\ when\ hypotenuse\ is\ 1\text{"}}$$

The scaling factor is cancelled out each time.

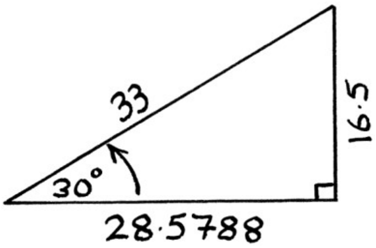
The length of the opposite side divided by the length of the adjacent side gives us what is called the “gradient” or “slope” of the hypotenuse side. [The term “gradient” is used more commonly in the United Kingdom among other places, while “slope” is used more commonly in the United States]. Understanding the concept of gradients can be useful for dealing with waves. Gradient can be thought of as being similar to an angle, in that the gradient also indicates the steepness of the hypotenuse, but instead of degrees, a gradient is a number between zero and infinity. A gradient of zero would be a horizontal line, a gradient of 1 would be an angle of 45 degrees, and an infinite gradient would be a vertical line, that is to say a line at an angle of 90 degrees.

Given the gradient, it is possible to know the exact shape of the triangle (but not its size), which in turn lets us calculate the angle. Every angle produces a hypotenuse with a particular gradient, and every gradient implies a particular angle that set it.
A triangle with a 30-degree angle and a hypotenuse of 1 unit would have an opposite side of 0.5 units (Sine 30), and an adjacent side of 0.8660 units (Cosine 30):



Its gradient would be 0.5 ÷ 0.8660 = 0.5774.

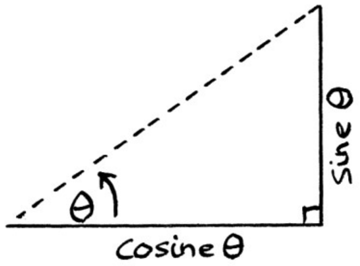
A triangle with the *same* angle (30 degrees) and a hypotenuse of 33 units would have an opposite side of 16.5 units (33 * Sine 30), and an adjacent side of 28.5788 units (33 * Cosine 30).



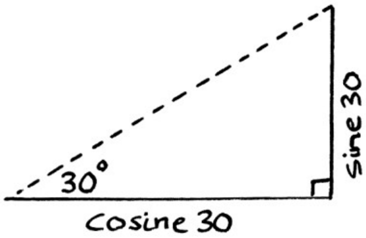
Its gradient would be 16.5 ÷ 28.5788 units = 0.5774, which is the same as before. The fact is that all right-angled triangles of the same shape have the same gradient, and therefore, *all* 30-degree right-angled triangles have a gradient of 0.5774. This should be apparent because all right-angled triangles with an angle of 30 degrees will have the same shape, and if they have the same shape, then the ratio of one particular side to another will always be the same.

Given that the shape of a right-angled triangle with a particular angle is always the same, and that the *size* of the triangle is actually irrelevant, we do not need to know the length of the hypotenuse to calculate the gradient. We just need the length of the adjacent and opposite sides, or more accurately, the length of the adjacent and opposite sides of *any* triangle with that shape. Actually, we need to know only the ratio – the opposite side divided by the adjacent side. We can work out the ratio using just the angle, and by treating the triangle as if the hypotenuse were 1 unit long.

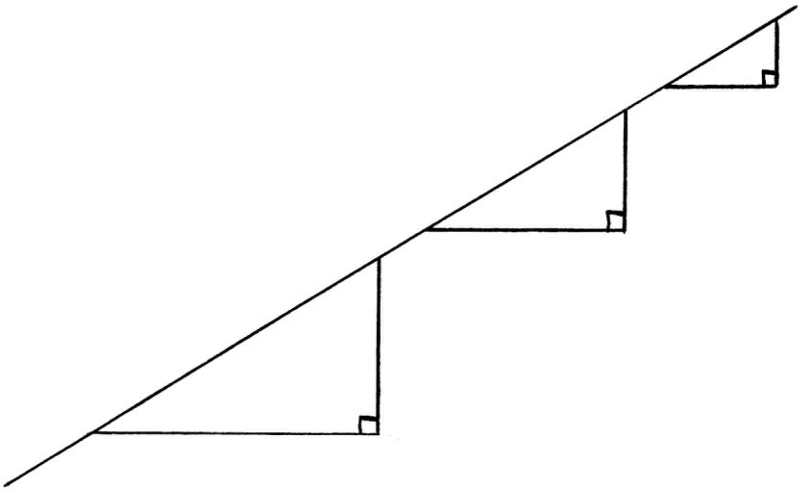
The ratio will be the Sine of the angle divided by the Cosine of the angle:
sin θ ÷ cos θ.



A right-angled triangle with an angle of 30 degrees might have any size whatsoever – it might have a hypotenuse of a centimetre, a metre, or even a kilometre, but its shape will always be the same, and therefore the ratios of its sides will always be equal to sin 30 ÷ cos 30:

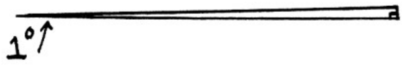


One way to be sure that it is the shape of the triangle that dictates the gradient, rather than the actual size of the triangle is to draw a line at a particular gradient, and then draw triangles of various sizes against it. Every triangle is a different size, but they all have the same shape. We know every hypotenuse is at the same angle because they are all along the length of the line. Therefore, each hypotenuse has the same gradient because their gradients must be the same as the line:

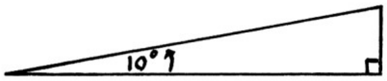


Some examples of gradients and angles

A right-angled triangle with a 1-degree angle has a gradient of:
sin 1 ÷ cos 1 = 0.01746.

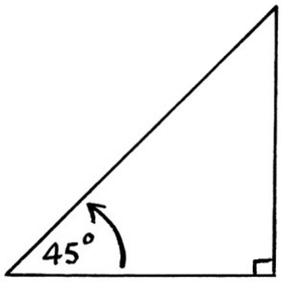


A right-angled triangle with a 10-degree angle has a gradient of:
sin 10 ÷ cos 10 = 0.1763.

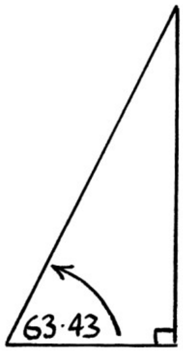


A triangle with a 45-degree angle has a gradient of:
sin 45 ÷ cos 45 = 1.

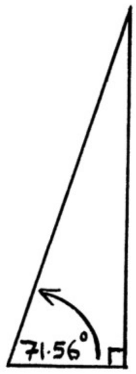
Another way of thinking about this is that the adjacent and opposite sides are the same length so one divided by the other will result in the number 1.



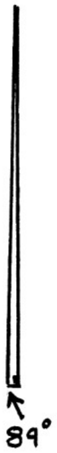
The hypotenuse of a right-angled triangle with an angle of 63.43 degrees has a gradient of: sin 63.43 ÷ cos 63.43 = 1.9996. In other words, the opposite side is approximately twice as long as the adjacent side.



A hypotenuse at an angle of 71.56 degrees has a gradient of approximately 3. The opposite side is roughly three times as long as the adjacent side.



A hypotenuse at an angle of 89 degrees has a gradient of 57.2899.



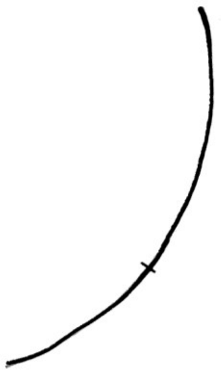
A hypotenuse at an angle of 89.9999999999 (ten nines after the decimal point) degrees has a gradient of 573,000,000,000. A hypotenuse at an angle of 0.0001 degrees has a gradient of 0.000001745.

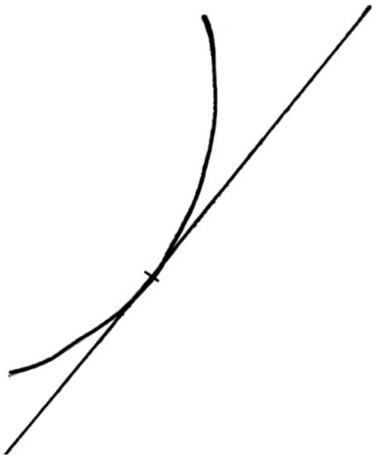
From these examples, we can see that low angles give low gradients, and high angles give *extremely* high gradients. In fact, there is no limit to how high a gradient can be. If a triangle could have a 90-degree angle, its gradient would be infinitely high. As the angle of a right-angled triangle increases, the adjacent side becomes smaller and the opposite side becomes larger. This makes the division to calculate the gradient consist of a larger number divided by a smaller number. As the adjacent side becomes closer to zero, the gradient becomes closer to infinite.

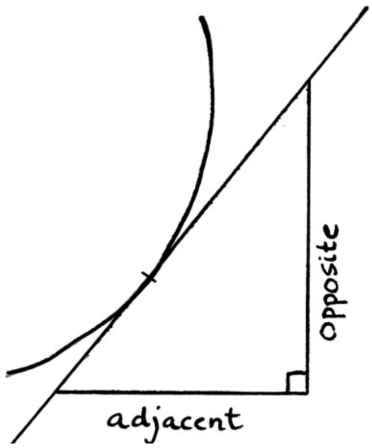
If a triangle could have a zero degree angle, its gradient would be zero. One can see that this is true because it would have a zero length opposite side, and dividing zero into any number of pieces still results in zero.

Generally, the formula “sin θ ÷ cos θ” is treated as one entity: “Tangent θ”. Tangent θ is shorthand for “sin θ ÷ cos θ”. If we have a right-angled triangle with a particular angle, then the Tangent of that angle is the ratio between the opposite and adjacent sides. In other words, the Tangent of the angle is the length of the opposite side divided by the length of the adjacent side. The Tangent button on a calculator just calculates the Sine of a number divided by the Cosine of that number. If the Tangent button on a calculator breaks, we can just use the Sine and Cosine buttons. In a way, having these extra names and buttons is useful – it saves a lot of time and pressing – the downside is that it can make us forget what is actually happening and how things are related to each other.

The word “Tangent” is often shortened to “Tan”. The word “Tangent” is ultimately derived from the Latin word “tango” which means “I touch”, “I border on” or “I am immediately next to”. In the context of gradients, the word “Tangent” refers to the way that the gradient of a point on a *curved* line can be calculated by placing a straight line tangentially to the line – that is touching it, or next to it. That straight line can then be treated as the hypotenuse of a right-angled triangle of a particular shape (the size is not important), of which the opposite and adjacent side can be used to calculate the gradient. In this way, it is possible to calculate the gradients of points on curves.







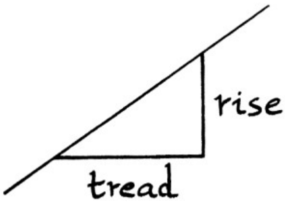
Remembering how to calculate the gradient

With a right-angled triangle, you need to remember that the gradient is calculated by the length of the opposite side divided by the length of the adjacent side. Repeatedly saying the phrase, “opposite over adjacent” will help you to remember the order of the division.

$$\frac{opposite}{adjacent}$$

When it comes to a line that is not part of a triangle, we can place the hypotenuse of a right-angled triangle next to it, and then divide the opposite side by the adjacent side to obtain the line’s gradient. For this, remembering the phrase “opposite over adjacent” is still useful.

Another way to think of the calculation of a line’s gradient is that it is the vertical change of any section of the line divided by the horizontal change for that section. This is straightforward, but sometimes it can be difficult to remember which way around to perform the division. A quick way of remembering the order is to imagine the line lying on a staircase. The calculation is “the rise” divided by “the tread”, or “rise over tread” [where a “tread” is the part of a step that one treads on, and the “rise” or “riser” is the vertical part that connects the treads].

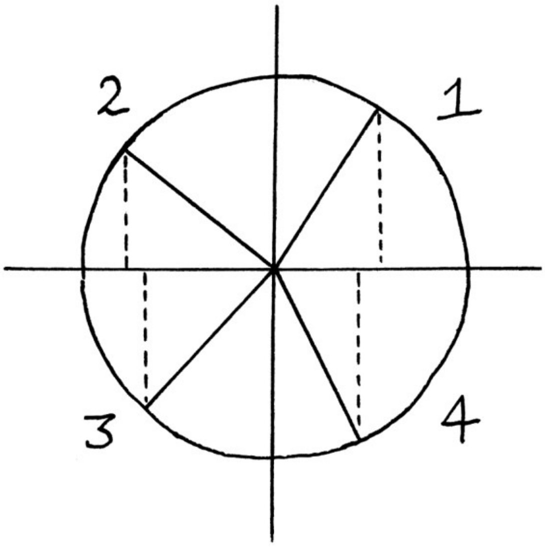


$$\frac{rise}{tread}$$

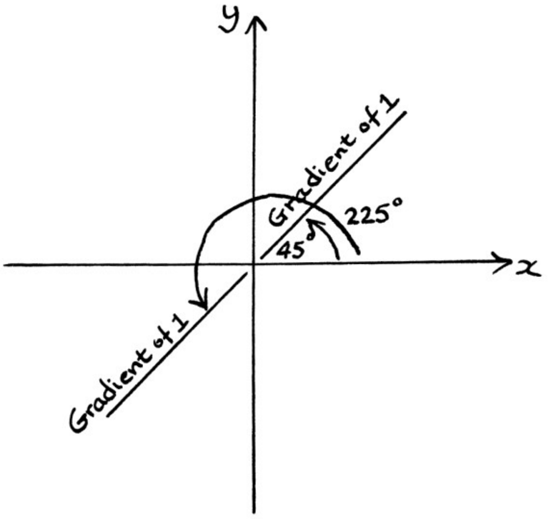
If you say the phrase “rise over tread” enough times, you will always remember which way around to perform the division.

More on gradients

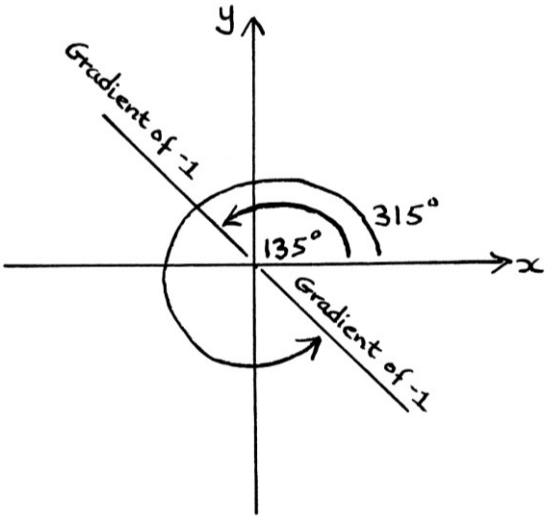
So far, we have looked at the gradients of right-angled triangles that are the “correct” way round – in other words, the opposite side is on the right. This allows only for angles from just over 0 to just under 90 degrees. For other angles, it is necessary to use a circle to demonstrate gradients. If we divide the circle into 4 quarters or quadrants, we can see how the gradients of the four types of triangle look. Generally, the quarters are numbered anticlockwise to match how angles increase anticlockwise.



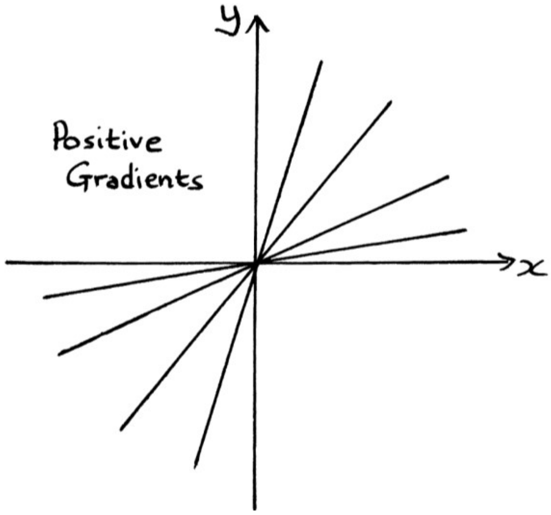
One thing to know about gradients is that they are either positive (when the slope goes up to the right and down to the left), or they are negative (when the slope goes up to the left and down to the right). Although it would be possible to distinguish four types of gradients by indicating their direction, this is not generally done. This means that any particular angle and an angle 180 degrees higher or lower will both produce lines that have the same gradient. For example, lines drawn at 45 degrees and 225 degrees will both have gradients of 1.



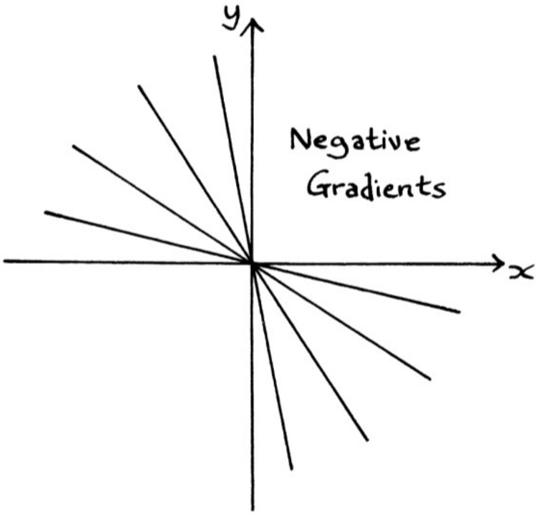
Lines drawn at 135 degrees and 315 degrees will both have gradients of −1.



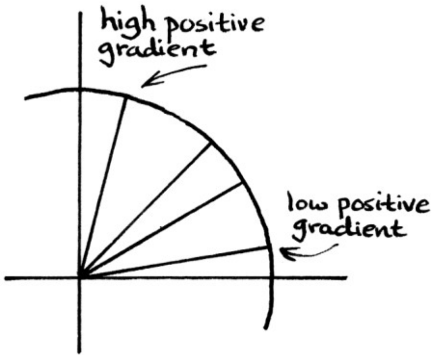
In the following picture, all the lines have positive gradients:



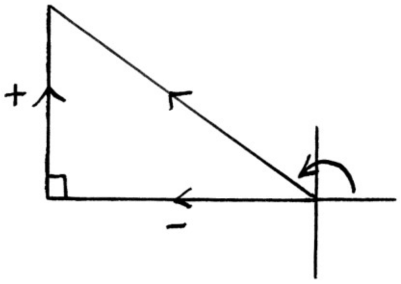
In the following picture, all the lines have negative gradients:



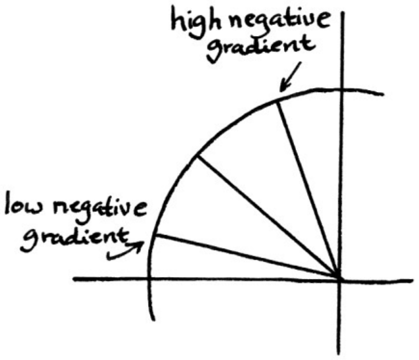
If we go through each quadrant of the circle, we can see how the gradients will appear. In Quadrant 1, for angles from 0 to 90 degrees, as discussed already, the lines (or the hypotenuses of the right-angled triangles) will have gradients from 0 to infinity.



In Quadrant 2, for angles from just over 90 degrees up to 180 degrees, the adjacent side of any right-angled triangle would be negative, while the opposite side would be positive. Therefore, the result of opposite ÷ adjacent would be negative, and the gradient would be negative.



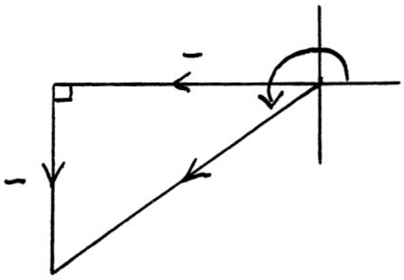
A negative gradient just means that the hypotenuse slopes from top left to bottom right, or from bottom right to top left, depending on how you want to phrase it. Quadrant 2's gradients go from just under negative infinity down through the negative numbers to zero, as the angle increases.



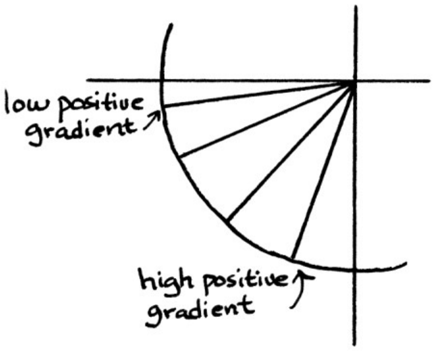
[Note that positive infinity and negative infinity are theoretical concepts used to portray the maximum and minimum values possible. They cannot exist in nature, but the concepts are useful as tools in maths. Strictly speaking, infinity and negative infinity are fictitious values that can be approached but never reached. In a sense, they are not really numbers, but ideas. Given that, many people would argue that a number divided by zero is undefined, rather than infinity. In this book, to keep things simple, I am not as pedantic about the meaning of infinity as I could be.]

From the point of view of *gradients*, a gradient of positive infinity and a gradient of negative infinity are the same thing – they both represent a vertical line. If we were distinguishing direction in gradients, then they would be opposites, but we are not, so they are not. We could just as easily say that the gradients from Quadrant 2 go from positive infinity, down through negative gradients, and down to zero, but that can be harder to comprehend.

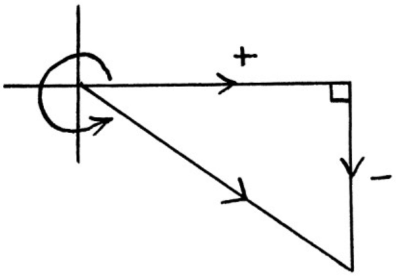
In Quadrant 3, for angles from just over 180 up to 270 degrees, the adjacent side of any right-angled triangle is negative, and the opposite side is also negative. This means that the opposite side divided by the adjacent side will be *positive.*



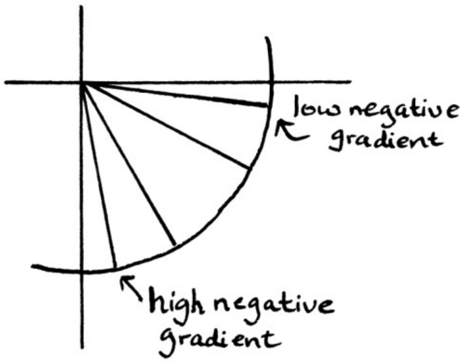
Although a triangle would be pointing downwards, the gradient is still going from top right down to bottom left (or bottom left up to top right – they mean the same thing when dealing with gradients). Quadrant 3's gradients go from zero to infinity as the angle increases.



In Quadrant 4, for angles from just over 270 degrees up to 360 degrees, the adjacent side of any right-angled triangle is positive, and the opposite side is negative. Therefore, the gradient will be negative. Such a triangle would be pointing to the right and downwards.



Quadrant 4's gradients go from negative infinity up through the negative numbers to zero as the angle increases. We could just as well say that the gradients go from *positive* infinity, through the negative gradients up to zero, but that is harder to visualise.



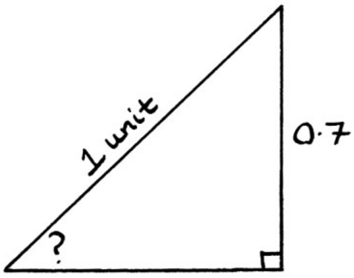
Inverse Sine, inverse Cosine, inverse Tangent

As we have seen, given the angle of a unit-hypotenuse right-angled triangle, the Sine function finds the length of the opposite side, the Cosine function finds the length of the adjacent side, and the Tangent function finds the result of the opposite side divided by the adjacent side (the gradient). However, if we only have the length of the opposite side, the length of the adjacent side, or the gradient, and we wanted to find the angle, then we would need the inverse versions of the Sine, Cosine and Tangent functions.

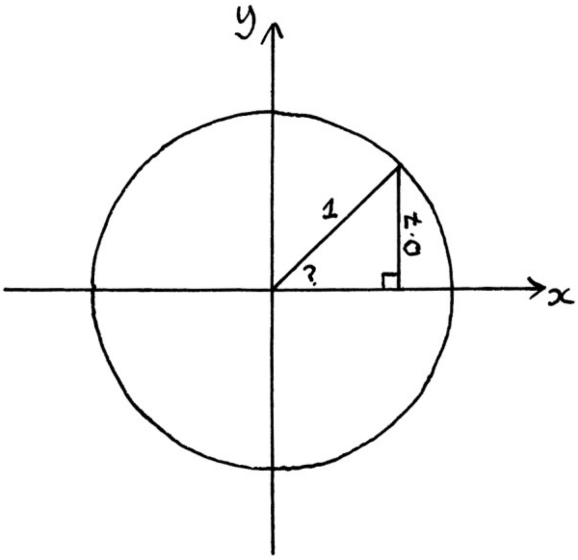
The simplest way to calculate all of these is to draw the triangle in question on a piece of paper and measure the angle with a protractor, but for more accuracy, it makes sense to use a calculator or a computer.

Inverse Sine

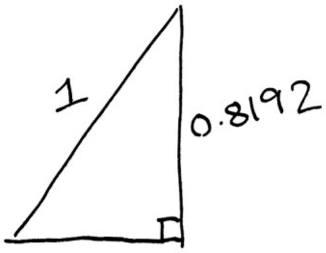
The inverse Sine function finds the angle of a right-angled triangle with a unit long hypotenuse, given just the length of the opposite side.



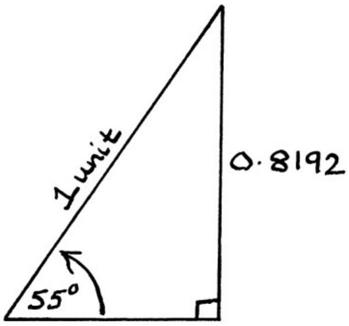
To put it another way, the inverse Sine function finds the angle of a line drawn from the centre of a unit-radius circle to the circumference, given the y-axis height of that point on a circle.



As an example, we will say we have been given a rough drawing of a triangle, and told that it has a unit long hypotenuse and an opposite side of 0.8192 units. We need to find the angle.

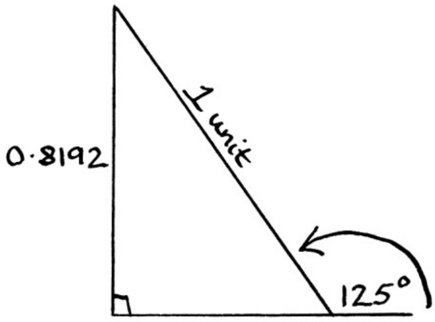


One way to find the angle is to make an accurate drawing, and to measure the angle. We will probably not be able to draw a triangle to a level of accuracy that warrants having 4 decimal places in the length of the opposite side, but we should be able to find out that the angle is around 55 degrees.

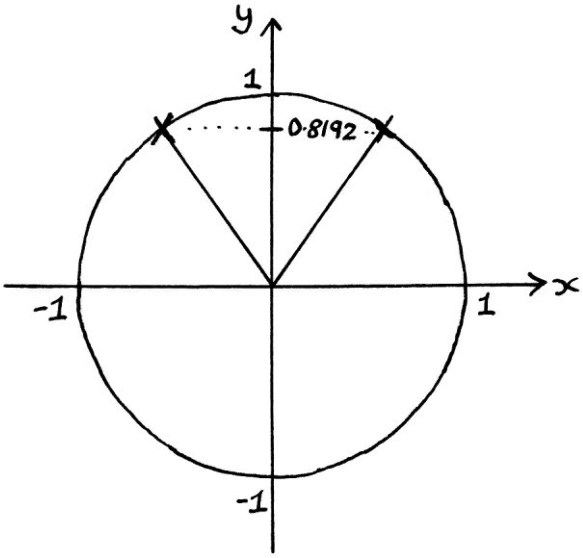


The second way to find the angle is to type 0.8192 into a calculator and press the “inverse Sine” button. It will give the same result.

For the particular triangle in that example, there is only one answer. However, it pays to be aware that a triangle drawn “the wrong way around” would also have the same height, yet would have an angle of 125 degrees:



That inverse Sine has two potential results is easier to visualise when thinking about circles. On a circle, we can see that there are two points on its edge where the height is 0.8192 units.



From the circle, it is easier to see that the inverse Sine of 0.8192 is both 55 degrees and 125 degrees. As mentioned in the last chapter, the Sine of an angle so many degrees below 90 will be the same as the Sine of an angle that number of degrees above 90.

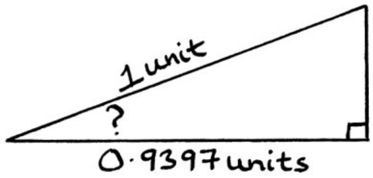
Technically, the inverse Sine of a number has countless solutions because the Sine of countless angles results in the same answer. Depending on how many times we go around the circle, we can see that the Sine of −305 (which is 55 − 360), −235 (or 125 − 360), 55, 125, 415 (or 55 + 360), 485 (or 125 + 360), 775 (or 55 + 720), 845 (or 125 + 720), and so on, all result in 0.8192. Generally, it is sufficient to be aware of the two positive angles under 360 degrees, and usually when dealing with an individual triangle, we will only need to know the result under 90 degrees. A calculator will usually give only the positive result closest to zero. All of this means that whenever we calculate the inverse Sine on a calculator, we need to think about the circle to know if we might need the other result instead.

The inverse Sine function has several alternative names that all mean the same thing: inverse Sine, inverse Sin, inv sin, arcsine, arcsin, $\sin^{-1}$, asin. You may see any of these being used in explanations. On a calculator, “arcsin” is probably the most common. The term “$\sin^{-1}$” looks as if it refers to some other mathematical calculation, but it is just shorthand for inverse Sine.

Inverse Cosine

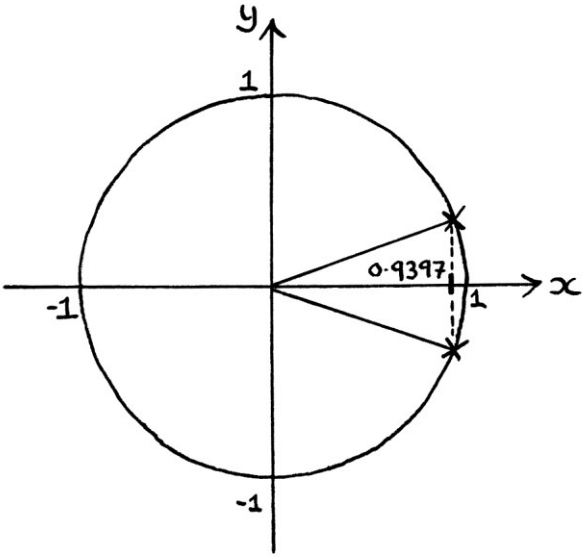
Inverse Cosine is similar to inverse Sine. It finds the angle in a unit-long hypotenuse right-angled triangle when all we have is the length of the adjacent side.

If, for example, we wanted to find the angle in a unit hypotenuse right-angle triangle with an adjacent side of 0.9397 units, we could achieve a reasonably good result by drawing the triangle and measuring the angle.



For a more accurate result, we could enter 0.9397 into a calculator and press the “inverse Cosine” button. Both methods would result in an angle of 20 degrees (if we round up the result).

If we are dealing with circles, or with triangles that are not necessarily the “right way round”, then, as with inverse Sine, there are two results to the inverse Cosine of a number. This is because there are two points on a circle where the horizontal distance from the origin has the same value. As mentioned in the last chapter, the Cosine of an angle so many degrees below 180 will be the same as the Cosine of an angle that number of degrees above 180. Therefore, in this case, the Cosine of 20 degrees will be the same as the Cosine of 340 degrees:



This means that the inverse Cosine of 0.9397 is both 20 degrees and 340 degrees.

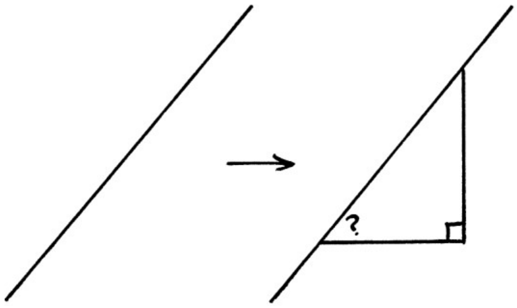
Again, if we think of angles that are more than one revolution around the circle, there are really countless possible results because the Cosine of a value added to 360 or multiples of 360, or subtracted from 360 or multiples of 360, will have the same result. For example, the Cosine of −340 (which is 20 − 360), −20 (which is 340 − 360), 20, 340, 380 (or 20 + 360), 700 (or 340 + 360), 740 (20 + 720), 1060 (340 + 720), and so on, all result in the value 0.9397. With inverse Cosine, it is sensible only to consider the two positive results less than 360 degrees, while being aware that the others exist. When calculating the inverse Cosine, a calculator will usually only give the lowest possible positive result.

The inverse Cosine function has several names that mean the same thing: inverse Cosine, inverse cos, inv cos, arccosine, arccos, $\cos^{-1}$, acos. On a calculator, it is frequently called "arccos".

Inverse Tangent

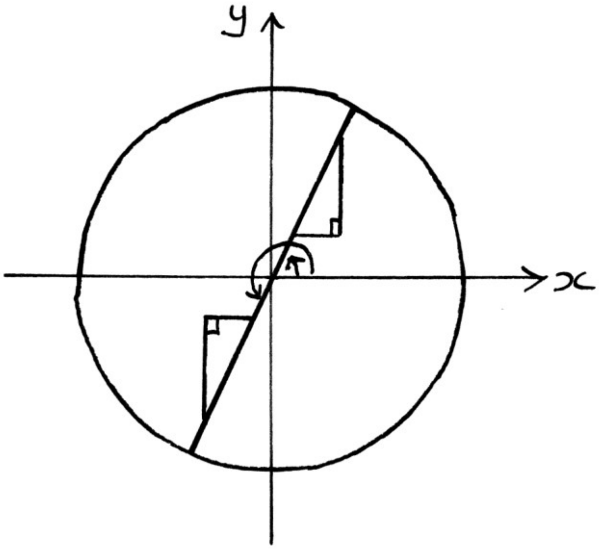
Inverse Tangent is, of course, the inverse of the Tangent function. Inverse Tangent finds the angle of a right-angled triangle based on the gradient. To put this another way, it finds the angle of a line with a particular gradient going through the origin of a circle. To put this yet another way, it finds the angle of a right-angled triangle based on the ratio of the opposite side to the adjacent side. Inverse Tangent will be very useful throughout this book. We can calculate inverse Tangent with a calculator, but we can also use a ruler and protractor to draw a right-angled triangle with an opposite and adjacent side that have the same ratio as the gradient, and then measure the angle.

A slightly convoluted way to calculate the angle of a line of a particular gradient is to draw a right-angled triangle of *any* size next to it, so that the hypotenuse is part of the line, and the opposite side is vertical and the adjacent side is horizontal. We then measure the angle of the triangle.

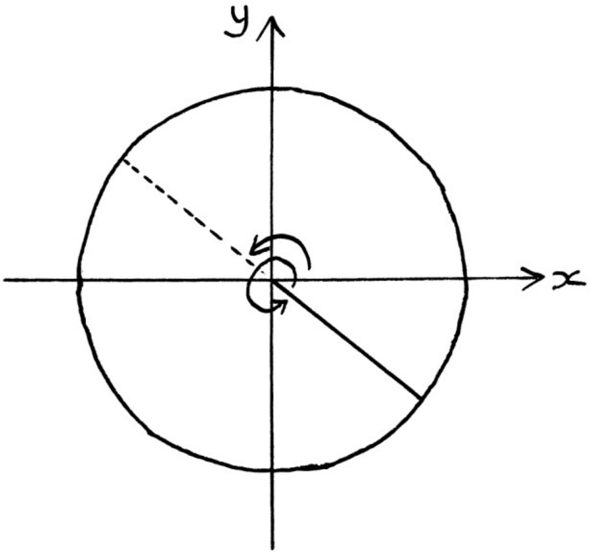


Of course, it would be quicker just to measure the angle of the line without drawing a triangle. However, seeing how a triangle fits against the line helps to reinforce how gradient and triangles relate to each other. Another consequence is that by drawing a triangle, we will end up with an opposite side and an adjacent side in exactly the correct proportions for that gradient, should we require such a thing. Whatever the size of the right-angled triangle that we draw against the line, it will be the same shape.

As with inverse Sine and inverse Cosine, inverse Tangent has two possible results within a circle. Two angles will produce the same gradient:



For example, within a circle, the inverse Tangent of 1 is both 45 degrees and 225 degrees. In other words, a line with a gradient of 1 can be said to be at an angle of 45 degrees or at an angle of 225 degrees. Again, there are really countless results depending on how far around a circle we want to go. The angles −315, −135, 45, 225, 405, 585, 765, and so on, all have the gradient of 1. Therefore, the inverse Tangent of 1 could be any of these values or more. For inverse Tangent, it is only worth considering the two possible positive results below 360, while being aware that the others exist. A calculator will only give one result. It is up to the user, first, to be aware that there are two potential results and that the supplied result might not be the desired one, and second, to be able to calculate the other angle. Calculating the other angle is simple as it will be 180 degrees around the circle. Therefore, we either add or subtract 180 degrees.



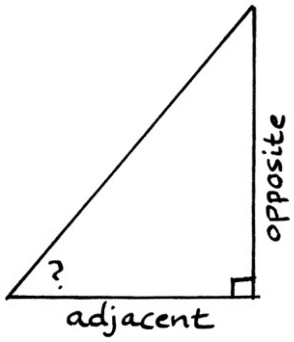
Many calculators give the result to inverse Tangent as an angle in the *right-hand* half of the circle. This means that they will give the result as a value between +90 degrees and −90 degrees. For example, where we might expect a positive angle between 270 and 360 degrees, we will be given a negative angle between 0 and −90 degrees. To change a negative result to a more useful positive angle, we add 360 degrees to it. For example, −65 degrees is also –65 + 360 = 295 degrees.

It always pays to think about whether the result is the desired angle, or whether we actually need the other angle in the circle that has the same gradient. The way to tell is to imagine how the gradient looks on the circle. If the gradient and the given angle point into the same quarter of the circle, then it is the correct angle. If they do not, then we want the other angle, which will be 180 degrees around the circle. We add or subtract 180 degrees to find the other angle. It is much easier to make mistakes with inverse Tangent if we do not check the result, than it is with inverse Sine or inverse Cosine.

Although inverse Tangent finds one of the two angles for a particular gradient, we will often use it for finding the angle in a right-angled triangle where we only have the opposite and adjacent sides. In other words, we will be using:

“θ = inverse tan (opposite ÷ adjacent)”

In such cases, we are still ultimately finding the angle for the gradient, but people tend not to think of it in that way.



When writing computer programs, it pays to be careful about division by zero when calculating the angle for inverse Tangent using opposite ÷ adjacent. Such a calculation will result in an error. Besides that, if we have a zero in the division for the gradient, then we do not need to bother using inverse Tangent at all, as we can intuitively know what the angle is by thinking about the circle. If the opposite side of a "right-angled triangle" is zero, then the angle will be 0 or 180 degrees; if the adjacent side of a "right-angled triangle" is zero, the angle will be 90 or 270 degrees.

The inverse Tangent function has several names that all mean the same thing: inverse Tan, Inv Tan, arctan, $\tan^{-1}$, atan. Calculators often refer to it as "arctan". In this book, I will generally refer to it as "arctan" in calculations from now on as it is easier to read.

Many programming languages have a similar function to inverse Tangent called "atan2". Whereas inverse Tangent operates on the gradient, "atan2" operates on the actual lengths of the opposite and adjacent sides of a right-angled triangle. We supply the command with the opposite and adjacent sides in the form: "atan2(opposite, adjacent)", and it finds the *only* angle that could have produced those sides. In this way, we do not need to think about whether the result is the one we want, because the result will always be the one we want.

When using inverse Tangent, it pays to remember that it works on gradients, and gradients are calculated by dividing the opposite side of a right-angled triangle by the adjacent side of a right-angled triangle. The further we fall into the world of waves, the easier it is to forget this. Try to remember the phrases "opposite over adjacent" and "rise over tread".

One easy mistake to make is to calculate the gradient the wrong way around when using inverse Tangent. If we do this, either we will see a result that is the wrong side of 45 degrees to the correct result, or we will see a result that is the wrong side of 45 degrees plus another 180 degrees. For example, if the two possible angles for a particular gradient are 10 degrees *above* 45 degrees (in other words 55 degrees) or 180 degrees more than that (235 degrees), calculating the gradient the wrong way around will either give us a result 10 degrees *below* 45 degrees (35 degrees), or it will give us a result 180 degrees more than that (215 degrees).

It is very easy to make this mistake. The mistake has this effect because the opposite and adjacent sides of a right-angled triangle become swapped around. If a triangle has an opposite side of 4 units and an adjacent side of 3 units, then using inverse Tangent the wrong way round will find the angle for a right-angled triangle with an opposite side of 3 units and an adjacent side of 4 units. [As well as saying the wrong result is the wrong side of 45 degrees, one can also think of it as being the correct result subtracted from 90 degrees as that means the same thing – for example, 90 – 359 = −269, which, after adding 360 to make it a positive angle, is 91 degrees. The angles 359 degrees and 91 degrees are either side of 45 degrees.]

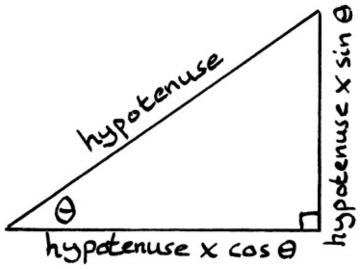
Calculating missing attributes of a triangle

Given that we can find the gradient of a hypotenuse at a particular angle using Tangent, and we can find the angle of a hypotenuse at a particular gradient using inverse Tangent, we can treat the angle and gradient as similar entities. Therefore, we can say that there are four main attributes of a triangle:

- angle or gradient.
- length of the hypotenuse.
- length of the opposite side.
- length of the adjacent side.

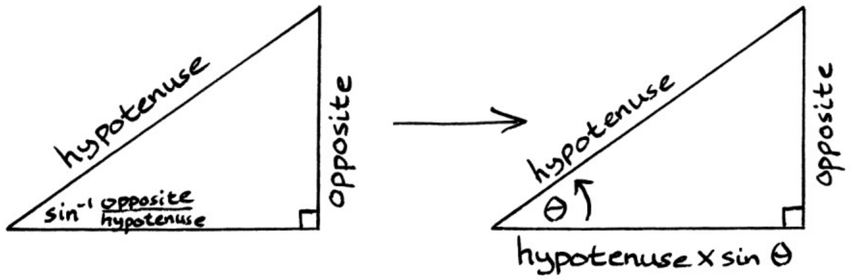
Given any two of these attributes, it is possible to calculate the remaining two either by drawing a right-angled triangle to scale or by using maths. This also means that any right-angled triangle can be defined or described using only two attributes.

• If we have the angle and the length of the hypotenuse, we can use Sine and Cosine to work out the adjacent and opposite sides, as we know:



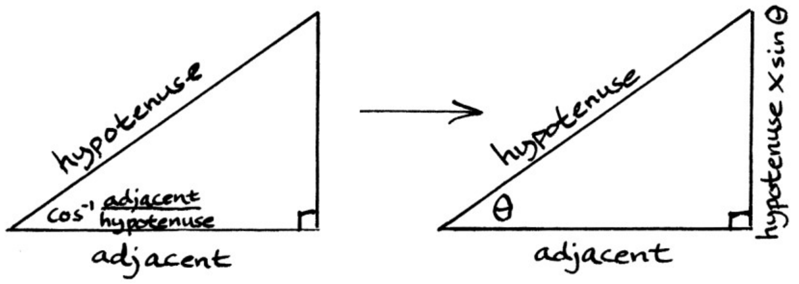
• If we have the length of the hypotenuse and the length of the opposite side, we can use inverse Sine to work out the angle and then use that angle and hypotenuse to work out the adjacent side using Cosine. Given the formula:
“hypotenuse * sin θ = opposite side”
... we can work out that:
sin θ = opposite ÷ hypotenuse
... and therefore:
θ = inverse sin (opposite ÷ hypotenuse)

As we are using inverse Sine, we need to check if we have the correct result out of the two possible results. [In this particular case, because we are dealing with a right-angled triangle that is the right way up and the right way around, we do not really need to check. However, it always pays to think about the two possible results when using the inverse functions.] Now we know “θ”, we can calculate the adjacent side with “hypotenuse * cos θ”.

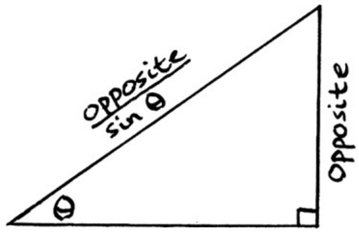


- If we have the length of the hypotenuse and the length of the adjacent side, we can use inverse Cosine to work out the angle, and then use that angle to work out the opposite side using Sine. With the formula:
“hypotenuse * cos θ = adjacent side”
... we can work out that:
cos θ = adjacent ÷ hypotenuse
... so:
θ = inverse cos (adjacent ÷ hypotenuse)

As we are using inverse Cosine, we need to check that the result is the correct one of the two possible ones. Then, as we now know θ, we can work out the opposite side: hypotenuse * sin θ.

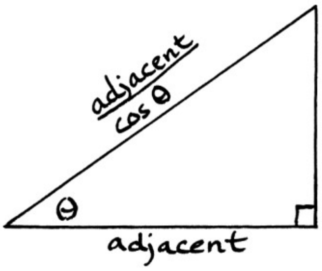


- If we have the angle and the opposite side, we can use the formula:
“hypotenuse * sin θ = opposite side”
... to show that:
opposite = hypotenuse * sin θ
... so:
opposite ÷ sin θ = hypotenuse
... so:
hypotenuse = opposite ÷ sin θ



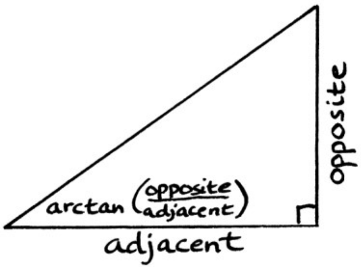
- If we have the angle and the adjacent side, we can use the formula:

"hypotenuse * cos θ = adjacent side"
... to show that:
adjacent = hypotenuse * cos θ
... so:
adjacent ÷ cos θ = hypotenuse
... so:
hypotenuse = adjacent ÷ cos θ

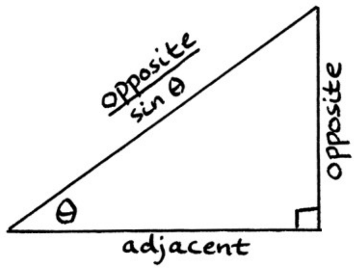


- If we have the length of the opposite side and the length of the adjacent side, we can use inverse Tan [arctan] to work out the angle:

tan θ = opposite ÷ adjacent
... so:
θ = arctan (opposite ÷ adjacent).


$\arctan\left(\frac{\text{opposite}}{\text{adjacent}}\right)$
opposite
adjacent

We check to see if we have the right angle out of the two possible ones for that gradient. Then, now we have θ, we can use the formula:
"opposite = hypotenuse * sin θ"
... to work out that:
hypotenuse = opposite ÷ sin θ
... and so calculate the hypotenuse. (We could also have used hypotenuse = adjacent ÷ cos θ)



Pythagoras’s theorem

A simpler way to calculate the length of a side given the lengths of two other sides is to use Pythagoras’s theorem. Pythagoras’s theorem states that the sum of the squares of the two shorter sides of a right-angled triangle (the opposite and adjacent sides) is equal to the square of the longer side (the hypotenuse). In other words:

$\text{hypotenuse}^2 = \text{opposite}^2 + \text{adjacent}^2$

... or:

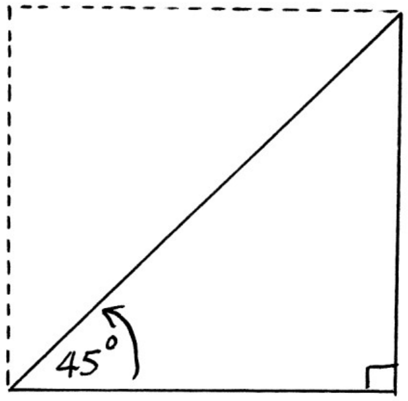
$\text{hypotenuse} = \sqrt{\text{opposite}^2 + \text{adjacent}^2}$

If, for example, a triangle has an opposite side of 3 units and an adjacent side of 4 units, then the hypotenuse will be $\sqrt{3^2 + 4^2}$ = 5 units.

If a triangle has a *hypotenuse* of 2 units and an adjacent side of 1.5 units, then we can rearrange the formula to be: $\text{opposite}^2 = \sqrt{\text{adjacent}^2 - \text{hypotenuse}^2}$. The opposite side would be $\sqrt{2^2 - 1.5^2}$ = 1.3229 units.

The formula works for any right-angled triangle, although if the triangle is upside down or the wrong way around, we need to check if the result should be made negative or not.

Pythagoras’s Theorem is easiest to visualise with a right-angled triangle with an angle of 45 degrees, when the opposite and adjacent sides will be the same length. If the opposite and adjacent sides are the same length, then we are really dealing with a square cut in half diagonally.

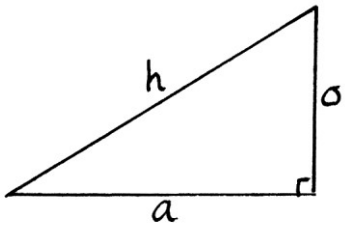


If the hypotenuse is 1 unit long, the formula will be:
$1^2 = x^2 + x^2$ where "x" is the length of the adjacent and opposite sides. This gives:
$1 = 2x^2$ which means:
$2x^2 = 1$
$x^2 = 0.5$
$x = \sqrt{0.5} = 0.7071$

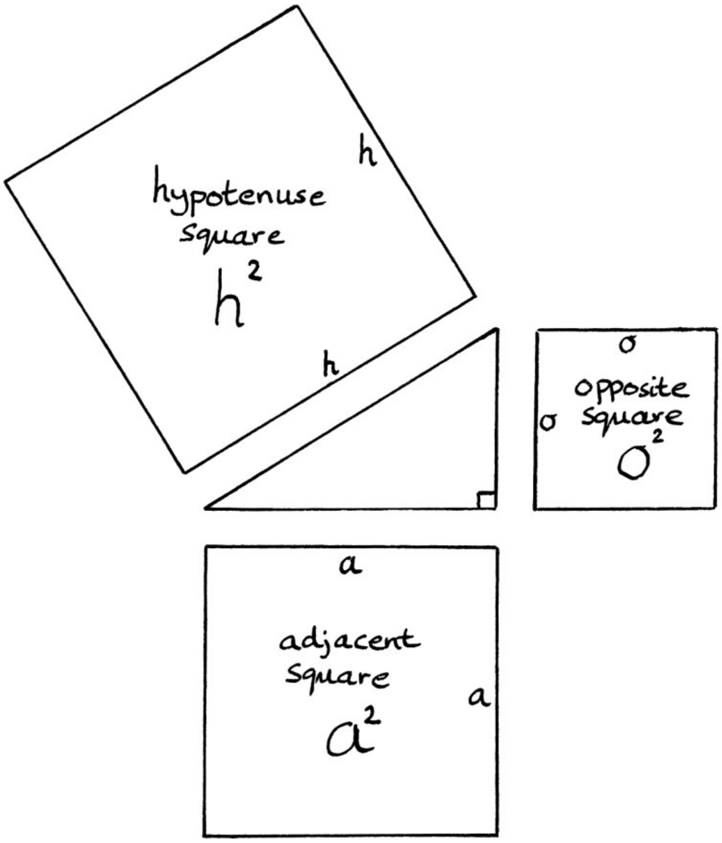
Therefore, the opposite and adjacent sides will both be 0.7071 units long, which also happens to be the Sine and Cosine of 45 degrees, which is something that should be expected if you think about such a triangle.

If the adjacent and opposite sides are both 1, then the hypotenuse will be:
$\sqrt{1^2 + 1^2} = \sqrt{2} = 1.4142$

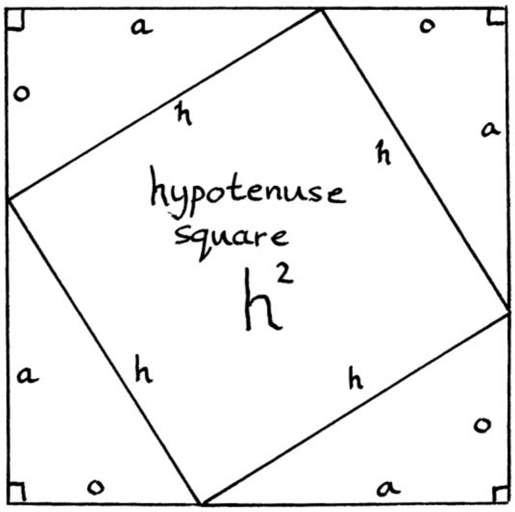
If no one had ever come up with Pythagoras's theorem, the fact that the Sine and Cosine of 45 degrees result in the square root of an integer would be a likely inspiration for research that would eventually discover it. Although it might take a long time to devise the formula for Pythagoras's theorem yourself, it is reasonably straightforward to prove that it is correct. First imagine the following triangle, with the sides labelled "h", "a" and "o" for hypotenuse, adjacent and opposite. The actual lengths of the sides do not matter:



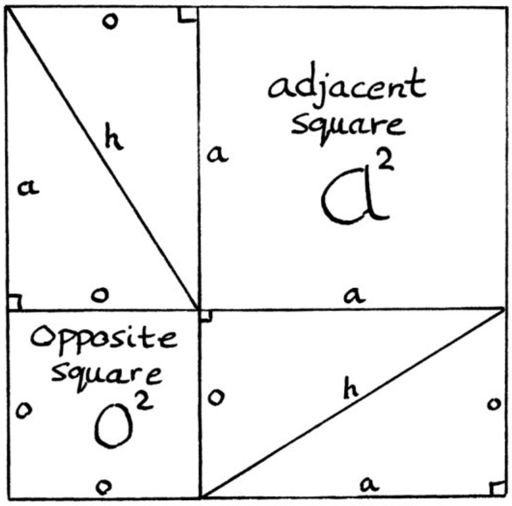
In the following picture, the same triangle has squares drawn around it. The squares have sides equal to the lengths of the sides of the triangle. In other words, the first square has sides equal to the length of the hypotenuse. Its area is the square of the hypotenuse. The second square has sides equal to the length of the opposite side. Its area is the square of the opposite side's length. The third square has sides equal to the length of the adjacent side. Its area is the square of the adjacent side's length.



We will focus on the triangle and the "hypotenuse square". If we push the square against the hypotenuse of the triangle, and then put three more copies of the triangle around the square, we end up with a larger square. This larger square's sides are the length of the opposite side added to the length of the adjacent side of the original right-angled triangle. The larger square contains the hypotenuse square and four copies of the right-angled triangle.



Next, we can make a second large square of the same size, and this one will contain both the "opposite square" and the "adjacent square". To pad out the empty space, it also contains four copies of the original triangle.



We can see that the two larger squares have exactly the same dimensions – they both have sides equal to the sum of the lengths of the adjacent and opposite sides of the original right-angled triangle. This means that the two larger squares have the same area. The two larger squares also each have four copies of the original right-angled triangle within them. This means that the area inside each of the two large squares that is not taken up with the four right-angled triangles must be the same. This means that the area of the "hypotenuse square" must be equal to the areas of the "adjacent square" and the "opposite square" added together. This, in turn, means that the square of the hypotenuse of the right-angled triangle is equal to the sum of the squares of the adjacent and opposite sides of the right-angled triangle. To put it more mathematically, $\text{hypotenuse}^2 = \text{opposite}^2 + \text{adjacent}^2$. It does not matter what dimensions the right-angled triangle has, as this will always be the case.

Delving deeper into Pythagoras's theorem, we can come up with other observations. As all triangles of the same shape have the same proportion in their sides, for the purposes of Pythagoras's theorem, any right-angled triangle with a hypotenuse that is other than 1 can be scaled to have a hypotenuse of 1.

As the square of 1 is 1, Pythagoras's theorem can be rewritten to be:

"The sum of the squares of the adjacent and opposite sides of a *unit*-hypotenuse right-angled triangle will always be 1".

... or:

adjacent^2 + opposite^2 = 1 (for unit-hypotenuse right-angled triangles).

This shows an interesting aspect of right-angled triangles.

If we know the angle of the unit-hypotenuse triangle involved, we can use the fact that the adjacent side is Sine θ and the opposite side is Cosine θ to create this formula:
$(\sin \theta)^2 + (\cos \theta)^2 = 1$

Note that $(\sin \theta)^2$ is often written as $\sin^2 \theta$. Similarly, $(\cos \theta)^2$ is often written as $\cos^2 \theta$. Therefore, the above formula could be written as:
$\sin^2 \theta + \cos^2 \theta = 1$.

Mathematical formulas for Sine and Cosine

When the relationship between the angle of a unit-hypotenuse right-angle triangle and its opposite and adjacent sides was first thought about, hundreds of years ago, the Sine and Cosine of an angle would have been calculated by accurately drawing and measuring triangles or circles. Mathematicians eventually came up with formulas that could calculate the Sine and Cosine of angles much more accurately than could ever be possible by drawing pictures. The standard way of accurately calculating the functions nowadays is with a "Taylor series", which is an infinite sum of ever-decreasing values. The more of the sum that is completed, the more accurate the result. However, it is impossible to achieve a completely accurate result, as there is always more of the sum to be calculated.

Obviously, nowadays, most people calculate Sine or Cosine by just pressing a button on a calculator, but if we want to know how to calculate them by hand, it pays to know about the Taylor series.

The basic formula for a particular Taylor series is devised in ways that are too complicated for the level of maths in this chapter. Fortunately, we do not need to know how a Taylor series was created to use it. We just need to fill in values.

The Taylor series for Sine and Cosine use radians instead of degrees. I explain radians in Chapter 22. For now, it is enough to know that radians are a way of dividing a circle into pieces in much the same way as degrees divide a circle into 360 pieces. Radians, however, divide the circle into 2π pieces (roughly 6.28 pieces), which, as I will explain in Chapter 22, is a way that is more suited to dealing with maths with circles. To convert an angle from degrees into radians, it is first necessary to divide it by 360 to give the portion of a circle that that angle represents, and then to multiply the result by 2π.

Sine Taylor series

The Taylor series for Sine is as follows:

$$\sin\theta = \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \frac{\theta^7}{7!} + \frac{\theta^9}{9!} - \frac{\theta^{11}}{11!} \dots$$

...where "θ" is any angle in radians and "!" means the factorial of a number, so 3! is equal to 3 * 2 * 1 = 6, and 5! is equal to 5 * 4 * 3 * 2 * 1 = 120.

The basic rule in this formula is that each step is alternately negative or positive, and consists of the angle being raised to ever-higher consecutive odd numbers and divided by the factorial of those numbers.

If we want to find the Sine of 45 degrees, we first convert the degrees into radians: 45 degrees = (45 ÷ 360) * 2π = 0.25π radians or 0.785398163397 radians to 12 decimal places. In this particular case, it is easier to use the value "0.25π" instead of writing out "0.78598163397" each time. Next, we calculate as much of the following sum as we need to achieve the accuracy we desire.

$$\sin 0.25\pi = 0.25\pi - \frac{(0.25\pi)^3}{3!} + \frac{(0.25\pi)^5}{5!} - \frac{(0.25\pi)^7}{7!} + \frac{(0.25\pi)^9}{9!} - \frac{(0.25\pi)^{11}}{11!} \dots$$

Depending on how many steps we go through, we will find the following results:

- After one step, the result is 0.78539816 (rounded to 8 decimal places)
- After two steps, the result is 0.70465265 (rounded to 8 decimal places)
- 3 steps: 0.70714305 (rounded to 8 decimal places)
- 4 steps: 0.70710647 (rounded to 8 decimal places)
- 5 steps: 0.70710678 (rounded to 8 decimal places)
- 6 steps: 0.70710678 (rounded to 8 decimal places)

If we only wanted 8 decimal places of accuracy, we could stop now as we have reached a level of accuracy after which the decimal places have stopped changing.

After just five steps, we have achieved a very good approximation to the Sine of 45 degrees. A calculator will give the result of Sine 45 as 0.70710678 (rounded to 8 decimal places), which shows how quickly the Taylor series for Sine works.

At each step, we achieve more accuracy. It is important to note that each step is only accurate to so many decimal places (although that number of decimal places increases with each step). Therefore, we need to keep going until we reach the step at which the number of decimal places we require is the same as the step before. We will always need to round up or down the answer to the number of decimal places we want. This point becomes most apparent when finding the Sine of an angle such as 90 degrees, 180 degrees or 270 degrees, which one can know intuitively by visualising a circle or a triangle. If we use the Taylor series to find the Sine of 90 degrees, after five steps, we end up with 1.00000354 (to 8 decimal places), and after six steps, we have 0.99999994 (to 8 decimal places). We will not end up with the correct answer, 1, unless we round up or down the answer at some point. Generally, if we want the Sine of say 0, 90, 180 or 270 degrees, it is quicker to remember how the relevant circle or triangle would look than to use the Taylor series.

A pocket calculator or a computer processor will probably not be using a Taylor series, but instead what is called a "CORDIC algorithm". This is conceptually more complicated than a Taylor series, but is more suited to the range of instructions available in a simple processor. A computer *program* on the other hand, should not need to use a Taylor series to calculate Sine or Cosine as it can rely on the processor to do it. When computer programs do not need much accuracy, or are repeatedly calculating a set group of the same values, it might be quicker and less effort for them to use a pre-calculated list as a lookup table, than to wait for the processor to give the results. When wanting the Sine of an angle of say 12.345 degrees, a program that needs to be quick might look up the Sine of 12 degrees instead in a pre-calculated table. A program where accuracy really does not matter might look up the Sine of 10 in a pre-calculated table and work with that instead.

Cosine Taylor series

The Taylor series for the Cosine of a number is as follows, where "θ" is an angle in radians:

$$\cos\theta = 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \frac{\theta^6}{6!} + \frac{\theta^8}{8!} - \frac{\theta^{10}}{10!} \ldots$$

This is similar to the Taylor series for Sine. The basic rule in this formula is that each step consists of the angle being raised to ever-higher consecutive *even* numbers and divided by the factorial of those numbers.

Tangent Taylor series

The Tangent of a number can be worked out either by using the Taylor series for Sine and Cosine and then calculating Sine θ ÷ Cosine θ, or by using the Taylor series for Tangent. The Tangent Taylor series still requires the angle to be in radians. Unlike with Sine and Cosine, this Taylor series requires the angle to be between, but not including, 0.5π and −0.5π. In other words, the equivalent angle in degrees must be less than +90 degrees and more than −90 degrees. To put it another way, it only works with angles in the right hand half of the circle that would produce gradients less than positive or negative infinity. The way the series progresses is not particularly intuitive.

$$\tan\theta = \theta + \frac{\theta^3}{3} + \frac{2 * \theta^5}{15} + \frac{17 * \theta^7}{315} + \frac{62 * \theta^9}{2835} + \frac{1382 * \theta^{11}}{155925} \ldots$$

Inverse functions

There are also Taylor series for inverse Sine, inverse Cosine and inverse Tangent. Each of these will give just one of the two possible results under 360 degrees.

Summary

In everyday life, the Taylor series for Sine, Cosine, Tangent and their inverses will not be of much use, but it is good to know that the formulas exist – partly because they demonstrate how there is no concise way to describe the relationship between an angle of a right-angled triangle and the length of its sides.

Conclusion

The contents of this chapter will have given you a greater insight into angles, triangles, circles and Sine and Cosine. Whereas Chapter 1 contained a lot that you could have worked out yourself reasonably easily, this chapter contains some topics that would be much harder to figure out without spending a lot of time on the subject.

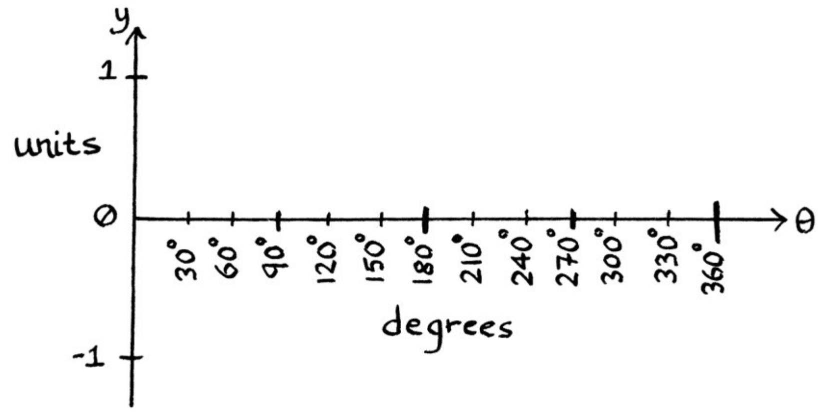
Chapter 3: Waves

The results of the Sine or Cosine of a range of angles can be plotted on a graph.

Sine graph

First, we will make a graph of the Sine function. The x-axis of our Sine graph will represent the degrees around the circle from 0 to 360. To make it clearer as to what we are showing, and to help avoid confusion, instead of calling it the x-axis, we will call it the θ-axis. The y-axis will represent the Sines of the angles represented by the θ-axis, and will be measured in units. The formula for this graph will be "y = sin θ".

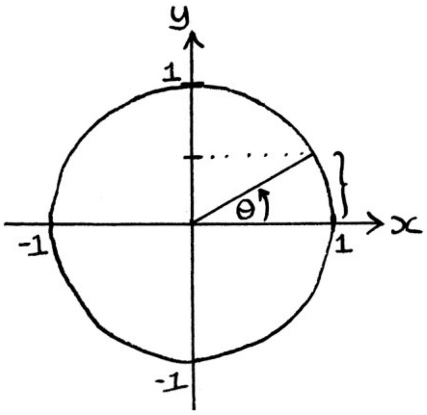
The axes look like this:



To make the graphs clearer, I will not usually write all the angles on the θ-axis in graphs.

The least technical way to draw a Sine graph is to draw a unit-radius circle, and then to measure the y-axis values of points on the circumference at evenly spaced angles around it. [This is the same as measuring the heights of points at evenly spaced intervals around the circumference]. Each measurement will be the Sine of that angle. We then plot those points on the graph.

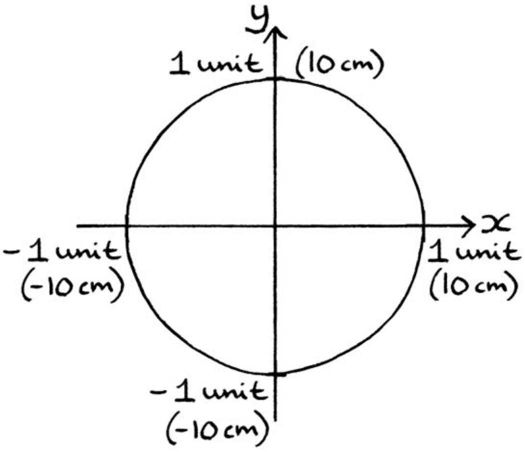
Therefore, we will draw lines from the centre of a circle out to its edge at angles of 0 degrees, 30 degrees, 60 degrees, 90 degrees, and so on, up to 360 degrees, and then measure the heights of where those lines meet the edge on the circle's y-axis.



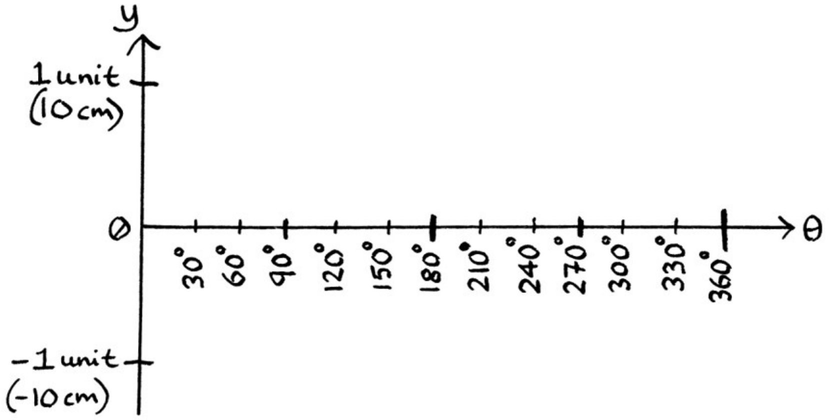
We will then plot those values as the y-axis values on our Sine graph, placed along the θ-axis at the relevant angle.

In practice, it is easiest to draw a circle with a radius of 1 unit if we create a unit system where 10 centimetres is one unit. Therefore, the circle will have a radius of 10 centimetres, and when measuring the heights of points on the circle's edge in centimetres, we will need to divide by ten to find the number of units. Smaller radiuses would produce results that were less accurate; larger radiuses would not fit on a typical piece of paper. It is also simplest to have the resulting Sine graph have a y-axis where 10 centimetres is also the length of 1 unit. In this way, we can copy the y-axis heights from the circle and plot them without any change on to the Sine graph.

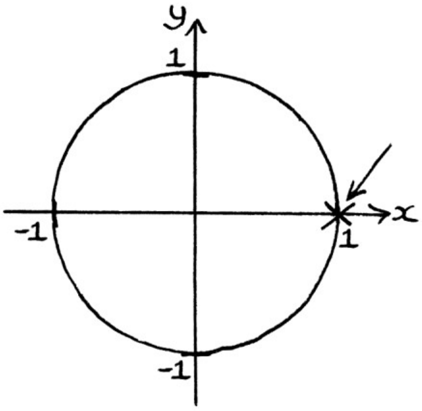
First, we draw a circle with a radius of 10 centimetres on x and y-axes. We mark the units of −1 and +1 on each axis:



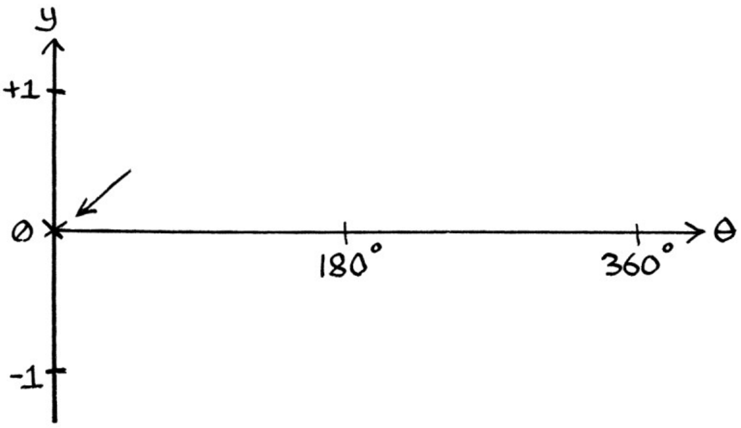
Next, we draw axes for our future Sine graph, where the y-axis has units from −1 to +1, and each unit takes up 10 centimetres on the piece of paper. The θ-axis units will go from 0 to 360 degrees. We can set the length of the θ-axis to anything that will fit on a piece of paper, and the layout of the angles will be scaled accordingly.



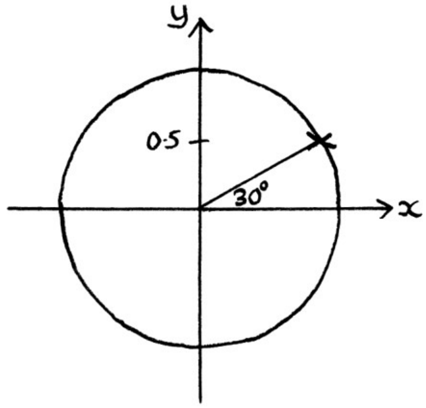
The first angle we will read from the circle will be 0 degrees. This has a y-axis height of 0 units.



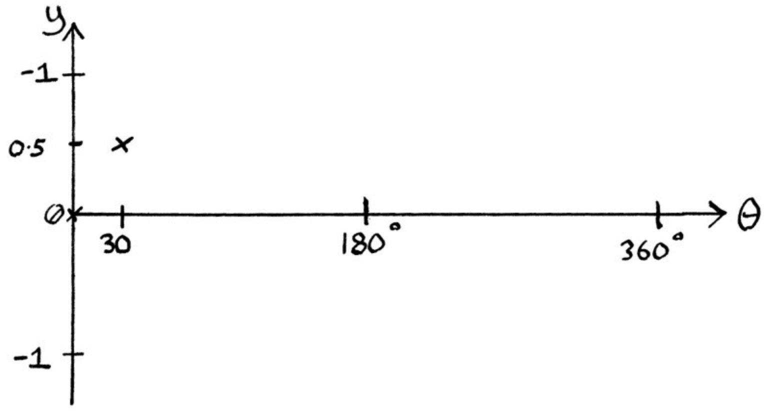
We can mark that height on the Sine graph:



Next, using a protractor we mark the edge of the circle at the point at 30 degrees from the origin. We measure the height of that point in centimetres. We then divide that by 10 to find the number of units (because we are using a system where 10 centimetres is one unit). The measurement will be a height of about 5 centimetres, which works out as 0.5 units.

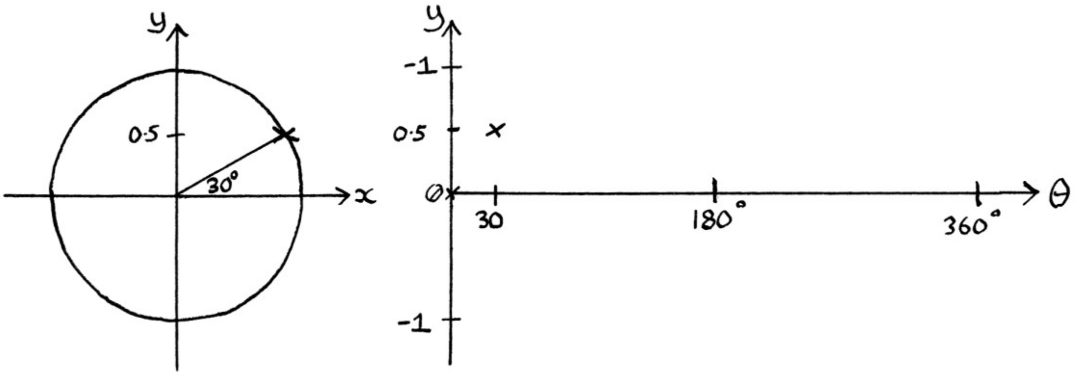


We then mark that height of 0.5 units on the Sine graph for the θ-axis value of 30 degrees.

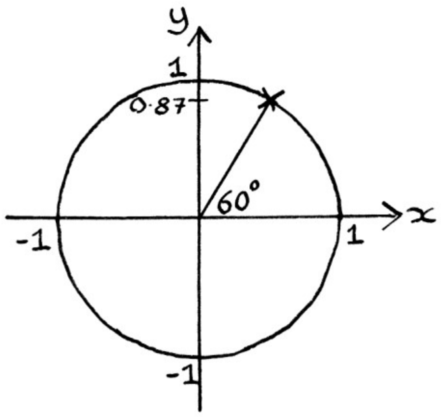


[As the scale for the circle and the Sine graph are the same, we can skip converting centimetres to units and back to centimetres, and just mark the point on the graph where the y-axis value is 5 centimetres from zero, and where "θ" is 30 degrees.]

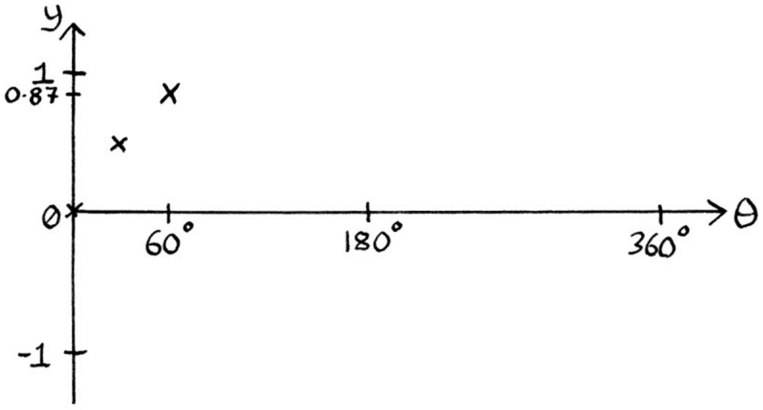
If we put the circle and the graph next to each other, we can see how the y-axis values are the same for the circle and the graph. For all of the construction of the Sine graph, the points along it will be at the same height as the corresponding points on the circle.



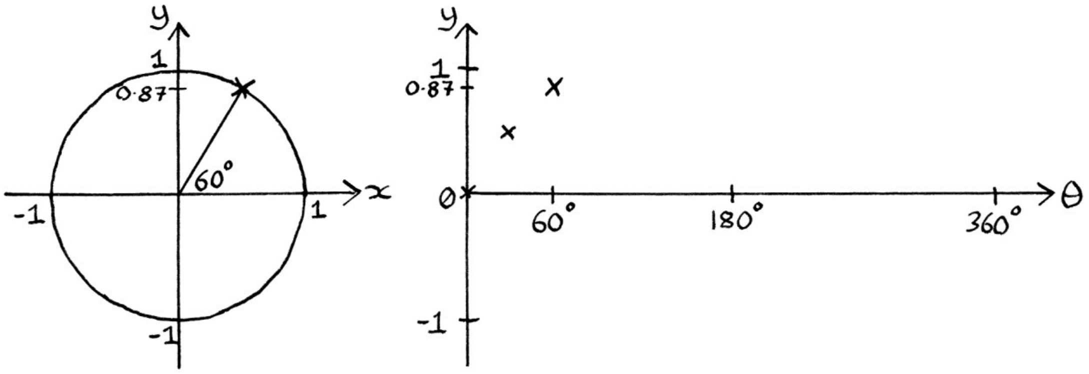
Next, we mark the edge of the circle at the point 60 degrees from the origin, and measure that height. It will be about 8.7 centimetres, which represents 0.87 units.



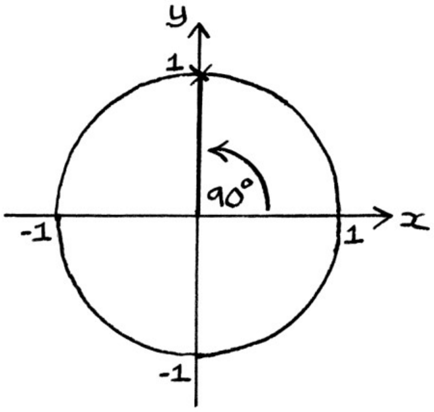
We mark that on the Sine graph:



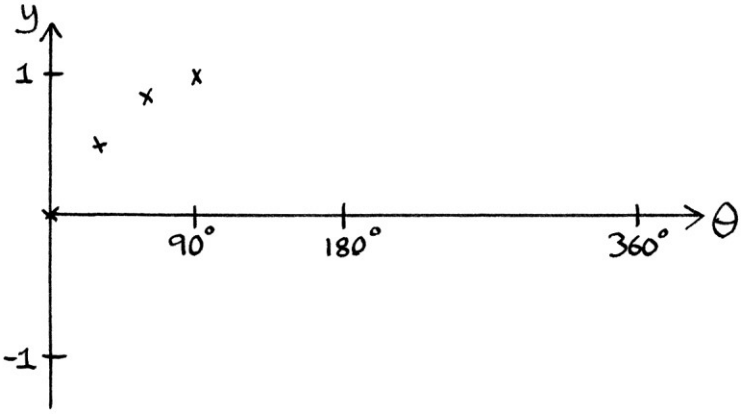
Again, if we put the circle next to the graph, we will see how the points on both relate to each other:



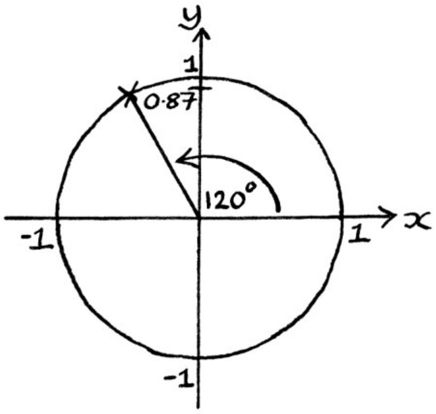
Next, we measure the y-axis position of the point on the circle's edge at 90 degrees. It will be 10 centimetres, which is 1 unit.

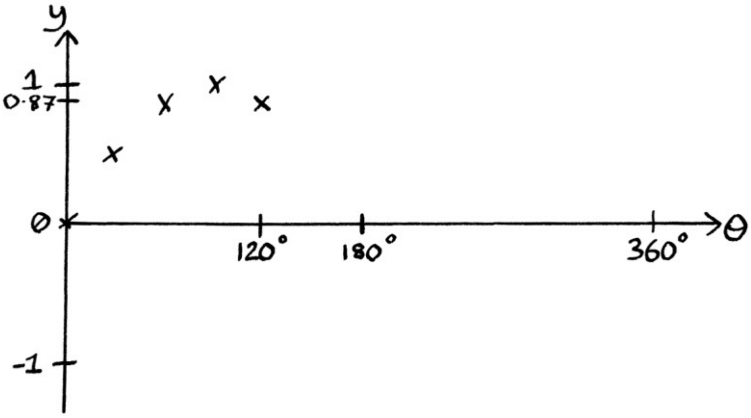


We mark that on the Sine graph:

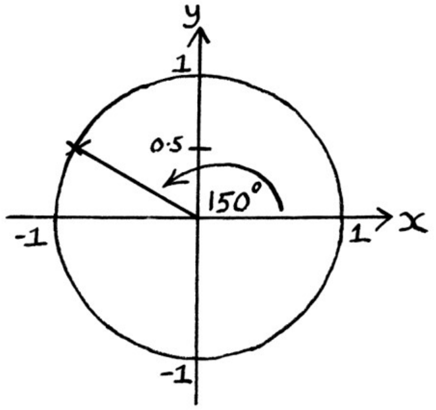


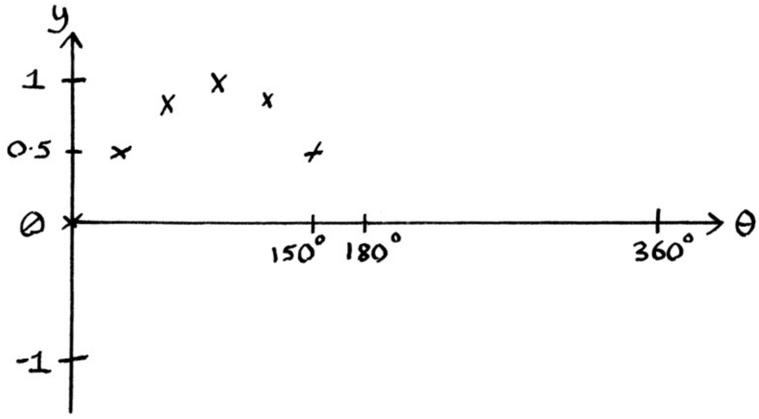
At 120 degrees on the circle, we will measure 8.7 centimetres, which is 0.87 units.



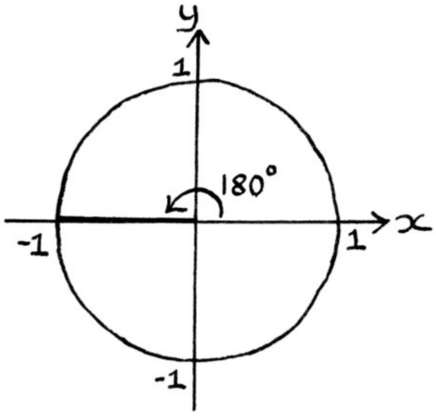


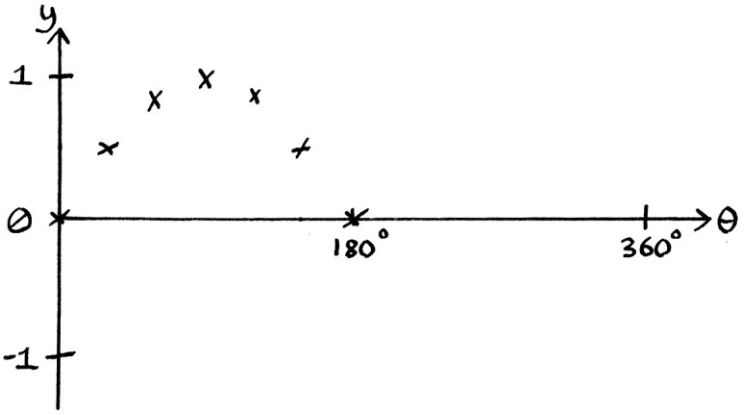
At 150 degrees, we will measure 5 centimetres, which is 0.5 units.



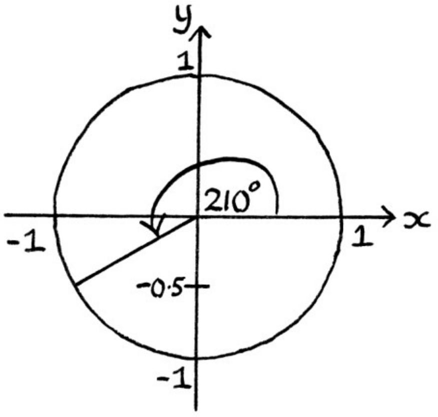


At 180 degrees, we will measure 0 centimetres, which is 0 units.

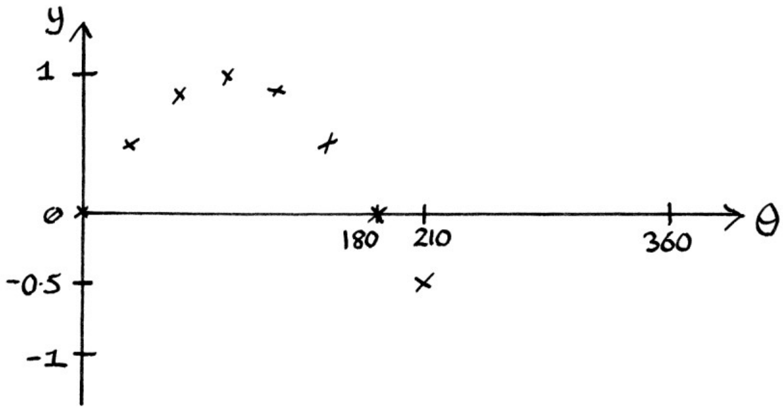




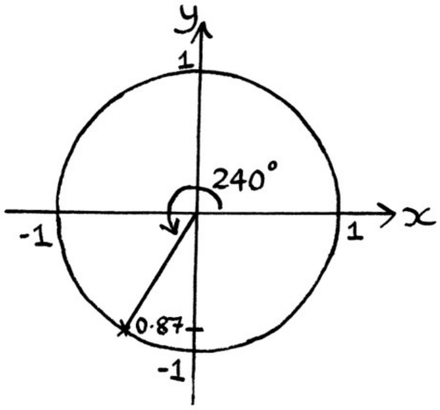
At 210 degrees, the y-axis height on the circle will be −5 centimetres, which represents −0.5 units.

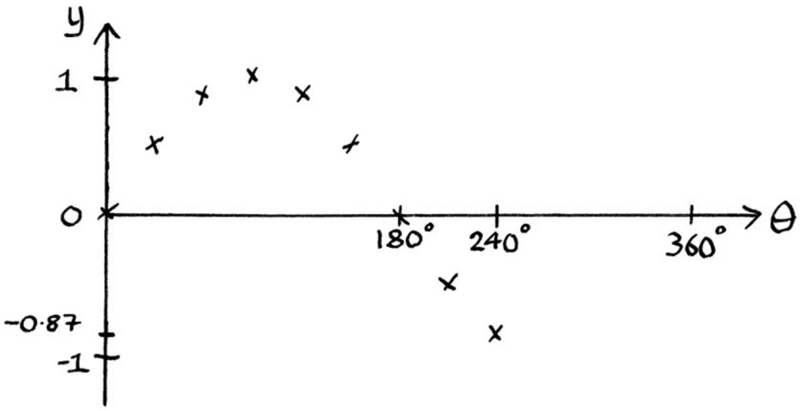


Although this is negative, we continue as before. We mark this on the Sine graph as −0.5 units.

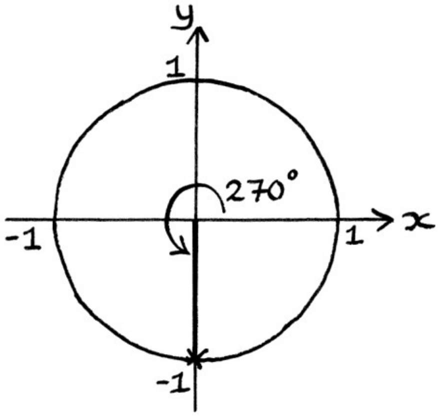


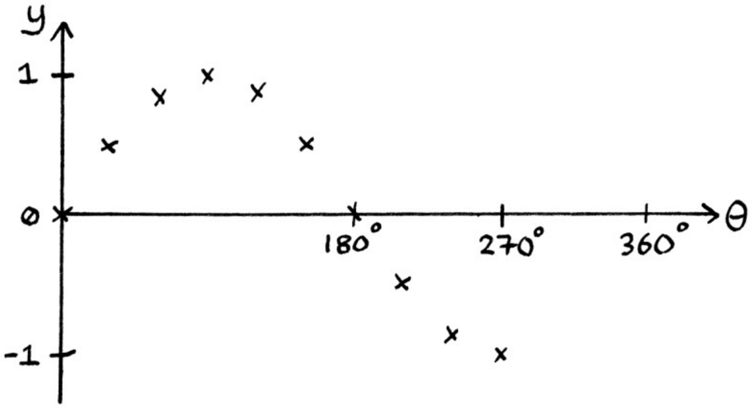
At 240 degrees, we will measure −8.7 centimetres, which is −0.87 units.



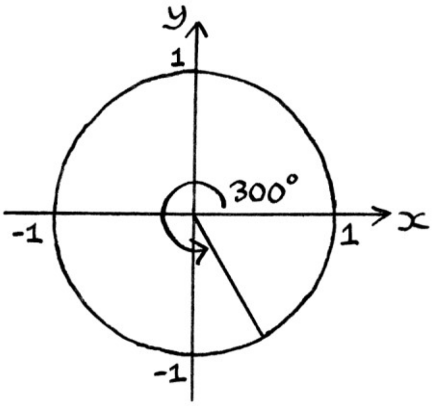


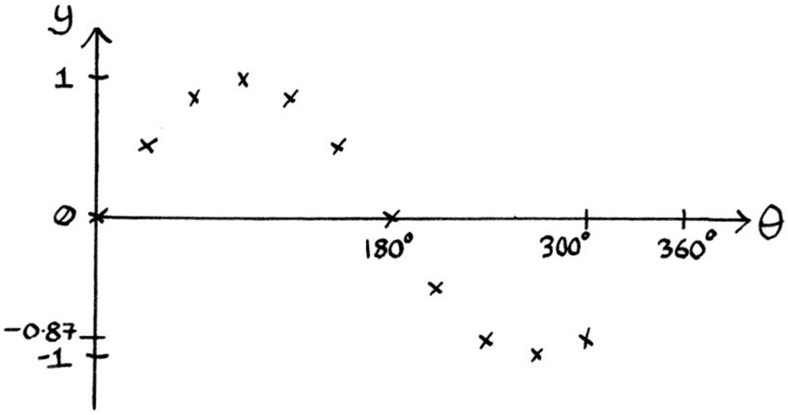
At 270 degrees, we read −10 centimetres, which is −1 units.



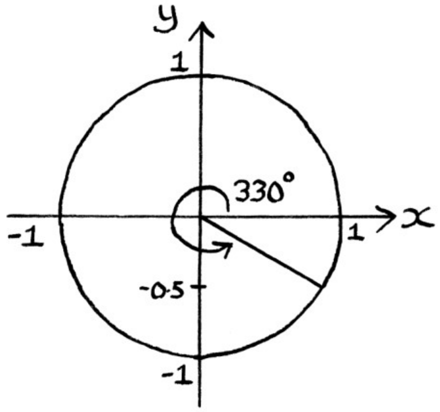


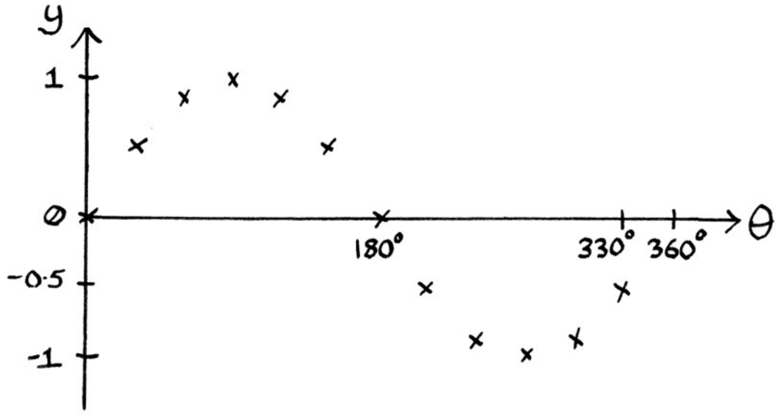
At 300 degrees, we read −8.7 centimetres, which is −0.87 units.



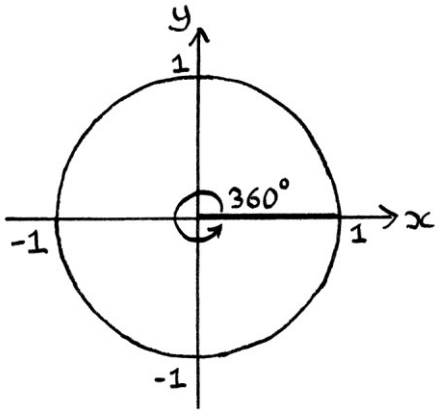


At 330 degrees, we read −5 centimetres, which is −0.5 units.

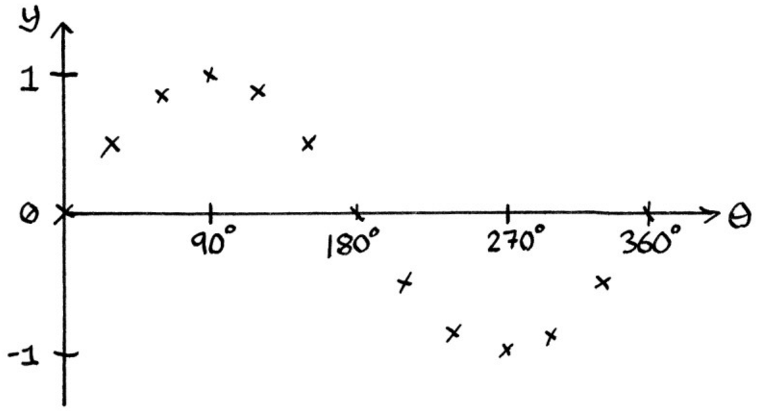




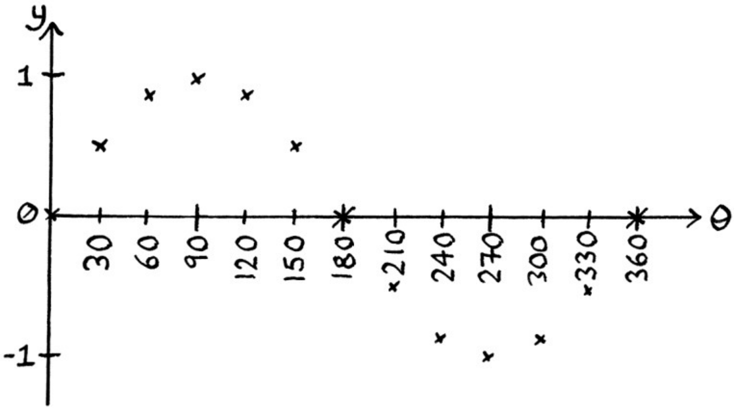
At 360 degrees, we read 0 centimetres, which is 0 units. This is the final reading.



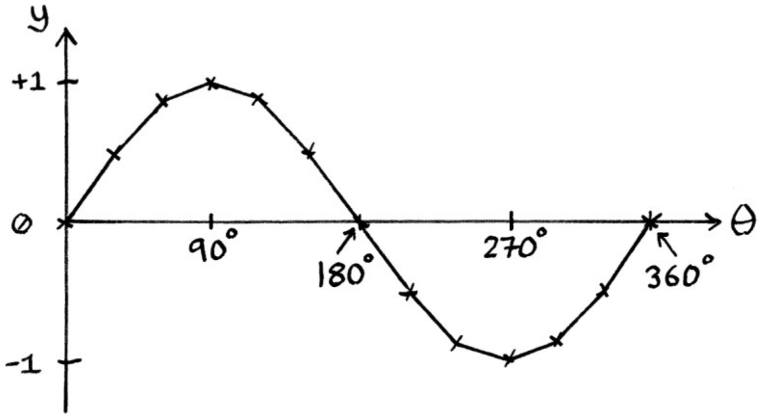
The Sine graph at this point should look like this:



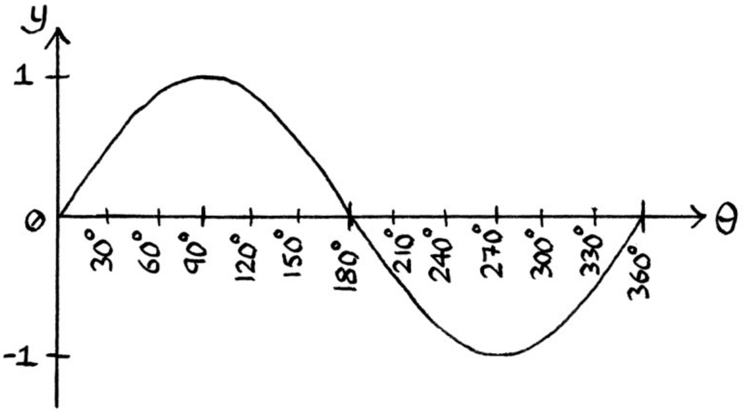
Or, with the θ-axis more fully labelled:



We then join up those points:

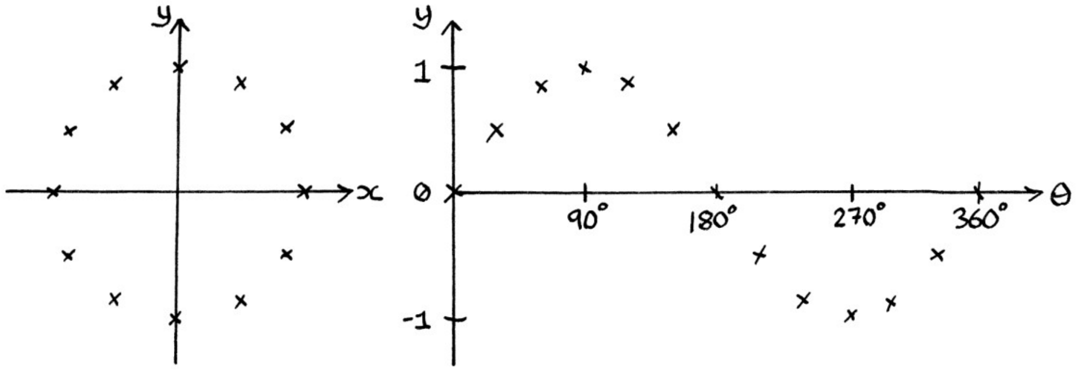


If we plotted points at smaller steps around the circle, the Sine graph would be smoother, but it would take a lot more time to do. If we plotted an infinite number of points, and we could read the heights from the circle with perfect accuracy, the Sine graph would be perfect.

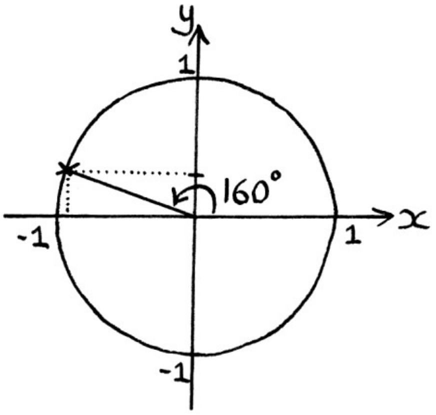


The perfect graph turns out to be a wave, and, by definition, it is a Sine wave.

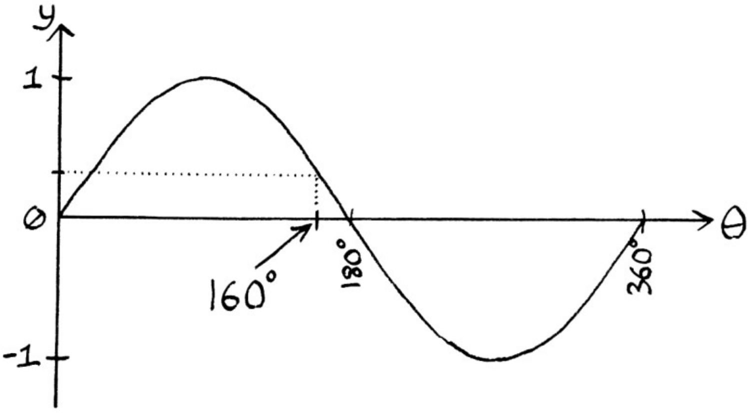
Every point in the Sine wave graph has a corresponding point in the circle. The circle and the graph are two different ways of showing the same information – the Sine of numbers. In the circle, the numbers being Sined are the angles of the lines going out to the edge of the circle; in the Sine wave graph, those same angles are laid out along the horizontal θ-axis. In the circle, the results of the Sined angles are the heights of the points on the circle's edge; in the Sine wave graph, those same numbers are portrayed as the heights of points on the curve. Those numbers are also the y-axis heights of points on both the circle and the curve.



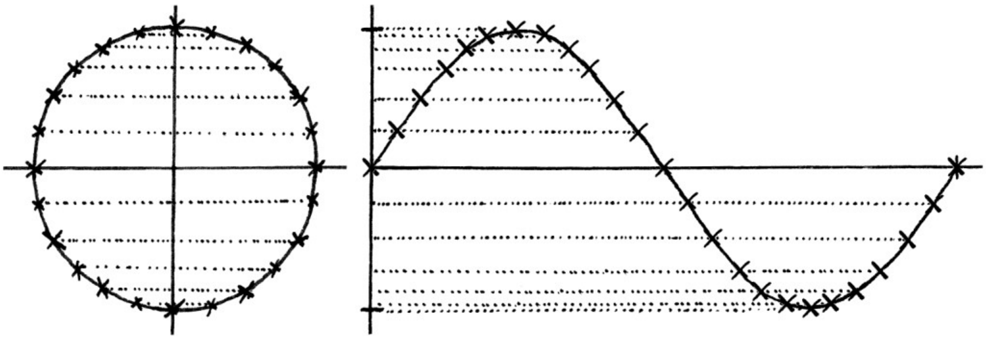
As an example, the Sine of 160 degrees on the circle is the y-axis point where a line drawn at 160 degrees meets the circle's edge:



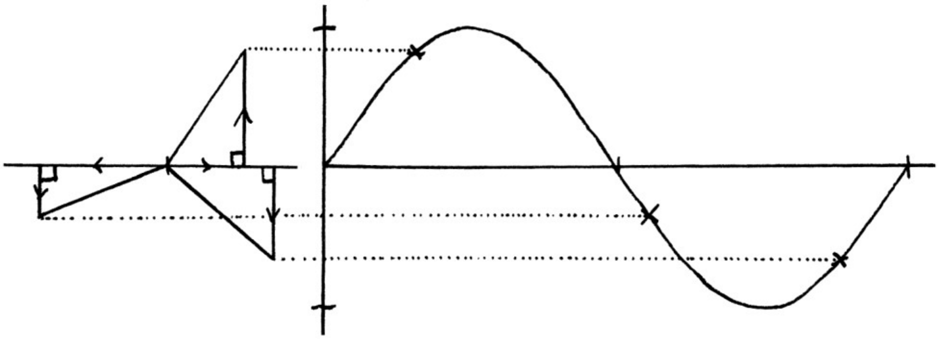
On the graph, the Sine of 160 degrees is the y-axis point where the θ-axis value is 160 degrees:



The Sine wave is really just a graph showing the heights of points around a circle that are at equally spaced angles from the origin.

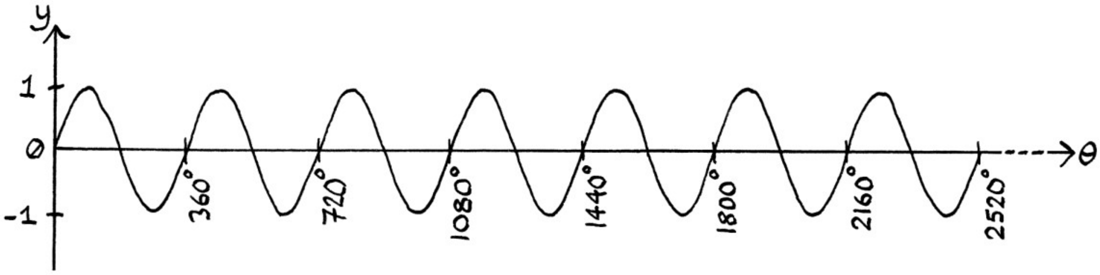


We could also say that a Sine wave is a graph that shows the lengths of the opposite sides of countless right-angled triangles.



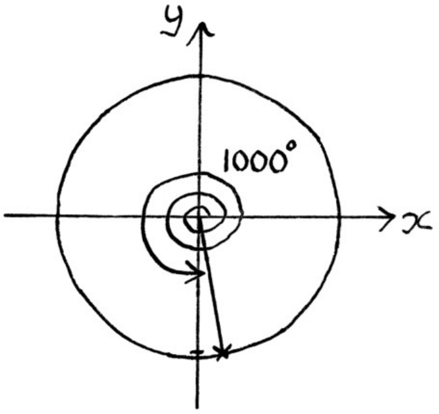
We can use both the circle and the Sine wave graph to calculate the Sine of numbers. On the circle, we draw a line from the origin at the angle for which we want the Sine, and then measure the y-axis value of where that line meets the edge of the circle. On the Sine wave graph, we just read the y-axis value at the angle on the θ-axis for which we want the Sine.

Often, a Sine wave on a graph is drawn as if it were going to be endlessly long. This is an arbitrary choice – it does not have to be drawn in that way. It all depends on what that particular Sine wave is portraying.

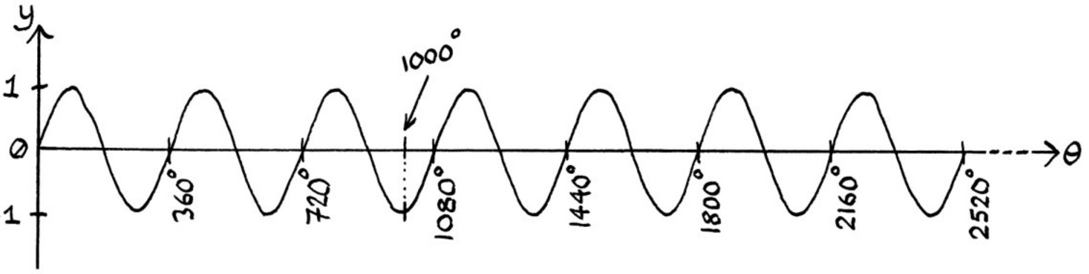


An unending Sine wave still relates to the circle, but one has to imagine the angles in the circle wrapping around and around forever.

For example, on the circle, the Sine of 1000 would be twice around the circle and then another 280 degrees:

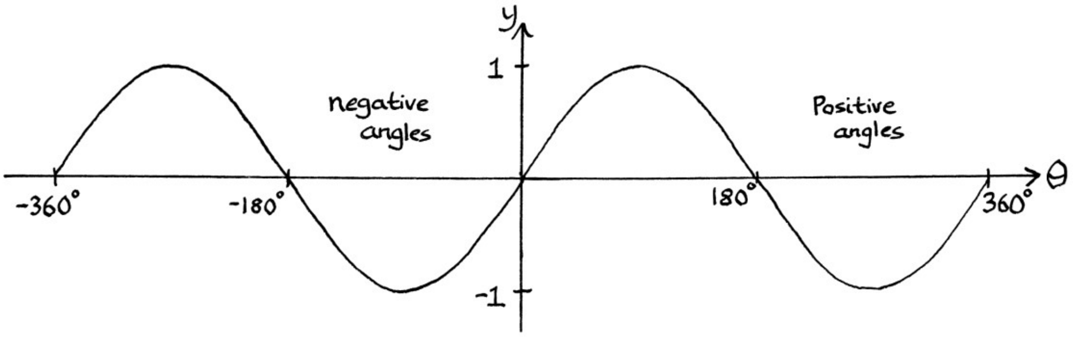


On the wave, the Sine of 1000 appears when the θ-axis is 1000 degrees:



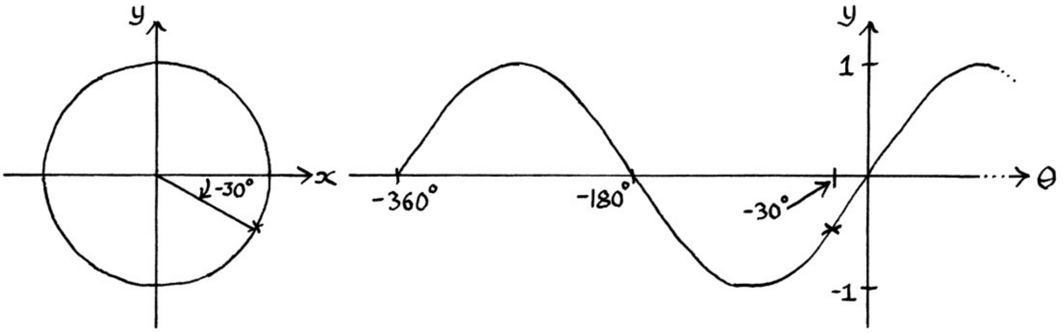
Notice how the shape of the graph repeats itself. It repeats itself because the heights of the points around the circle's edge repeat themselves as the angles increase and wrap around the circle.

A Sine wave graph might take into account negative angles too, in the sense that it could include angles going *clockwise* around the circle, in which case it would extend into the negative half of the graph.



By convention, angles are treated as increasing anticlockwise around the circle. Those positive angles on the graph stretch out towards the right on the θ-axis. Angles measured *clockwise* around the circle will be negative. Those negative angles stretch outwards to the left along the θ-axis.

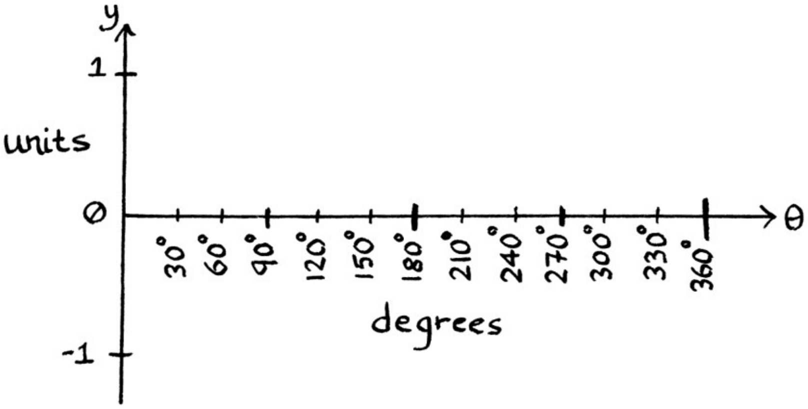
As an example, on the circle, the angle of −30 degrees has a height of −0.5 units, so on the Sine wave graph the corresponding point is marked at −30 degrees on the θ-axis at a height of −0.5 on the y-axis. [To fit the picture on the page, the right half of the wave has been truncated.]



Cosine graph

Drawing the Cosine graph is very similar to drawing the Sine graph. The x-axis on the Cosine graph, which for clarity, we will call the θ-axis, represents the degrees from 0 to 360, and the y-axis represents the Cosine of that number. The axes are exactly the same as those for the Sine wave graph. The formula for this graph will be: “y = cos θ”.

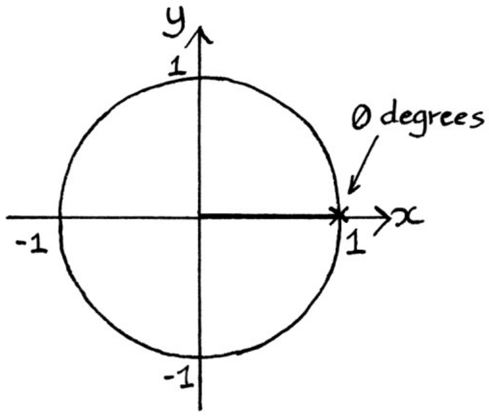
The blank axes look like this:



The least technical way of drawing a Cosine graph is to read the Cosine values of equally spaced angles off a unit radius circle, and to plot those on the graph. In other words, we measure the *x-axis* values on a circle where lines drawn outwards from the origin at evenly spaced angles meet the edge. In this example, for every 30 degrees from 0 to 360, we will read the Cosine of that angle using the circle, and then we will plot those values as the y-axis values on our Cosine graph, with "θ" being the angle at which we draw the lines.

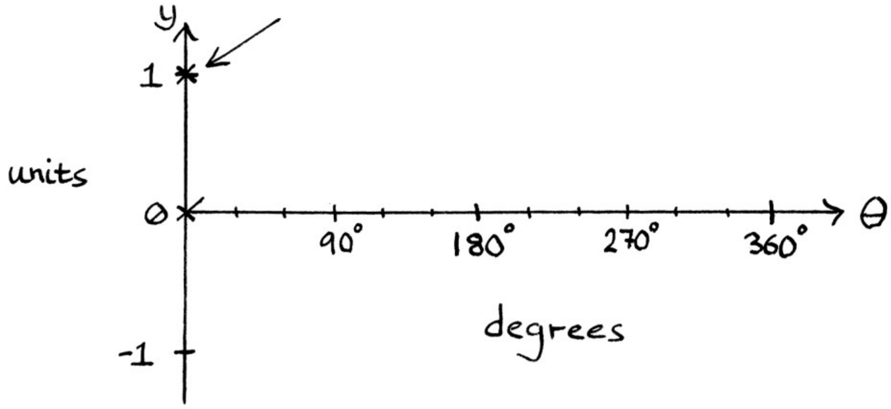
We will go through the steps one by one. First, we draw a circle with a radius of 1 unit. We will say 10 centimetres is 1 unit, and so we will draw a circle with a radius of 10 centimetres. All future measurements in centimetres will be divided by 10 to find the number of units. [If we wanted, we could actually use the same circle from the Sine wave graph example to save time, and we would be able to use the same points too.] We draw the axes for our Cosine graph with one unit as 10 centimetres on the y-axis and the degrees from 0 to 360 spaced evenly along the θ-axis.

First, we mark the point on the circle's edge at an angle of 0 degrees from the origin. We then measure the *x-axis* value of this point, which is the same as measuring the horizontal distance of that point from the origin of the axes:

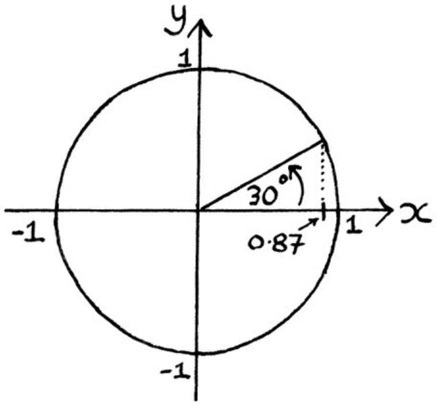


In this case, the x-axis value will be 10 centimetres, which represents 1 unit.

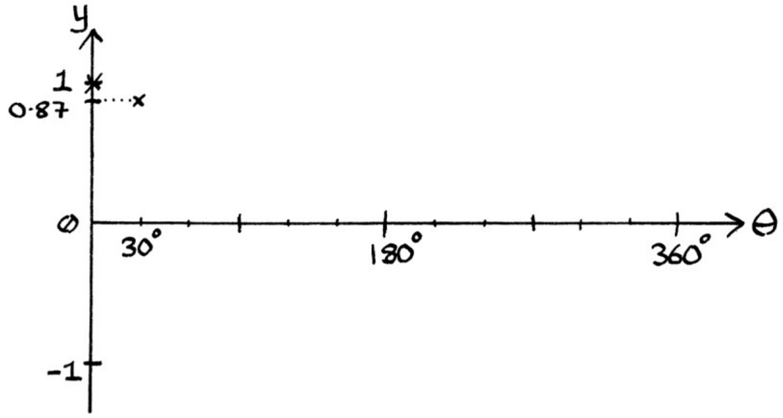
We then mark that value on our Cosine graph as the y-axis value at the point when “θ” is zero:



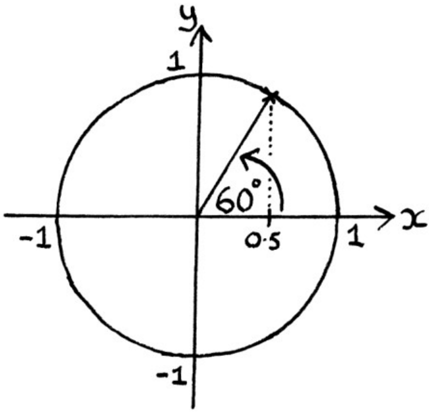
Next, we mark the point on the circle’s edge at an angle of 30 degrees from the origin. We then measure the horizontal distance of this point from the origin (which is also the x-axis value of this point). This will be roughly 8.7 centimetres, which is 0.87 units.



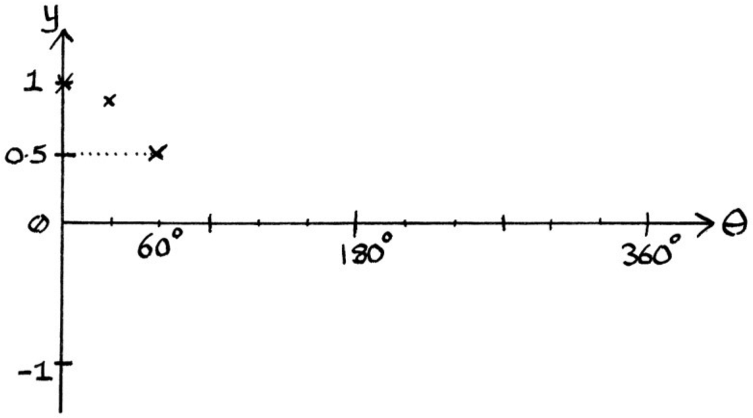
On our Cosine graph, we mark the point that is 30 degrees along the θ-axis and at 0.87 units on the y-axis.



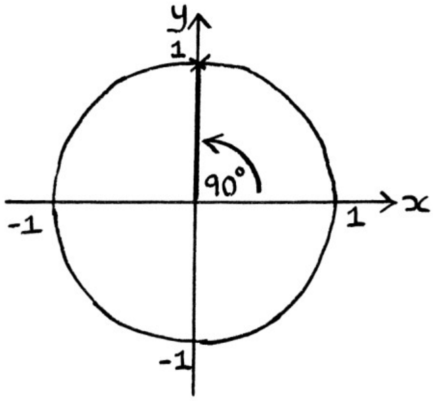
Next, we mark the point at 60 degrees on the circle, and measure the x-axis value of that point. It will be roughly 5 centimetres, which represents 0.5 units.

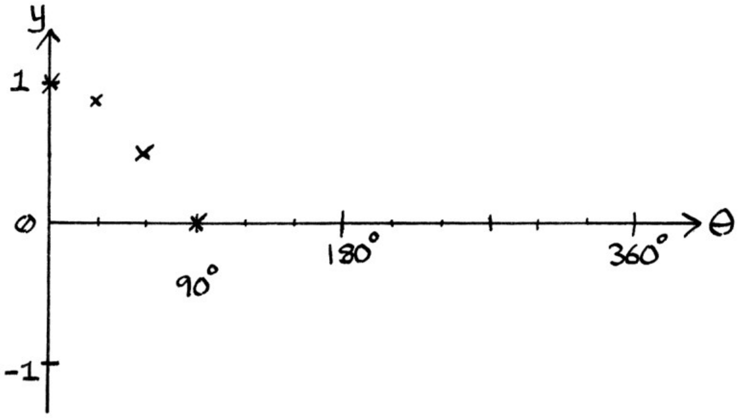


We then plot a new point on the Cosine graph 60 degrees along the θ-axis and 0.5 units (5 centimetres) up the y-axis.

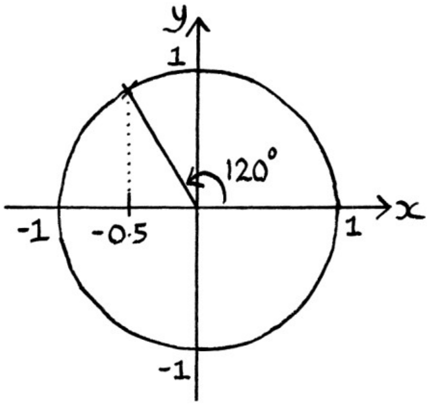


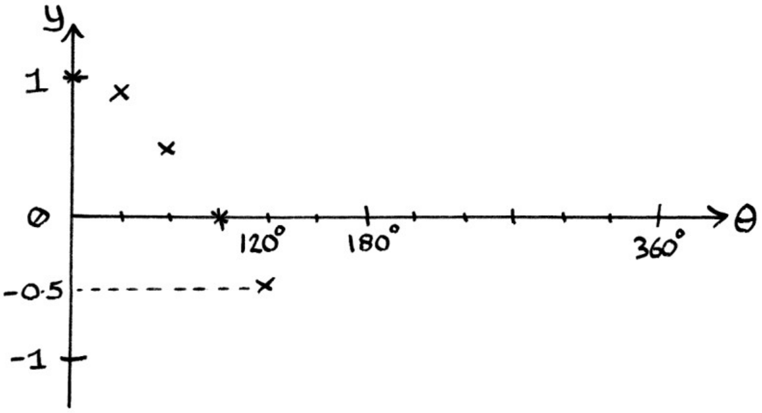
Next, we do 90 degrees:



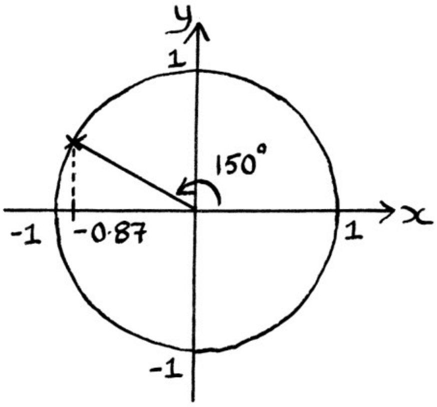


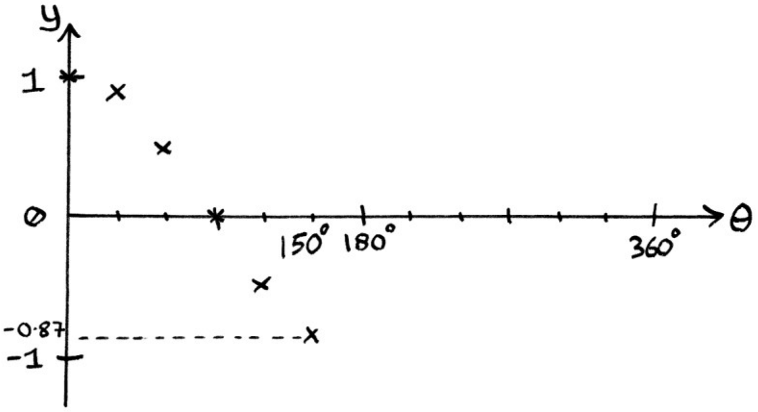
Then 120 degrees:



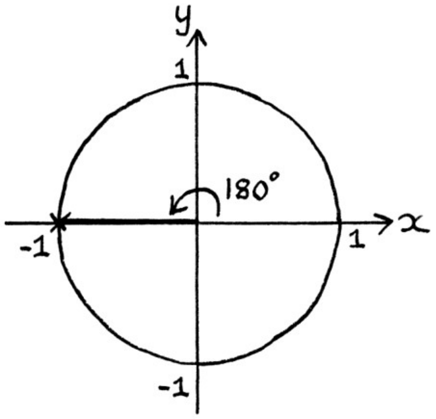


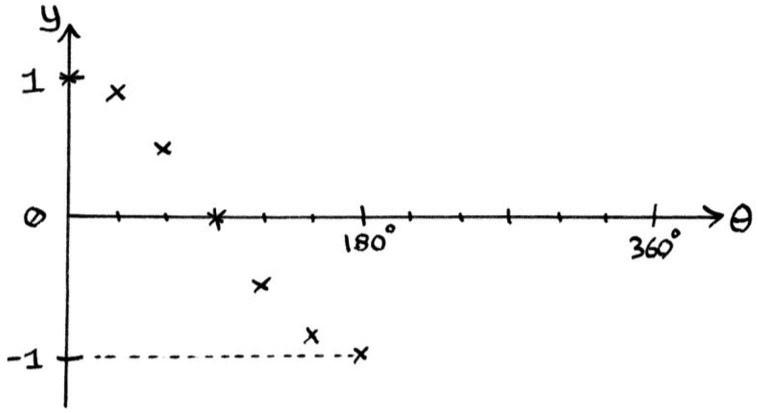
Then 150 degrees:



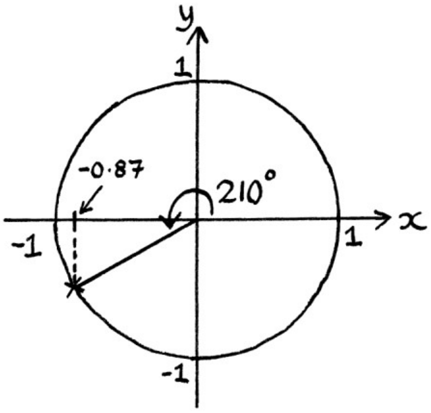


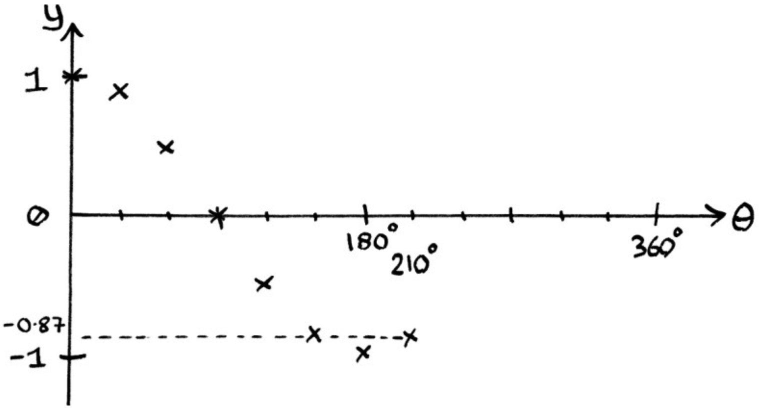
Then 180 degrees:



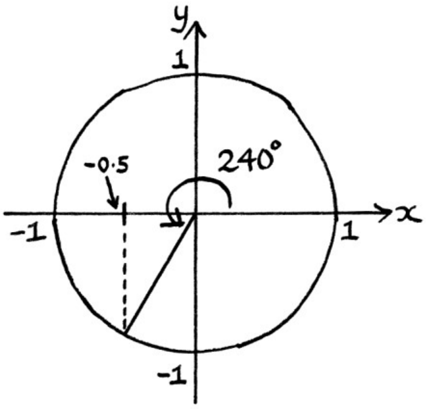


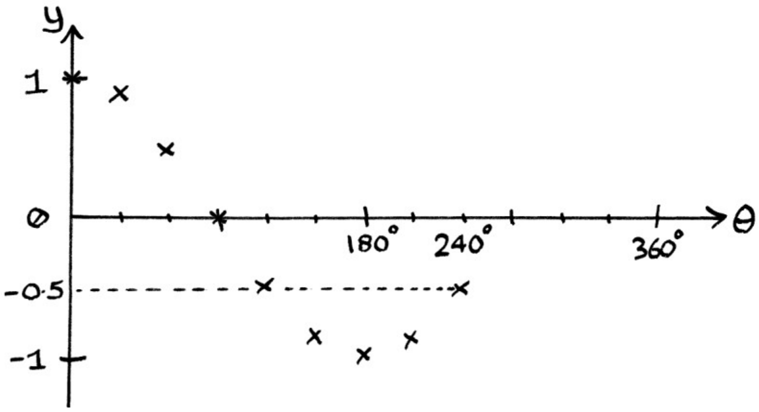
Then 210 degrees:



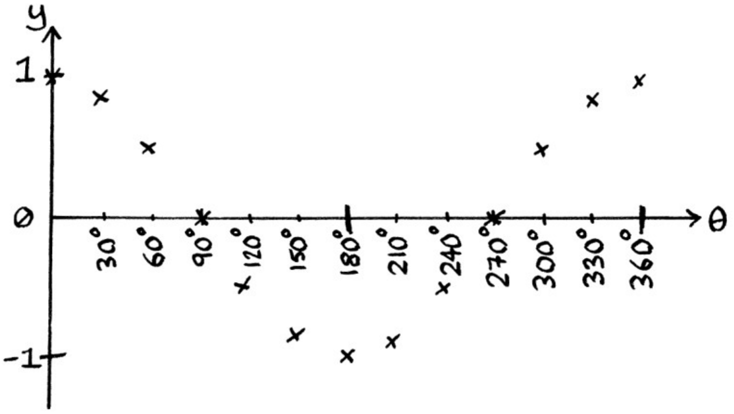


Then 240 degrees:

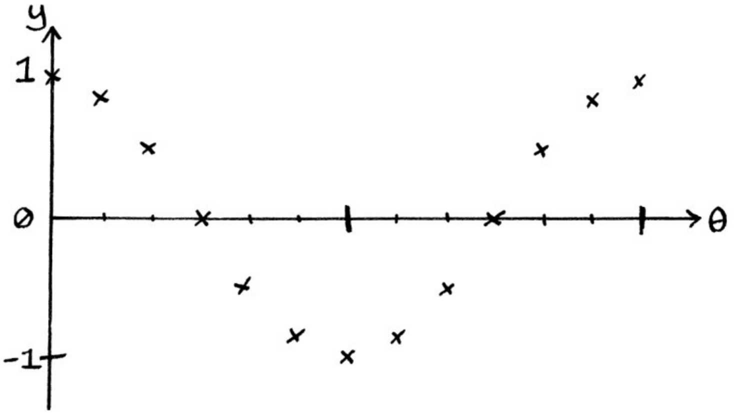




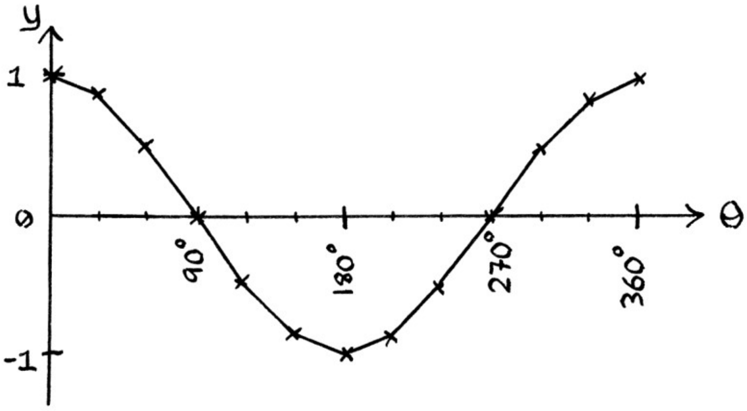
We continue around the circle until we have plotted the x-axis values for all the angles from 0 to 360.



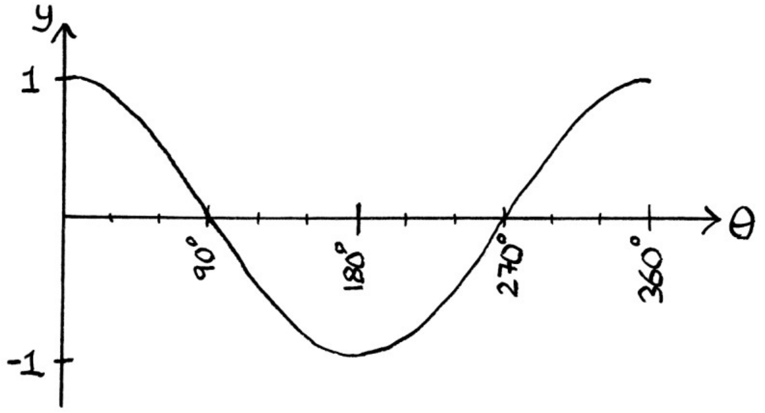
The graph is clearer without numbering on the θ-axis:



We then join up the points.



If we had picked angles that were closer together, the curve would be smoother. If we had picked an infinite number of angles and could read from the circle with perfect accuracy, the wave would be perfect:

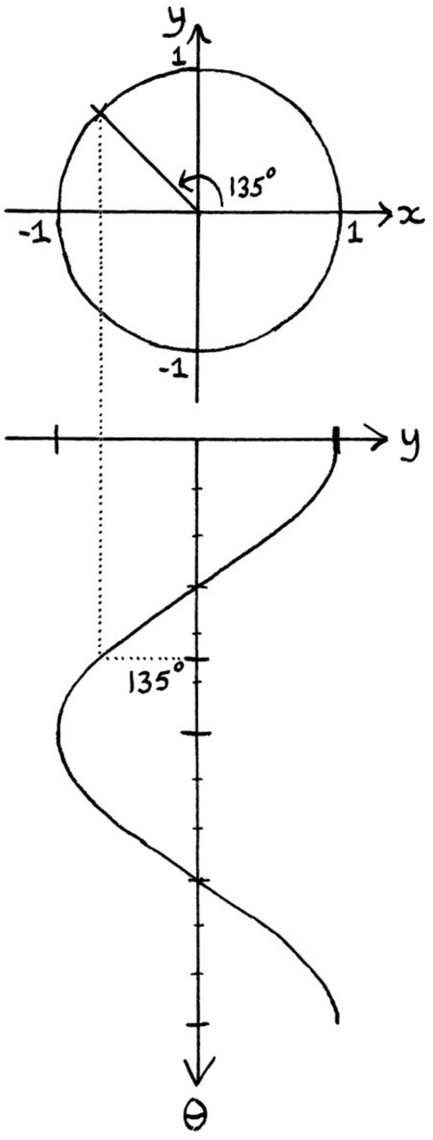


The graph turns out to be a wave, and by definition, it is a Cosine wave.

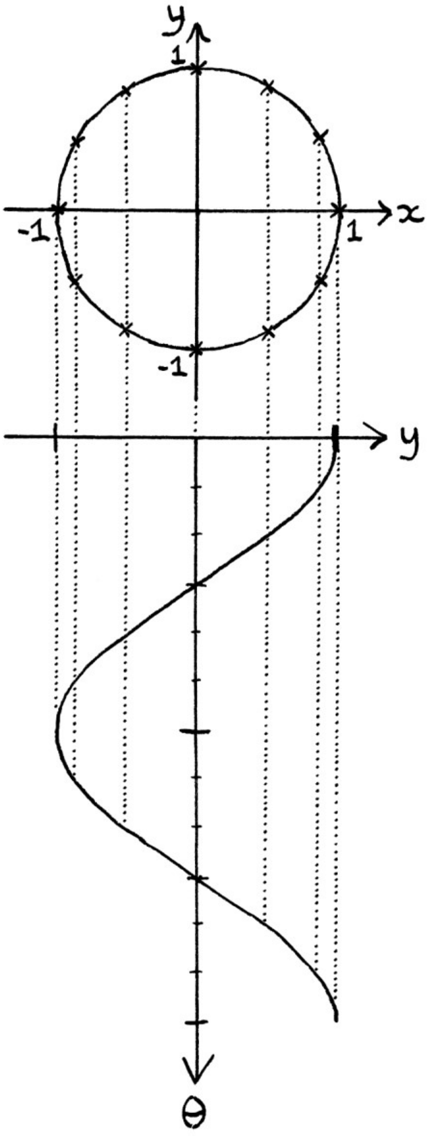
The Cosine wave is really just a graph showing the x-axis values of points on a circle's edge at evenly spaced angles from the origin.

When we had completed a Sine wave, we could place it next to the circle from which it was derived and the y-axis heights of the corresponding points on the circle and graph were the same. For the Cosine wave, this cannot be done as the Cosine of angles appears as the y-axis on the *wave*, but the x-axis on the *circle*. However, the Cosine wave can be rotated, and then it aligns with the x-axis of the circle, as we will see on the next page.

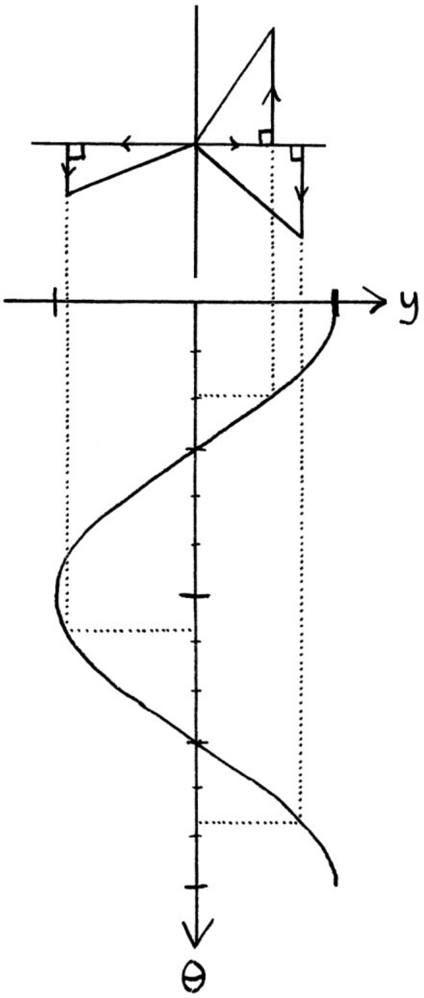
The following picture shows an example angle of 135 degrees marked on both the circle and the graph.



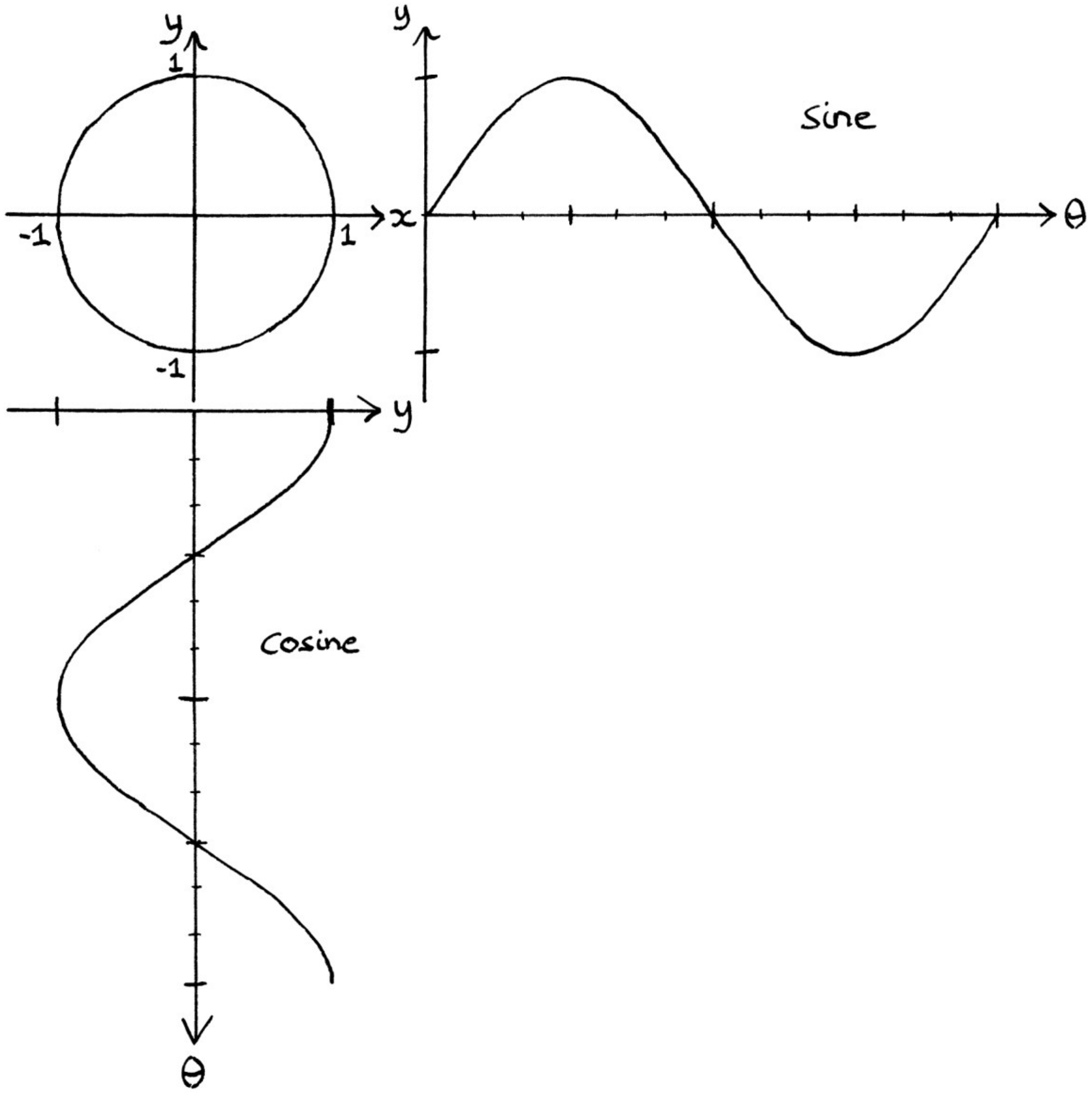
The rotated Cosine wave makes it easier to see how all the Cosine values on the circle have corresponding places on the Cosine wave:



The Cosine wave graph can also be thought of as representing the adjacent sides of an infinite number of right-angled triangles, facing to the right and left, and upwards and downwards:



Here are the circle, the Sine wave and the Cosine wave together:

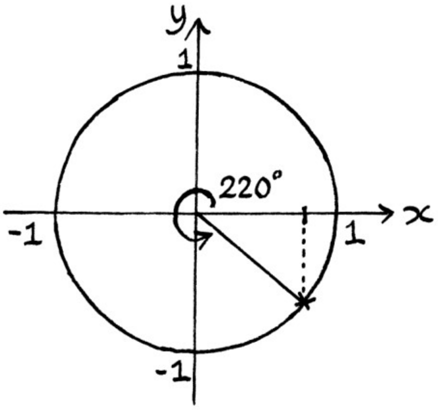


In the above picture, the x-axis values on the circle align with the y-axis values on the Cosine wave, and the y-axis values on the circle align with the y-axis values on the Sine wave.

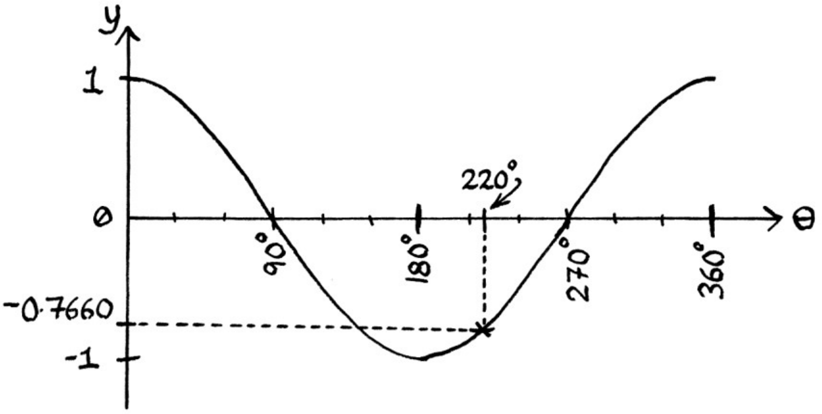
Generally, the Cosine wave would be drawn horizontally, and not vertically.

We can use both the circle and the Cosine wave graph to calculate the Cosine of numbers. On the circle, we draw a line from the origin at the angle for which we want the Cosine, and then read off the x-axis value where that line meets the circle's edge. On the Cosine wave graph, we read the y-axis value at the angle on the θ-axis for which we want the Cosine.

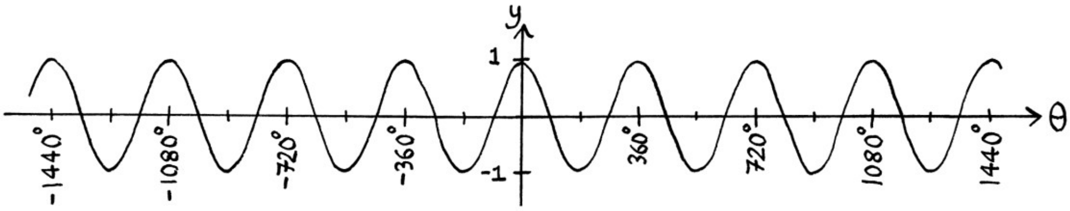
For example, to find the Cosine of 220 degrees on the circle, we draw a line at that angle until it meets the edge of the circle. Presuming we could read the circle very accurately, we would find that x-axis value of that point is −0.7660 units:



To find the Cosine of 220 degrees on the wave graph, we would read off the y-axis value of the curve where the θ-axis is 220 degrees. If we could read the Cosine wave very accurately, we would also see the result as −0.7660 units.



As with the Sine wave graph, the Cosine wave graph can be drawn showing the Cosine of angles over 360 degrees, and the Cosine of negative angles:

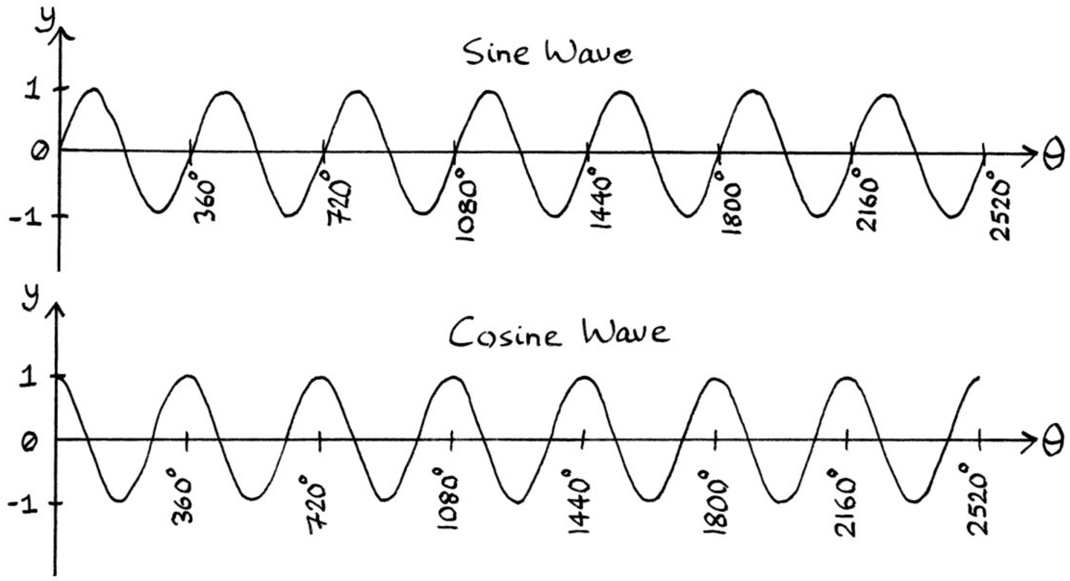


When drawing the Cosine wave in this way, it should be very apparent how similar it is to the Sine wave.

In these examples for drawing Sine and Cosine graphs, we could have had graphs that were more accurate, and more quickly, if we had typed the angles into a calculator instead of measuring a circle. However, if we had started in that way, it would not have reinforced the relationship between the circle and the wave graphs. One of the most important things to understand is what Sine and Cosine really are, and early use of a calculator when learning about them makes them much harder to understand.

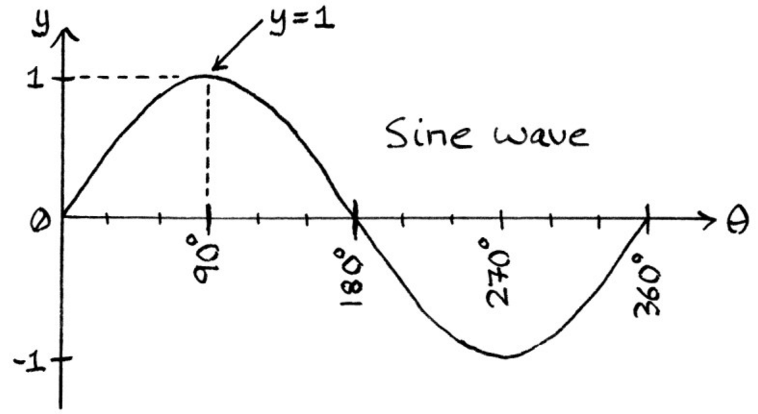
Sine and Cosine graphs

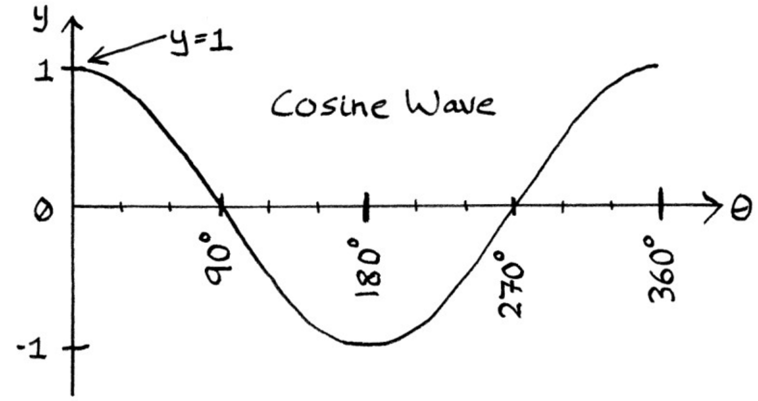
The Sine wave graph and the Cosine wave graph are identical in shape, although the curves sit at slightly different points along the θ-axis.



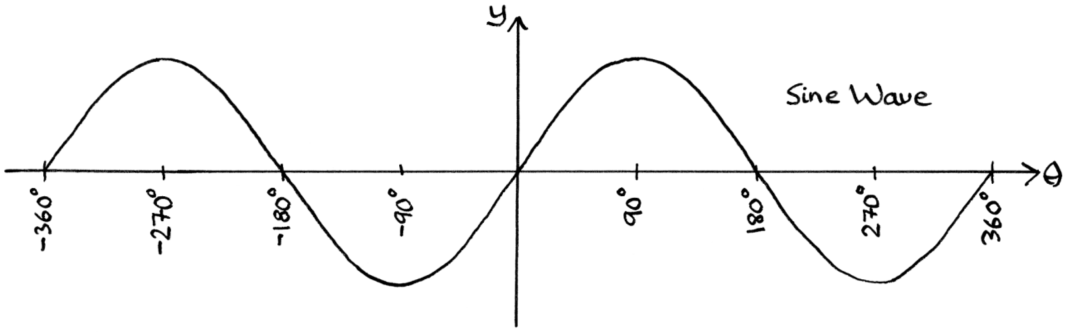
They have the same shape because we are actually measuring the height or the width of the circle at different points around its edge, and the circle has the same shape all the way around. If the circle were rotated 90 degrees, then its height would become its width and vice versa.

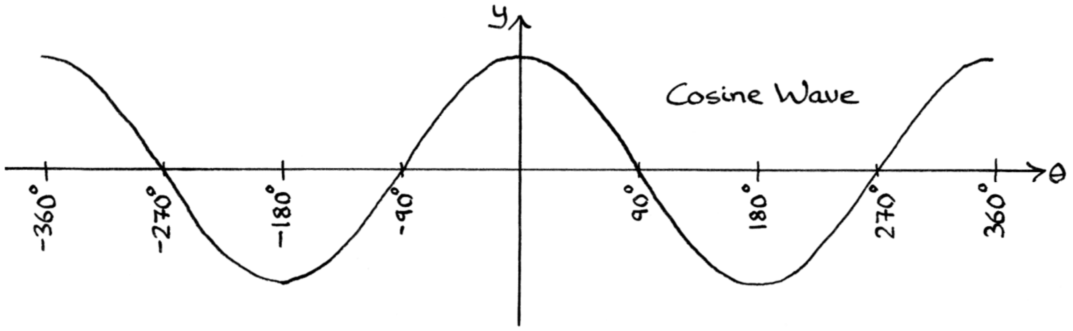
The Cosine wave graph is identical to the Sine wave graph in shape, but the y-axis points occur 90 degrees along the θ-axis sooner. For example, on the *Sine wave* graph, the y-axis value is 1 when the θ-axis is at 90 degrees. However, on the *Cosine wave* graph, the y-axis value is 1 when the θ-axis is 0 degrees. The y-axis value is 1 on the Cosine wave graph 90 degrees before it is 1 on the Sine wave graph.





As well as saying that the *Cosine wave*'s points occur 90 degrees sooner along the θ-axis than those of the Sine wave, we could just as easily say that the *Sine wave*'s points appear 270 degrees sooner than those of the Cosine wave. They are both true statements, and they mean the same thing as we are dealing with points around a circle's edge, and a circle's edge has no true beginning or end.





Just as the x and y-axes of the circle are at 90 degrees to each other, so are the two waves 90 degrees different from each other along the θ-axis. The 90-degree difference is a significant fact to observe.

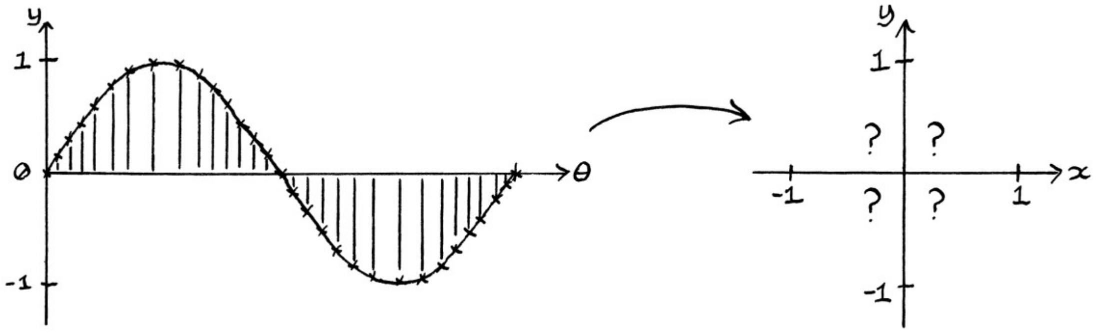
One point to note from all this is that the characteristics of our circle, and the angles of lines going from the origin to its edge, have resulted in two separate graphs. Although they are similar, being Sine and Cosine graphs, the graphs are each portraying a different aspect of our circle. The Sine wave is portraying the y-axis values of points around the circle's edge that are at equally spaced angles from the origin; the Cosine wave is portraying the x-axis values of points around the circle's edge that are at equally spaced angles from the origin. To put it another way, the Sine wave is portraying the y-axis values of equally spaced points around the edge of the circle; the Cosine wave is portraying the x-axis values of equally spaced points around the edge of the circle. As lines drawn out from the origin at evenly spaced angles divide the circle's edge into evenly sized sections, these are two ways of saying exactly the same thing.

[It is always worth remembering that the heights or widths indicated by Sine and Cosine are based on points at evenly spaced *angles* around the circle (or to put it another way, at evenly spaced points around the *edge* of the circle). The points are not evenly spaced along the x-axis or y-axis. If they were, then such a graph would just be one semicircle followed by an upside down semicircle]

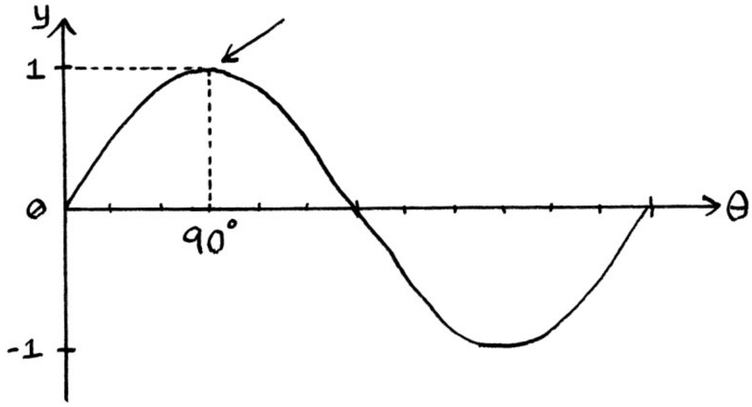
Although we might see Sine and Cosine waves in maths and signal processing without any reference to circles, it *always* pays to think of the circles that the waves either came from or *could* have come from. The relationship between Sine, Cosine and the circle is the most important concept to take on board in thinking about waves and signals. It is a concept that subtly underlies *everything* in the world of waves, yet is also one that is often overlooked.

Forgotten circles

Suppose, for some strange reason, we forgot how to draw a circle. Perhaps, we could use our Sine wave graph to recreate one. The Sine wave graph is really a portrayal of the y-axis heights of points on the circle's edge at evenly spaced angles from the origin, so that might be enough information to draw a circle. Perhaps we could take evenly spaced points along the θ-axis of our Sine wave graph and read off the relevant y-axis values, then plot those as heights on a new set of axes.

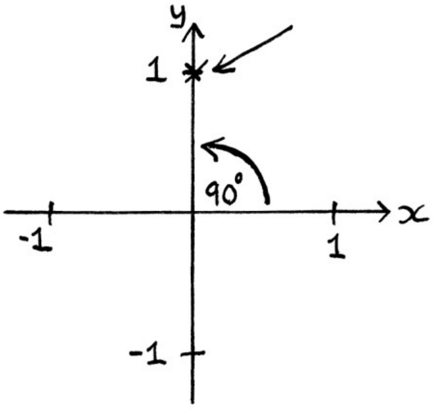


To demonstrate how this *might* be possible, we will look at our Sine wave graph, and read the y-axis value when the θ-axis value is at 90 degrees: when "θ" is 90 degrees, the y-axis value is +1 unit.

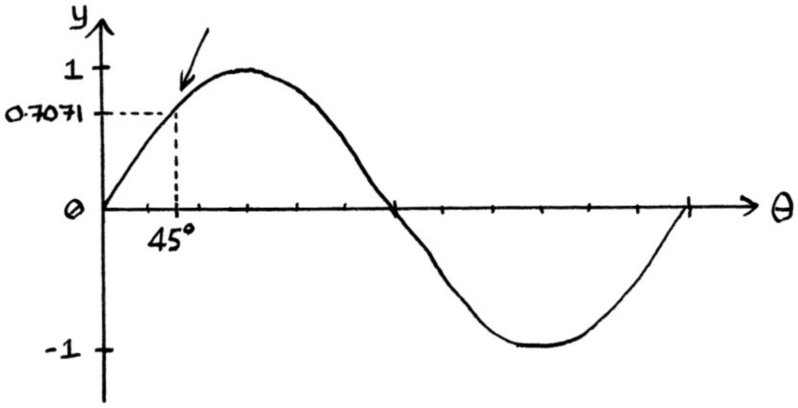


Therefore, on our future circle chart, we know that there is a point at 90 degrees that has a height of +1 unit.

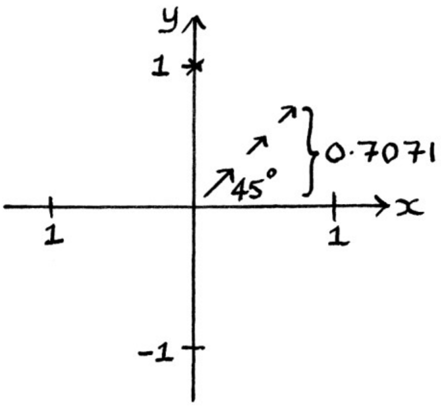
Therefore, we can mark such a point on our future circle chart:

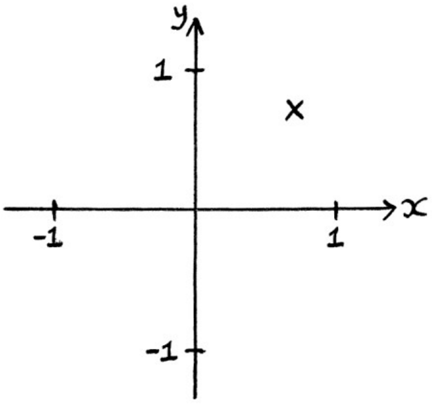


We can then read off more values from the Sine wave graph. From our Sine wave graph, we can see that when "θ" is 45 degrees, "y" is +0.7071 units.

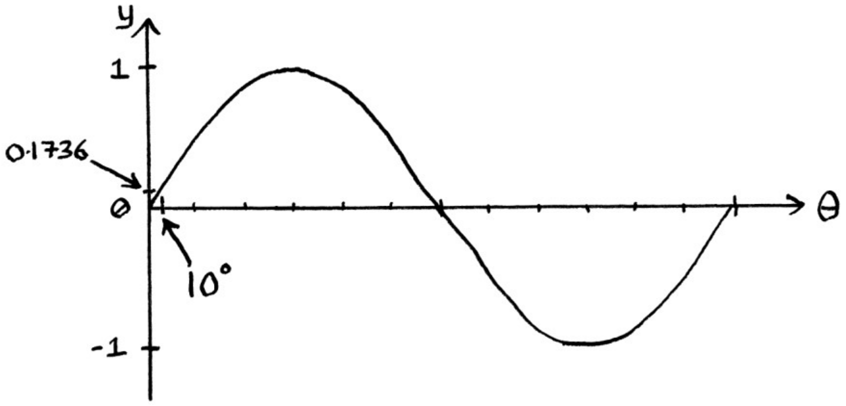


This means that we will draw a line outwards at 45 degrees from the origin on our future circle chart until the height of the end is +0.7071 units high. This gives us a second point.

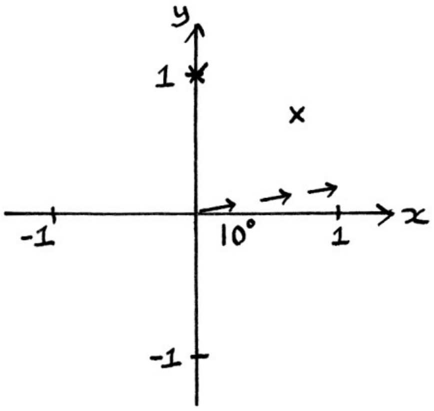


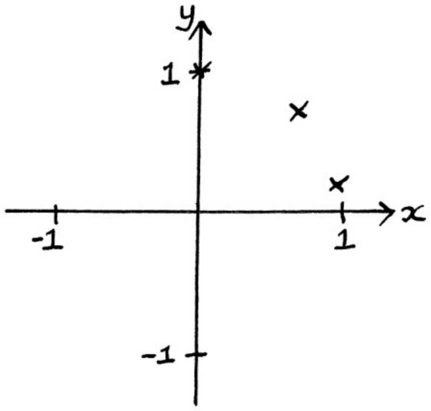


As an example of another point, on the Sine wave graph, when "θ" is 10 degrees, "y" is +0.1736 units.



Therefore, we draw a line at 10 degrees from the origin on our future circle chart until the height of the end of the line is +0.1736 units high. That gives us a third point.

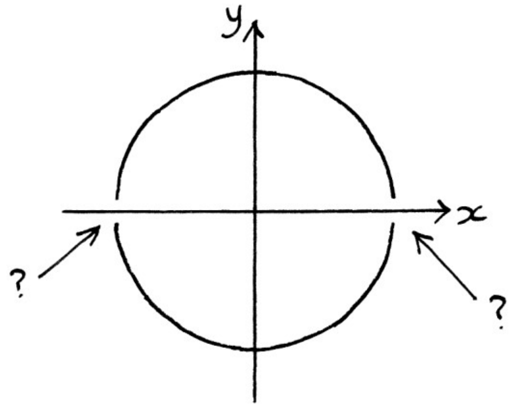




We can continue plotting points in this way – reading heights and angles from our Sine wave graph, and then plotting points around the future circle chart that are at that angle and have that height.

The process still works for the negative values on the Sine wave graph – we still draw the line at the appropriate angle on the circle chart until the end of the line is at the height of the y-axis value from the Sine wave graph.

However, we will meet a problem when we have an angle on the θ-axis for which the y-axis value is 0 units. This happens at 0 degrees and 180 degrees (and 360 degrees). It is not possible to mark a point on the future circle chart by drawing a line at an angle until the height of its end is zero units, because the height of any point on that line will be zero to start with. This means that either the point ends up at the origin (0, 0), which cannot be correct, or it ends up literally anywhere along the zero degree line. This problem means that we cannot complete the circle. We will end up with gaps at 0 degrees and 180 degrees:



We could read points off the Sine wave graph at 0.00001 and 359.99999 degrees or 179.99999 and 180.00001 degrees, and plot those, but we would still have a gap for the exact points at 0 and 180.

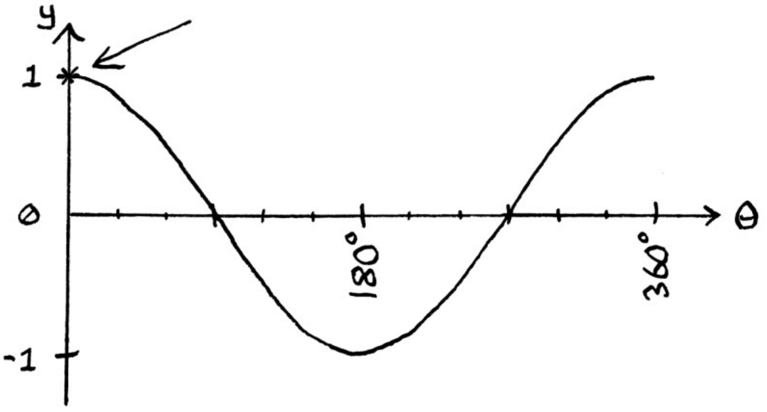
Obviously, in the real world, we might guess that the circle is joined up over those gaps, so this would not be a problem. We could also guess that given that all the other points around the circle are exactly 1 unit away from the origin, the points at 0 and 180 degrees are also likely to be 1 unit away, but in this example, we cannot know this for sure. We have completely forgotten what a circle looks like, so we cannot assume anything about those gaps.

Given this problem, we can see that the Sine wave graph on its own does not contain enough information to reproduce a completely accurate circle.

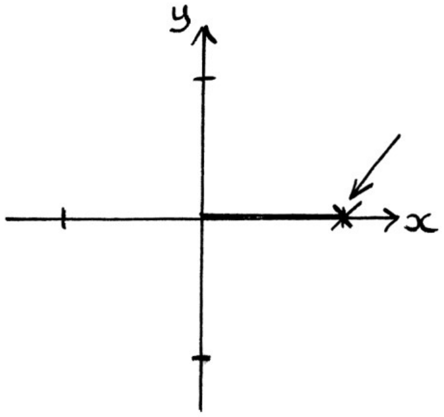
Cosine

When trying to draw a circle using a Cosine wave graph, instead of drawing lines at angles with the ends at particular heights, we draw lines at angles with the ends at particular *horizontal* distances away from the origin of the axes (in other words away from the y-axis or x = 0).

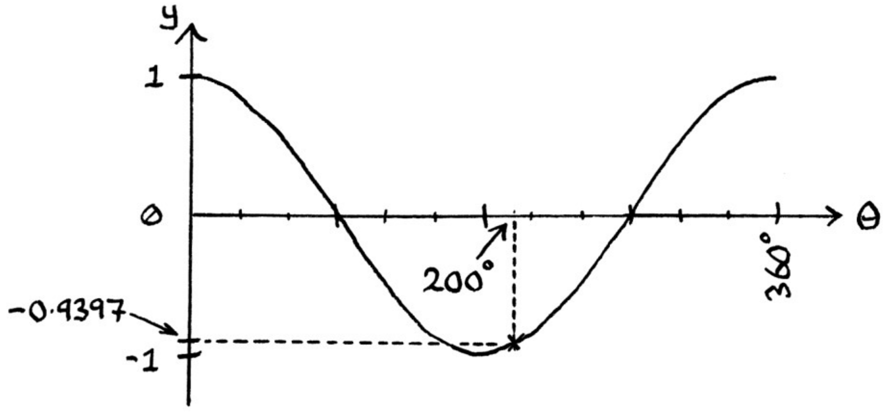
For example, on the Cosine wave, the x-axis value for an angle of 0 degrees is +1 unit:



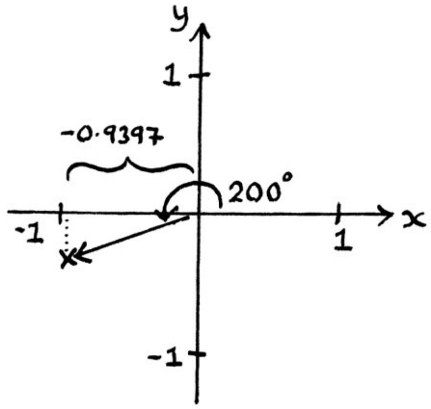
Therefore, on the future circle chart, we draw a line at 0 degrees, until its end is +1 units *horizontally* away from the y-axis.



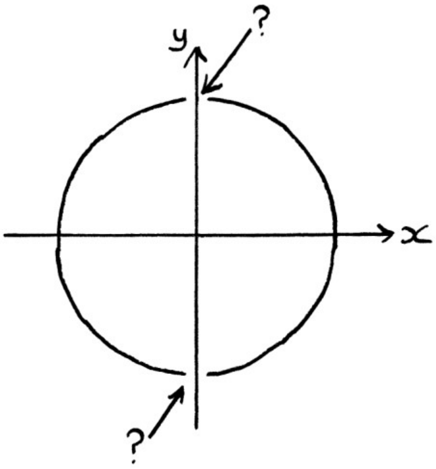
As an example of drawing another point, on the Cosine wave graph, the y-axis value when "θ" is 200 degrees is −0.9397 units:



Therefore, on the future circle chart, we draw a line at 200 degrees, until it is −0.9397 units away from the y-axis.



We can continue marking points around the future circle chart, but we would end up with a similar problem to the one we had with Sine – we would have gaps. However, now the gaps will be at 90 degrees and 270 degrees.

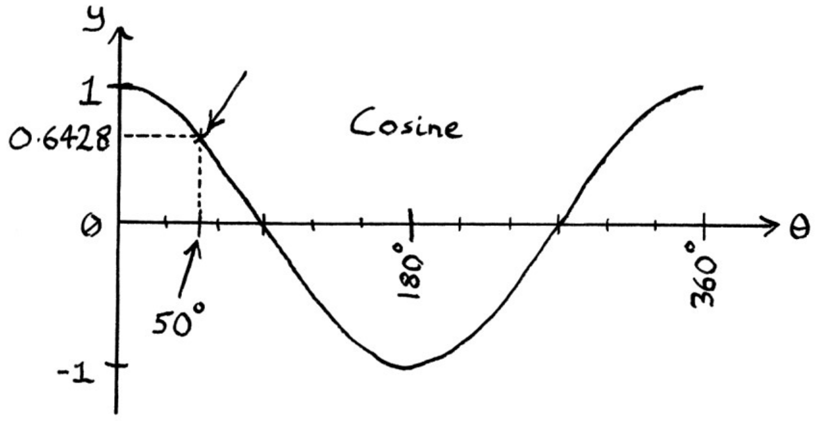


Therefore, the Cosine wave graph does not contain enough information to recreate the circle completely either.

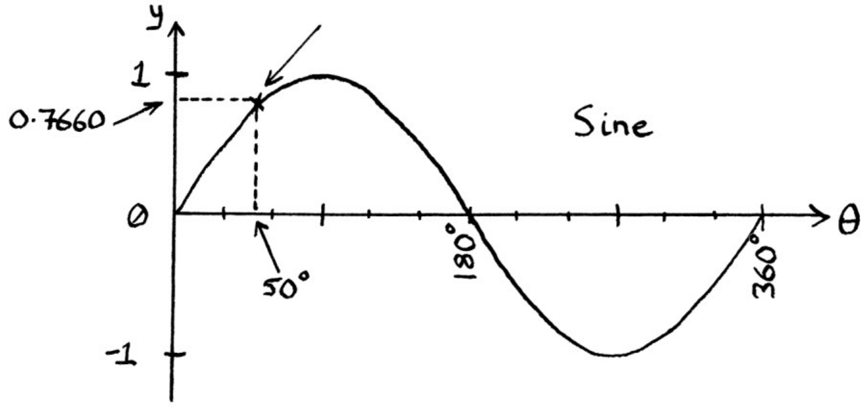
Sine and Cosine

It turns out that if we have both the Cosine and Sine parts of the circle (in the form of the Cosine and Sine graphs) we can recreate the circle perfectly. By reading off the y-axis values from both the Sine and Cosine graphs for the same angles, and then using them as coordinates on our future circle chart, with the Cosine part as the new “x”, and the Sine part as the new “y”, it is very easy to draw out a complete circle.

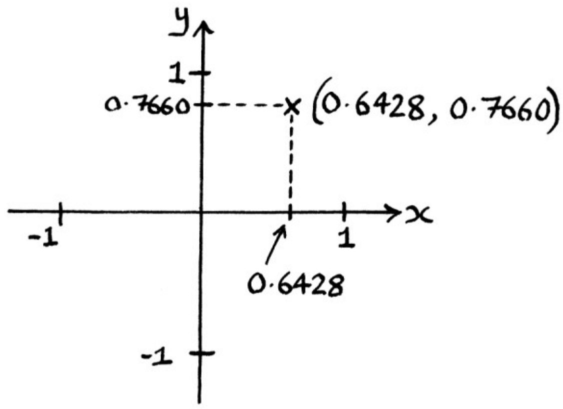
As an example, we will read off the y-axis value on our *Cosine* wave graph for when “θ” is 50 degrees. This gives “y” as +0.6428 units:



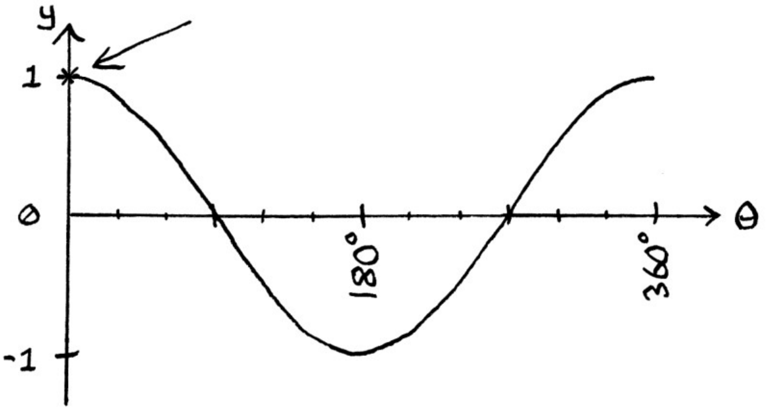
Then we will read the y-axis value on our *Sine* wave graph for when “θ” is 50 degrees. This gives “y” as +0.7660 units.



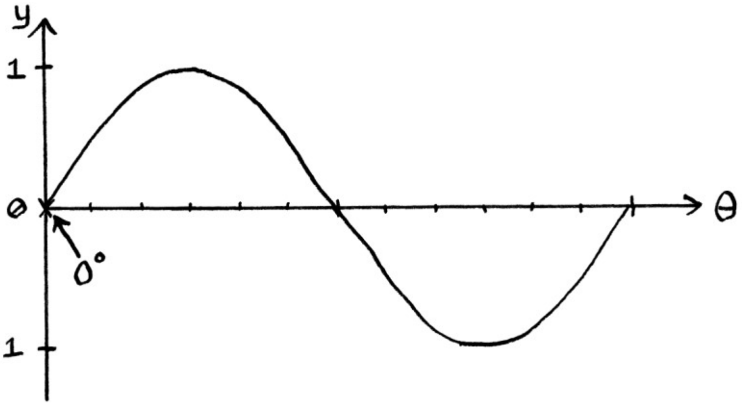
We will then plot a point on our future circle chart, with the “x” coordinate as 0.6428, and the “y” coordinate as 0.7660.



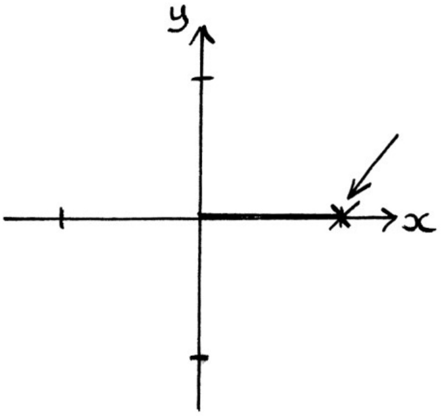
Every point on the circle can be plotted in this way, including the "missing" points from before. If we read the y-axis values on our Cosine and Sine graphs for when "θ" is 0 degrees, we will have 1 unit on the Cosine wave graph...



... and 0 units on the Sine wave graph.

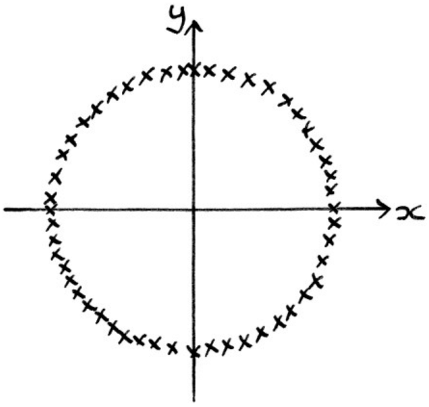


This gives us the coordinates of (1, 0), which we can plot on our future circle chart:

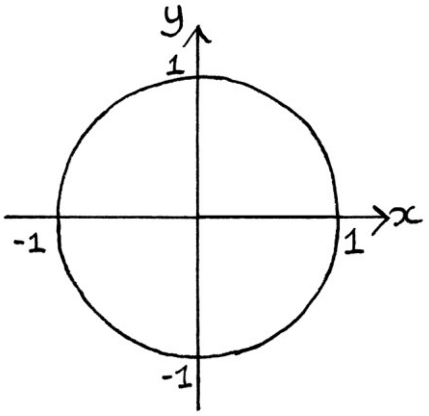


We can obtain points for our future circle chart at all the places where having just the Cosine graph or just the Sine graph would not have been enough. For example:
When "θ" is 90, we have the coordinates 0, 1
When "θ" is 180, we have the coordinates −1, 0
When "θ" is 270, we have the coordinates 0, −1.

We can keep plotting points from the two waves until our circle begins to take shape.



If we continued forever and could read the two waves with perfect accuracy, we would end up with a perfect circle.



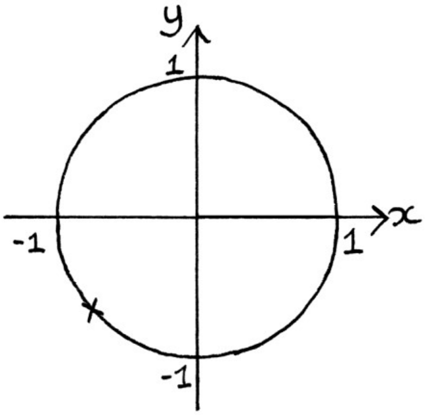
Although drawing a circle by reading the points from the Sine and Cosine waves is not particularly useful here, it helps reinforce the connection between the circle and the two waves. Within the original circle, there is the information about the construction of both the Sine wave and the Cosine wave. In either the Sine wave or the Cosine wave alone, there is only *most* of the information of the circle.

There can be situations when significant attributes of the circle are lost completely if only the Sine graph or only the Cosine graph is used alone – for example, if the circle is not centred on the origin of the axes.

The Cosine wave gives the circle its width; the Sine wave gives the circle its height. Supposing we were working with a Sine wave "in the wild", for example a radio wave, we would be able to understand it and work with it more easily if we could, first, treat the wave as one dimension of a circle, and second, have a corresponding wave to act as the second dimension of the circle. In other words, if we treated the wave as the Sine or height part of the circle, then having the Cosine or width part of the circle would be to our benefit. In the real world, radio and sound waves do not generally have a corresponding part to make up a circle, and for much of life this is not a problem – one wave alone is still sufficient to analyse, record and create signals. However, treating the wave as if it were derived from a circle, and thus treating it as if it were one of a pair of waves can be very useful.

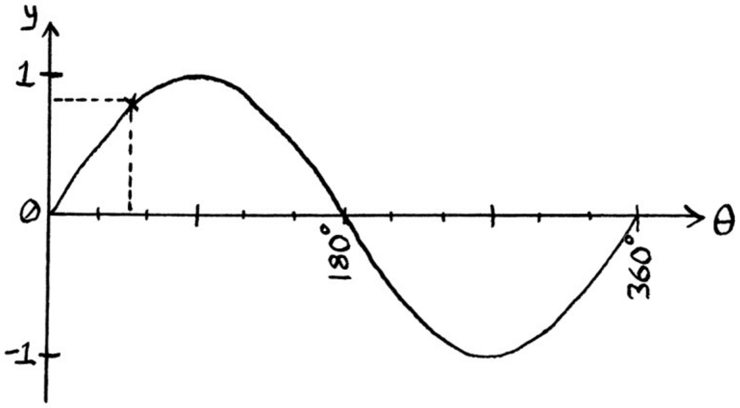
Remembered circles

Even though we cannot recreate the *entire* circle from a Sine wave on its own or a Cosine wave on its own, the fact is that we *do* know what a circle looks like. Therefore, unlike in the last example, when we had forgotten what a circle looked like, in the real world, we could use our knowledge of a circle to know what the values in the gaps should be. For example, we know that a circle's radius does not change – if a single point on the circle's edge is 1 unit away from the origin, then every point will be 1 unit away from the origin. We also know that a circle's edge does not have any gaps in it.

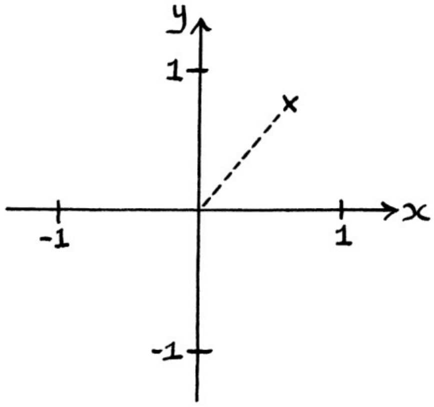


In practice, if we know the circle is centred on the origin of the axes, we can recreate the entire circle, as we are using it here, with just *one* non-zero point taken from either wave graph.

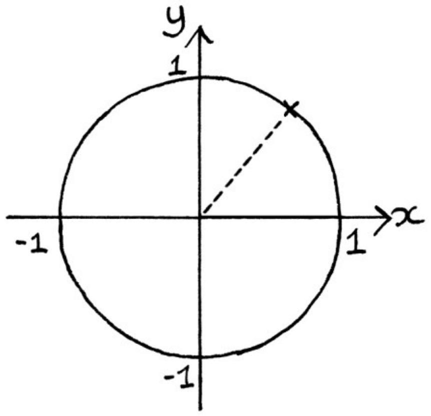
Supposing we are reading from a Sine wave graph, we read off the y-axis value at *any* value of “θ” as long as that y-axis value is not zero:



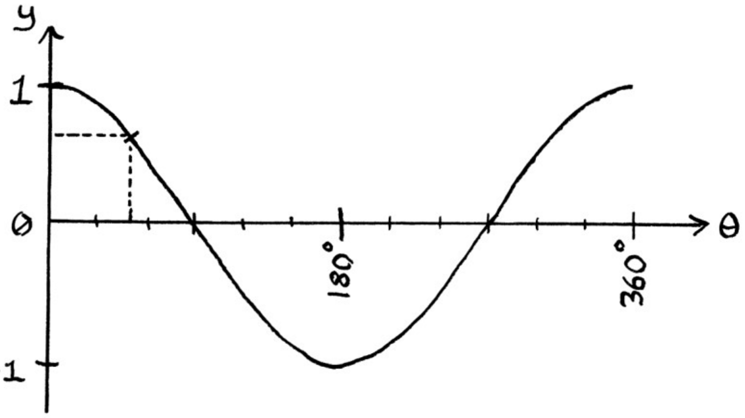
Then, on our blank circle chart, we plot that y-axis value by measuring out from the centre of the axes at that angle until we reach a point that has the height of the value we read from the graph:



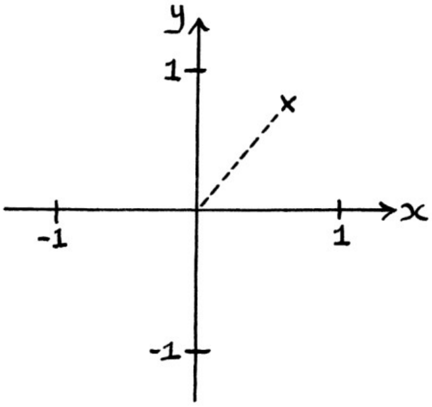
The circle’s edge will pass through this point. As all points on a circle are the same distance from its centre, if we have one point, we know enough to draw the full circle. Therefore, we just need to draw a circle, centred on the origin, that passes through this point.



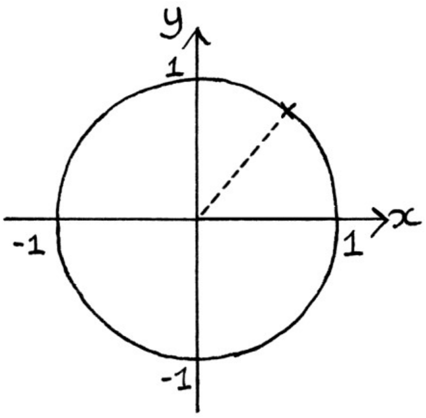
If we were reading from a Cosine wave graph, we would read off one non-zero y-axis value at any value of "θ" from the graph...



... and then plot that point by drawing a line at the angle of "θ" so that the *horizontal* distance to the end is equal to that Cosine wave's y-axis value.



Then we draw a circle that goes through that point.



Note that this method of recreating the circle from just one point from a wave might not work for waves that have more complicated attributes.

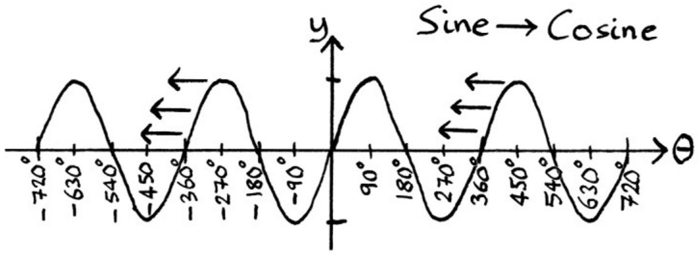
Recreating missing waves

As we know, a circle with a unit radius implies the existence of a particular Sine wave and a particular Cosine wave. In other words, if someone gives us a unit-radius circle, we could work out what the Sine and Cosine waves would look like. Another point is that those Sine and Cosine waves are unique for a circle with that size of radius. Given that, if we see a particular Sine wave, we know what its circle would look like (if that circle is centred on the origin of the axes), and if we see a particular Cosine wave, we know what its circle would look like (if that circle is centred on the origin of the axes). This means that if we are given a Sine wave, we can work out what its circle looks like, and then work out what that circle's Cosine wave looks like. Similarly, if we are given a Cosine wave, we can work out what its circle looks like, and then work out what that circle's Sine wave looks like.

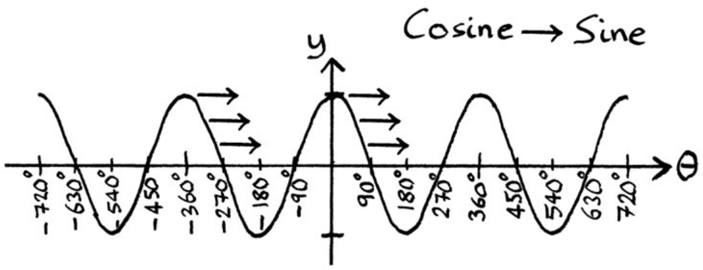
In other words, a particular Sine wave or Cosine wave implies the existence of a particularly sized circle, centred on the origin, and that circle then implies the existence of a particular Cosine wave or Sine wave. Therefore, a particular Sine wave implies a particular Cosine wave, and a Cosine wave implies a particular Sine wave. This means that if we only have a Sine wave, we can recreate its Cosine "twin", and if we only have a Cosine wave, we can recreate its Sine "twin".

There is also a another way of finding the missing Cosine wave or missing Sine wave derived from the same circle. If we presume that the circle from which the waves are derived is centred on the origin of the x and y-axes, then another way of finding the missing "twin" wave is by shifting the wave we have by 90 degrees along the θ-axis.

If we have the Sine wave from a circle, we shift it 90 degrees to the *left* to find the Cosine wave. This means we take every single point along the wave and redraw it so it appears 90 degrees earlier.



If we have the Cosine wave from a circle, we can shift it 90 degrees to the right to find the Sine wave. In other words, we take every single point along the curve and redraw it so it appears 90 degrees later.



When we have both waves, we can recreate the circle that they would have come from by using the points from each wave as coordinates.

The shifting idea is another way of realising how Sine (θ + 90) is equal to Cosine θ, and how Cosine (θ − 90) is equal to Sine θ, as discussed in Chapter 2.

On some occasions, we might not know whether the wave we have is a Sine wave or a Cosine wave. In such situations, the best we can do is treat the wave as if it were either a Sine wave or a Cosine wave, and shift it 90 degrees up or down the θ-axis to create the other wave. Doing this is acceptable if the result does not need to be faithful to reality, and there is no other choice.

Note that if we are dealing with waves that are not "pure" Sine or Cosine waves [as defined later in this book – essentially waves that could not be said to be derived from a circle], then this will not work. In that case, a wave must be treated as the sum of various Sine or Cosine waves, and it will need to be split into its constituent parts first. Then, each part can have its corresponding Sine or Cosine wave recreated, and there will be one circle for each part. This concept is looked at in more depth later in this book.

If the circle from which the waves are derived was not centred on the origin of the axes (that is to say the point where x = 0 and y = 0), the position of the reconstructed circle might not be in the right place on the axes.

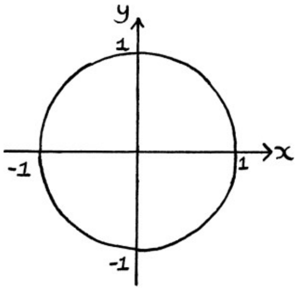
The concept behind recreating missing parts is really quite simple, but also quite significant. For waves that are derived from, or could have been derived from, circles centred on the origin, any Sine wave implies exactly how its corresponding Cosine wave will look, and therefore, it implies how its circle will look. Similarly, any Cosine wave implies exactly how its corresponding Sine wave will look, and therefore implies how its circle will look too.

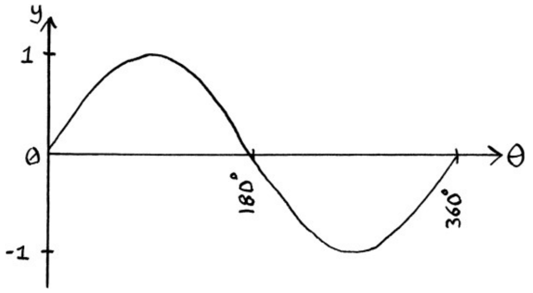
Because it is possible to recreate missing corresponding waves and the circle, it is also the case that any way of describing an individual wave is also a way of describing its corresponding wave and its circle (if the circle was originally centred on the origin). Also, any way of describing a circle is also a way of describing both its waves.

Circles with radiuses other than 1 unit

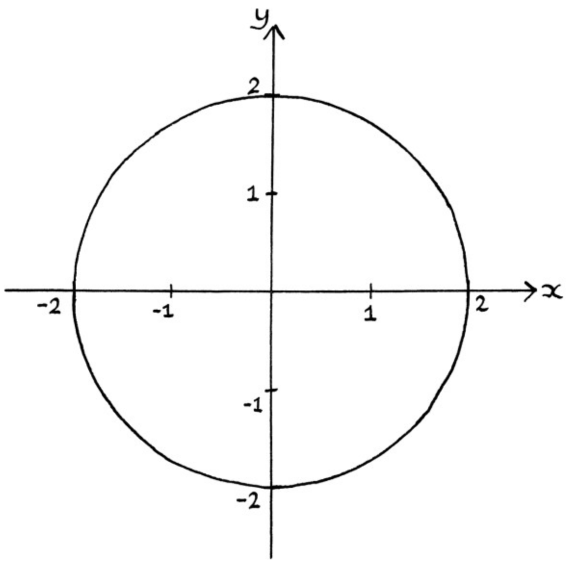
A circle that has a radius of a length other than 1 unit will produce a Sine wave and a Cosine wave that are scaled in *height* to the same proportions. In other words, if a circle has a radius of 2 units, then the Sine and Cosine waves will have peaks and dips that are twice as high as if the circle had a radius of 1 unit. If a circle has a radius of 0.5 units, the Sine and Cosine waves will have peaks and dips that are half as high as if the circle had a radius of 1 unit.

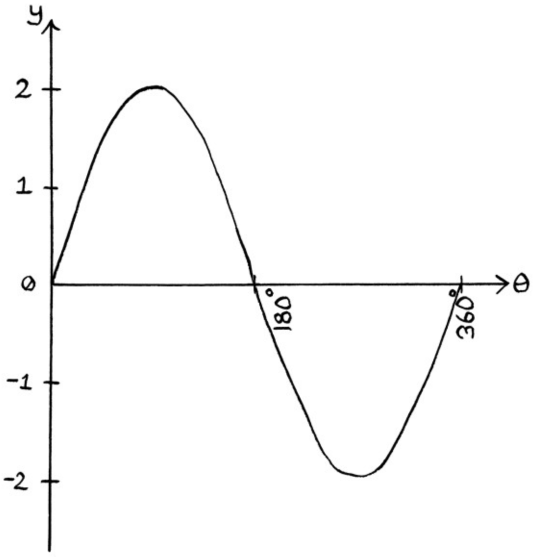
A radius of 1 unit:



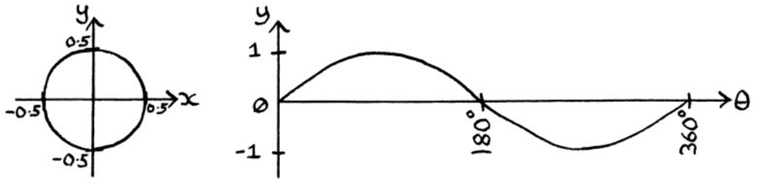


A radius of 2 units:





A radius of 0.5 units:



You might be able to guess that this would be the case by how, if the circle has a particular radius, the maximum and minimum points on the resulting wave will be equal to the maximum and minimum points on the circle. The maximum y-axis value on a circle will match the maximum y-axis value on a Sine wave, and the maximum x-axis value on a circle will match the maximum y-axis value on a Cosine wave.

Another way of thinking about this is by realising how the hypotenuses, adjacent sides and opposite sides of right-angled triangles all scale in proportion. Scaling the radius of a circle is the same as scaling the hypotenuses of the triangles that make up that circle, and so the opposite and adjacent sides will scale by the same amount. Therefore, the Sine wave derived from that circle, which represents the opposite sides of those triangles, will scale in proportion, and the Cosine wave, which represents the adjacent sides of those triangles, will scale in proportion too.

Conclusion

The main points to remember from this chapter are as follows:

A Sine wave graph portrays the y-axis values of points around a circle’s edge that are at evenly spaced angles from the origin. This is the same as saying a Sine wave graph portrays the lengths of the opposite sides of the right-angled triangles that make up that circle with reference to a range of evenly spaced angles of interest.

A Cosine wave graph portrays the x-axis values of points around a circle’s edge that are at evenly spaced angles from the origin. This is the same as saying a Cosine wave graph portrays the lengths of the *adjacent* sides of the right-angled triangles that make up that circle with reference to a range of evenly spaced angles of interest.

A circle contains the information for both a Sine wave and a Cosine wave.

A Sine wave and its corresponding Cosine wave together can be used to recreate the circle if corresponding points are taken from each and used as coordinates.

For circles that are centred on the origin of the x and y-axes:

- A Sine wave has an implied corresponding Cosine wave. In other words, a Sine wave can be used to recreate its corresponding Cosine wave.

- A Cosine wave has an implied corresponding Sine wave. In other words, a Cosine wave can be used to recreate its corresponding Sine wave.

- Therefore, a Sine wave or a Cosine wave on its own contains enough information to recreate the circle.

- Therefore, to describe a particular circle, its Sine wave, or its Cosine wave, it is only necessary to have the details of that circle, the details of that Sine wave, or the details of that Cosine wave, as each of the parts is enough to reconstruct the others.

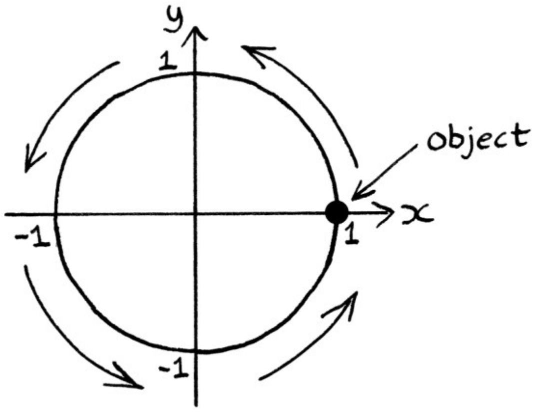
It is common to think of Sine and Cosine waves as entities in their own right, but it always pays to remember that they display attributes of a circle. Even if a given wave does not seem to have had any connection with a circle, it always makes sense to imagine the hypothetical circle that it *could* have come from. The relationship between Sine, Cosine and the circle is the most important aspect in this entire book.

Chapter 4: Time

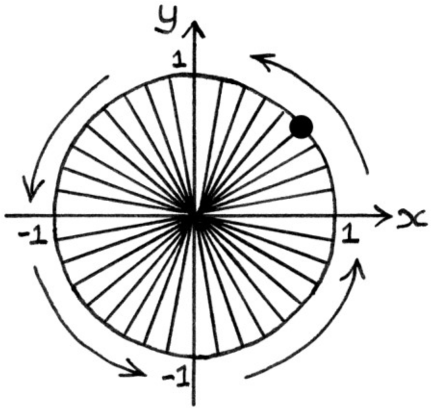
So far, we have dealt with degrees around a circle, and the corresponding Sine and Cosine waves. Now, we will introduce the concept of time. Everything will be similar to before, but more information will be represented.

The time circle

We will imagine an object that moves around the circumference of a unit-radius circle, so that it completes one revolution every second. It starts at an angle of 0 degrees, moves anticlockwise, and after one second, it is at an angle of 360 degrees.



In other words, if we think of angles, it completes 360 degrees every second.



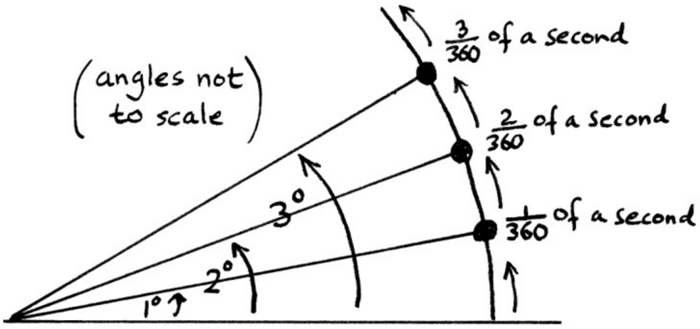
We can also say that it travels one degree every 360th of a second.

When the object starts, it is at 0 degrees in relation to the centre of the circle. We could also say that when the time is 0 seconds, the object is at 0 degrees.

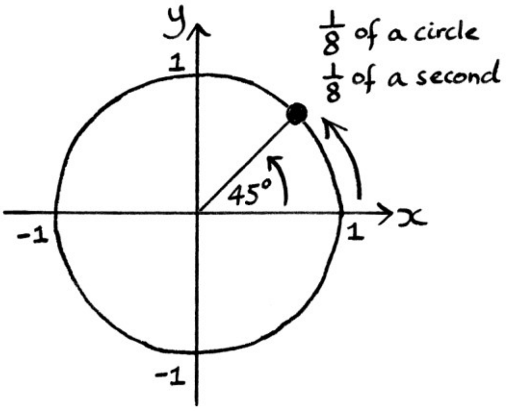
After one 360th of a second has passed, the object will have moved to be at 1 degree.

After two 360ths of a second, it will be at 2 degrees.

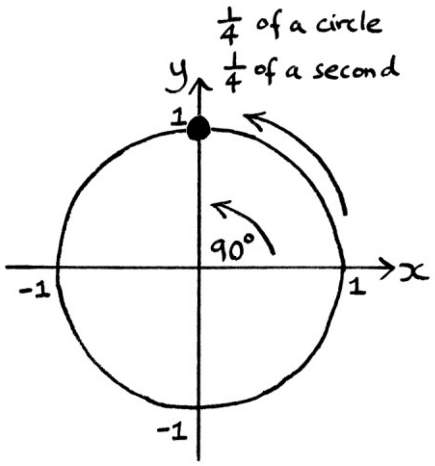
After three 360ths of a second, it will be at 3 degrees, and so on.



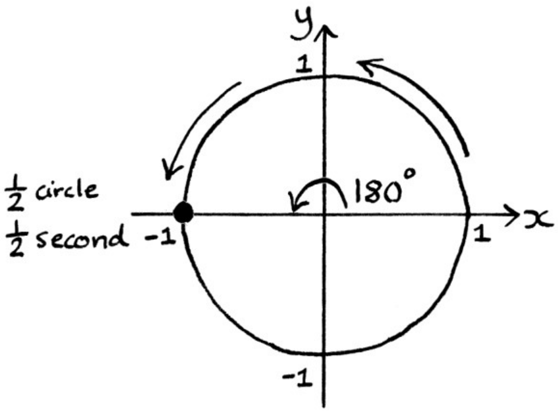
After forty-five 360ths of a second (an eighth of a second), it will be at 45 degrees. In other words, after an eighth of a second, it will have travelled an eighth of the way around the circle.



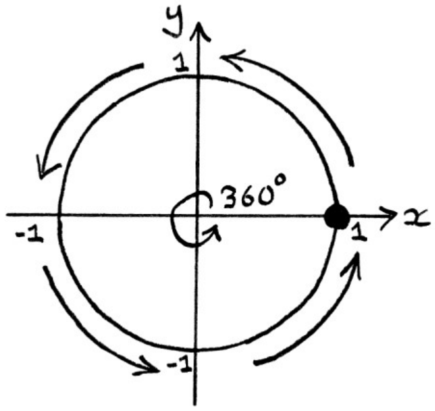
After ninety 360ths of a second (0.25 seconds), it will be at 90 degrees. In other words, after a quarter of a second, it will be a quarter of the way around the circle:



After one hundred and eighty 360ths of a second (half a second), it will be at 180 degrees, which is half way around the circle.

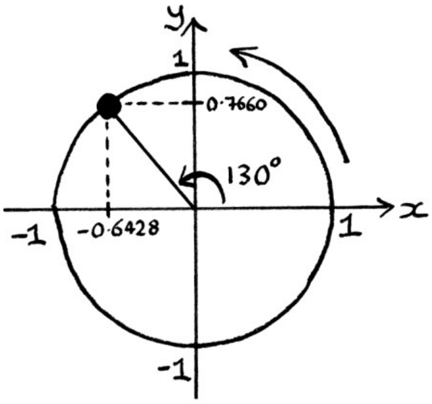


The object moves around the circle's edge, and after three hundred and sixty 360ths of a second (1 second), it will have completed 360 degrees. It will have completed one full rotation of the circle, and it will have achieved it in one second.

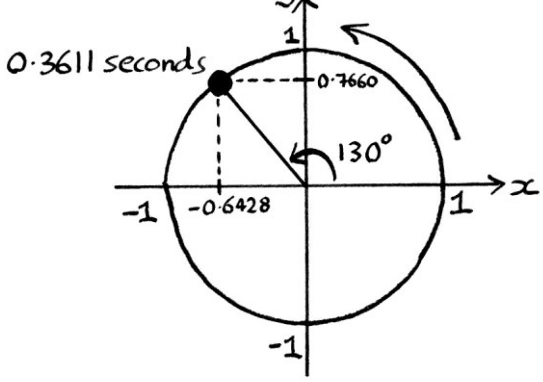


Supposing the object continued to rotate around the edge of the circle at the same speed, it would pass through exactly the same points on the second second, and on the third second, and on later seconds. It would travel the second revolution in the duration of the second second, the third revolution in the duration of the third second and so on.

For the purposes of Sine and Cosine, the circle still works in the same way. We can calculate the position of our object (its y-axis and x-axis values) in the same way as we did before we included time. For example, if our object is at 130 degrees, then the Sine of 130 degrees is 0.7660 units, and the Cosine of 130 degrees is −0.6428 units. Therefore, our object has a y-axis value of 0.7660 units and an x-axis value of −0.6428 units. Swapping these values around, we can give its coordinates as (−0.6428, 0.7660).



As we are now paying attention to time, we can also say that we are finding the object's position at a particular *time*. Given that the object is travelling at one full circle per second, which is 360 degrees per second, which is also 1 degree per one 360th of a second, we can tell that when the object is at 130 degrees, it will have been moving for 130 * (1 ÷ 360) = 0.3611 seconds.



In fact, when the object is at any chosen angle, the time in seconds is just that angle multiplied by 1/360, or to put it more succinctly, that angle divided by 360.

We can calculate when the object will be at a particular angle, and we can calculate at which angle the object will be at any particular time. The angle and the time are linked together.

Another way of thinking about all this is that the angle of the object at any moment can be said to be a percentage of the circle. At any percentage around the circle, the time will be the same percentage of a second. In other words, at 60% of a second, the object will be 60% of the way around the circle. At 10% of a second, the object will be 10% of the way around the circle. At 78.3% of the way around the circle, the time will be 78.3% of a second. As we are dividing a circle up into 360 divisions (i.e. degrees) instead of 100 divisions, it makes sense to divide the time up into 360 divisions too. Therefore, at four 360ths of a circle, in other words, at 4 degrees, the time will be four 360ths of a second, and so on.

Examples

At 0 seconds, the object will be at an angle of 0 degrees. Its x-axis value will be 1 unit, and its y-axis value will be 0 units. In other words, the Cosine of its angle will be 1 unit, and the Sine of its angle will be 0 units. Its coordinates will be (1, 0).

At one 360th of a second, the object will be at an angle of 1 degree. The Cosine of its angle, and therefore, its x-axis value will be 0.9998 units. The Sine of its angle, and therefore, its y-axis value, will be 0.01745 units. Its coordinates will be (0.9998, 0.01745).

At 0.38 seconds, the object will be 0.38 of the way around the circle. Therefore, it will be at an angle of 0.38 * 360 = 136.8 degrees. Its coordinates, when it is at 136.8 degrees, will be (cos 136.8, sin 136.8) = (−0.7290, 0.6845). Therefore, at 0.38 seconds, the object will be at (−0.7290, 0.6845).

When the object is at 102 degrees, the time will be 102 ÷ 360 = 0.2833 seconds. Its x-axis value will be Cosine 102 = −0.2079 units; its y-axis value will be Sine 102 = 0.9781 units. Its coordinates will therefore be (−0.2079, 0.9781).

At half a second, the object will be half way around the circle, and so will be at 180 degrees. Its x-axis value, and therefore, the Cosine of its angle will be −1 units. Its y-axis value, and therefore, the Sine of its angle, will be 0 units. Its coordinates will be (−1, 0).

At three quarters of a second (0.75 seconds), the object will be 0.75 of the way around the circle, and so be at 0.75 * 360 = 270 degrees. Its x-axis value and the Cosine of its angle will be 0 units. Its y-axis value and the Sine of its angle will be −1 units. Its coordinates will be (0, −1).

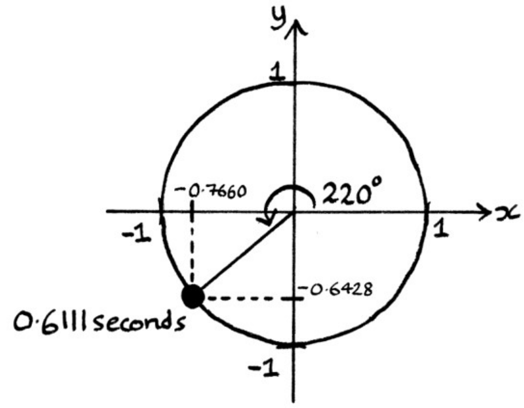
At 1 second, the object will be all the way around the circle, and so will be at 360 degrees (or zero degrees, as they are the same thing). Its coordinates will be (cos 360, sin 360), which is the same as (cos 0, sin 0), which works out as (0, 1). This is the place where it started.

If the object continues to rotate around the circle, at 1.5 seconds, the object will be 1.5 times around the circle, and so will be at 1.5 * 360 = 540 degrees. This is the same as 180 degrees. Its coordinates will be (cos 540, sin 540), which is the same as (cos 180, sin 180), which works out as (−1, 0).

At 5.7 seconds, the object will have travelled 5.7 times around the circle, and so be at 5.7 * 360 = 2,052 degrees. If we repeatedly subtract 360 from this number until we have a value between 0 and 360, we end up with 252, which means that 2,052 degrees is the same as 252 degrees. Therefore, the object is at 252 degrees on the circle. We could also have removed the whole seconds from 5.7 to produce 0.7, and then calculated the object’s angle as 0.7 * 360, which is also 252. The object’s coordinates at this time are (cos 252, sin 252) = (−0.3090, −0.9455).

When the object is at 220 degrees, it will be 220 ÷ 360 = 0.6111 of the way around the circle. If it is on its first revolution, it will have been travelling for 0.6111 of one second. Its position would be the same if it had been travelling for 1.6111 seconds, 2.6111 seconds, 3.6111 seconds and so on.

Its coordinates will be (cos 220, sin 220) = (−0.7660, −0.6428).



Examples using the coordinates

If we just have the coordinates of the object, we can use arctan [inverse Tangent] to work out its angle. Then we can use the angle to know when it was at those coordinates. As this is the first time we have used arctan since we first saw the concept in Chapter 2, we will go through the process very simply to start with.

Arctan finds the angle of a line given its gradient. To find out the gradient of a line connecting the object to the origin of the axes, we first imagine that the object is at the end of the hypotenuse of a right-angled triangle. The length of the opposite side will be the y-axis coordinate; the length of the adjacent side will be the x-axis coordinate. The gradient will be the opposite side divided by the adjacent side. [The phrase that will remind you which way around to perform the division is "opposite over adjacent" or "rise over tread".]

Once we have the gradient, we use arctan on it to find *one of the two angles* that would produce that gradient. Two angles on the circle will produce the same gradient, and these will always be 180 degrees apart. A calculator will only give one of the angles. We have to check that the result from a calculator is the one that we want, so we calculate the other result by adding or subtracting 180 degrees. We then think about which quarter of the circle the original coordinates are in, and choose the angle that points into that quarter.

This process might seem long, but it is much quicker to do than to explain.

Example 1

As an example, we will say that we want to know when the object will be at the coordinates (0.5, −0.8660).

First, we imagine that the object is at the end of the hypotenuse of a right-angled triangle. This triangle will have an opposite side of −0.8660 units and an adjacent side of 0.5 units. We calculate the gradient as the opposite side divided by the adjacent side:
−0.8660 ÷ 0.5 = −1.732

Once we have the gradient, we can work out the angle using arctan. We calculate this:
arctan (−1.732)

A typical calculator will give the result as −60 degrees, which most people would think of as +300 degrees [We add 360 to a negative angle to find the positive angle that refers to the same point on the circle].

There are two angles that would produce the gradient of −1.732. The first is 300 degrees, and the second will be 180 degrees around the circle, which is 300 − 180 = 120 degrees. [Sometimes we need to add 180 degrees to the angle to find the other result; sometimes we need to subtract 180 degrees – we do whichever will produce a result between 0 and 360 degrees.]

We have to choose which is the correct result. The coordinates of the object tell us that the object is in the bottom right quarter of the circle. Therefore, we want the angle that points into that quarter of the circle, which is 300 degrees. We can now say that when the rotating object is at the coordinates (0.5, −0.8660), it is at 300 degrees.

When the object is at 300 degrees, it is 300 ÷ 360 = 0.8333 of the way around the circle, which means it has been travelling for 0.8333 of a second. Of course, this presumes that it is on its first revolution – it would be in the same place if it had been travelling for 1.8333 seconds, 2.8333 seconds, or 3.8333 seconds and so on.

Example 2

As another example, we will find out when the object is at the coordinates (−0.7431, −0.6691) on its first revolution around the circle.

The gradient of this point is the y-axis coordinate divided by the x-axis coordinate:
−0.6691 ÷ −0.7431 = 0.9004.

One of the two angles that has this gradient is:
arctan(0.9004) = 42 degrees.

The other angle that has that gradient is 180 degrees around the circle, so is:
42 + 180 = 222 degrees.

The coordinates of the object place it in the bottom left-hand quarter of the circle. Therefore, the angle we want is 222 degrees because that is also in the bottom left-hand quarter of the circle.

Because the object moves around the circle at 1 degree every 360th of a second, it will take:
222 ÷ 360 = 0.6167 seconds to reach that angle.

Therefore, the object will be at the coordinates (−0.7431, −0.6691) at 0.6167 seconds.

Example 3

As another example, we will find out at what time the object is at the coordinates (−0.9848, 0.1736) on its first revolution.

The angle for this point *might be*:
arctan(0.1736 ÷ −0.9848) = −9.9974 degrees.

As this angle is negative and we want positive angles, we add 360 degrees to it and obtain:
−9.9974 + 360 = 350.0026 degrees.

This angle points into the bottom right hand quarter of the circle, but our coordinates are in the top left hand quarter of the circle. Therefore, we want the angle that is 180 degrees around the circle. This is:
350.0026 − 180 = 170.0026 degrees.

The object takes 170.0026 ÷ 360 = 0.4722 seconds to reach this angle. Therefore, when the object is at (−0.9848, 0.1736), the time is 0.4722 seconds.

Sine and Cosine of the time

As we know, to find the y-axis position of the object rotating around the circle, we calculate the Sine of its angle from the origin of the axes. To find the x-axis position we calculate the Cosine of its angle from the origin of the axes. We do not take the Sine or Cosine of the *time* itself. If we were to take the Sine and Cosine of the *time in seconds*, instead of the *angle*, we would calculate the wrong result.

From the point of view of the Sine or Cosine functions, if we give them a value of, say, 1 second, that value will be treated as if it were in degrees, and therefore a 360th of a circle. Therefore, the Sine of 1 second will just result in the Sine of 1 degree, which is not what we want. If our object takes one second to rotate around the circle, the Sine of 1 would not be its y-axis value after one second, but its y-axis value after 1/360th of a second. The point I am making is that we cannot use Sine or Cosine directly on the time in seconds. This should be obvious, really, but the idea is important. You might wonder why we would even want to be able to use Sine and Cosine on the time itself, and the reason is that it would make a lot of future mathematical processes and notation simpler.

If we have an object rotating around a circle at 360 degrees per second, then feeding the time in seconds into Sine and Cosine will give us a result that is equivalent to taking the Sine and Cosine of the wanted angle divided by 360. This is because the number representing the time in seconds is always 360 times less than the number representing the angle in degrees. The angle of 360 degrees is equivalent to 1 second. For example, when the time is 0.5 seconds, the angle is 0.5 * 360 = 180 degrees.

This means that we cannot directly use Cosine and Sine on the time to work out the x-axis and y-axis values of the object unless we first multiply that time by 360. The time must be "corrected" so that one second represents 360 degrees. To put this another way, the time must be treated as if it, too, were divided into 360 portions, so that one portion of time is equivalent to one portion of the circle. The way to do this is to multiply the time in seconds by 360. Then, we can use Sine and Cosine on the "corrected time", without needing to worry about the actual degrees.

This corrected time could be called "degree time" as it matches the degrees at a particular time. In practice, "angle time" or "*angular* time" is more apt as there are other ways to divide a circle than into 360 pieces [For example, radians divide a circle into 2π pieces, as discussed in Chapter 22]. The term "angular time" is a good introduction to the concept of "angular frequency", which I explain in Chapter 6.

To summarise this, we cannot use Sine and Cosine on the time in seconds, but we can use Sine and Cosine on a corrected form of the time, which I will be calling "angular time". The reason why we would want to do such a thing will make more sense as we progress through this and later chapters. We "correct" the time by multiplying it by 360.

Examples

When the time is 0.4 seconds, the y-axis position of the object moving around the circle will be Sine (360 * 0.4) = Sine (144) = 0.5878 units. The number 144 is the "angular time" at 0.4 seconds. It is also the case that 144 is the angle in degrees at this time too. The x-axis position will be Cosine (360 * 0.4) = −0.8090 units.

If the time is 0.21 seconds, the y-axis position of the object will be Sine (360 * 0.21) = 0.9686 units; the x-axis position would be Cosine (360 * 0.21) = 0.2436 units.

If the time is 1 second, the y-axis position of the object will be Sine (360 * 1) = 0. The x-axis position will be Cosine (360 * 1) = 1.

Coordinates on the circle

When we were dealing with just angles, we could give the coordinates of any point on a unit-radius circle's edge using the Sine and Cosine of the angle.

For example, the coordinates of the point on a unit-radius circle's edge at 45 degrees from its centre is (cos 45, sin 45). This works out as (0.7071, 0.7071). Sometimes, particularly if we want to have very accurate coordinates, it can be better to keep the coordinates in terms of Sine and Cosine.

Now we are dealing with time, we can give the coordinates of the position of the rotating object at any moment in *time* using Sine and Cosine. For example, if the object has been travelling for 0.125 seconds, then its coordinates will be:
(cos (360 * 0.125), sin (360 * 0.125)).
... which are the same coordinates as (0.7071, 0.7071).

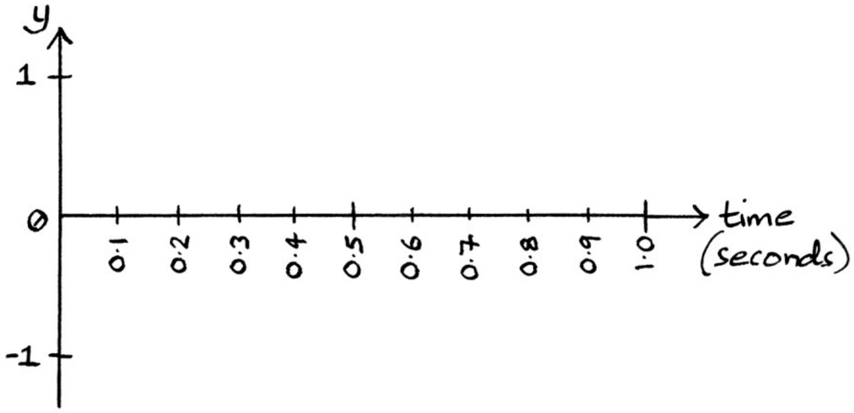
If the object has been travelling for 0.9 seconds, then its coordinates will be:
(cos (360 * 0.9), sin (360 * 0.9)).
... which are the same coordinates as (0.8090, −0.5878).

At any moment in time, the coordinates of the rotating object will be:
(cos 360t, sin 360t)
... where t is the time in seconds since it started travelling.

Sine and Cosine time waves

Now we will look at the Sine and Cosine graphs derived from our "time circle". The graphs will still be Sine and Cosine waves, but they will now be showing the y-axis and x-axis points of the object as it rotates around the circle *at any particular moment in time*. (Previously, the Sine and Cosine graphs showed the position of static points of the edge of the circle at particular *angles*.). To clarify this, the time-based Sine and Cosine waves will not be showing the stationary points on the edge of the circle at a particular time – such a concept is not particularly helpful because the circle itself is the same at all times. Instead, the time-based Sine and Cosine waves will be showing the position of an object rotating around the edge of the circle with respect to time.

Whereas previously, the units for the x-axis (the θ-axis) in the Sine and Cosine graphs were degrees, now the x-axis units will be *seconds*. We will now start calling the x-axis, the "time axis" or "t-axis".

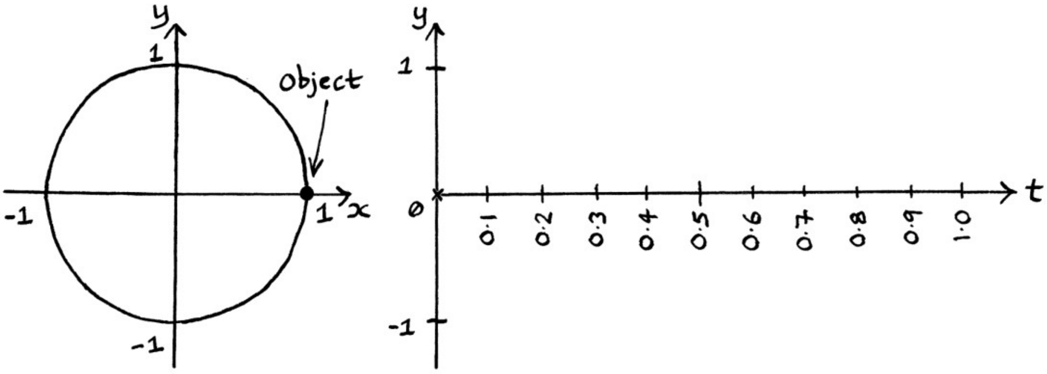


We can think of the Sine wave time graph as showing the object's y-axis position on the circle at any moment in *time* since the object started moving. We can think of the Cosine wave time graph as showing the object's x-axis position on the circle at any moment in *time* since the object started moving. As there is a fixed relationship between the time since the object started moving and the angle of the object, the waves will have an identical shape to when they portrayed angles.

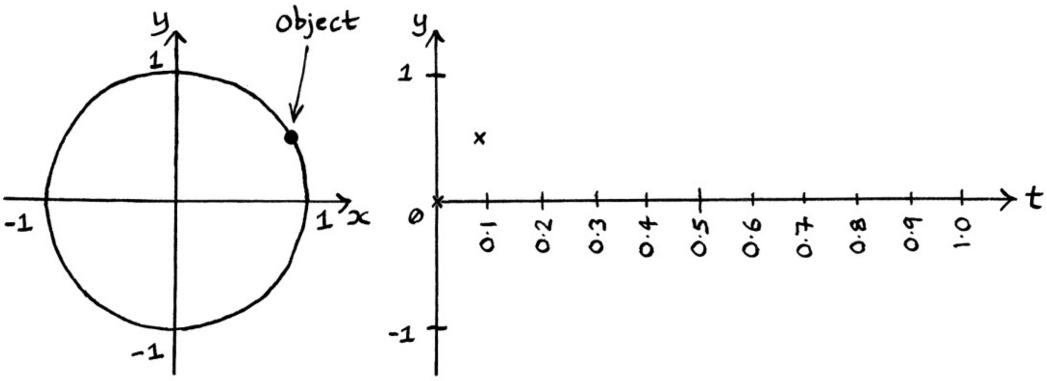
Sine wave time graph

If we wanted to draw a Sine wave graph that related to time, we could proceed as follows. The object rotates around the circle at one revolution per second, so we will read its y-axis positions on the circle graph at evenly spaced moments *in time* and plot them on the new Sine wave time graph. To make this easier to do, we will read the y-axis position of the object once every twelfth of a second. This is the same as making one reading every thirty 360$^{\text{ths}}$ of a second.

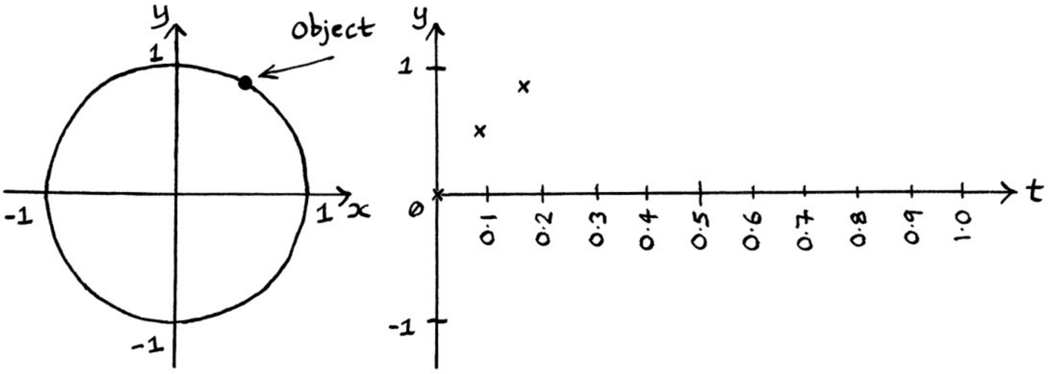
When the object starts, at t = 0, its y-axis height on the circle is 0. We mark that on the future Sine time graph:



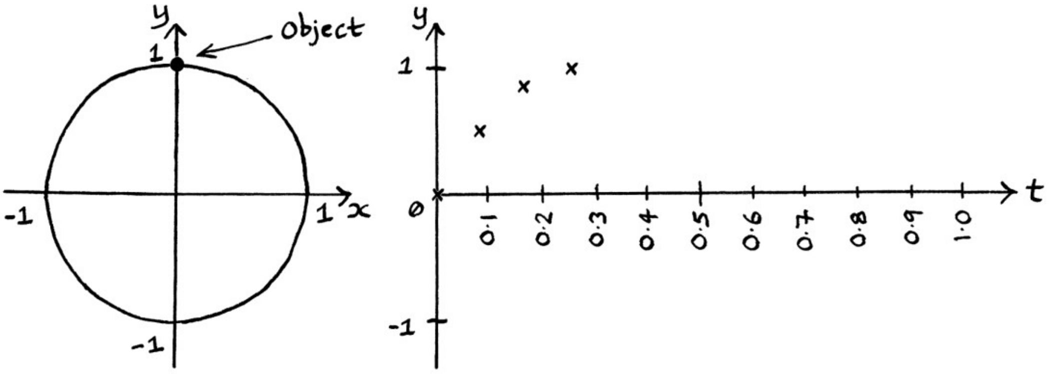
At thirty 360ths of a second, which is 0.08333 seconds or one twelfth of a second, the object's y-axis position on the circle's edge is Sine (360 * 0.08333) = Sine 30 = 0.5 units. If we had not been dealing with time, we would have just said that the y-axis value of the point at 30 degrees on the circle's edge is the Sine of 30 degrees, which amounts to exactly the same thing in this situation. However, to become used to dealing with time it is better to think of the Sine of "thirty 360ths of a second multiplied by 360", where the multiplication by 360 allows us to use Sine on the "angular time". Doing this is a preparation for later chapters where slightly more complicated concepts are introduced. We mark that y-axis value on our graph where the time-axis value is 0.08333 and the y-axis value is 0.5.



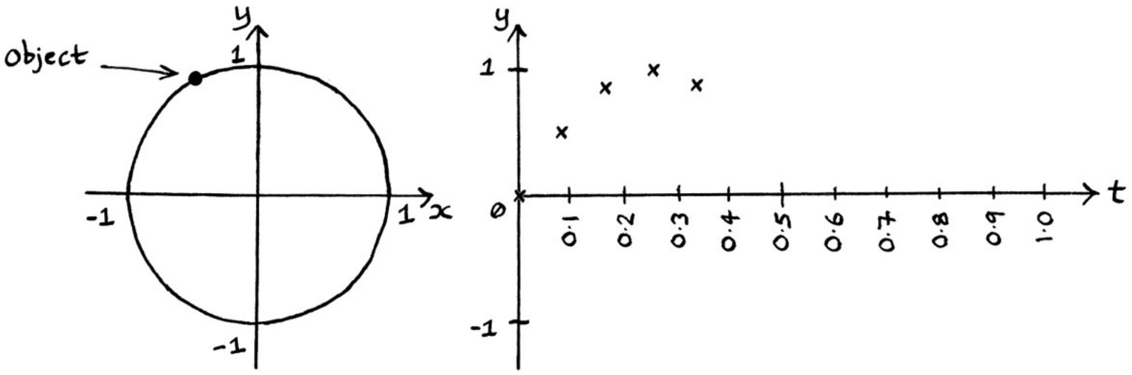
At sixty 360ths of a second, which is 0.1667 seconds, the object's y-axis position on the circle's edge is Sine (360 * 0.1667) = Sine 60 = 0.8660 units. We mark that point on our graph:



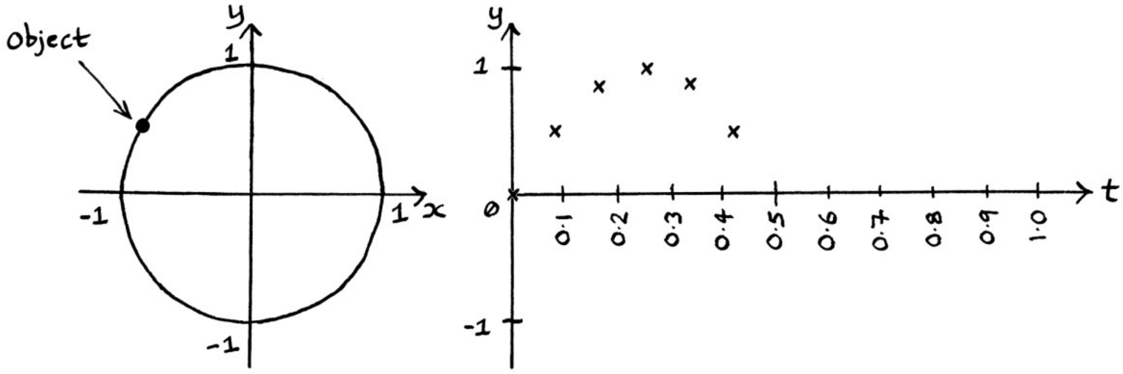
We can continue following the object's y-axis positions around the circle at particular times, and marking them on the time wave graph until the object has completed one revolution of the circle. At ninety 360ths of a second, which is 0.25 seconds:



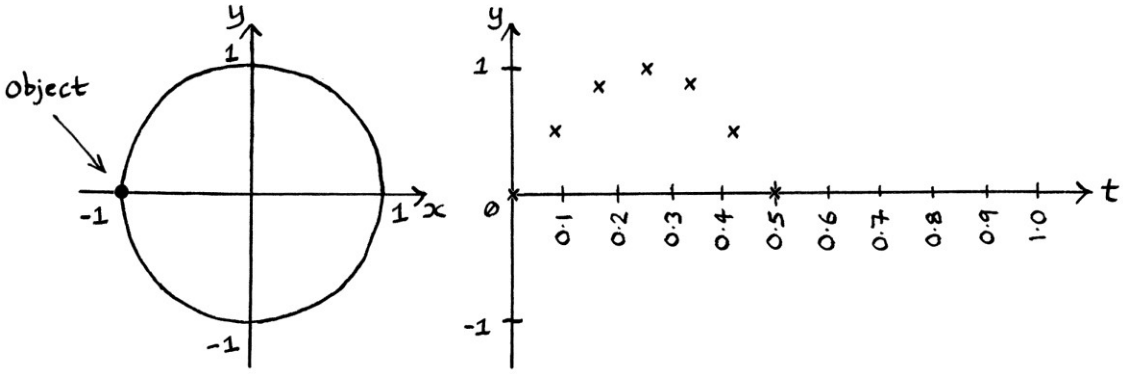
At one hundred and twenty 360ths of a second, which is a third of a second:



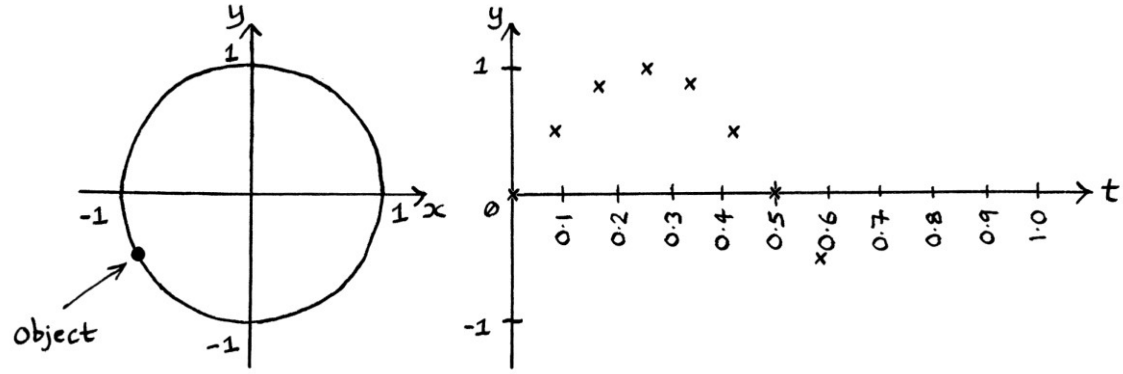
At one hundred and fifty 360ths of a second, which is 0.4167 seconds:



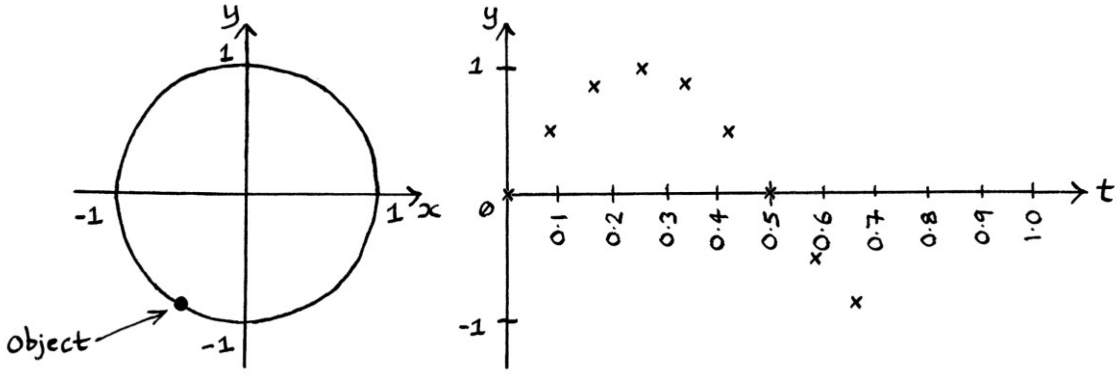
At one hundred and eighty 360ths of a second, which is 0.5 seconds:



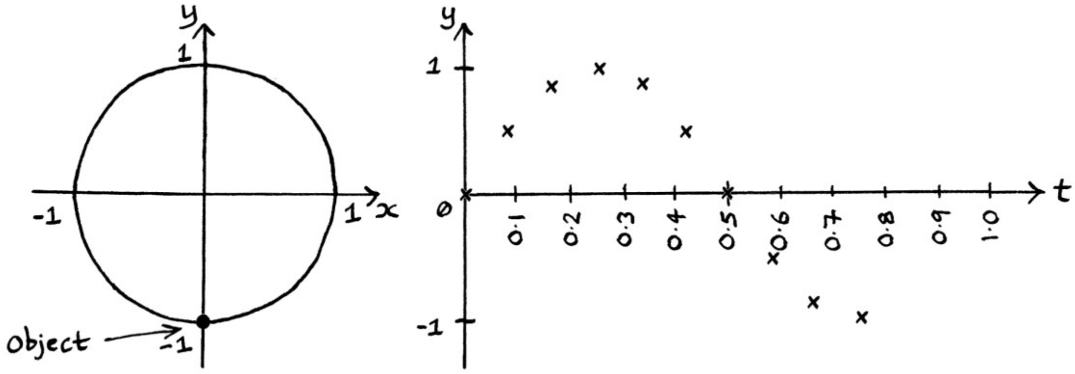
At two hundred and ten 360ths of a second, which is 0.5833 seconds:



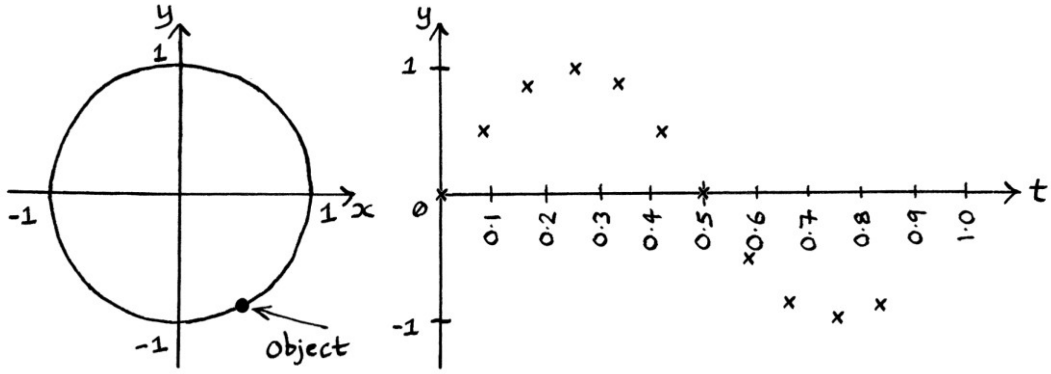
At two hundred and forty 360ths of a second, which is 0.6667 seconds:



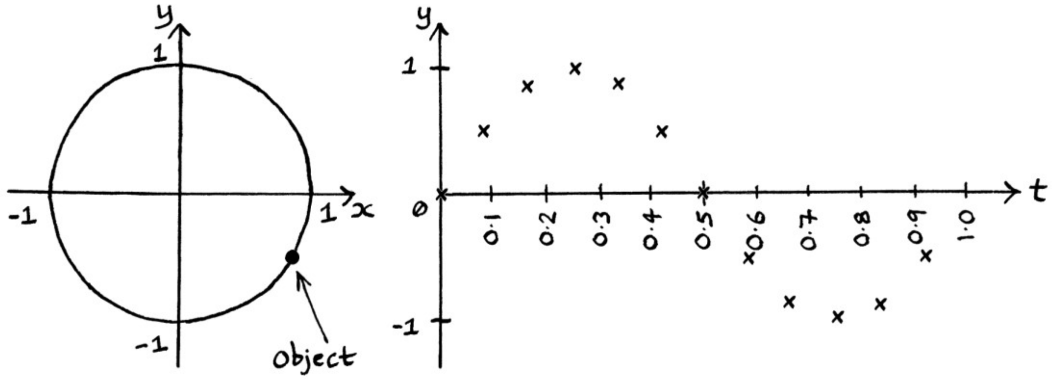
At two hundred and seventy 360ths of a second, which is 0.75 seconds:



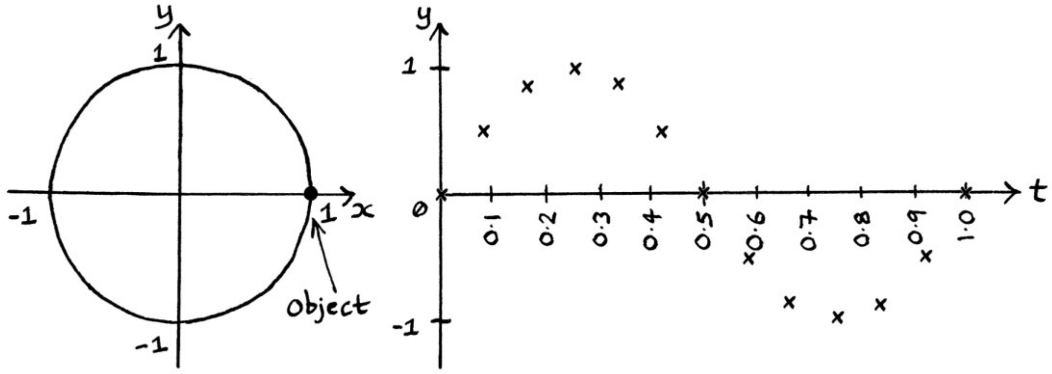
At three hundred 360ths of a second, which is 0.8333 seconds:



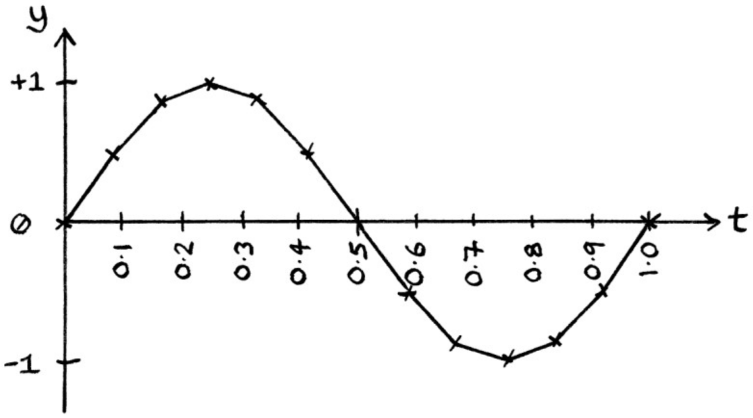
At three hundred and thirty 360ths of a second, which is 0.9167 seconds:



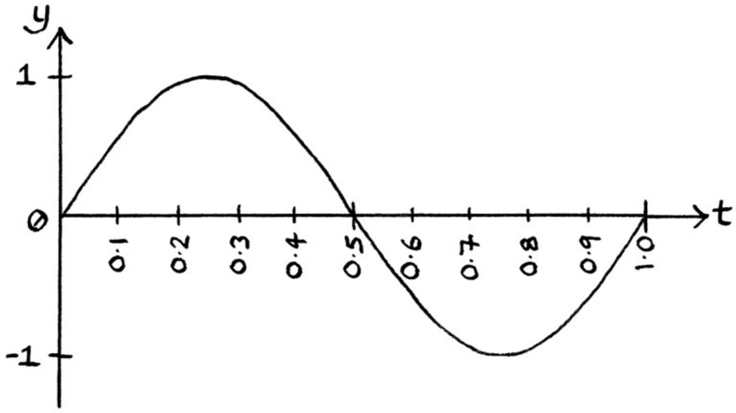
At three hundred and sixty 360ths of a second, which is 1 second:



After the object has completed one revolution of the circle, we can join up the points, and the result is this graph:



If we could measure the object's x and y-axis values at infinitely small time intervals, and if we could read values from the circle with perfect accuracy, we would end up with the following smooth Sine wave time graph:

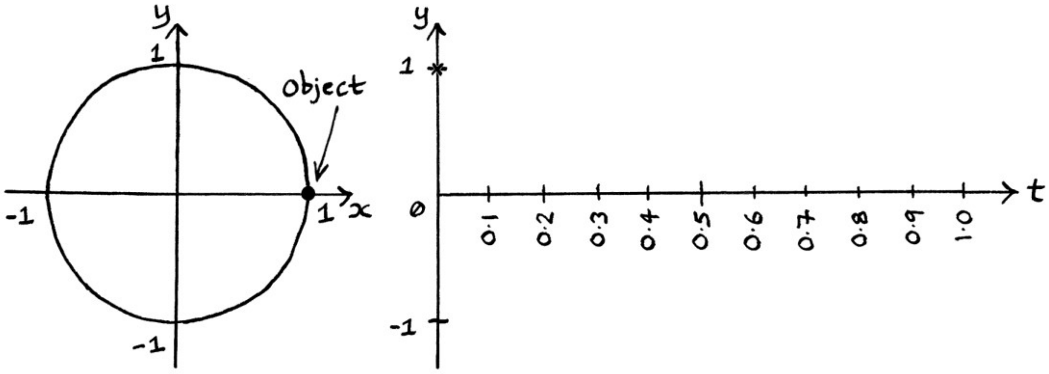


We will look at this in more detail later in this chapter.

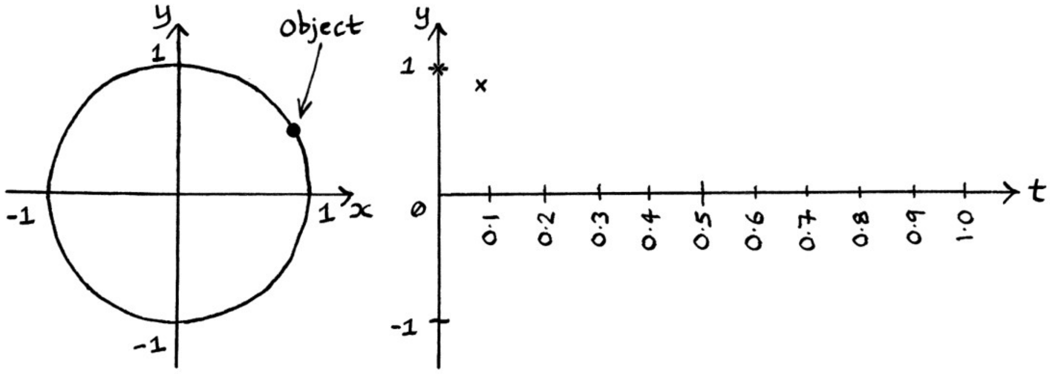
Cosine wave time graph

To draw the Cosine wave time graph, we make a note of the object's x-axis position at regularly spaced moments in time as it rotates around the circle. We then plot those x-axis positions as the y-axis positions for those times on the new Cosine wave time graph.

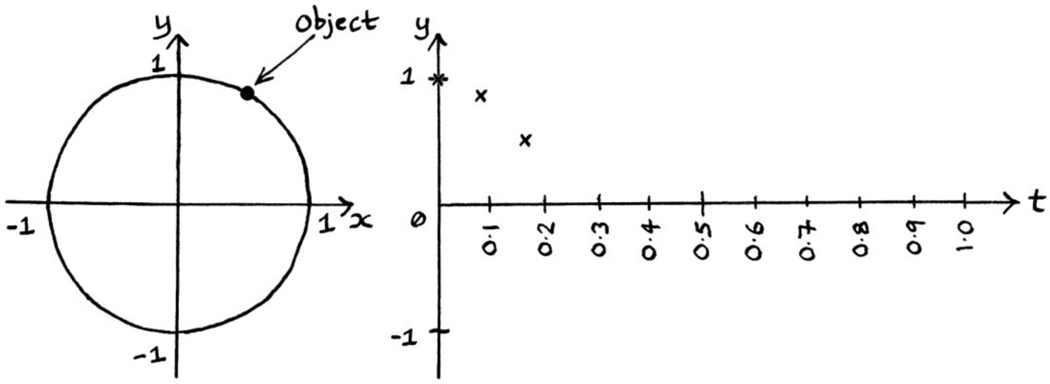
At 0 seconds, the object will be at 0 degrees on the circle. Its x-axis position will be 1 unit. We therefore mark the point at 0 seconds and 1 unit on the Cosine time wave graph.



At thirty 360ths of a second, which is 0.08333 seconds, the object's x-axis position on the circle will be Cosine (360 * 0.08333) = cos 30 = 0.8660 units. On the Cosine time wave graph, we mark the point where the y-axis is 0.8660 units and the time is 0.08333 seconds.

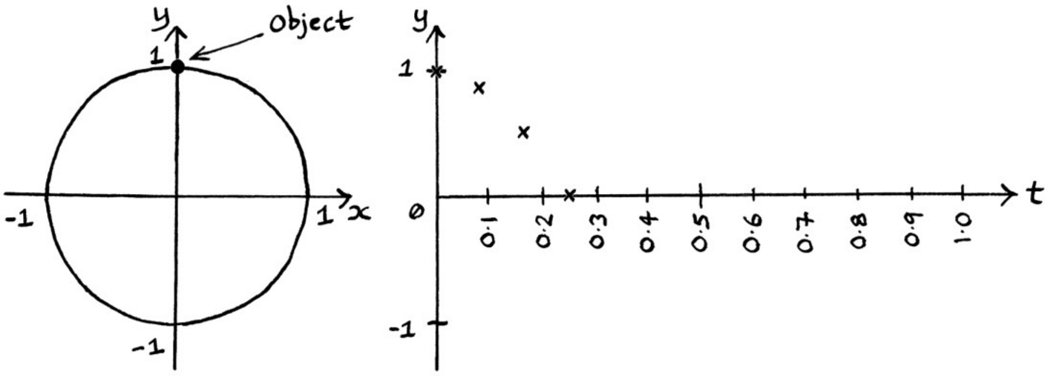


At sixty 360ths of a second, which is 0.1667 seconds, the object's x-axis position on the circle will be Cosine (360 * 0.1667) = cos 60 = 0.5 units. Therefore, we mark the point on the Cosine wave graph where the y-axis is 0.5 units and the time axis is 0.1667 seconds.

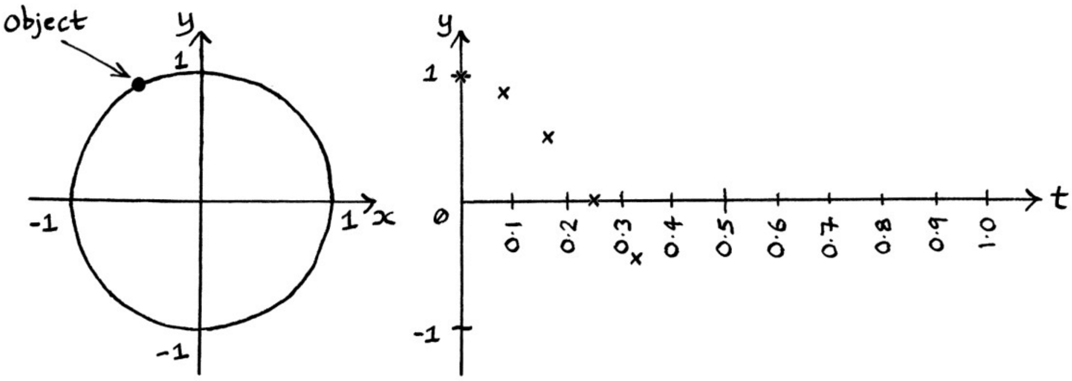


We continue in this way, observing where the object is at various times and plotting those x-axis positions on the time graph.

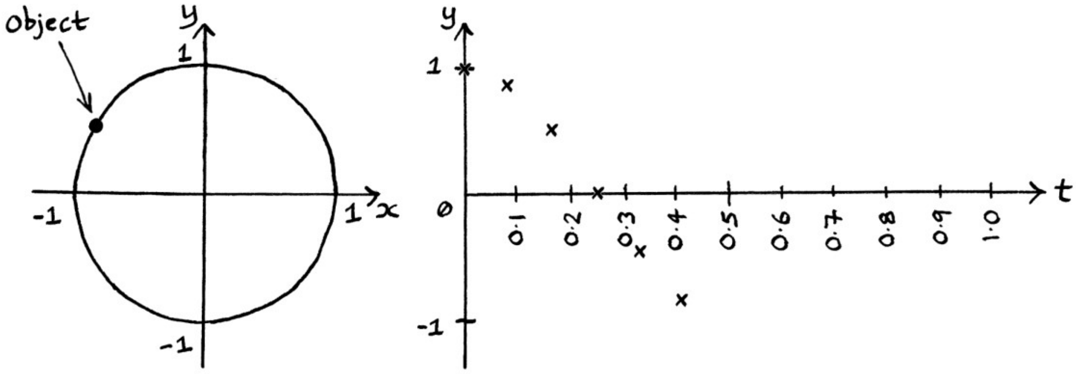
At ninety 360ths of a second, which is 0.25 seconds:



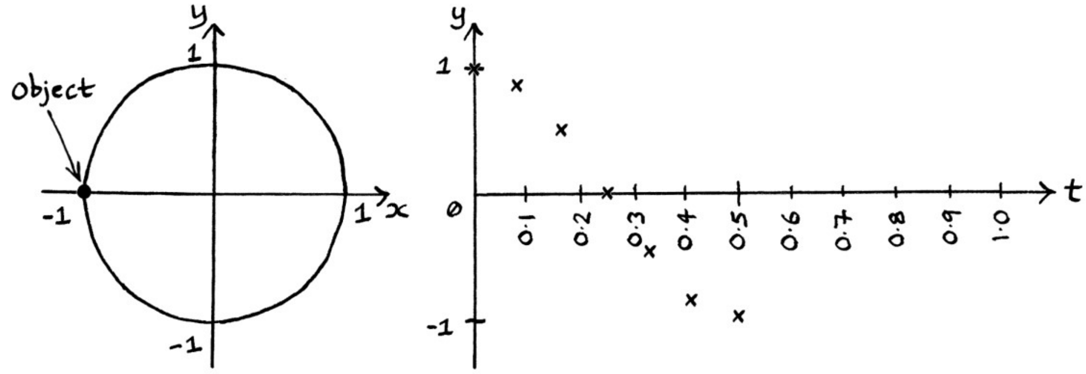
At one hundred and twenty 360ths of a second, which is a third of a second:



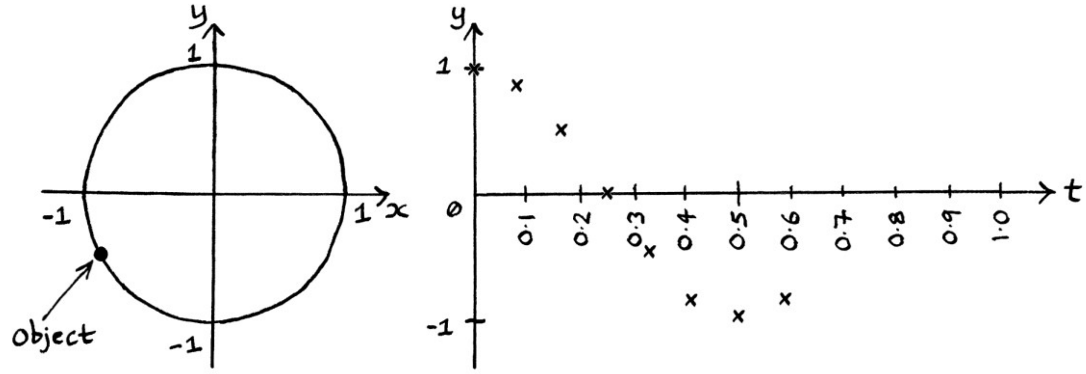
At one hundred and fifty 360ths of a second, which is 0.4167 seconds:



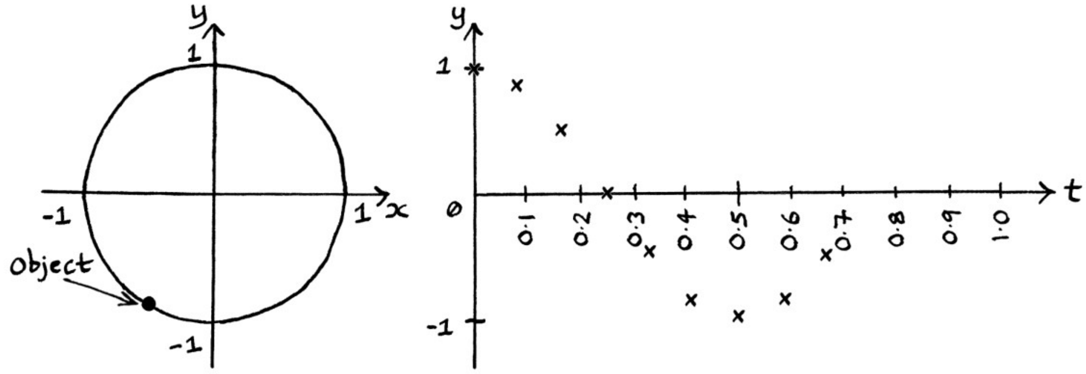
At one hundred and eighty 360ths of a second, which is 0.5 seconds:



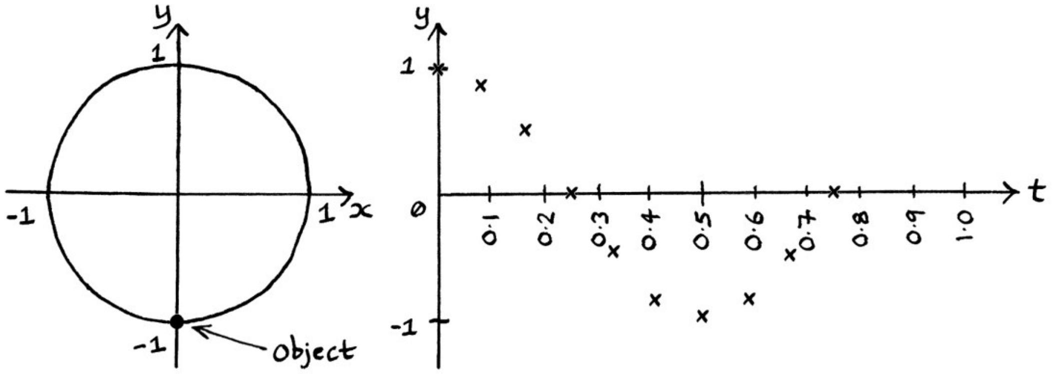
At two hundred and ten 360ths of a second, which is 0.5833 seconds:



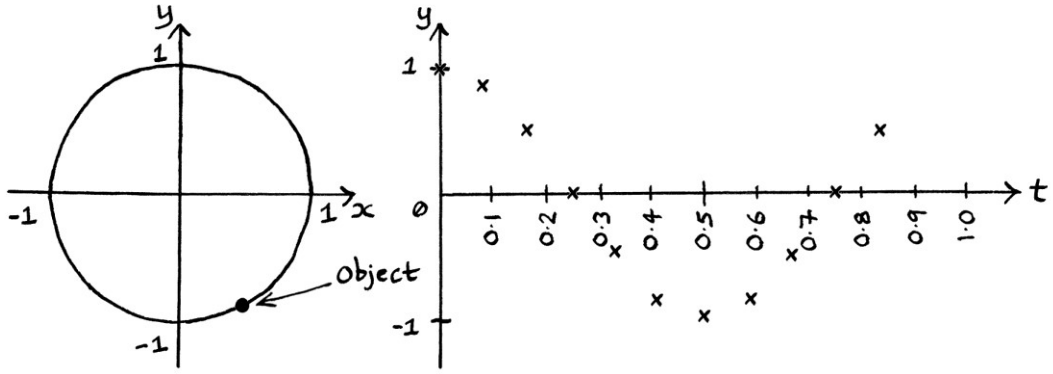
At two hundred and forty 360ths of a second, which is 0.6667 seconds:



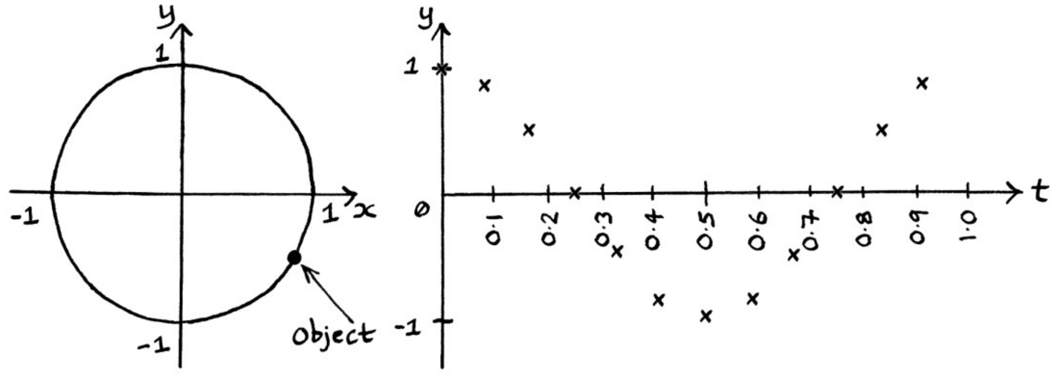
At two hundred and seventy 360ths of a second, which is 0.75 seconds:



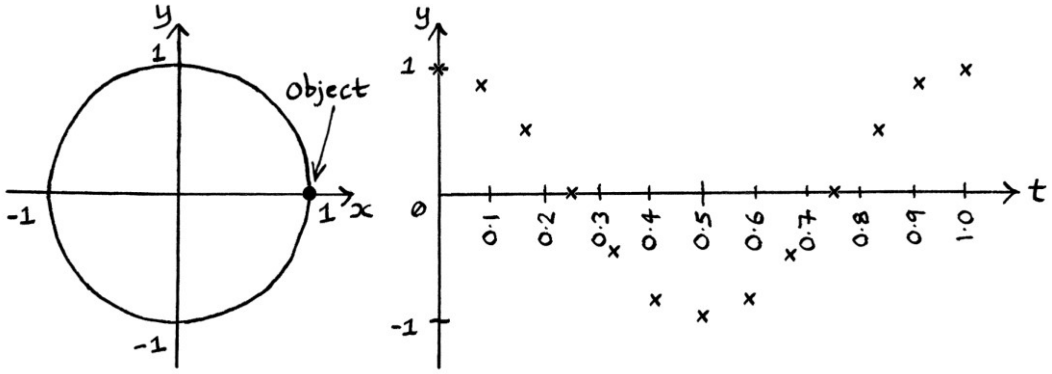
At three hundred 360ths of a second, which is 0.8333 seconds:



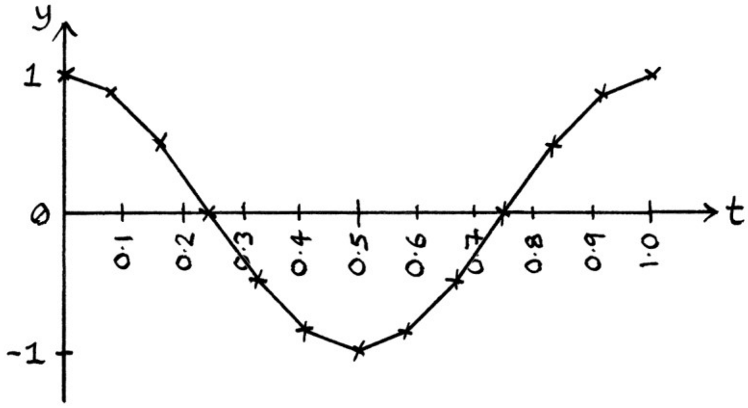
At three hundred and thirty 360ths of a second, which is 0.9167 seconds:



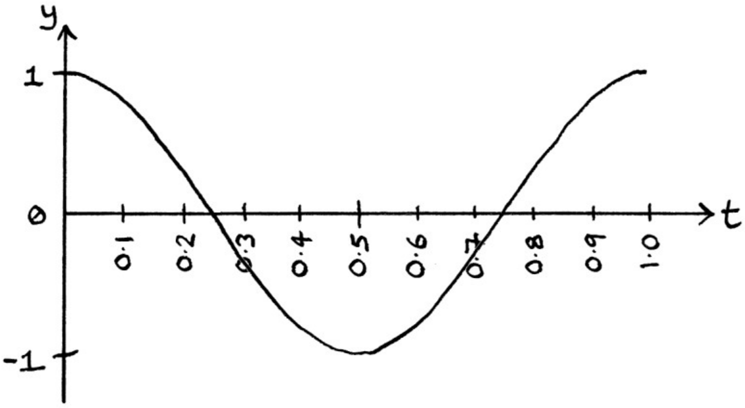
At three hundred and sixty 360$^{\text{ths}}$ of a second, which is 1 second:



After the object has completed one revolution of the circle, we can join up the points and the result will be this graph:



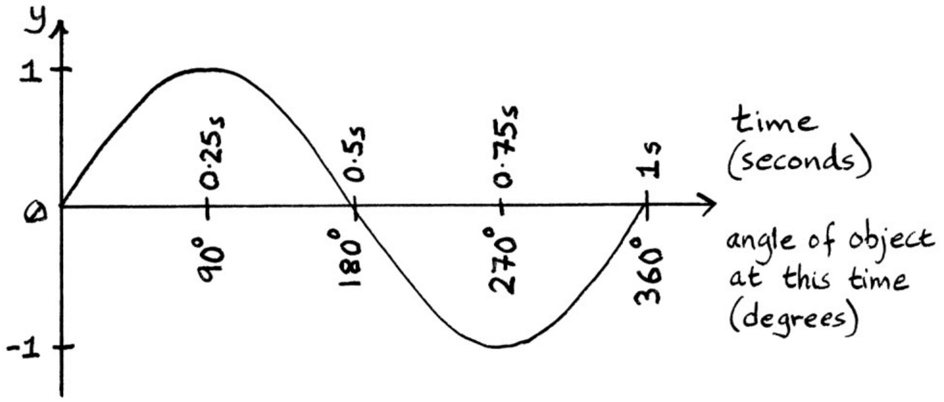
If we measured the object’s position around the circle at infinitely small moments of time, and we could read the circle with perfect accuracy, we would end up with this smooth Cosine wave time graph:

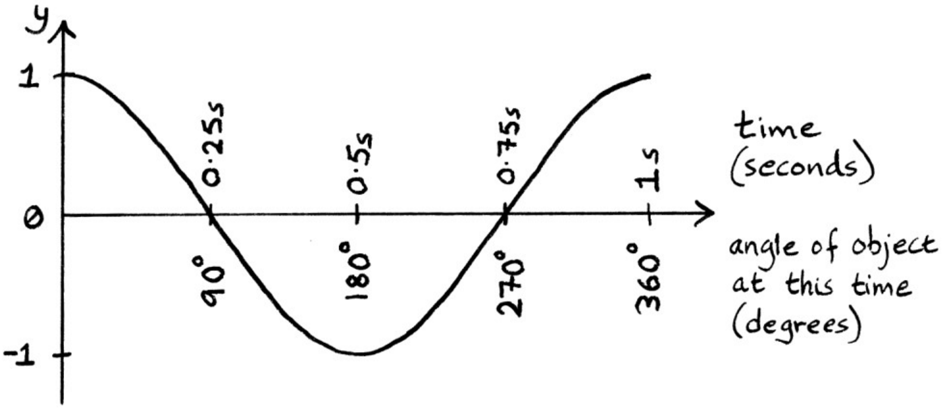


Sine and Cosine

As we can see from the finished Sine and Cosine time waves, the two graphs have exactly the same shape as the Sine and Cosine waves when we were dealing with just degrees. (It might be obvious that this would be the case, given that the object is following the curve of the circle). The only difference, and it is a *very* important difference, is that one repetition of each wave now completes in 1 *second*, whereas before, one repetition of each wave completed in 360 *degrees*. Essentially, the repetitions are still completed in 360 degrees, but we are now considering those degrees in relation to the time when the object rotating around the circle is at each of those degrees. This means that although the time graphs have exactly the same shape as the previous "angle" graphs, they are portraying more information.

In a way, the Sine and Cosine wave time graphs are really lookup charts as to when the object will be at a particular angle, or at a particular point on the circle's edge. Given that, we could, if we wanted, write the angle and the time along the x-axis at the same time. The angle will be the angle of the object on the circle at that particular time.





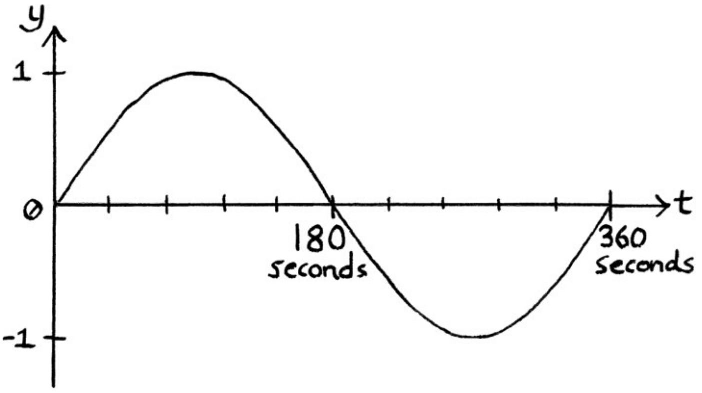
Generally, labelling the axes with both time and angle is not done. Reasons for not doing it include how it clutters up the axis, it can look confusing, and if you understand waves, you do not need to do it. If you know the shapes of an "angle" wave, you can tell when the object was at a particular angle such as 0, 90 degrees, 180 degrees, or 270 degrees on a "time" wave by the ups and downs of the curve. For example, when a Sine time wave is at its highest y-axis point, the object rotating around the circle will be at its highest point too, so will be at 90 degrees on the circle. When a Cosine time wave is at its highest point, the object rotating around the circle will be at its maximum positive x-axis position on the circle, so will be at 0 degrees. When a Sine time wave has a y-axis value of 0, the object rotating around the circle will be at y = 0 on the circle, so will be at either 0 or 270 degrees. You can tell which of the two it is by whether the y-axis value on the wave is increasing or decreasing at the time.

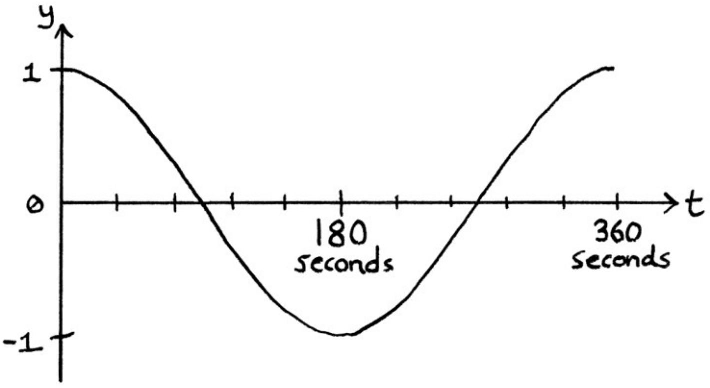
It always pays to remember that the time in seconds is directly related to the angle of the object moving around the circle at that time. The y-axis value of the object moving around the circle is related to both its angle from the origin of the axes, and the time since it started moving. The x-axis value of the object moving around the circle is related to both its angle from the origin, and the time since it started moving.

The graph formulas

One thing to notice is that the Sine and Cosine time graphs are showing the y-axis and x-axis values of the object moving around the circle as they are at any particular time. This is different from the actual results of the Sine of the time in seconds and the Cosine of the time in seconds, which is something I explained in the last section.

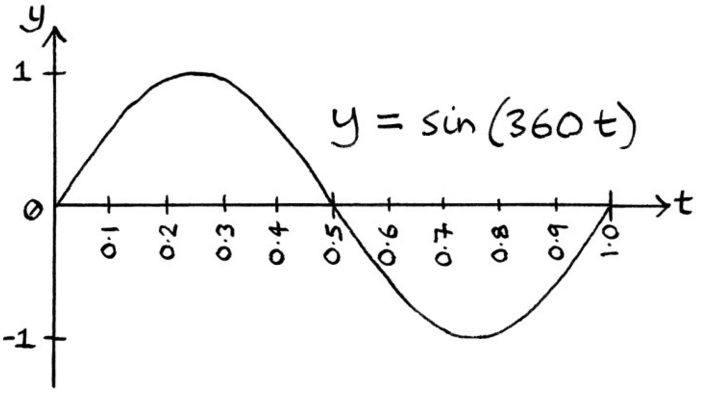
If the graphs were based on the formula for the functions of Sine and Cosine as performed on the time in seconds, i.e. “y = sin t” and “y = cos t”, they would look like this:



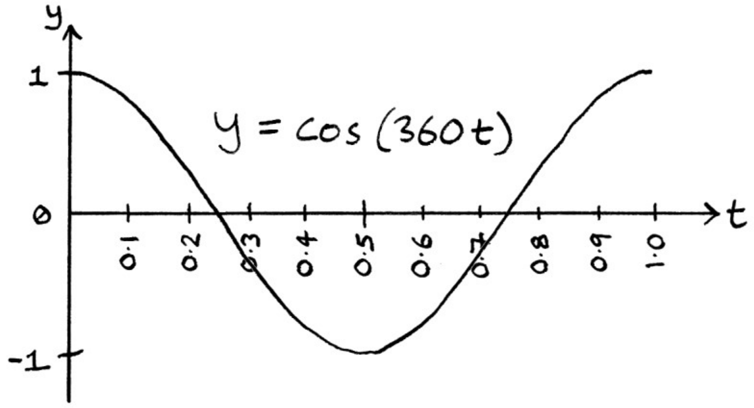


The time needs to reach 360 seconds for the waves based on “y = sin t” and “y = cos t” to repeat. Therefore, as I said before, if we want the graphs for Sine and Cosine’s effect on time to match the coordinates of the object moving around the circle at the same time, we need to consider “angular time”, and so first multiply the time by 360 in the formulas. Our formulas for the graphs of the object rotating around the circle at 1 revolution per second are therefore:

“y = sin (360 *t)”:



... and “y = cos (360 * t)”



By multiplying the seconds by 360, we are really adjusting the time to represent the angles in the circle. The “360 * t” part in the formula means that there is a one-to-one relationship between our adjusted time and the angle.

One potentially confusing aspect as to how these waves are drawn is that the multiplication by 360 is taken into account in the formula, but not seen in the numbering of the axes on the graph. For example, the standard Sine wave time graph picture would be given by the formula “y = sin (360 * t)”, despite the graph not mentioning multiplication by 360. This is because the graph shows the literal time for the position of the object moving around the circle. At say, 0.25 seconds, the object is at a height of 1 unit. The wave graph does not need any time correction. However the actual *formula* for this graph has to have the time corrected with a multiplication by 360. To accurately give the height of the object on the circle at, say, 0.25 seconds, we need to calculate sin (360 * 0.25) = sin 90.

If we wanted to have a graph with axes that incorporated the multiplication by 360, then the x-axis would not be *time*, but *angular time.* In other words, it would be divided into 360$^{\text{ths}}$ of a second, and go from zero up to three hundred and sixty 360$^{\text{ths}}$. In practice, there is no need for such an axis, and it is much easier to have the time in uncorrected seconds.

Formulas

The formulas for our two time-based waves are:
“y = sin (360 * t)”
... and:
“y = cos (360 * t)”

It is possible to write these formulas in slightly different ways, which all mean the same thing:

- With brackets and the multiplication sign: “y = sin (360 * t)”
- With the multiplication sign removed: “y = sin (360t)”
- With the brackets removed: “y = sin 360t”
- With brackets, multiplication sign, and an unneeded one: “y = sin (360 * 1t)”

Which is the best way to write a formula depends on which is the clearest and least ambiguous for a particular situation. Brackets help remove ambiguity, but on the other hand, too many can make a formula harder to read. Generally, people remove any superfluous “ones”, so they would have “t” instead of “1t”. However, having superfluous “ones” and brackets in a formula might be appropriate if we had a list of formulas that required brackets and values such as 2t, 3t, 4t and so on, and we wanted to maintain a pattern to make everything clearer.

Measuring time circles for the Sine and Cosine graphs

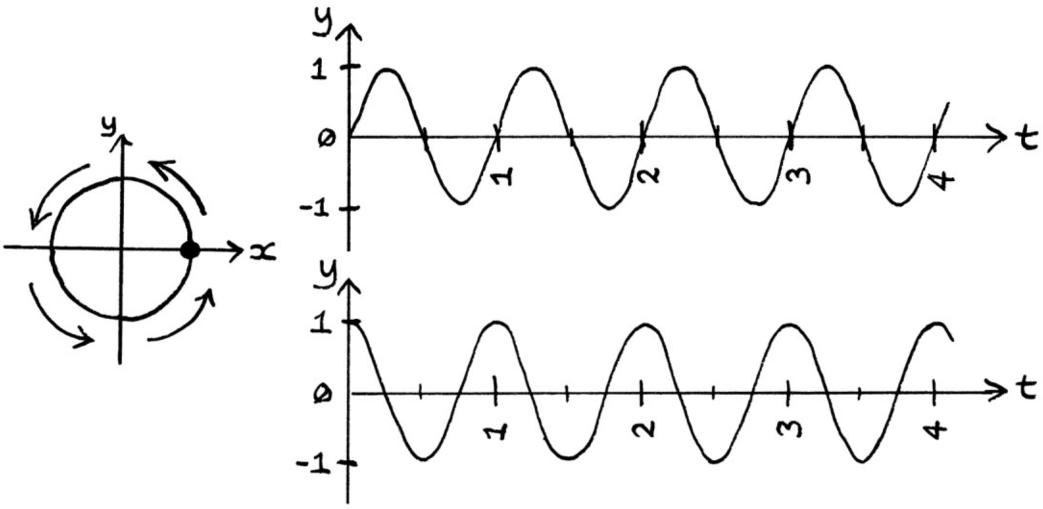
Previously, when we were creating the Sine and Cosine wave graphs from circles based on angles, we measured the y-axis values of points around the circle at evenly spaced *angles* to create the Sine wave graph, and the x-axis values at evenly spaced *angles* to create the Cosine wave graph. Now we are dealing with time, we need to measure the y-axis values of points around the circle at evenly spaced *times* to create the Sine wave graph, and the x-axis values at evenly spaced *times* to create the Cosine wave graph.

When dealing with an object that rotates around a circle at 1 rotation a second, the fraction of a full circle and the fraction of the time will be identical, so it does not really matter if we measure x and y-axis values at angles or times. However, in other situations, they might not be identical, and when we combine circles together [as in Chapter 14], reading the angle instead of the time will lead to confusing mistakes.

One thing to note is that if the object were moving faster or slower than 1 revolution per second, a static drawing of a circle would give no clues as to where the object was at any particular time. We would need other information so that we could draw the two waves.

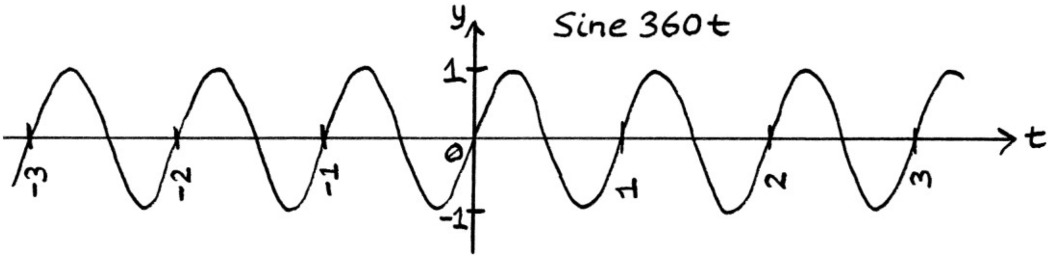
Forever time circles

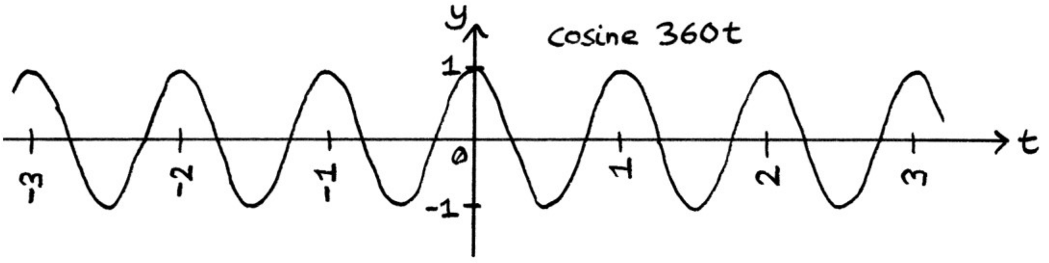
If the object continues to move around the circle after completing one revolution, then the Sine and Cosine time waves would continue further:



Note that whether a time-based wave graph continues for more than one revolution or not depends on what the graph is showing. If the object only completes one revolution, the wave graph would stop at 1 second. If the object rotates forever, the graph would continue forever.

We can consider the time before zero seconds, too:





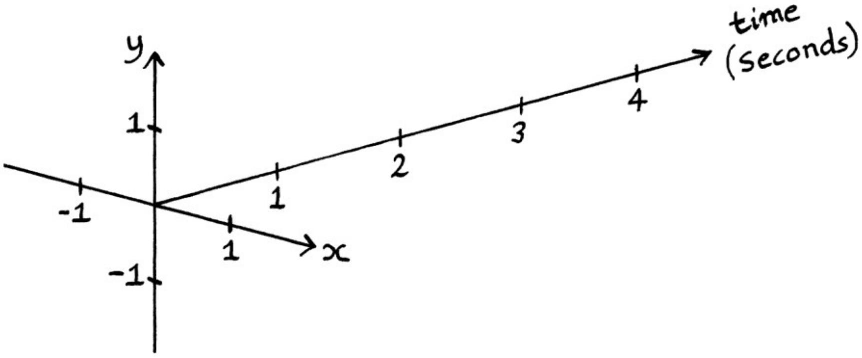
The easiest way to think about what such graphs mean is to think of them as portraying the position of the object at times before we started observing it, with the idea that we started observing it at 0 seconds.

The time helix

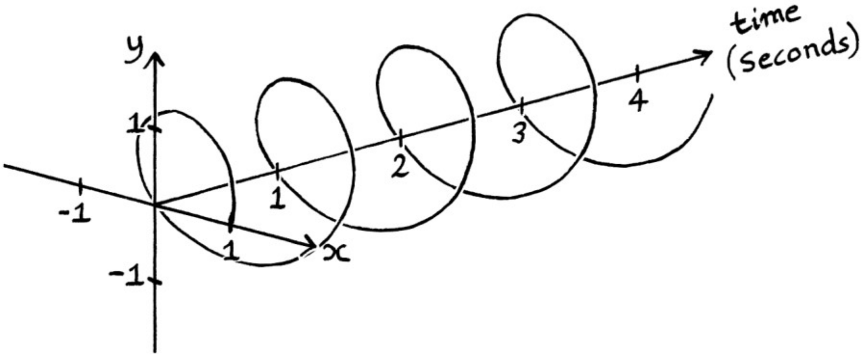
So far, we have illustrated Sine and Cosine with respect to time using graphs and the circle. The circle lets us visualise an overall view of angles and time, while the time wave graphs show how the y-axis and x-axis values of the object rotating around the circle vary according to time. The circle and graphs each have their own advantages depending on which aspect we are trying to observe.

With our circle, once the object passes 360 degrees or 1 second, it starts going around again. The position of the object at 360 degrees or 1 second is exactly the same as when it started. If the object travels for 2 seconds and completes 720 degrees, it will be in the same place too. From a visualisation point of view, this makes it hard to see what happens after one revolution – it is difficult to portray later journeys around the circle if they are drawn as being in the same place. Therefore, the circle is not as suitable for portraying changes over time as the graphs are.

We can improve how we visualise both the circle and waves, by changing the circle from a two dimensional chart into a three dimensional chart. We do this by giving the circle a third axis – that of time. Objects still move around as before, but they step outwards, away from the observer down the time axis as they move.



For our object that moves at one degree every 360th of a second, for every degree it moves around the edge of the circle, it will also move out one 360th of a second down the time axis. To put this another way, for every 360th of a second it moves out down the time axis, it rotates one degree around the edge of the circle. However we want to think of this, it results in an anticlockwise helix, instead of a circle.



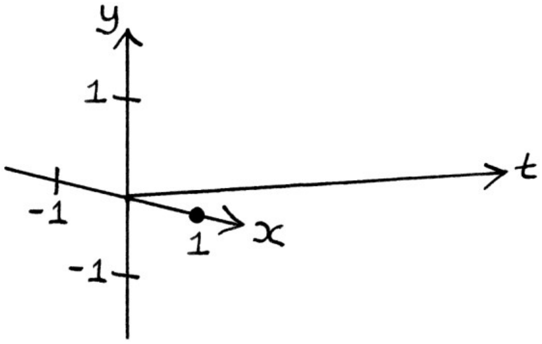
After one second, the object will have completed one revolution, and will have moved 1 second down the time helix's time axis. After two seconds, it will have completed 720 degrees, and have moved down the time helix's time axis by 2 seconds.

From this, we can distinguish the different repetitions of the Sine wave and the Cosine wave on what was previously the circle.

We can give the position of our object on our three-dimensional helix at any particular time in two different ways. We can either describe it by its angle around the origin and the distance along the time axis, or by its "x" and "y" position and its distance along the time axis.

Some positions using the angle and its time axis position are as follows:

At t = 0, the object's position is 0 degrees, 0 units along the time axis.

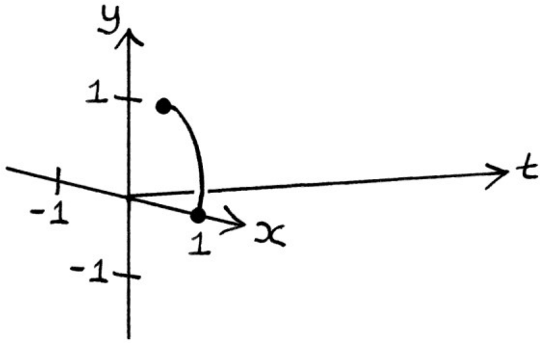


When the time is one 360th of a second, the object's position is one degree around, and one 360th of a second along the time axis.

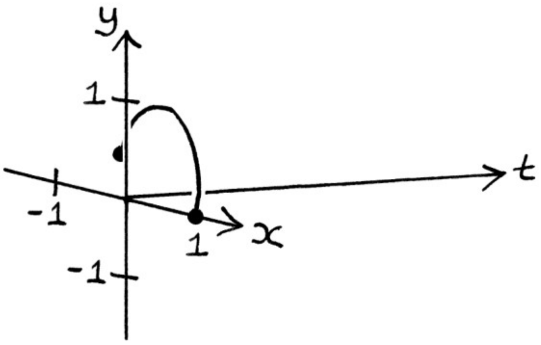
When the time is two 360ths of a second, its position is 2 degrees around, and two 360ths of a second along the time axis.

At t = 0.125 seconds, its position is 45 degrees around, and 0.125 seconds down the time axis.

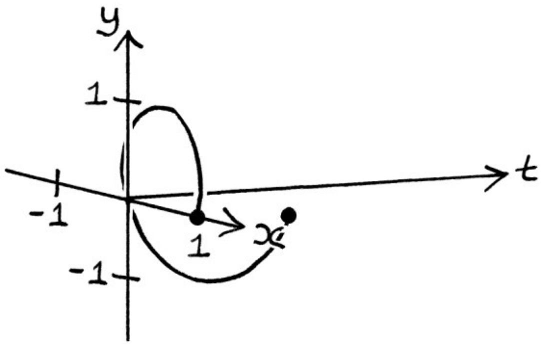
At t = 0.25 seconds, its position is 90 degrees around, and 0.25 seconds down the time axis.



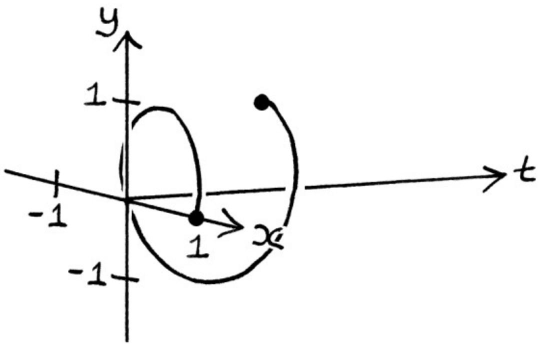
At t = 0.5 seconds, the object's position is 180 degrees, and it is half a second down the time axis.



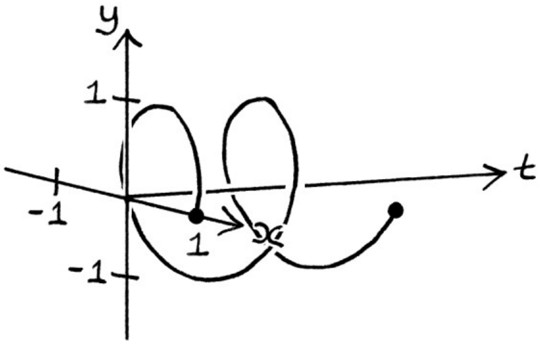
At t = 1 second, the object's position is at 0 degrees again, but now it is 1 second down the time axis.



At t = 1.25 seconds, the object's position is at 90 degrees, and the object is 1.25 seconds down the time axis.



At t = 2 seconds, the object’s position is at 0 degrees again, and the object is 2 seconds down the time axis.



When t = 2.5 seconds, the object’s position is 180 degrees, and it is 2.5 seconds down the time axis.

If we give the coordinates of the object on our 3-dimensional helix as (x, y, t), the “x” and “y” coordinates are the same as in our original circle, but the “t” coordinate is the one going off out into the third dimension.

When t = 0, the coordinates are (cos 0, sin 0, 0) = (1, 0, 0).

When t = one 360th of a second, the coordinates are (cos 1, sin 1, 0.002778) = (0.9998, 0.0175, 0.002778).

When t = two 360ths of a second, the coordinates are (cos 2, sin 2, 0.005556) = (0.9994, 0.0349, 0.005556).

When t = 0.25 seconds, the coordinates are (cos 90, sin 90, 0.25) = (0, 1, 0.25).

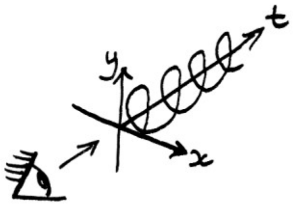
When t = 0.5 seconds, the coordinates are (cos 180, sin 180, 0.5) = (−1, 0, 0.5).

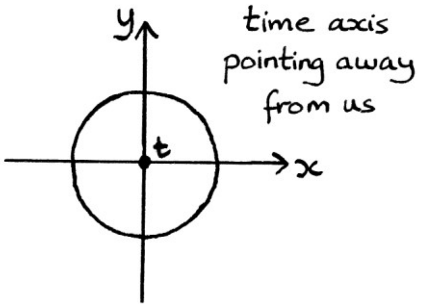
When t = 1 second, the coordinates are (cos 360, sin 360, 1) = (1, 0, 1).

When t = 2.5 seconds, the coordinates are (cos 900, sin 900, 2.5) = (−1.0, 0, 2.5).

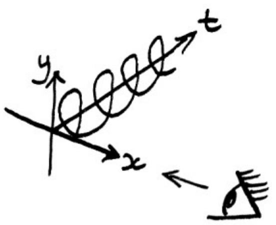
How the time helix relates to the circle, sine and cosine

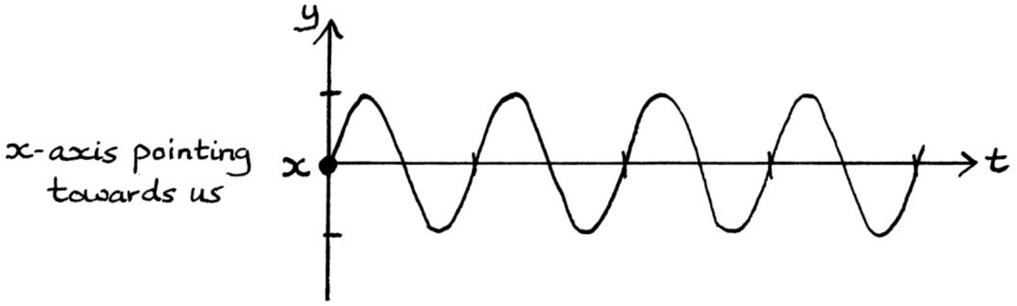
If we were to look at the helix end on, in other words, so that the time axis were pointing directly away from us, the helix would look exactly like our old circle (assuming there is no perspective):





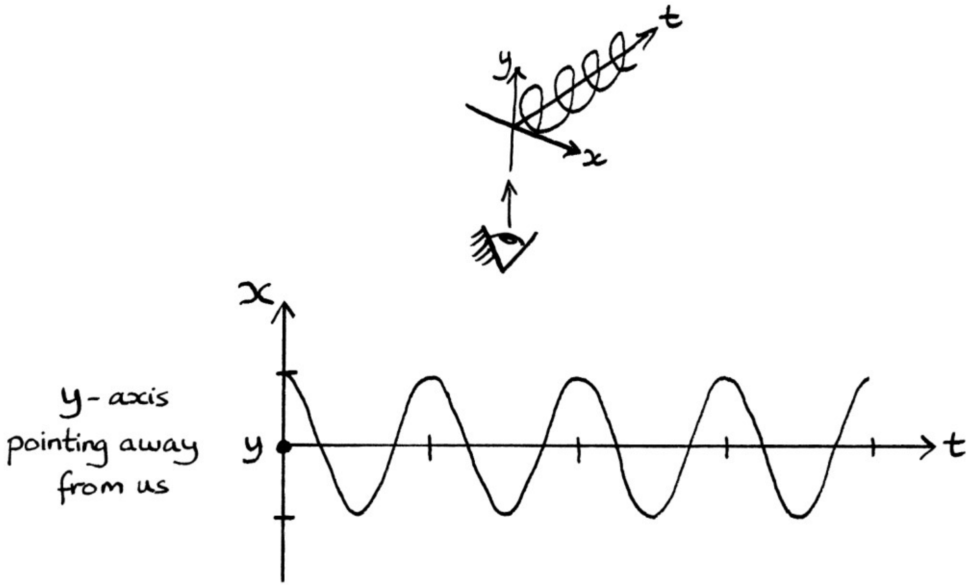
If we look at the helix side on, so that the x-axis is pointing directly at us, the helix will look exactly like our Sine wave graph:





In a way, this is obvious, as we are really seeing the y-axis values of the circle (which represent the Sines of the angles) as portrayed over time.

If we look at the helix from underneath, that is with the y-axis pointing away from us, it looks exactly like our Cosine wave graph.



Again, this is to be expected as we are seeing the x-axis values of the circle (which represent the Cosines of the angles) as portrayed over time.

[If these pictures do not make these ideas clear, hold a spring and look at it end on, side on, and from underneath.]

Our helix is really the missing link between the circle and the Cosine and Sine graphs. It contains all the information about all three, and strictly speaking, it *is* all three.

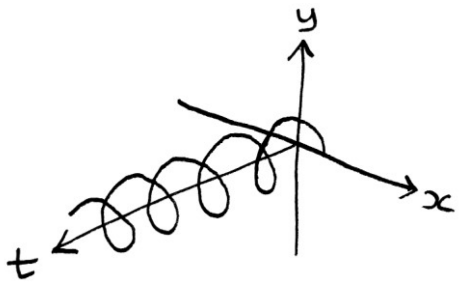
The helix can be a helpful way of thinking about the relationship between the circle, the Cosine wave graph and the Sine wave graph over time. The downside to the helix is that it is difficult to draw well or accurately, and it is difficult to read meaningful values from. Although I gave coordinates for the object's position on the helix, in practice, there probably is not much point in doing such a thing.

The helix is mainly useful just as a way of visualising how waves and circles fit together. Few people would actually use a drawing of a helix for any purpose other than explaining or teaching general concepts. However, the concept is an important one to consider. In practice, the time circle is a much more useful concept, but it pays to remember that the information portrayed in the circle could also be portrayed as a helix.

Possible sources of confusion

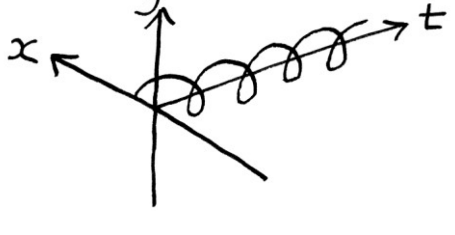
There are numerous ways of thinking about the time helix and the direction of the axes.

Some people think of the time helix with the time axis moving out of the page *towards* the viewer. In this way, with every rotation of the circle, the object moves one second towards the viewer. This changes how the helix looks.



There is nothing wrong with this idea of the time helix, but in my view, the helix looks better with time going into the page. It is a matter of choice. Note that if the helix is portrayed with the time axis coming out of the page, then looking at it side on or from underneath will not give the Sine and Cosine waves in the same way as before. Instead, it is necessary to look from the far side for the Sine wave, and from the far side and the top for the Cosine wave.

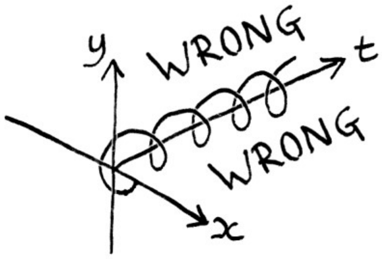
If someone portrays the time axis moving out of the page towards the viewer, then if the helix is turned around to face the other way, it will look like this:



The above picture looks confusing, but really, we are looking at the axes from the other direction. The time axis is moving into the page, but the x-axis has been flipped from one side to the other. This is still portraying the same information as the helix with the time-axis coming out of the page, but it is much harder to visualise what is going on. The object moving around the circle and the helix *seems* to rotate in the wrong way, but in reality, it is rotating the proper way for a helix with time coming out of the page, but the way it is drawn is confusing.

If we untangle the axes and draw them in the way that most people would expect to see them drawn, then the actual rotation is much more apparent. Some people will draw the helices in this way, and some people will draw helices in even more confusing ways. It pays to become used to untangling the drawings in your mind.

Sometimes books will have diagrams where the author, or their publisher, has mistakenly drawn the helix rotating the wrong way around:



The helix should rotate anticlockwise as the object moves outwards down the time axis (i.e. into the paper or the screen) because an object moving around the circle moves anticlockwise, or to put it another way, angles are measured as increasing anticlockwise. [Of course, the convention that angles are measured as increasing anticlockwise is a social construct, so technically there is no true correct or incorrect way of drawing the helix. However, if we say that angles increase anticlockwise, and the time axis is drawn as moving away from the observer, then the helix must rotate anticlockwise.]

As few people actually use a drawing of a time helix for obtaining accurate measurements, strictly speaking, it does not matter if people draw it incorrectly. The time helix is really just a way of illustrating a concept, so that people can learn about the subject of waves more easily. Therefore, the greatest minds in the world could think about it the wrong way around, and it would not affect their work in any way whatsoever. However, drawing it incorrectly confuses people trying to learn about waves.

The difference in time positions

When dealing with *angles* on the circle, every revolution is the same as the last. The results of the Sine of an angle, and the Sine of an angle a revolution later (360 degrees higher), are identical. The same is true for Cosine. When dealing with *time* and thinking of the waves or the helix, every revolution will produce the same results for the Sine or Cosine of the *angles*, but the position of an object will be different on the *time* axis. An object's up or down position will be the same, but its time position will be different. Therefore, technically, the first revolution of an object is not the same as the second revolution because the "time-position" of the object is different. This distinction is something to think about from the point of view of understanding the theory of waves.

Conclusion

The introduction of time is an important stage in learning about waves. Time makes waves slightly more complicated, but that complexity is reduced if you think about the circle or helix from which a Sine or Cosine wave are derived. A common theme in this book is that everything is more straightforward if you think about circles.

Chapter 5: Amplitude

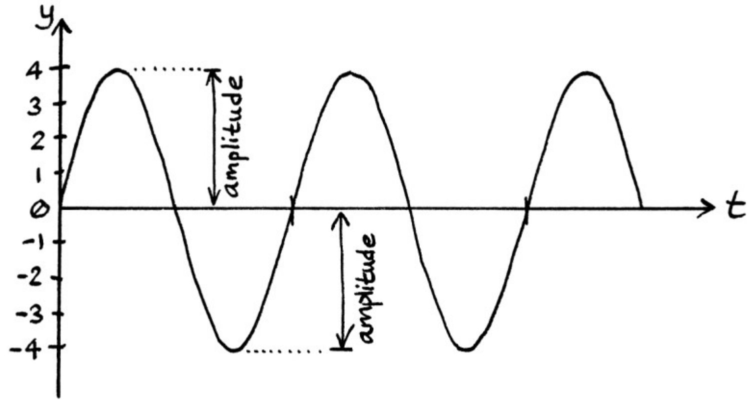
Over the next few chapters, we will look at the different attributes of Sine and Cosine waves relating to time. There are four distinct attributes of a Sine or Cosine wave that help us describe it. These are the amplitude, the frequency, the phase, and the mean level.

Amplitude

"Amplitude" is the name given to how far the points on a wave reach either up or down the y-axis from the centre of the wave. Another way of saying this is that the amplitude of a wave is the distance from the highest point on the wave's curve to its middle. This will always be the same as the distance from the lowest point on the wave's curve to its middle. Another way of defining amplitude is by saying that it is half the distance from the highest point on the curve to the lowest point on the curve. These definitions all rely on a wave showing at least one repetition of its shape.

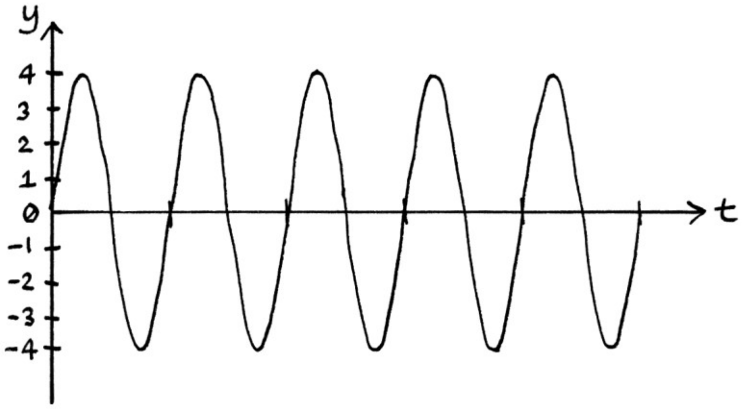
Examples

This wave has an amplitude of 4 units. Its maximum point is at +4 units. Its minimum point is at −4 units. The distance from its highest y-axis value or its lowest y-axis value to its middle is 4 units.

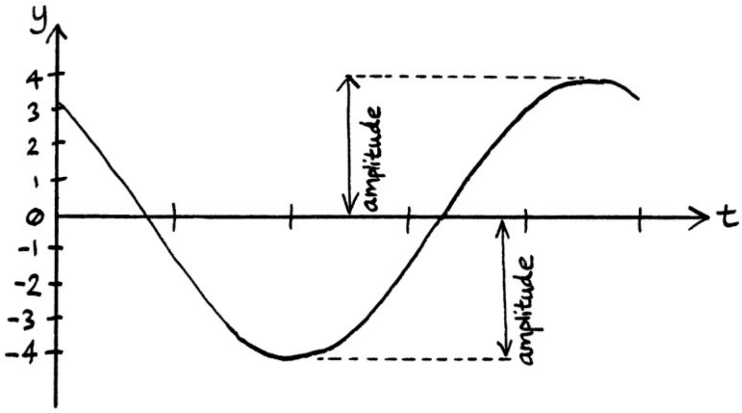


I did not number the time axis on this graph to emphasise how amplitude is only relevant to the y-axis.

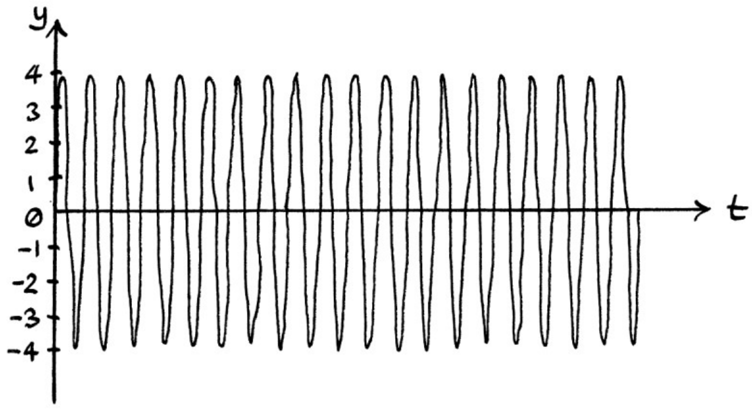
The following wave, despite having a slightly different shape, also has an amplitude of 4 units:



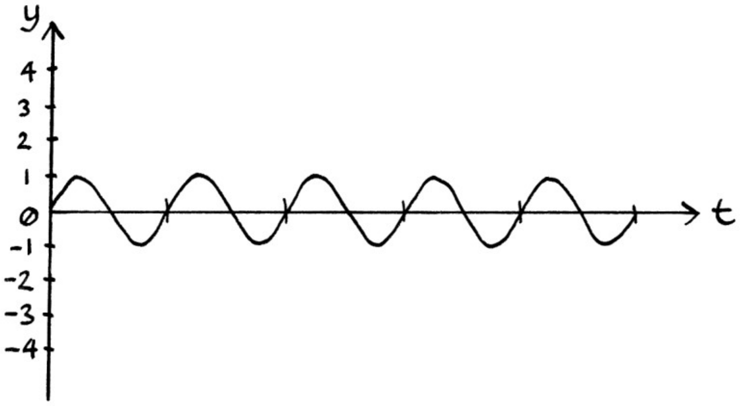
This wave, although it does not start in the same way as the Sine or Cosine waves that we have seen before, also has an amplitude of 4 units:



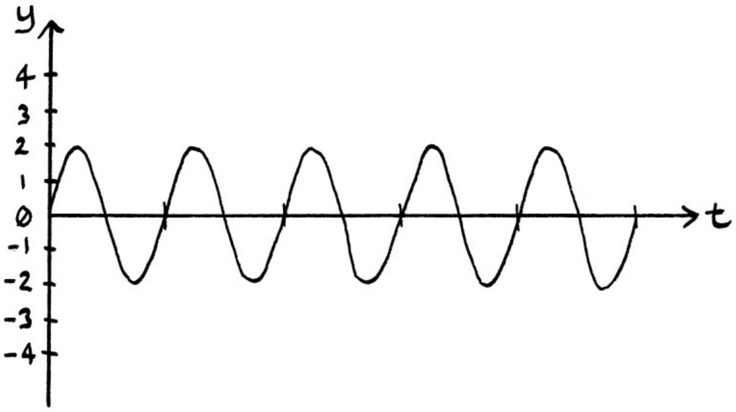
... as does this wave:



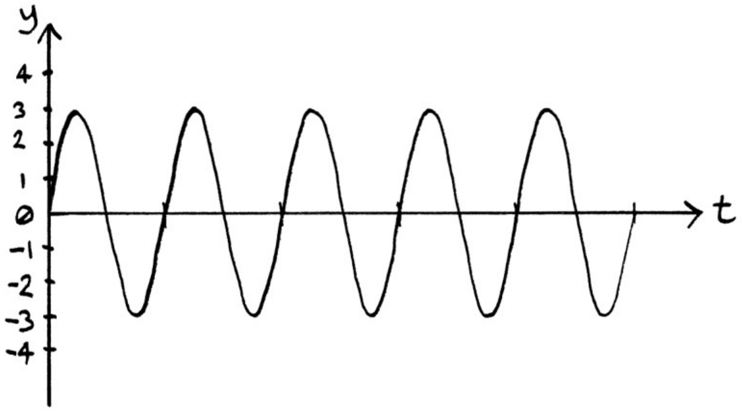
The following wave has an amplitude of 1 unit. Its maximum y-axis value is 1 unit above its centre. Its minimum y-axis value is 1 unit below its centre.



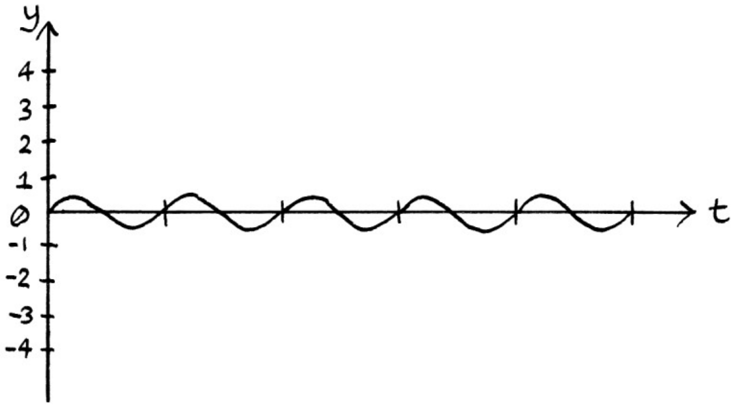
This wave has an amplitude of 2 units:



This wave has an amplitude of 3 units:

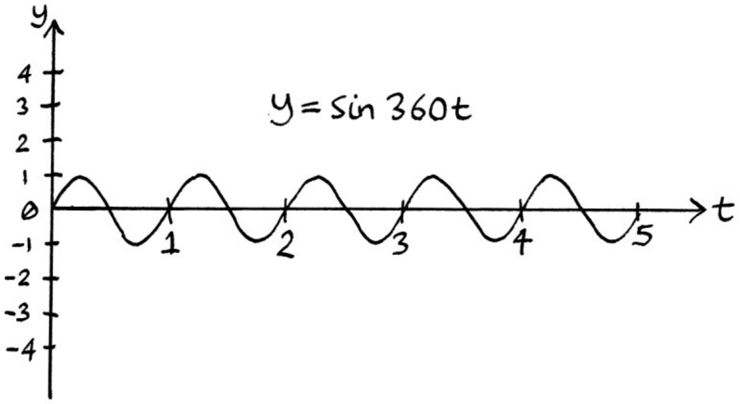


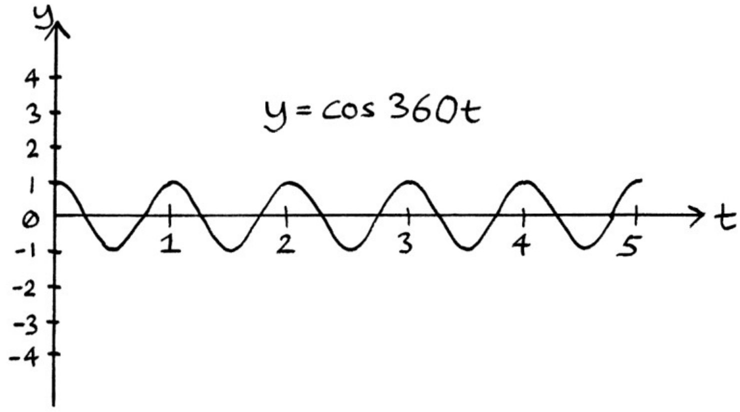
This wave has an amplitude of 0.5 units:



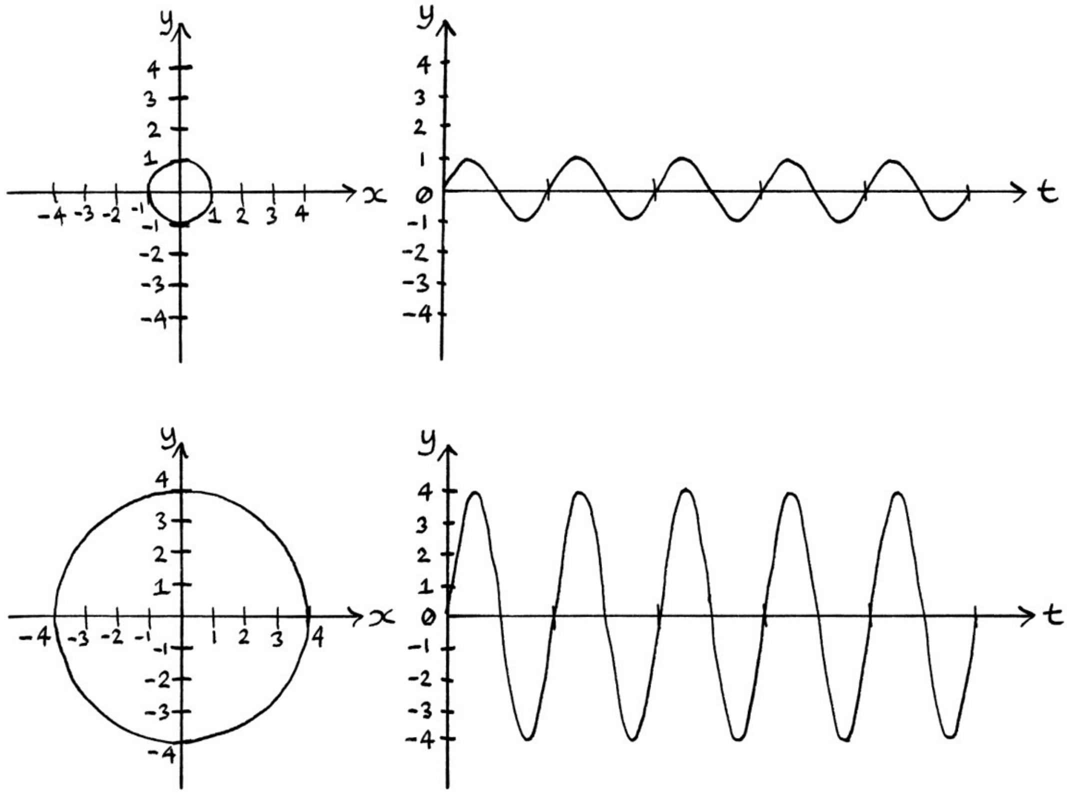
Amplitude and circles

The amplitudes of our “y = sin 360t” and “y = cos 360t” waves are both 1 unit. In other words, the maximum and minimum points reached by “y = sin 360t” or “y = cos 360t” are +1 and −1. The maximum distance that the points on each wave reach in either direction from their centres is 1.



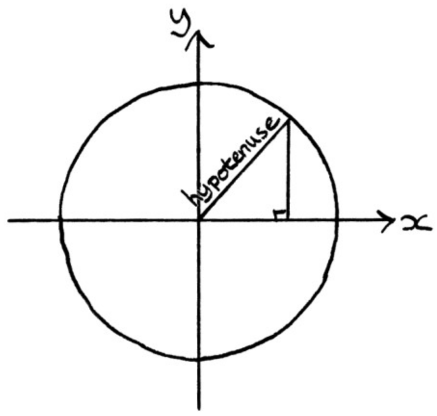


The amplitude of a Sine wave or a Cosine wave is the same as the radius of the circle from which it is derived.



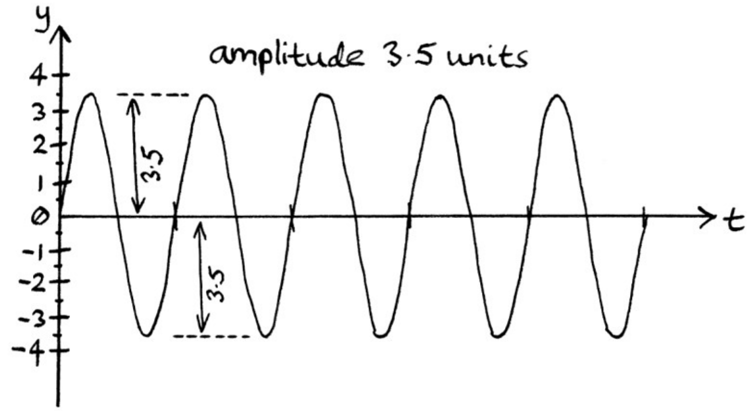
We can tell this would be so by how the highest point on a Sine wave will be the same as the highest point on the circle's edge, and the highest point on a Cosine wave will be the same as the right-most point on the circle's edge. Both these values are equal to the radius of the circle.

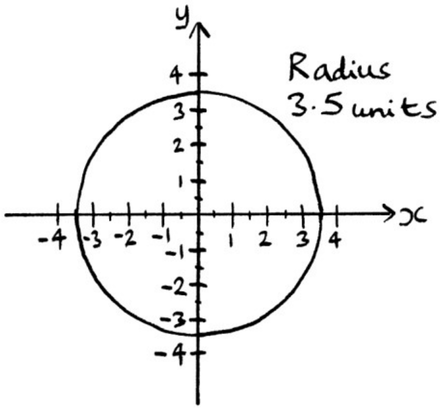
The amplitude of a Sine wave or a Cosine wave is also equal to the length of the hypotenuses of the countless right-angled triangles that make up that circle. The hypotenuses of the right-angled triangles are equal in length to the radius of the circle.

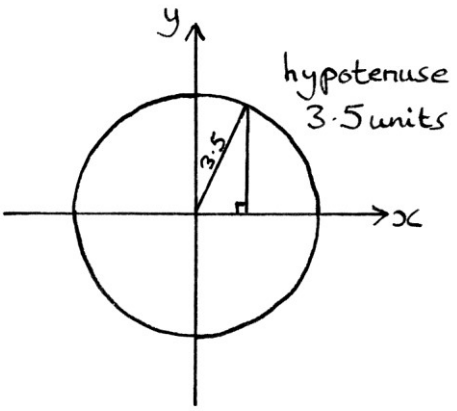


We could also say that those hypotenuses are equal to the length of the stick that corresponds to the hypotenuses of those triangles.

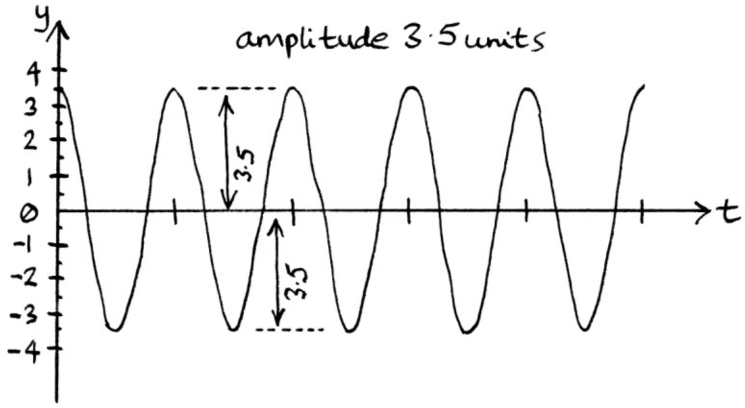
As an example of radius and amplitude being connected, if the amplitude of a Sine wave is 3.5 units, then the maximum and minimum y-axis values will be +3.5 units and −3.5 units. The circle from which such a Sine wave is derived would have a 3.5 unit radius, and the triangles that make up that circle would have hypotenuses that are 3.5 units long.



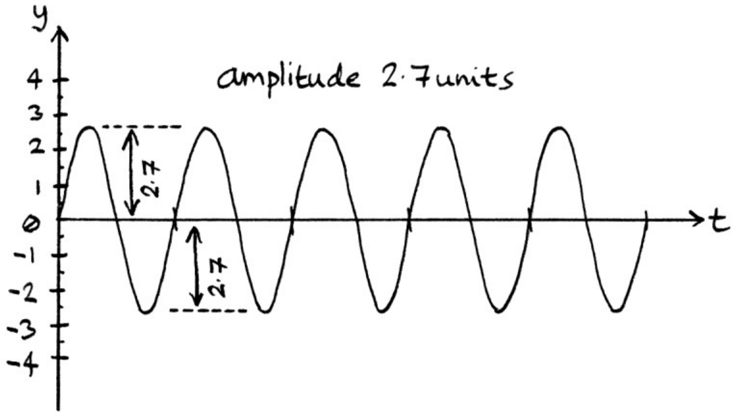


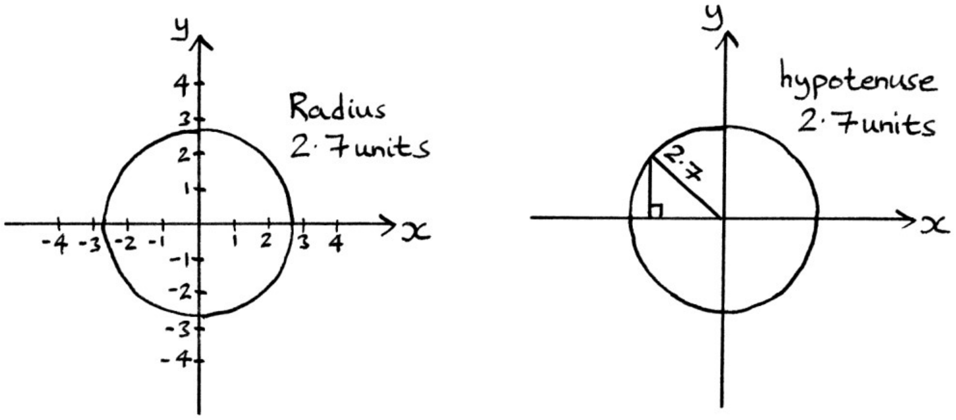


The Cosine wave derived from that circle would also have an amplitude of 3.5 units.

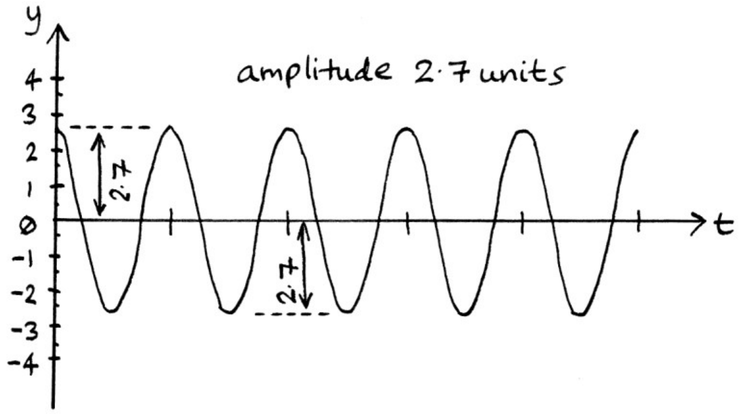


If the amplitude of a Sine wave is 2.7 units, then the maximum and minimum y-axis values will be +2.7 units and −2.7 units. The circle from which such a Sine wave is derived would have a 2.7 unit radius, and the triangles that make up that circle would have hypotenuses that are 2.7 units long.

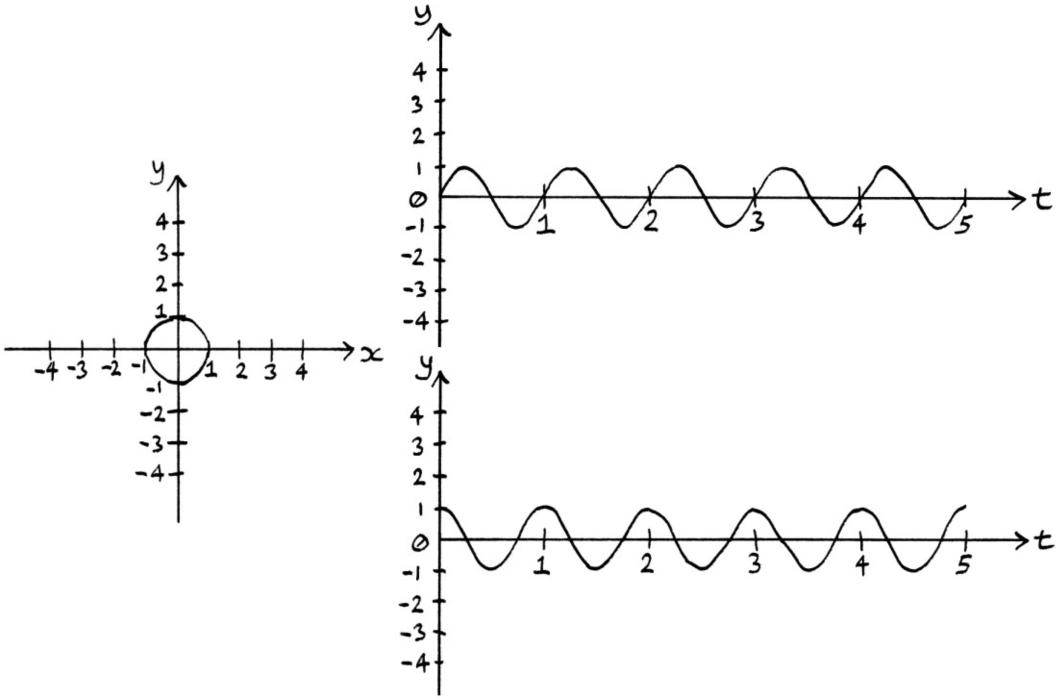


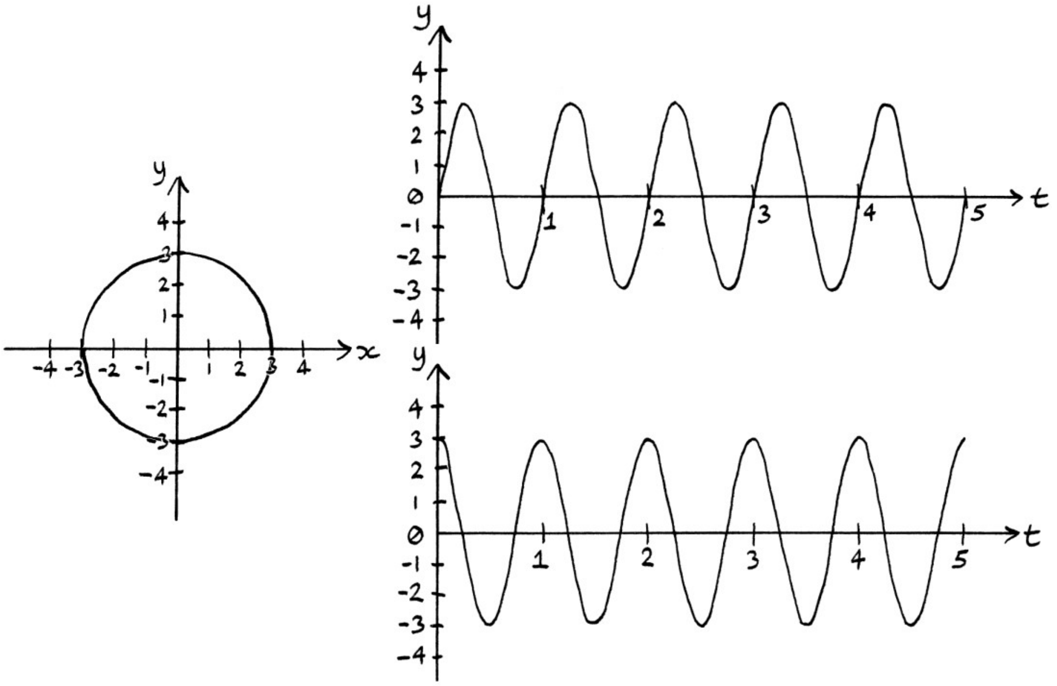


The Cosine wave derived from that circle would also have a 2.7-unit amplitude:



Another way of describing the amplitude of a wave is by saying it refers to how much the radius of the circle, from which the wave is derived, has been scaled in relation to a unit-radius circle. [If we wanted a way of remembering the term "amplitude", we could say it refers to how much the radius of the circle has been *amplified*].

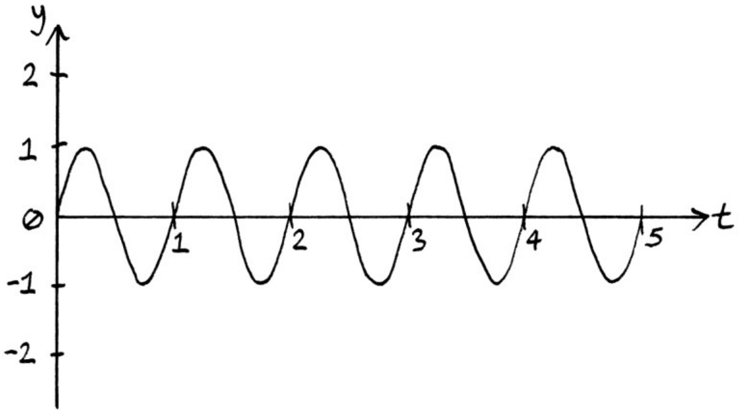




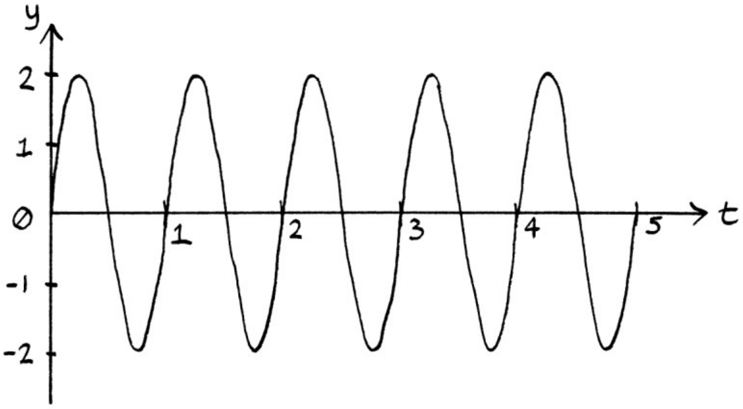
Formulas

For our original “y = sin 360t” wave, if we double the overall amplitude, our formula becomes “y = 2 * (sin 360t)”. This would more commonly be written without brackets as “y = 2 sin 360t”. The result of the Sine of every value becomes multiplied by 2 afterwards. In other words, for every point on the t-axis, “y” will be double what it used to be. The peaks of our graph double in size, and reach up to +2 and down to −2. They have been scaled by 2.

“y = sin 360t”:



“y = 2 sin 360t”:

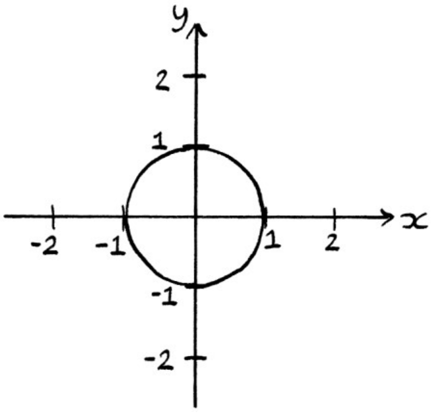


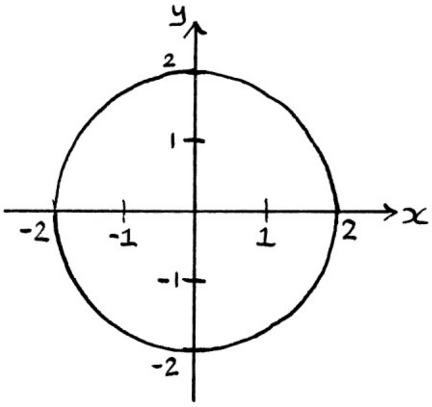
Note that the “y = 2 sin 360t” wave still repeats every second. It is only the height of the peaks and dips that has changed.

When the “y = 2 sin 360t” formula is portrayed by the circle from which it is derived, we will have a circle with a radius of twice the size, which is 2 units.

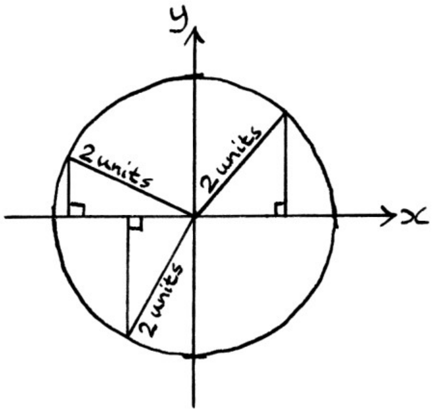
Circle for “y = sin 360t”:

Circle for “y = 2 sin 360t”



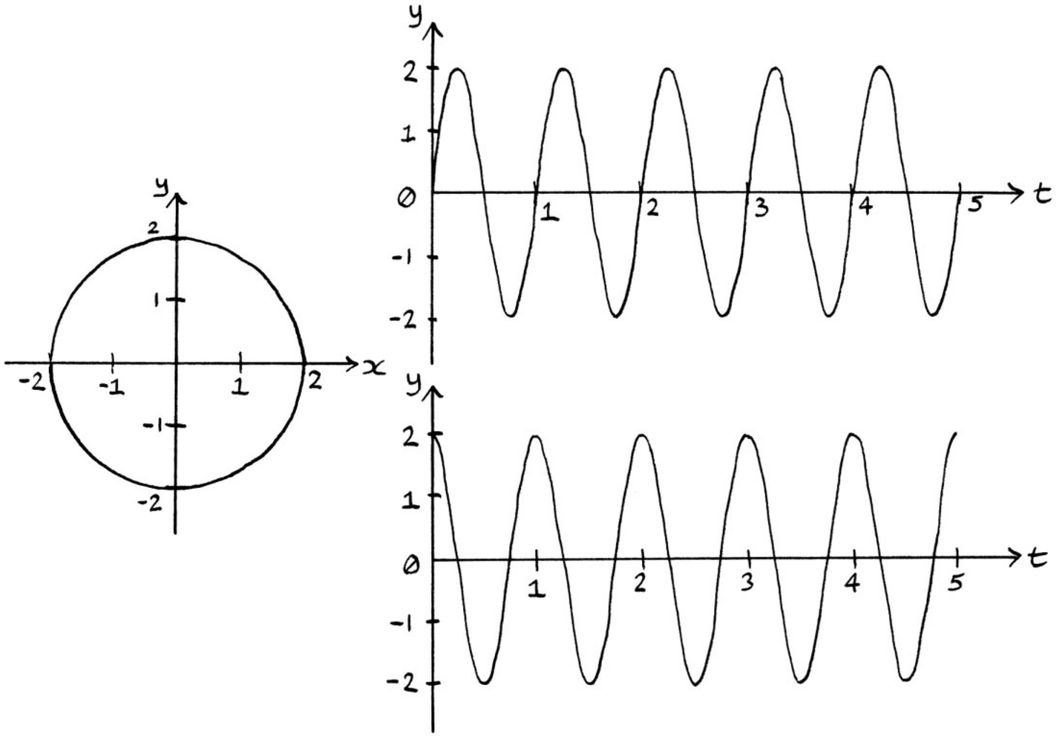


If we draw some of the right-angled triangles that make up the circle, we can see that the hypotenuses of each one has doubled in length:



If we went back to the stick example at the start of this book, the stick would be 2 metres long instead of 1 metre long.

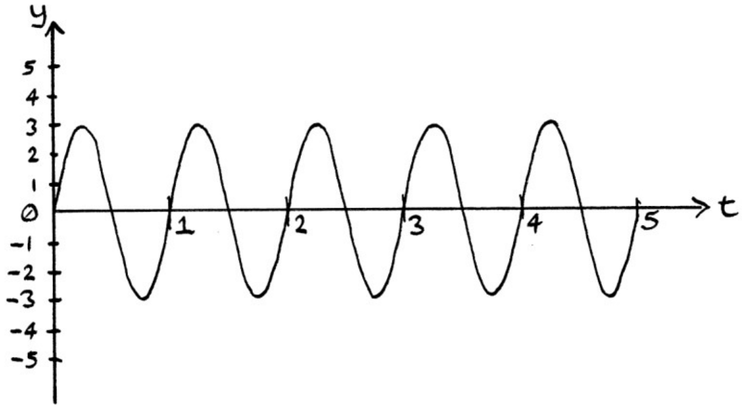
As the circle from which our "y = sin 360t" wave came has doubled in radius, the Cosine wave from that circle will also have doubled in amplitude. Therefore, "y = 2 sin 360t" implies a corresponding Cosine wave of "y = 2 cos 360t":



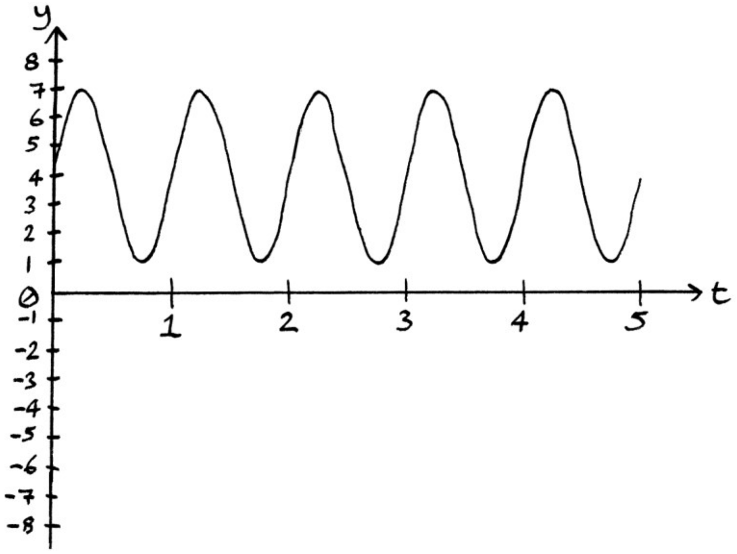
Waves not centred at y = 0

Supposing the wave is not centred at “y = 0” (which is the same as saying it is not centred on the x-axis or t-axis), the amplitude will still be measured as the distance from either the highest or the lowest point of the wave to the wave’s centre. Therefore, the following three waves all have the same amplitude – an amplitude of 3 units.

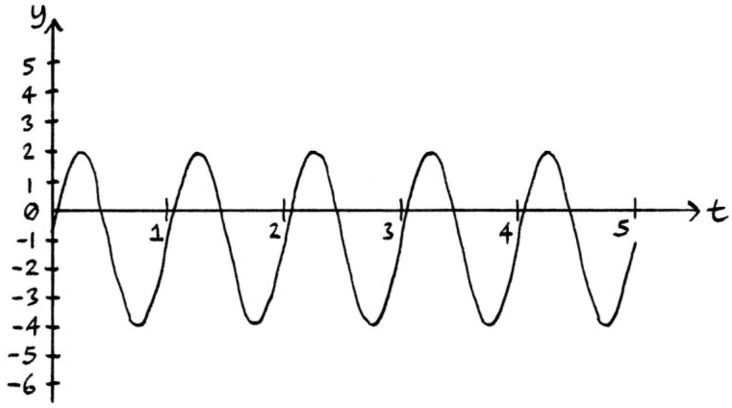
The following wave’s maximum point is at +3 units; its minimum point is −3 units. Its centre is at y = 0. Therefore, its amplitude is 3 units.



This wave’s maximum point is at +7 units; its minimum point is at +1 units. Its centre is at +4 units. We can calculate its amplitude in two ways: 7 − 4 = 3 units, or 4 − 1 = 3 units.



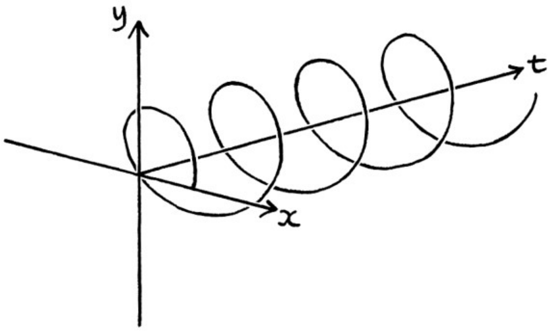
The following wave’s maximum point is at +2 units; its minimum point is at −4 units. Its centre is at −1 units. Its amplitude is 2 − −1 = 3 units. We could also have worked this out as −1 − −4 = 3 units.



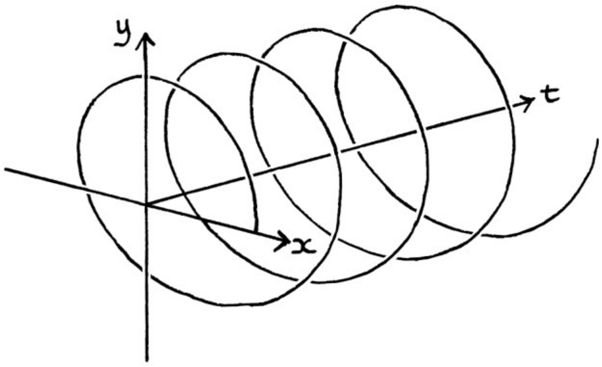
I will explain how waves that are not centred at “y = 0” appear on the circle when I discuss mean levels in Chapter 8.

Amplitude on the helix

When a circle and its pair of waves are portrayed as a helix, the amplitude indicates what is essentially the radius of the helix. This might be obvious, given that the helix is the same size as the circle and the amplitude is the radius of the circle.



A bigger amplitude means a bigger circle, and therefore a bigger helix.



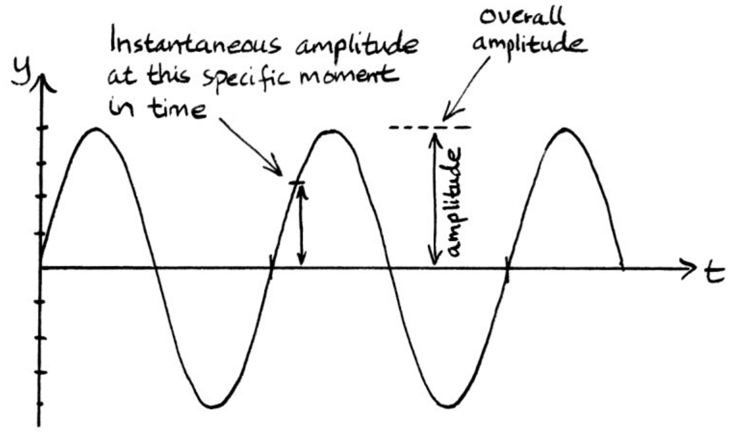
Note that the helix shape will still repeat in the same amount of time as one with a lower amplitude. This is because amplitude only affects the radius of the helix.

A potential source of confusion

Note that the term “amplitude”, as I am using it here, refers to the overall amplitude of the wave – in other words, the scaling factor for the wave as a whole in relation to a wave that derives from a unit-radius circle. Sometimes, people will also use the term “amplitude” to refer to the y-axis value of a wave graph at a *particular* point. They might say such things as “the amplitude when t = 0.4 seconds is 7 units”. In such a case, it might be better if they used a different term such as “instantaneous amplitude” to avoid confusion. Some people treat the word “amplitude” as mainly meaning the “instantaneous amplitude” and they might refer to “overall amplitude” with terms such as “peak amplitude”. [Peak amplitude is used because it refers to the “peaks” or the maximum and minimum points of the graph].

In this book, when I use the word “amplitude” on its own, I will *always* be referring to “overall amplitude”, which is the same as “peak amplitude”. If I want to refer to the y-axis value at a particular point on the graph, I will refer to “instantaneous amplitude”. As you become more used to waves, the possible confusion from other people’s use of words diminishes because you will recognise the meaning from the context. However, it pays to be aware of the potential ambiguity.

The different meanings of "amplitude" are shown here:



Formulas for amplitude

The general formula for a Sine wave relating to time that takes into account amplitude is:
y = A sin 360t
... where "A" stands for "amplitude" and "t" stands for the time in seconds.

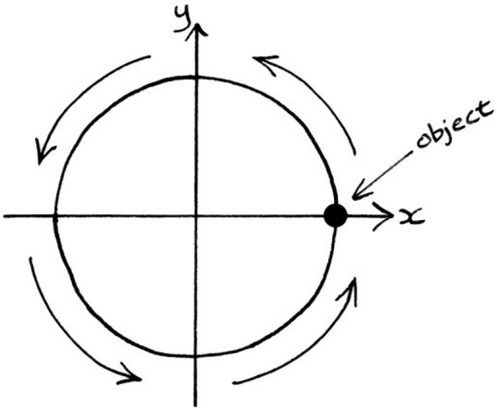
The formula for a Cosine wave that takes into account amplitude is:
y = A cos 360t

If we were dealing with waves that had no mention of time, and just referred to angles, amplitude would still have the same meaning and the same effect as it does here. There would be no mention of time and so no need for the multiplication by 360 to correct the time. The formulas would be:
y = A sin θ
... and:
y = A cos θ

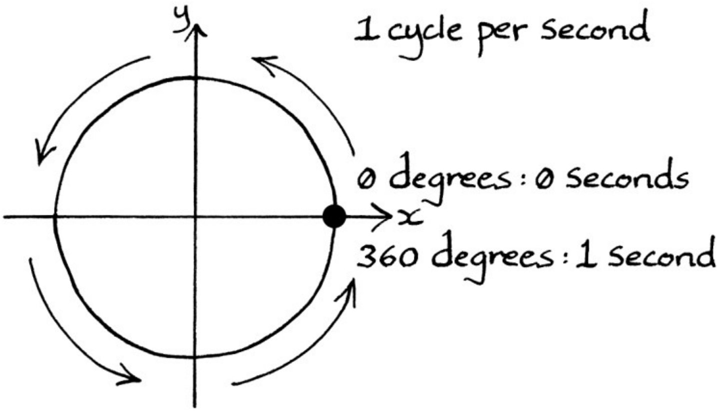
Chapter 6: Frequency

Frequency

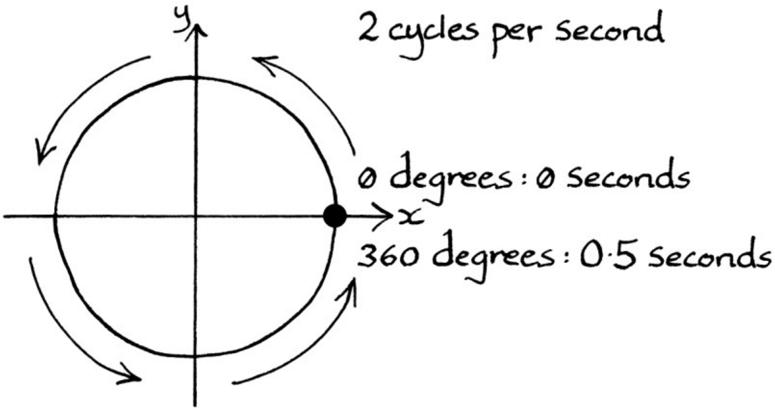
When thinking of the circle, frequency refers to the number of times per second that an object rotating around it completes one full revolution.



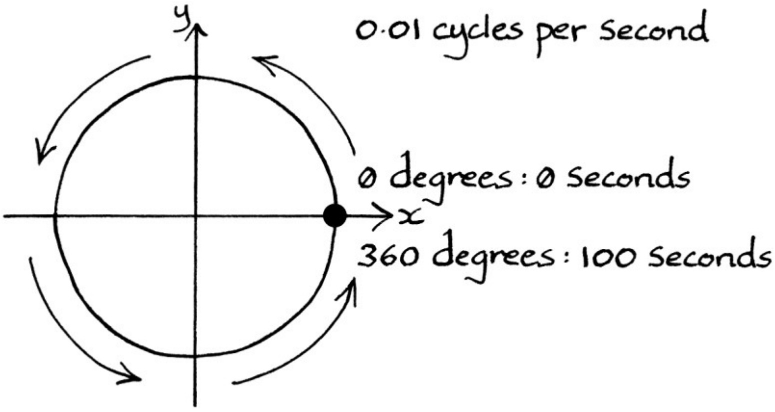
If an object completes one rotation of the circle in one second, that is to say, it takes one second to complete one revolution, then its frequency is said to be "1 cycle per second". It completes one cycle in one second.



If an object completes 2 rotations around the circle in one second, that is to say, it only takes 0.5 seconds to do 1 revolution, then its frequency is 2 cycles per second.

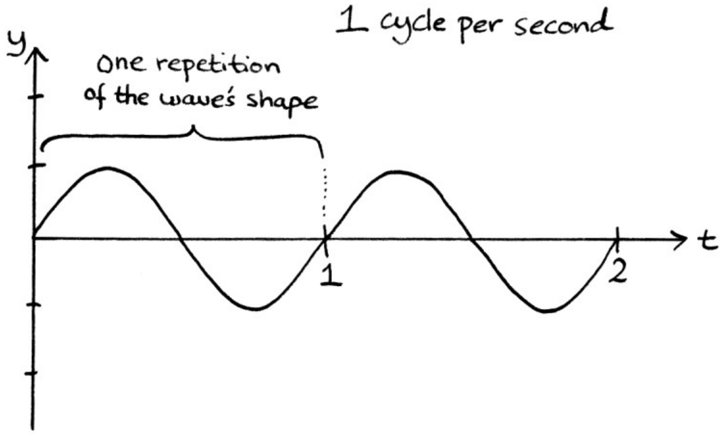


If an object completes 0.01 revolutions per second, that is to say, it takes 100 seconds to rotate around the circle, then its frequency is 0.01 cycles per second.

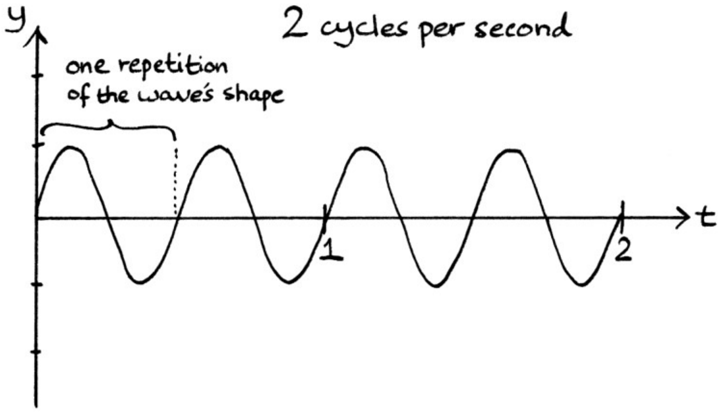


When thinking of the waves derived from a circle, the frequency of an object can be seen in the number of times per second that the wave completes one repetition of its *shape*.

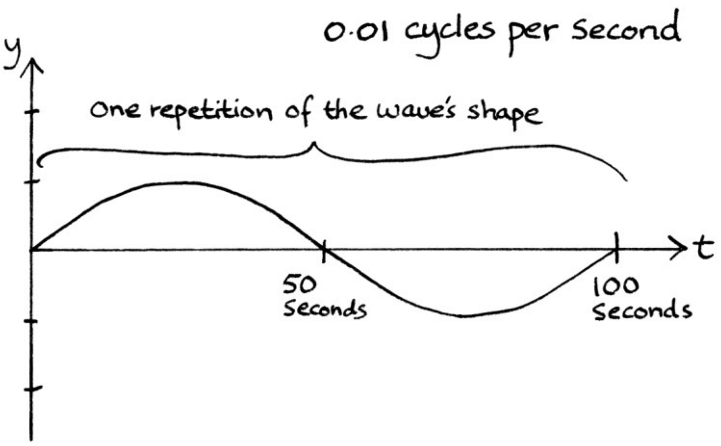
If the wave's shape repeats once every second, it has a frequency of 1 cycle per second:



If the wave's shape repeats once every half second (it repeats twice per second), it has a frequency of 2 cycles per second:



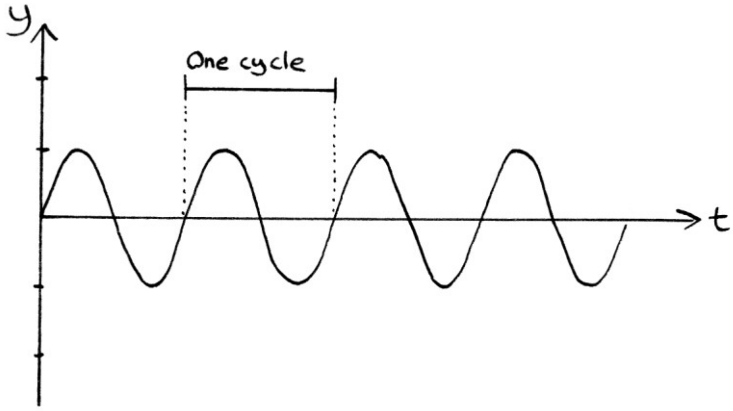
If the wave's shape repeats once every 100 seconds (it repeats 0.01 times per second), then it has a frequency of 0.01 cycles per second:



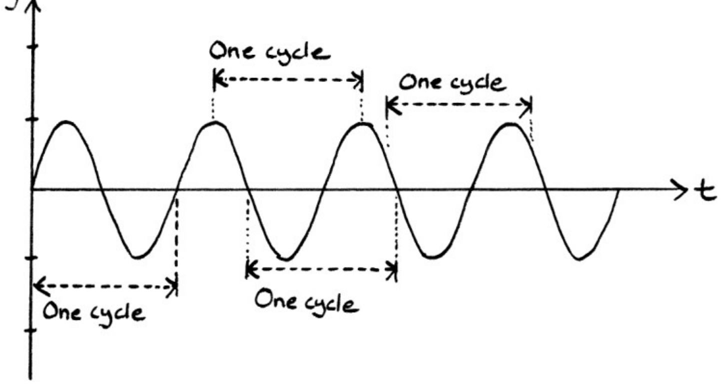
[The above wave repeats after such a long time that I changed the scale on the graph's time axis so that it fits on the page.]

Given that a wave is derived from a circle, it should be apparent that the frequency of an object rotating around a circle will be the same as both the frequency for the Sine wave and the frequency for the Cosine wave that describe the position of that object.

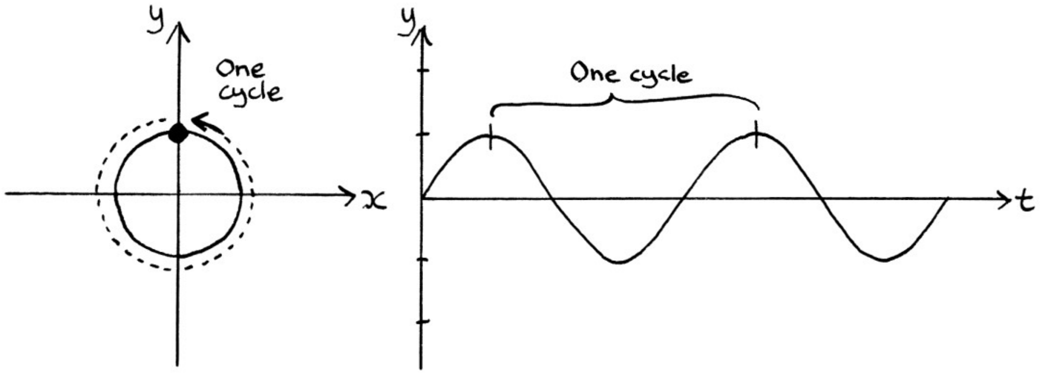
The proper name for a section of the wave before its shape starts again is "a cycle", which is a term that matches the movement of an object around the circle.

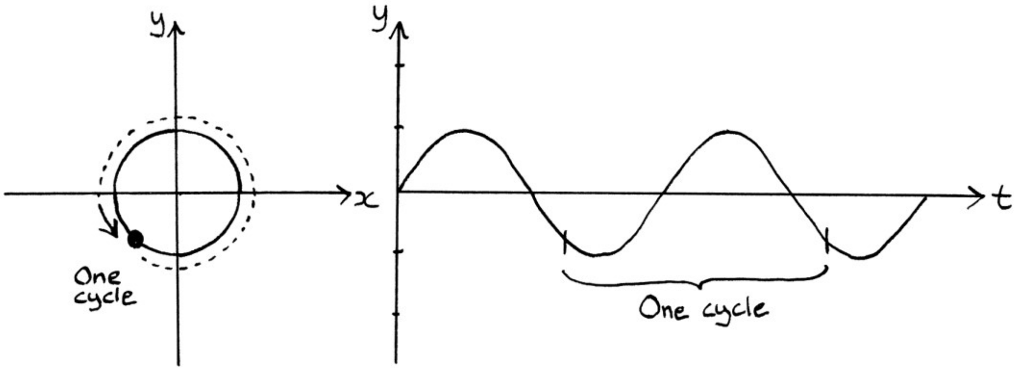


We can think of one cycle as being a section of the wave between two peaks, or a section between two dips, or a section between two parts rising from zero, or in fact, any two places that are the nearest corresponding points of a wave's shape. These all amount to exactly the same thing when defining a cycle on a particular wave. A cycle is just one section of a repeating wave.

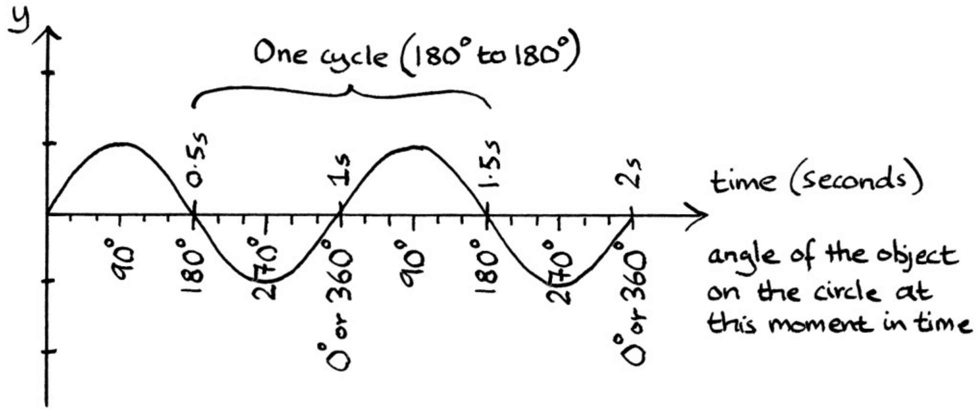


A cycle on a wave, as measured from one similar point to the next, is related to a cycle completed on the circle the wave is derived from, starting from any point around the circumference and finishing at that same point.



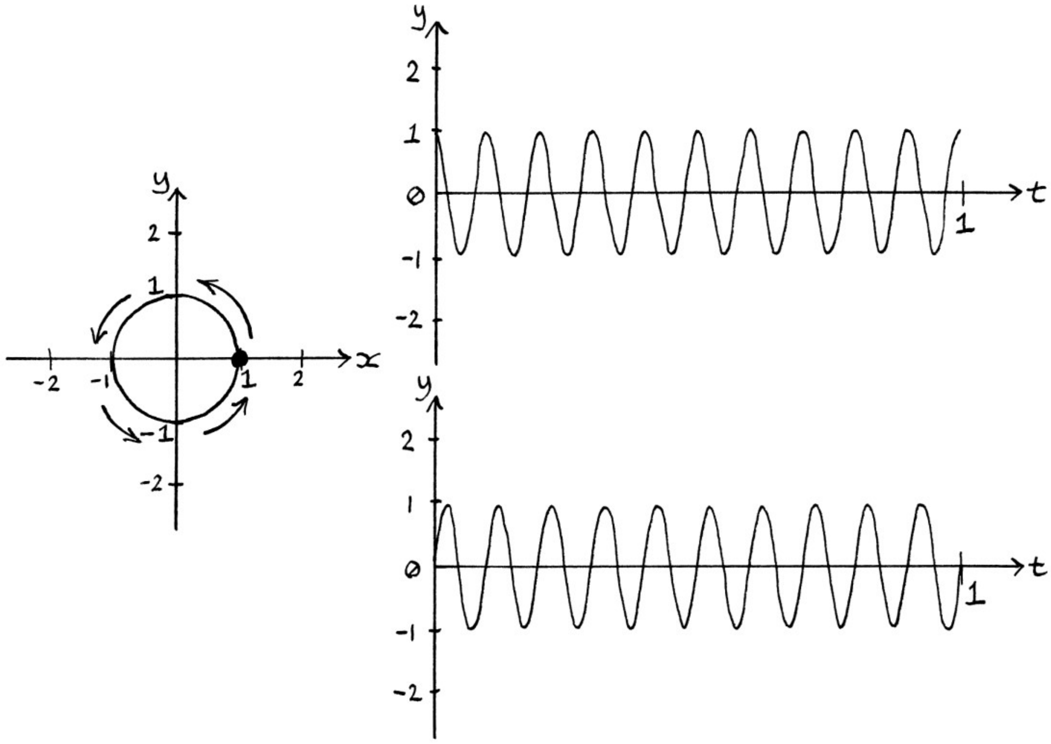


Another way of saying all this is that a cycle on a wave is a section where the object rotating around the circle will start and end at the same angle (the same place) on the circle. This idea is easiest to see if the angle of the object on the circle for each time is marked on the time graph. In the following graph, the x-axis is still time in seconds, but I have written the angles of the object on the circle at particular moments in time as well.



To summarise everything so far, frequency, when looking at a circle, is the number of cycles completed in one second. Frequency, when looking at a wave, is also the number of cycles completed in one second.

An object moving around a circle at a particular frequency will produce a Sine and Cosine wave that have the exact same frequency. As an example, an object rotating around the circle at 10 cycles per second, will produce a Sine wave and a Cosine wave that also repeat at a rate of 10 times per second. The full range of y-axis values of the circle (the Sine of the angles) and the full range of the x-axis values of the circle (the Cosine of the angles) will repeat 10 times per second.



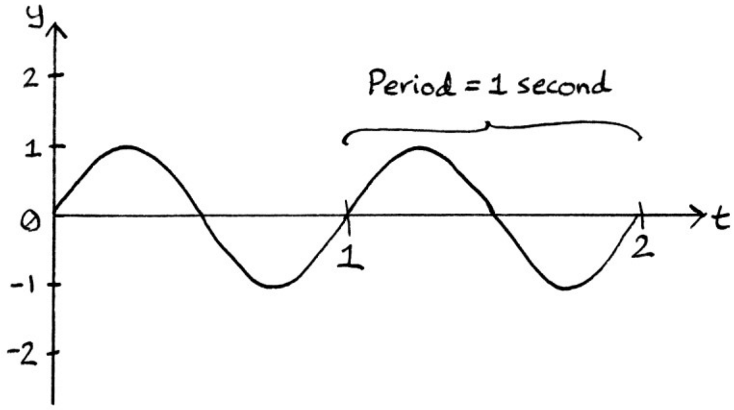
Frequency is measured in “cycles per second” (abbreviated to “cps”), or “hertz” (abbreviated to “Hz”), where 1 hertz is one cycle per second. Cycles per second and hertz are two terms that mean exactly the same thing. In this book, I will use the terms “cycles per second” and “hertz” interchangeably depending on the context. From the point of view of learning, “cycles per second” is a better term as it reinforces the idea that something is repeating a cycle every second, whether that cycle is a cycle of a wave or a rotation around a circle. Therefore, if I am trying to explain something related to frequency, I will use “cycles per second”, and otherwise I might use hertz. In modern mathematics books, “hertz” is the more commonly used term.

Note that the term "hertz" has a lower-case "h", even though it is named after a scientist with the name "Hertz". The abbreviation for "hertz" (Hz) has an upper-case "H", even though it is short for "hertz" with a lower-case "h". This is consistent with the generally accepted conventions for naming scientific concepts.

Period

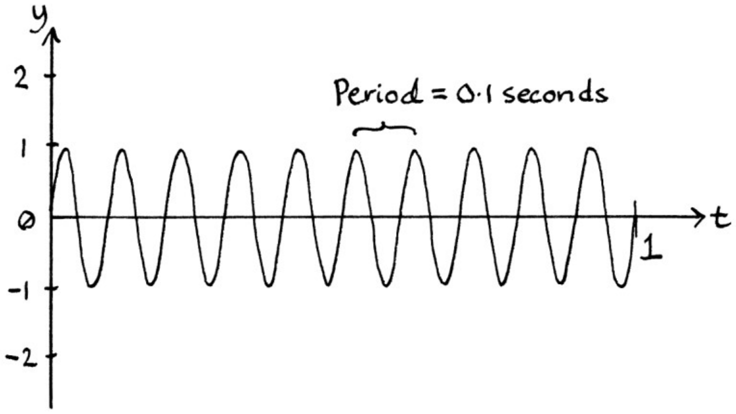
As I have said, frequency is the number of cycles completed per second. The term for the opposite of this, that is to say the number of seconds required to complete one cycle, is the "period". The period is the length of time that one cycle takes to complete – it is the *period* of time that one cycle takes to complete. It refers to the seconds per cycle. We could also say that the period is the duration of one cycle. All these definitions mean the same thing. The term "period" can be used to describe the duration of one cycle of a wave, and also the duration of one revolution on the circle.

For this wave, the period is 1 second. A cycle lasts one second.

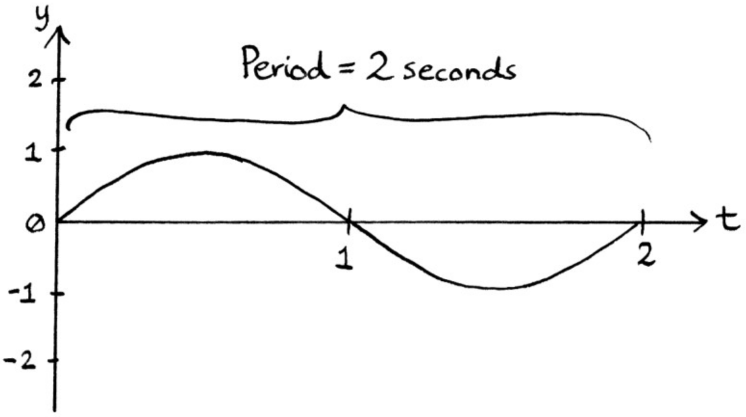


The frequency and period of a particular wave have a fixed relationship to each other – the period is the reciprocal of the frequency, and the frequency is the reciprocal of the period. In other words, 1 divided by the frequency gives the period, and 1 divided by the period gives the frequency. If we know the frequency, then we have enough information to calculate the period; if we know the period, we have enough information to calculate the frequency.

If a wave has a frequency of 10 cycles per second, then its period, in other words, the period of time required to complete one cycle, or in other words, the duration of one cycle, will be 1 ÷ 10 = 0.1 seconds.



If a wave has a period of 2 seconds, then its frequency will be 1 ÷ 2 = 0.5 cycles per second.



Side note: wavelength

Some real-world entities, such as radio and sound, travel over a distance as well as having characteristics that can be portrayed using waves. For such entities, it is possible to draw a wave graph showing the fluctuations over time, and it is also possible to draw a wave graph showing the fluctuations as they happen over *distances*. In such cases, there are really two types of waves that can describe the behaviour of an entity: time-based waves and distance-based waves. If the fluctuations over a distance are portrayed on a graph, the x-axis becomes the "distance axis", and will have units of distance such as metres. In such a case, the equivalent of frequency would not be measured in cycles per second, but in cycles per *metre*. The equivalent of period would be measured in *metres per cycle*, and would be called “wavelength” or “the wavelength”. Wavelength refers to the literal *distance* over which a moving wave repeats one cycle. This is the same as the distance between the start and end of a cycle. It is the length of a cycle.

Period refers to the *time* over which a cycle repeats; wavelength refers to the *distance* over which a cycle repeats. It is important to know that wavelength is only relevant to a graph if the x-axis of the graph is distance. Similarly, period is only relevant to a graph if the x-axis of the graph is time.

Not all entities that have characteristics that can be described using time-based waves have characteristics that can be described using distance-based waves. It pays to remember that there are more types of waves than just sound and radio waves. For example, a point on the outside of a spinning disc that is fixed to the ground can be portrayed using a time-based wave, but there would be little point portraying it with a distance-based wave because all the y-axis values would be placed on top of each other at x = 0 metres. On the other hand, the point on the top of a piston on a moving vehicle could be portrayed with a time-based wave and a distance-based wave.

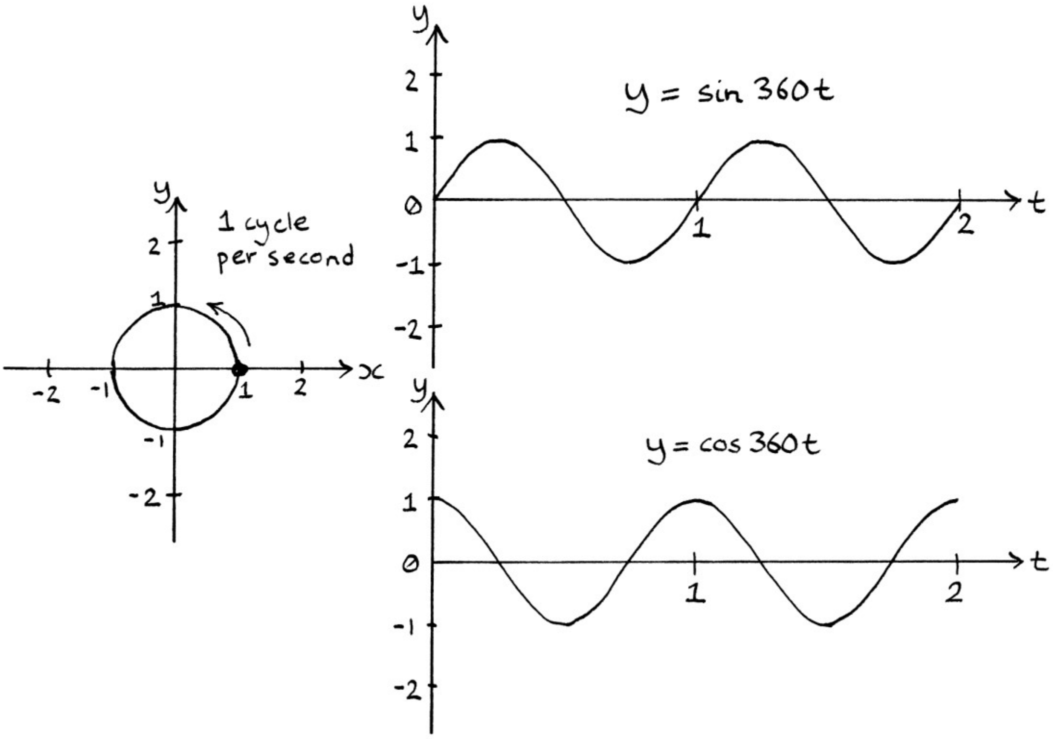
Although similar concepts, wavelength and period refer to different ideas. They are regularly confused with each other, even by people who should know better.

I am only mentioning wavelength here to stop any confusion with period. Ignore the idea of wavelength for now, as it will not be relevant until much later in this book. I will explain more about it in Chapter 31.

Frequency in formulas

If we think of an object rotating around a circle and taking 1 second to complete a revolution, then its frequency will be 1 cycle per second. Its period will be 1 second. The formula for the Sine wave derived from this circle will be:
“y = sin 360t”
[The multiplication by 360 is necessary so that the Sine function can be performed on the time in seconds, as explained in Chapter 4.]

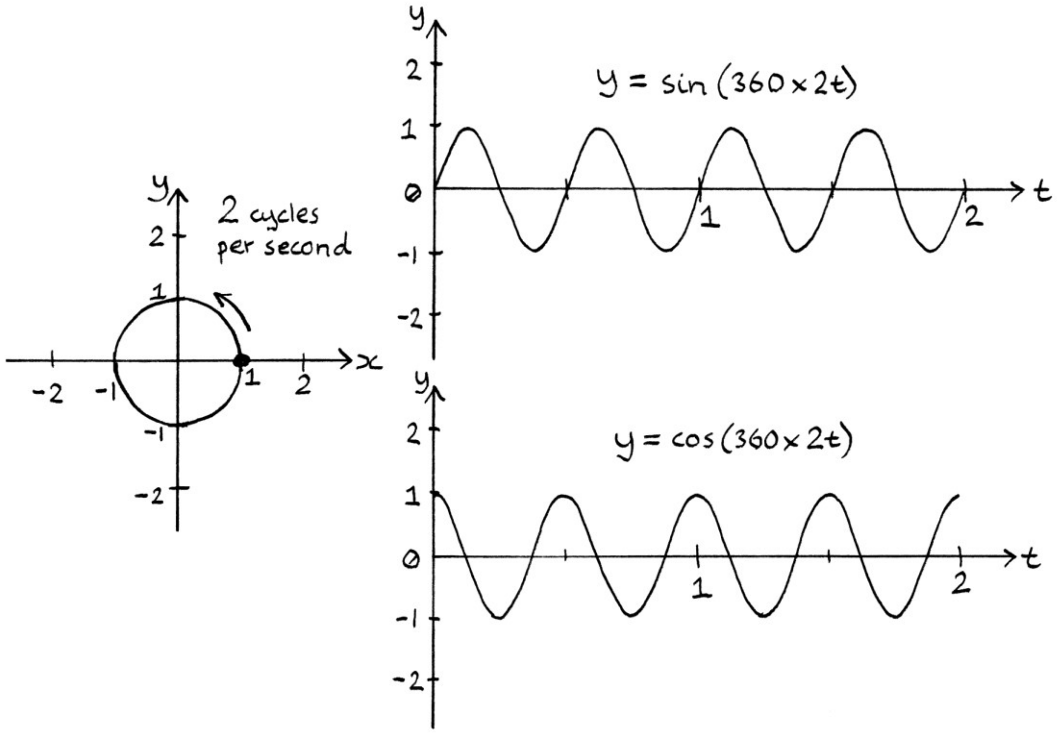
The formula for the Cosine wave derived from the same circle will be:
“y = cos 360t”



If we double the frequency of our object’s rotation around the circle, we will also be doubling the frequency of our Sine wave and our Cosine wave. The “y = sin 360t” and “y = cos 360t” formulas thus become:
“y = sin (360 * 2t)”
... and:
“y = cos (360 * 2t)”.

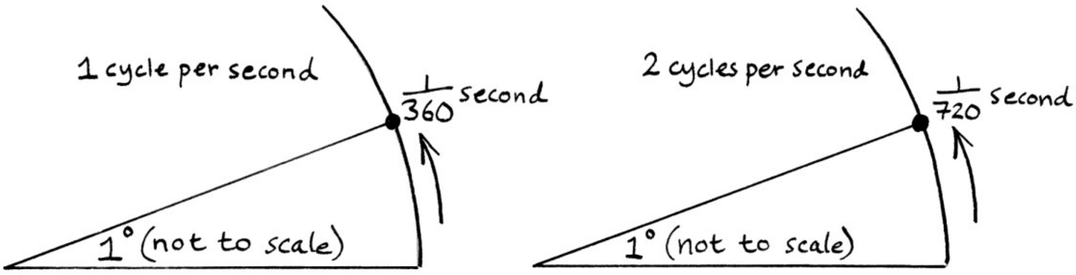
[Note that I could equally well write these as “y = sin 720t” and “y = cos 720t”, but for clarity’s sake, I am keeping the 360 and the frequency separate.]

The formulas show that “t” is doubled before it is multiplied by 360, and then subjected to the Sine and Cosine functions. Therefore, the amount being Sined or Cosined is double what it was before, and the y-axis values on the graphs reach the peaks and dips sooner than if the amount had not been doubled. This results in our waves speeding up and repeating their cycles twice as quickly. The peaks of the wave graphs become closer together.

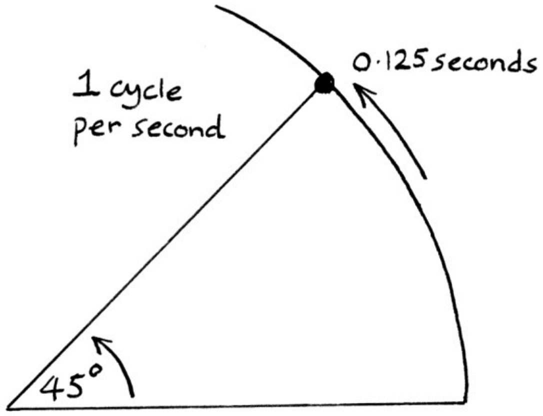


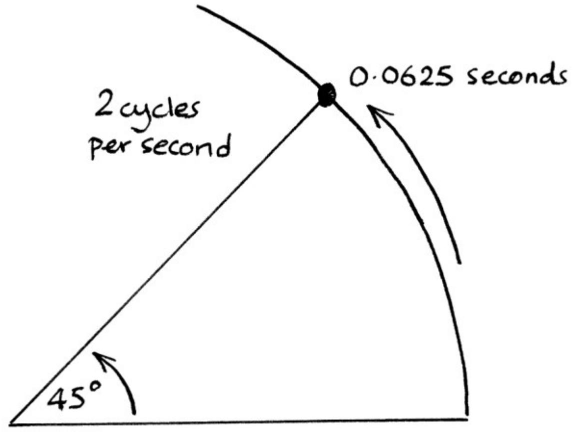
Note that the radius of the circle from which the two 2-cycles-per-second waves derive, is the same as the radius of the circle from which the original 1-cycle-per-second waves were derived. In other words, the amplitude of the two waves is the same as before. Changing the frequency of the circle or its waves does not affect anything else. We can know this is true by how a circle does not change its radius when something rotates faster around it.

If we were to look at the movement of the object around the circle's edge, we would see it moves more quickly. Whereas before, when the frequency was 1 cycle per second, it moved one degree in 1/360th of a second, now it moves 2 degrees in 1/360th of a second. Or to put it another way, it moves one degree in 1/720th of a second. It is moving at twice the speed.

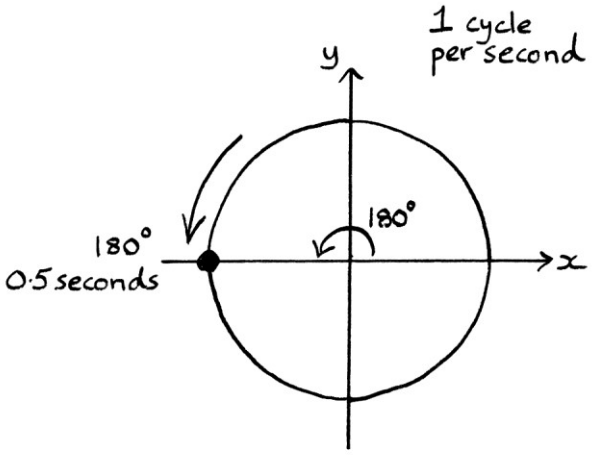


Whereas previously, it reached 45 degrees in forty-five 360ths of a second (0.125 seconds), now it reaches 45 degrees in only 22.5 360ths of a second (0.0625 seconds).

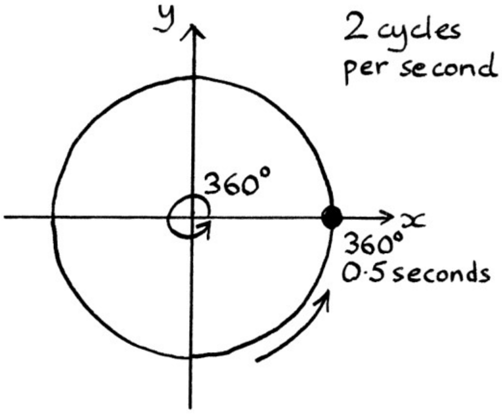




When the time is 0.5 seconds, the object rotating at 1 cycle per second will be at 180 degrees:

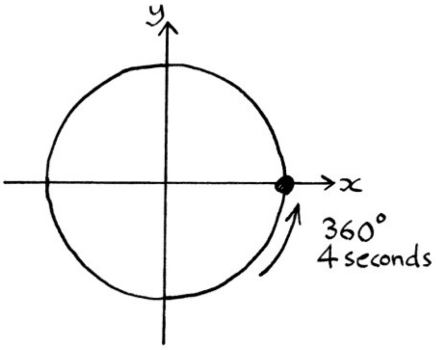


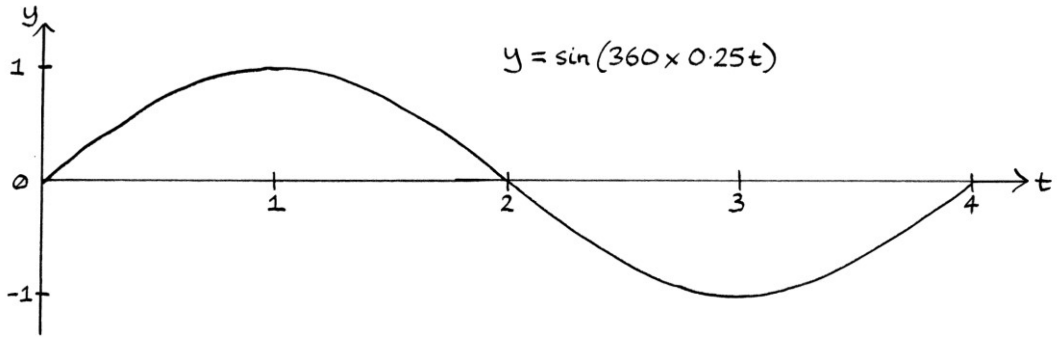
... but the object rotating at 2 cycles per second will be all the way around at 360 degrees.

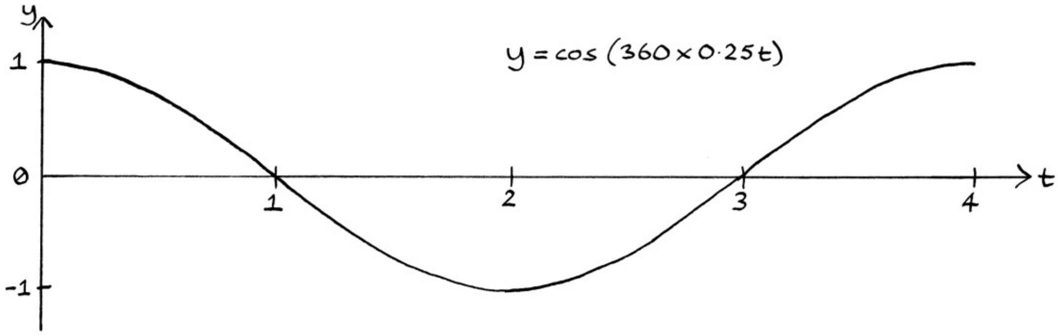


The whole circle is completed in 0.5 seconds, and each degree takes one 720th of a second to complete.

We will imagine the object moving around the circle at a slower speed. We will say it takes 4 seconds to complete one revolution. In this case, the period of the object is 4 seconds, and the frequency is: 1 ÷ 4 = 0.25 cycles per second. The formulas for the Sine and Cosine waves are:
"y = sin (360 * 0.25t)"
... and:
"y = cos (360 * 0.25t)"

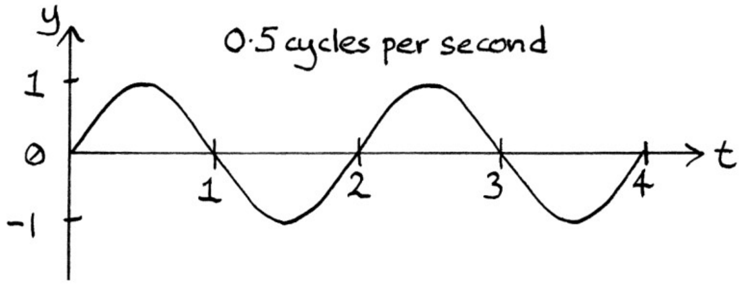


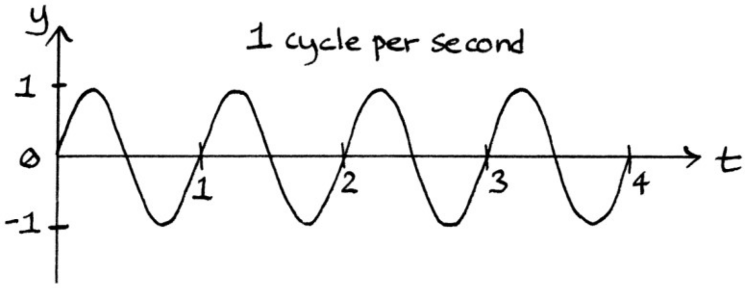


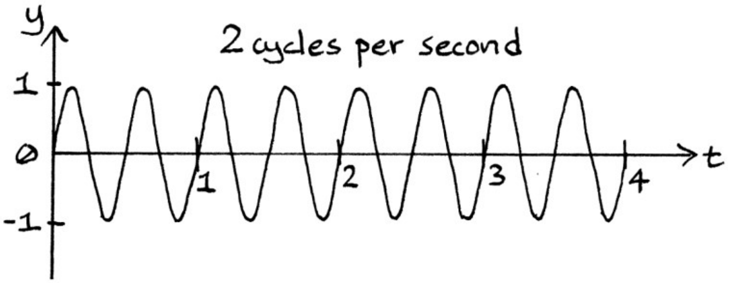


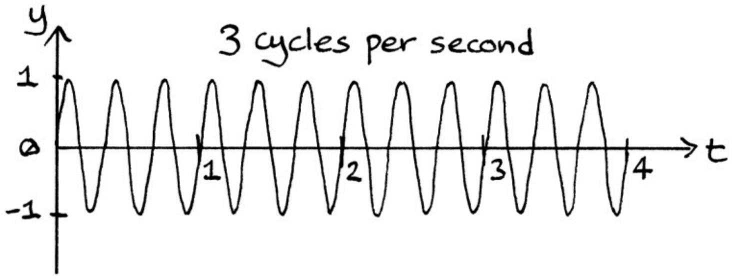
Note how the shapes of the two waves are shallower than before [This would be more apparent if the waves did not need to be scaled to fit on the page]. The waves repeat less quickly.

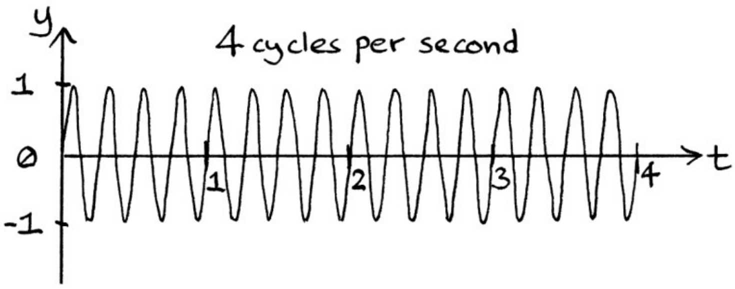
Examples of waves with different frequencies, drawn to the same scale are:

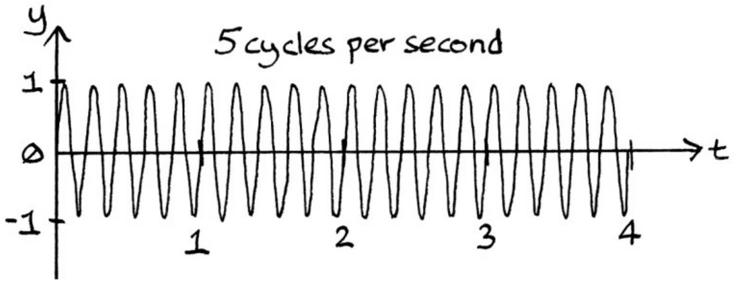






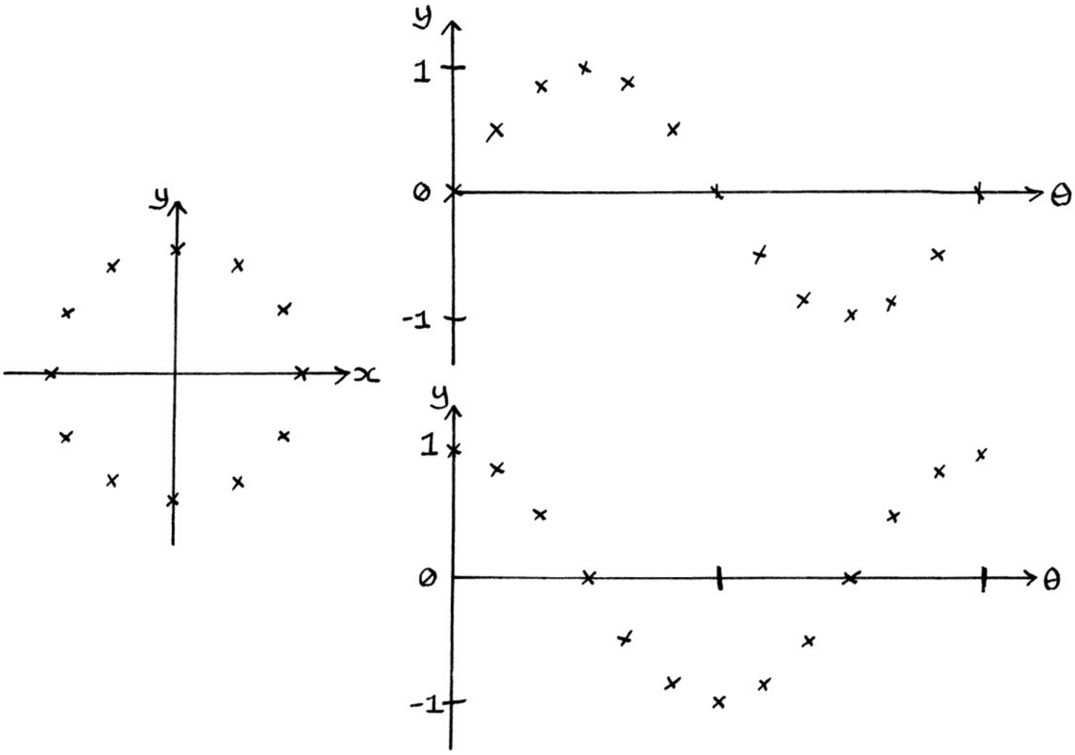






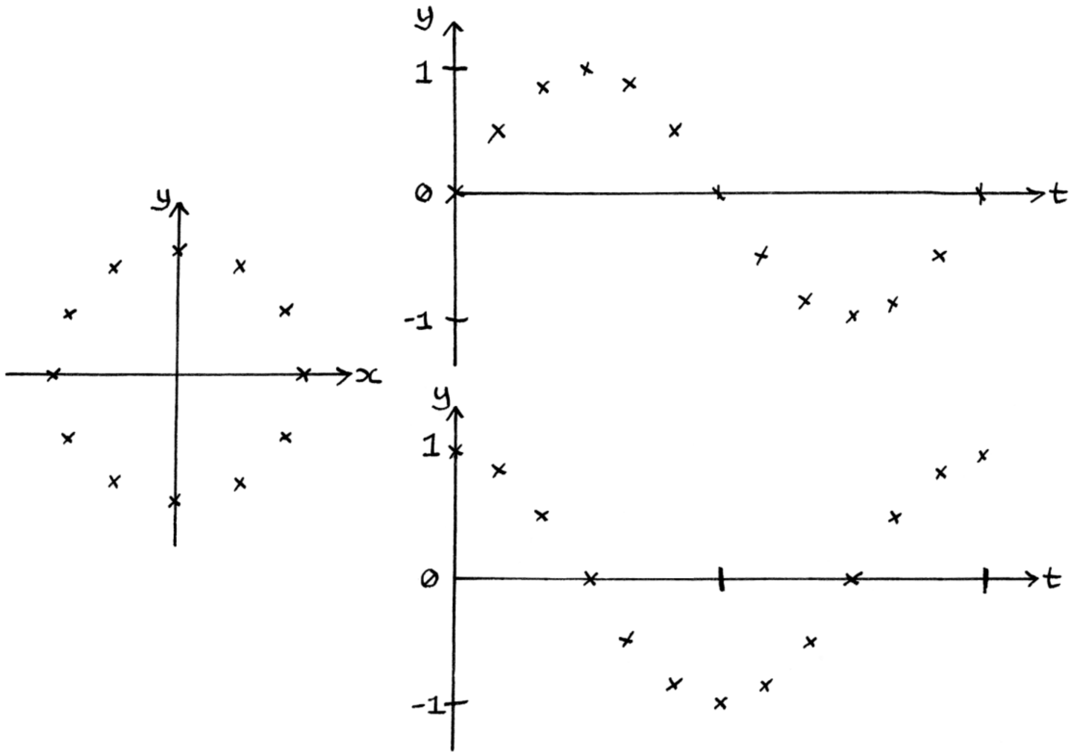
Deriving waves from the time circle

Before we dealt with time, we could derive the *angle-based* Sine and Cosine waves from the circle by reading off the y-axis and x-axis values of points around the circle's edge at evenly spaced angles.

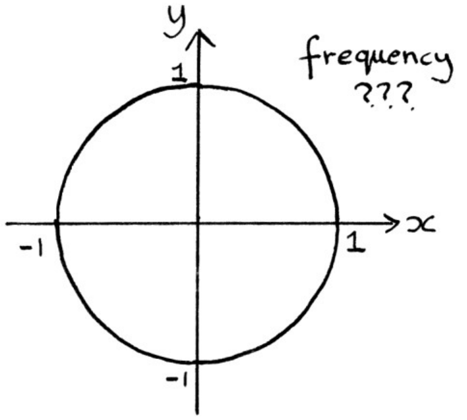


If we have an object rotating specifically at 1 cycle per second around the circle (and we know that for sure), we can derive the time-related Sine and Cosine waves in the same way – by measuring the y-axis and x-axis values at evenly spaced angles. The method still works because at any fraction of 360 degrees around the circle, the object will be at the same fraction of a second.

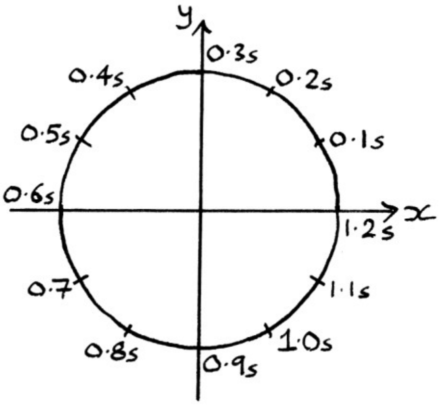
[The following picture is the same as the one above, but with the wave graphs having time as the x-axis.]



Things become more complicated when an object is rotating around the circle at a frequency other than 1 cycle per second. The reason for this is that the resulting circle will look identical if the frequency is 0.1 cycles per second, 1 cycle per second, 10 cycles per second, or indeed any frequency. The circles for those frequencies all look like this:



Of course, if we marked the position of the object after every, say, 0.1 seconds, we could determine the object's movement and therefore its frequency:

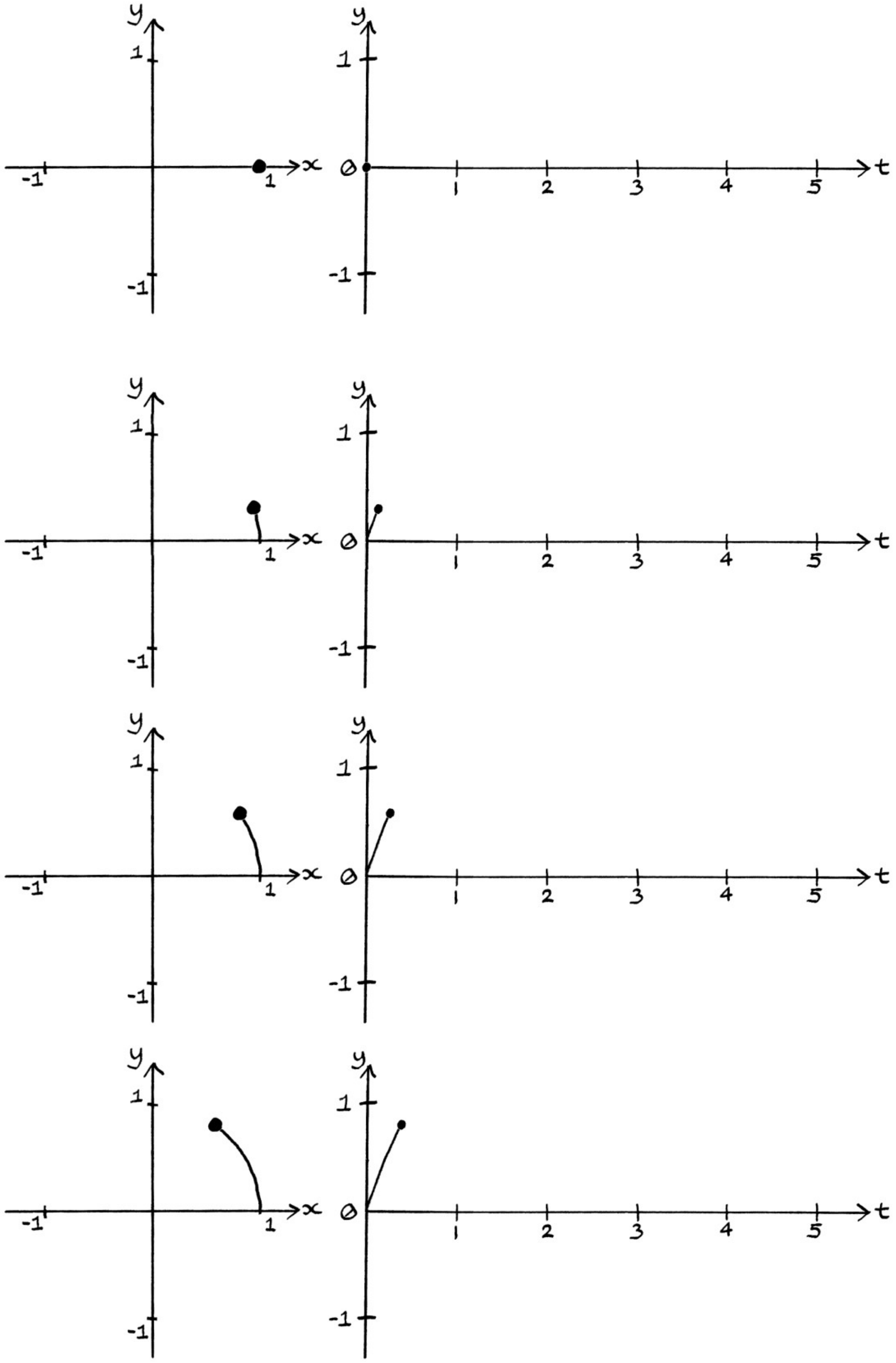


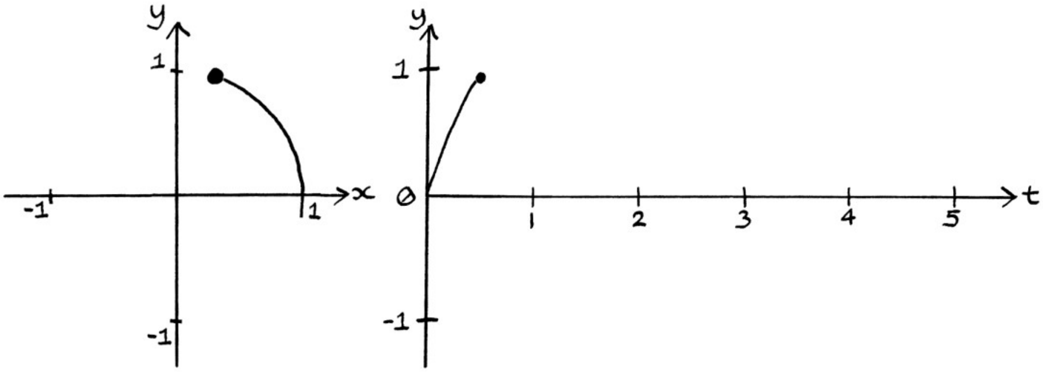
... but in general, there is no consistently good way of marking the frequency on a circle. Marking every 0.1 seconds helps for frequencies of 1 cycle per second or 10 cycles per second, but would not help for frequencies of 1,000,000 cycles per second or 0.0001 cycles per second.

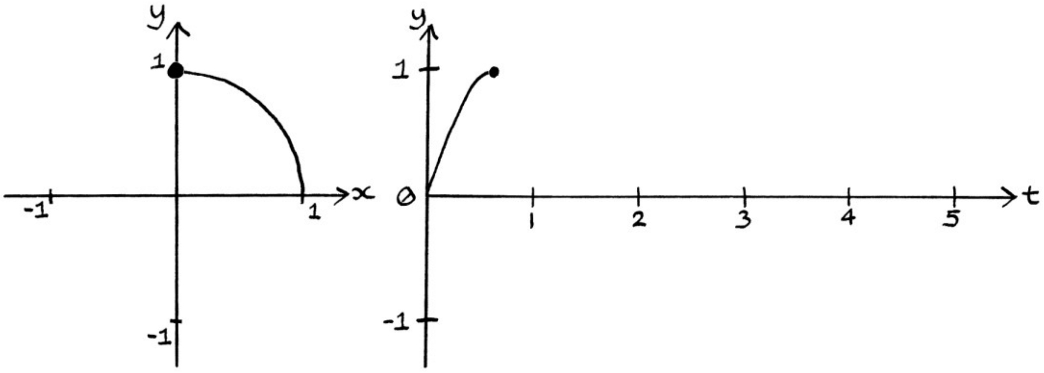
If we are presented with a time circle with no other information, it is impossible to know how fast the object is rotating around it. Therefore, it is impossible to know where it is at any particular time, and therefore, it is impossible to derive a time-related Sine and Cosine wave from it. It is still possible to use the Sine and Cosine waves for that circle to recreate the circle, but it is not possible to create those waves from the circle. Of course, if we know the frequency of the rotating object, then it *is* possible to derive the two graphs. Similarly, if we have the helix, and can accurately read values from it, then it is possible to create the waves (although if we have the helix, we essentially already have the waves).

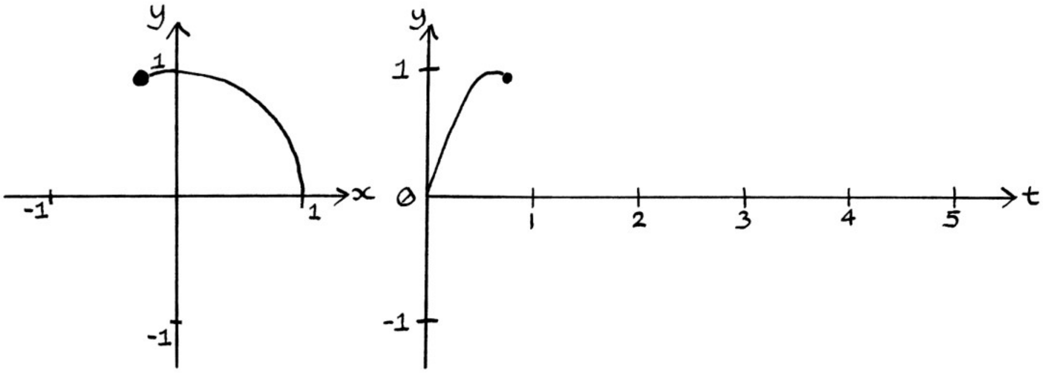
There is one sure way to know the frequency of the object though, and that is if we are lucky enough to witness the object rotating around the circle as it does it. In that case, we would be able to create the derived waves by plotting the relevant points on the derived Sine and Cosine waves.

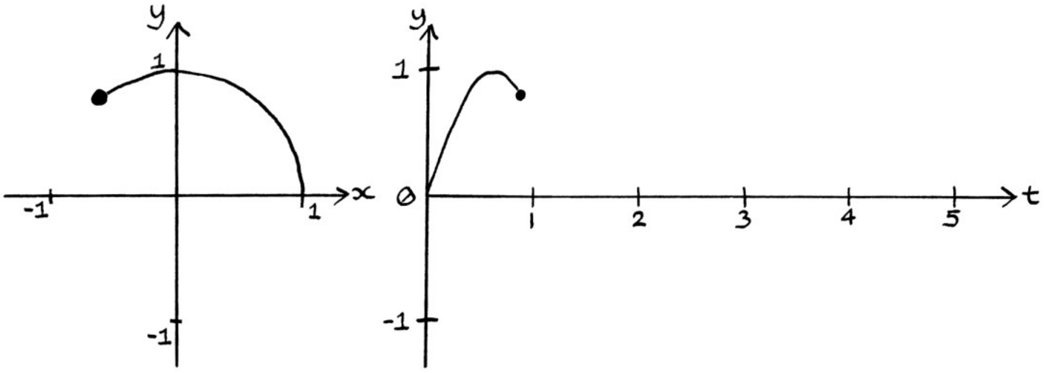
For example, if we observe an object rotating around the circle at the following particular rate, we can plot the relevant points on the Sine wave graph. We will plot them every 0.125 seconds:

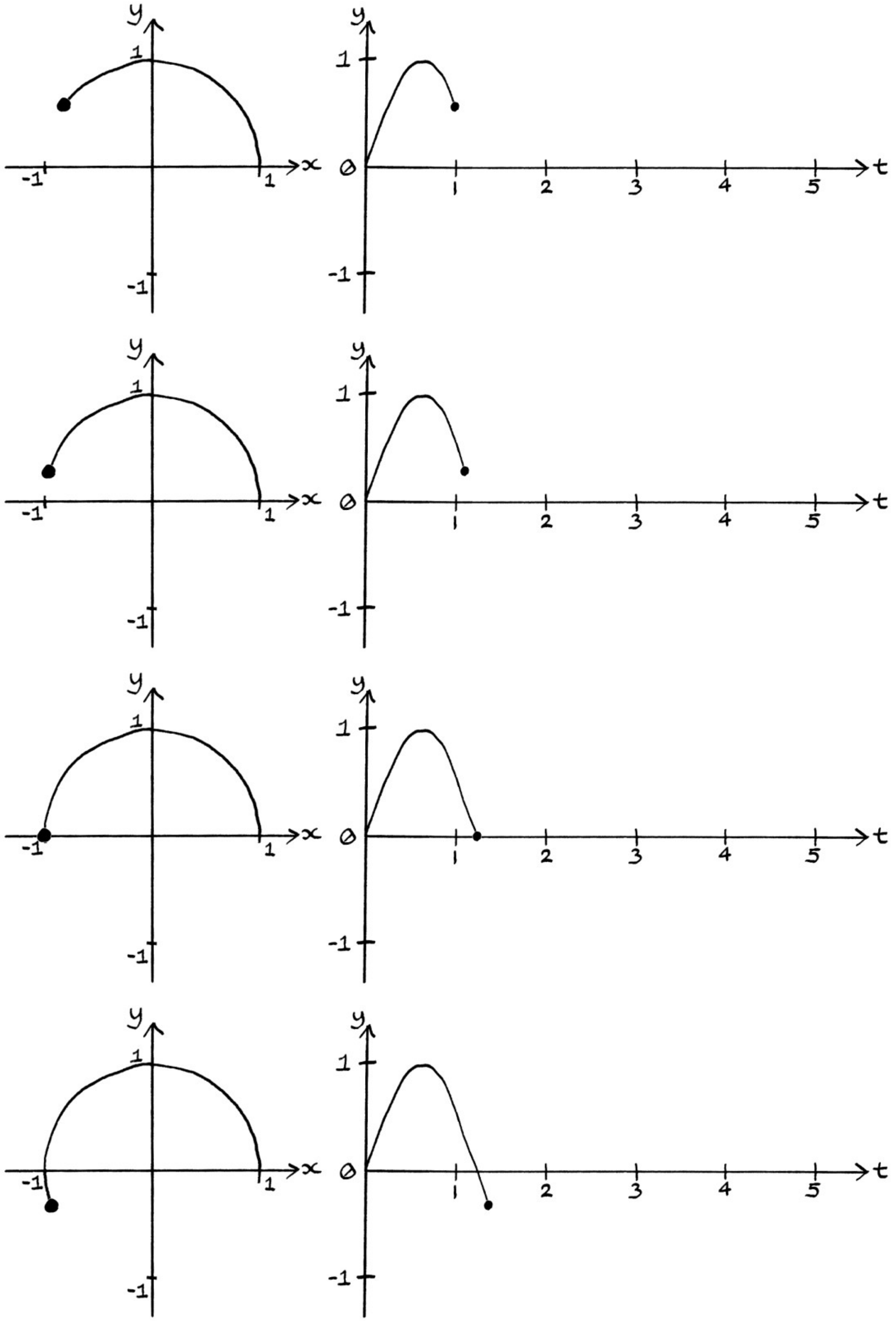


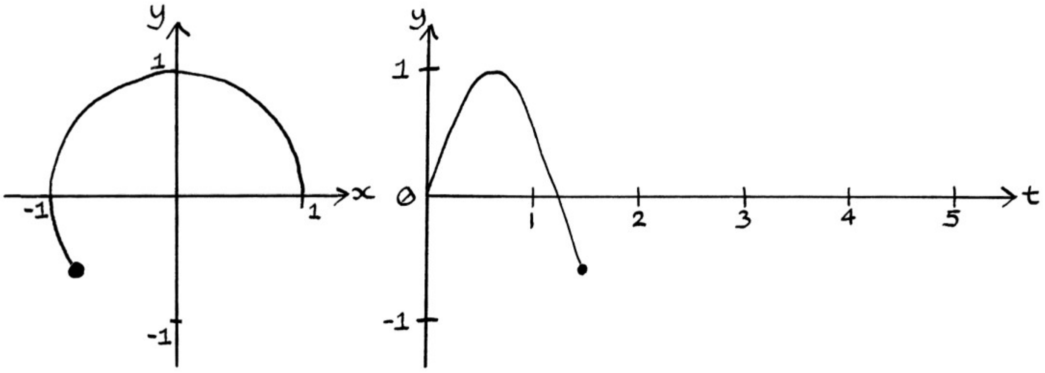

y
1
-1
1
x
-1
y
1
0
1
2
3
4
5
t
-1

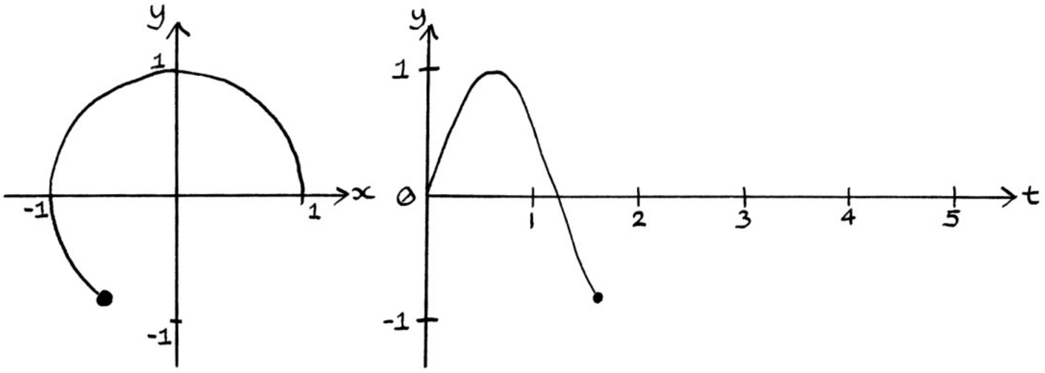

y
1
-1
1
x
-1
y
1
0
1
2
3
4
5
t
-1

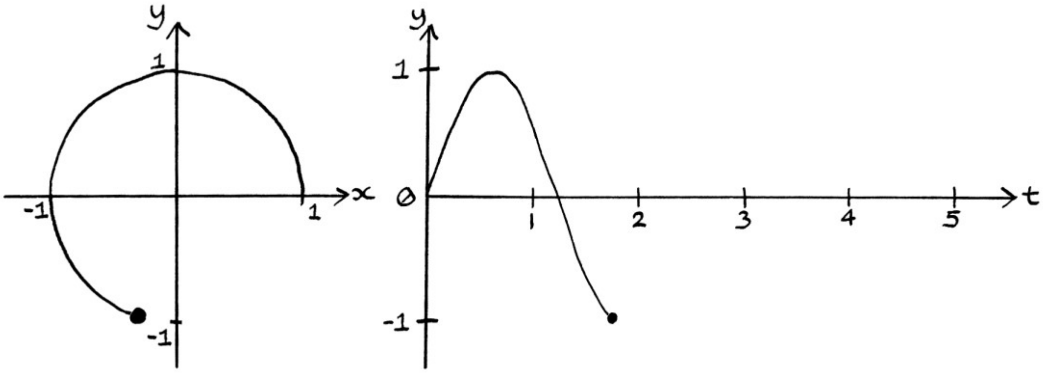

y
1
-1
1
x
-1
y
1
0
1
2
3
4
5
t
-1

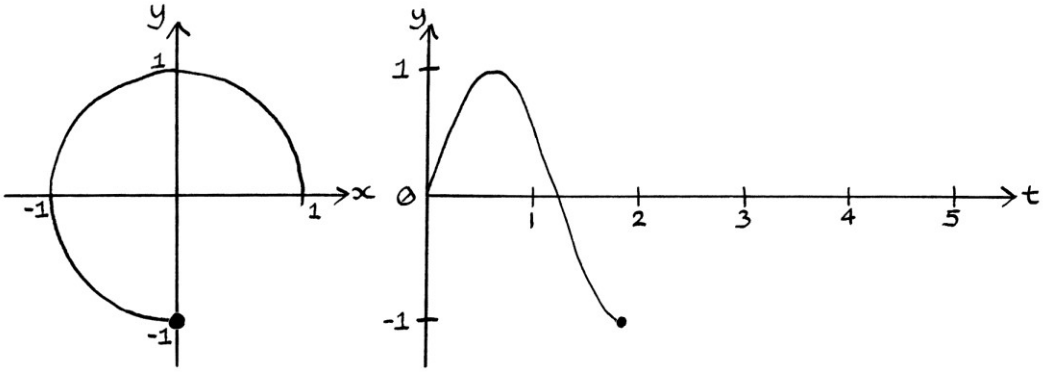

y
1
-1
1
x
-1
y
1
0
1
2
3
4
5
t
-1

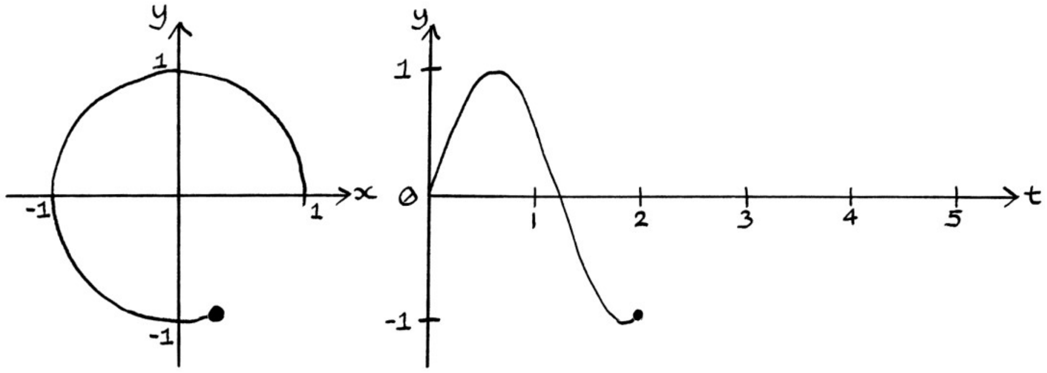

y
1
-1
x
0
t
1
2
3
4
5
-1

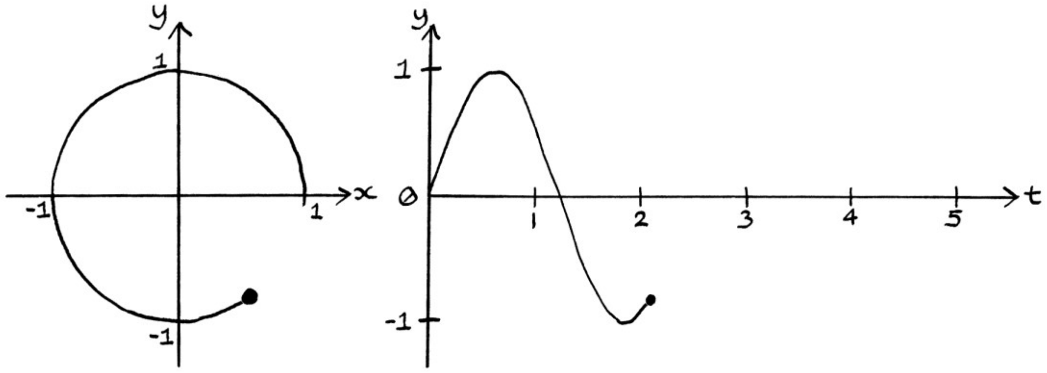

y
1
-1
1
x
-1
y
1
0
1
2
3
4
5
t
-1

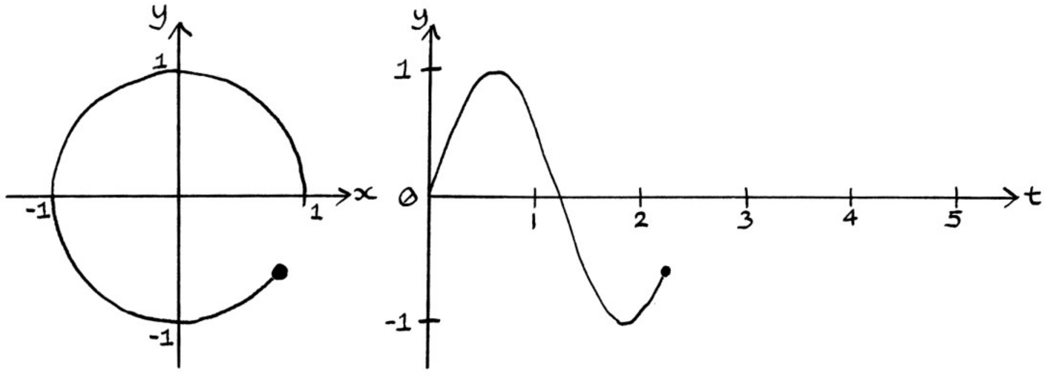

y
1
-1
1
x
-1
y
1
0
1
2
3
4
5
t
-1

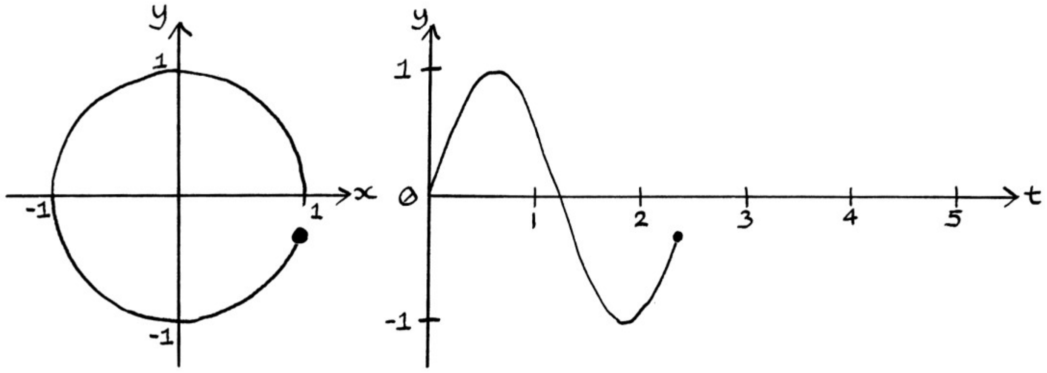

y
1
-1
1
x
-1
y
1
0
1
2
3
4
5
t
-1

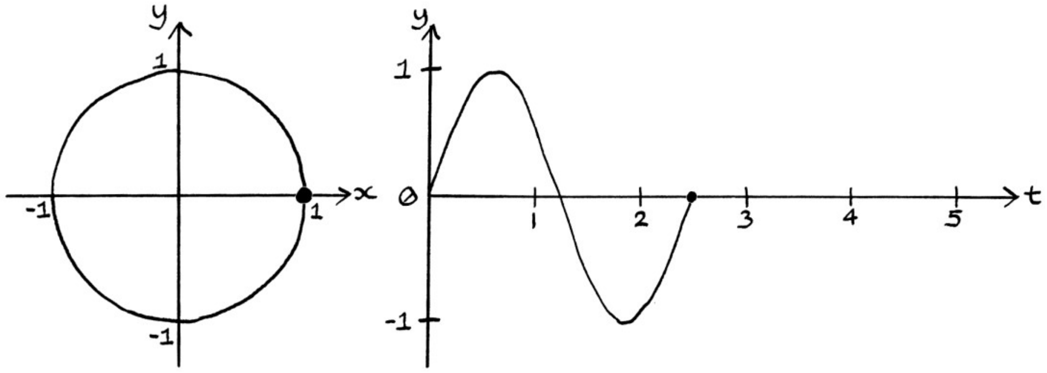

y
1
-1
1
x
-1
y
1
0
1
2
3
4
5
t
-1


y
1
-1
1
x
-1
y
1
0
-1
1
2
3
4
5
t

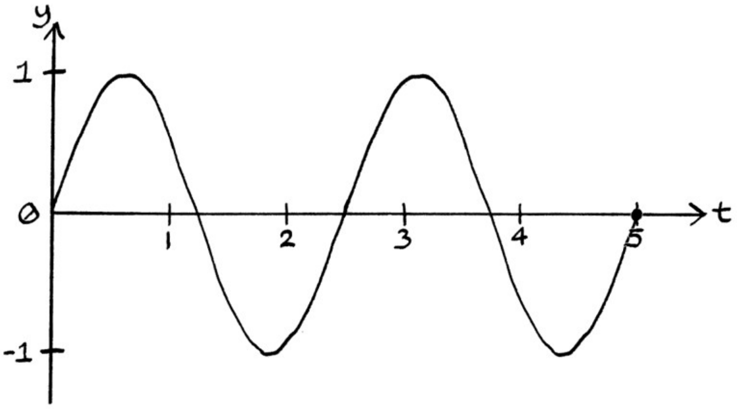

y
1
-1
1
x
-1
y
1
0
-1
1
2
3
4
5
t


y
1
-1
1
x
-1
y
1
0
-1
1
2
3
4
5
t


y
1
-1
1
x
-1
y
1
0
-1
1
2
3
4
5
t



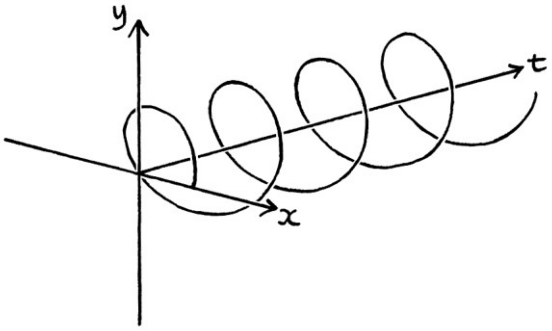
From observing the object as it rotates, we can see that it completes one revolution in 2.5 seconds. This is reflected in the resulting wave graph, where one cycle takes 2.5 seconds. The period of the wave is 2.5 seconds. Therefore, the wave's frequency, and also the frequency of the object rotating around the circle, is 1 ÷ 2.5 = 0.4 cycles per second. We could not have obtained this value from just looking at the static picture of the circle – it was necessary to observe the object moving around the circle. Not that it matters, but if we continued watching the object rotate around the circle for another rotation, we would end up with the following Sine wave graph:



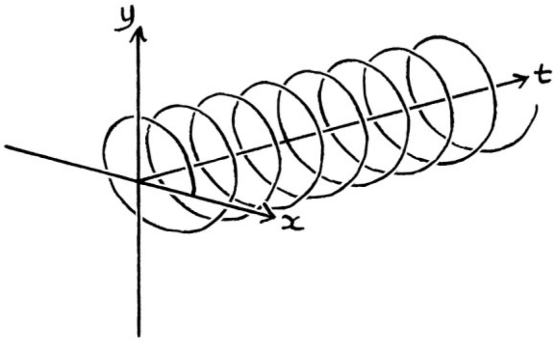
With time waves and the circle, the Sine and Cosine waves are still derived from the circle, but the practicalities of doing this oneself might be difficult or impossible, depending on the other information available to us. It still makes sense to refer to the Sine and Cosine waves as the "derived" waves, because that is what they are. Realising that frequency cannot be derived from just a static picture of a circle can be useful to remember.

Frequency on the helix

Frequency can be easy to understand on a helix. If we have an object completing so many cycles per second like this:



... and then we double the frequency, it will complete twice as many cycles every second:

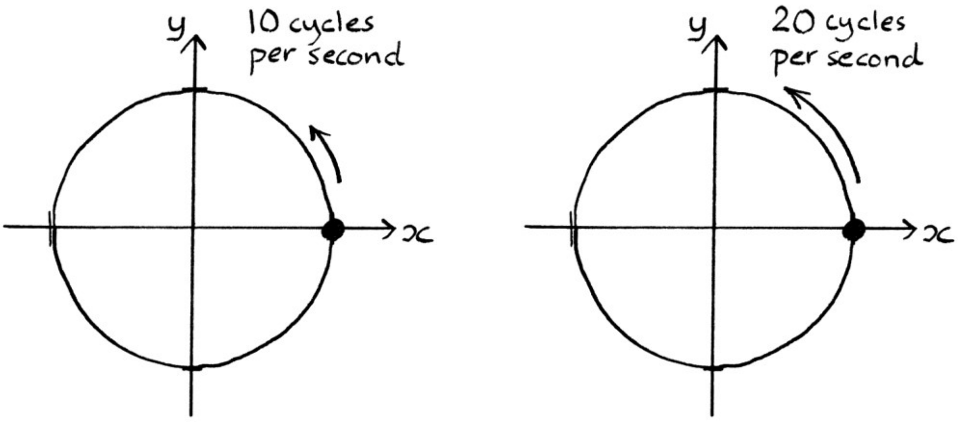


Frequency and speed

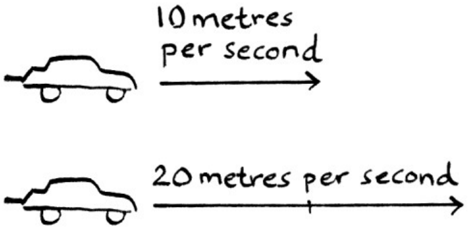
The concept of frequency, when discussing waves, is very similar to the concept of speed, when discussing objects moving in a straight line. Frequency refers to the number of cycles completed per second, while speed refers to the number of units of distance completed per second.

Some concepts to do with frequency in circles and waves are easier to understand if we imagine how the concepts would transfer to the speed of a vehicle or other object travelling in a straight line.

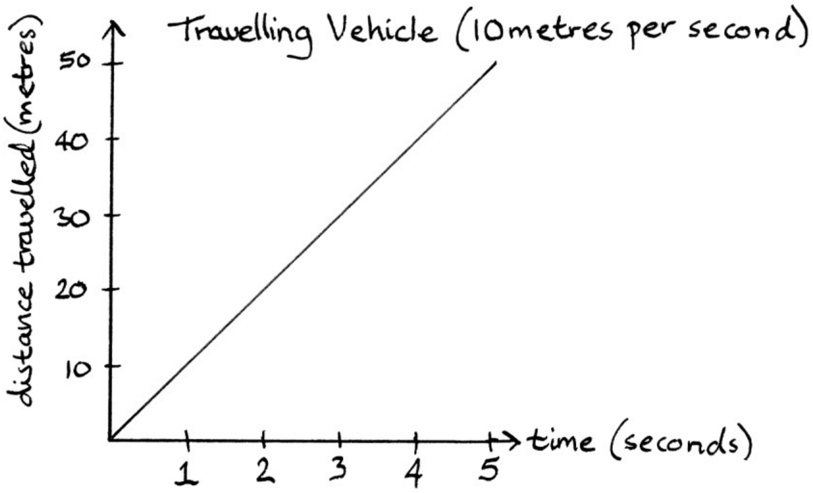
For an object rotating around a circle, if we double the number of cycles completed per second, the frequency doubles:



With a vehicle driving in a straight line, if we double the number of units of distance completed per second, the speed doubles:

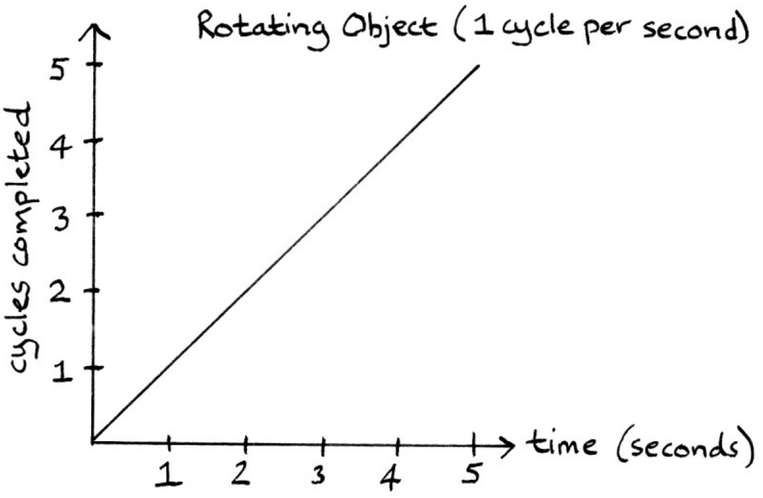


The most obvious way to illustrate the similarities between speed and frequency is on graphs. For the speed of a vehicle, we can plot a graph of distance travelled (which is the same as "metres completed") since the start of the journey against time in seconds.



The y-axis shows the distance travelled since we started observing the vehicle. The x-axis shows the time passed. The line on the graph indicates the position of the vehicle at any moment in time. The speed of the vehicle at any moment will be the distance travelled divided by the time it took to travel that distance. The speed is constant for this vehicle, which means the result will always be 10 metres per second.

For the frequency of an object rotating around a circle, we can plot a graph of cycles completed against time. The y-axis in this case shows the number of cycles around the circle that have been completed by the object. The x-axis shows the time passed. The line on the graph shows the number of cycles that have been completed at any particular time. The frequency at any moment in time will be the number of cycles completed divided by the time passed. As the frequency for this object is constant, the result will always be 1 cycle per second.



While speed and frequency do refer to different concepts, there are enough similarities in the ideas to be helpful in achieving a better understanding of what frequency really is, and knowing this will be useful when dealing with some concepts later on.

Period can be thought of as being analogous to the time that a vehicle moving in a straight line takes to complete a particular distance.

Angular frequency

When we first incorporated time into our Sine and Cosine formulas in Chapter 4, it was necessary to multiply the time by 360 for Sine and Cosine to give the same results as if we had been using degrees. I called this "corrected time" or "angular time" as the 360 multiplied by the time corresponded to the angles for that time. We were really splitting 1 second into 360 parts to match how a circle had been split into 360 degrees, and adjusting the formula for the Sine and Cosine waves to match that.

Note that it is easy to become confused, and think that we should have *divided* the time by 360 instead. However, this is the wrong way round. If we had left the time untouched, in other words used "y = sin t", then the wave would have taken 360 seconds to repeat, because "y = sin θ" would take 360 *degrees* to repeat. Therefore, we needed to speed up the wave – in other words, increase its frequency. Multiplying the time by 360 means the wave speeds up by 360 times, and therefore repeats in 1 second, instead of 360 seconds. If we had, instead, divided by 360, the waves would have *slowed down*, and therefore would have taken 360 * 360 = 129,600 seconds to repeat a cycle instead.

Given this fact, and given that we have seen examples of different frequencies in formulas, such as "y = sin (360 * 4 * t)", it might be apparent that the multiplication by 360 can be thought of as a *frequency* correction. It is speeding up the wave by 360 times, and therefore behaving as a frequency. This correction means that we can use "t" in Sine and Cosine formulas and still have a wave repeating once every second. Therefore, although I called the "360 * t" part of the formula, "angular time" earlier, it can just as easily be called "angular frequency". [I called it angular time because I had not introduced the concept of frequency at that point]. The idea is usually called "angular frequency".

Although I just described angular frequency as a correction to let Sine and Cosine work directly with the time, there is another way of thinking about the same concept. A more common way of thinking of angular frequency is that it refers to the number of degrees completed per second by an object rotating around the circle – in other words, the *angle* completed per second. If we had an object that rotated around the circle at 1 cycle per second, then it would also be travelling at 360 degrees per second. Its angular frequency would be 360 degrees per second. If the object rotated at 2 cycles per second, then it would also be travelling at 720 degrees per second. If an object rotated at 0.5 cycles per second, then it would also be travelling at 180 degrees per second. If we think of angular frequency this way, then it is analogous to measuring a vehicle's speed in either metres per second or centimetres per second. The value indicating speed for centimetres per second will always be 100 times higher than that indicating metres per second; the value indicating degrees per second will always be 360 times higher than the value indicating cycles per second.

To summarise this, there are two ways of thinking about the multiplication by 360:

1. As a frequency correction to enable Sine and Cosine to work on the time to produce waves that have a default frequency of 1 cycle per second (as otherwise the waves would have a frequency of 1/360th of a cycle per second).
2. As a way of indicating how many degrees per second around the circle an object is travelling.

These are different ways of thinking, but they amount to exactly the same thing. If we think of the multiplication as achieving the first of these, then we also end up with a value that indicates the degrees per second. If we think of the multiplication as achieving the second of these, then we also end up with a value that allows Sine and Cosine to work on the time.

Strictly speaking, the frequency correction is necessary to use time as if it were an angle, and so should be the primary reason for the multiplication. If we only thought of the second reason and one day decided not to think of degrees per second in formulas, but instead cycles per second, then none of our formulas would work. For many people, the reasons for the multiplication are lost in time, and they just accept it without questioning. Those who do think of it, generally only think of the second reason. However, those who think of it too much will eventually have to think of the first reason as well or else everything becomes very confusing.

The important thing to know is that if we are going to work with time in our formulas, then we *have* to multiply that time by 360 (if we are working in degrees).

Generally, in mathematics, when it comes to the formulas, the “360” is grouped, not with the time, but with the frequency. Therefore, a formula will not be given as so:
y = sin (3 * 360t)
... but as so:
y = sin (360 * 3t)

Of course, these mean exactly the same thing.

Examples

If we have a formula “y = sin (360 * 2t)”, then our cycles-per-second frequency is 2. The cycles repeat twice a second. Our *angular frequency* on the other hand is 360 * 2, which is 720 degrees per second. The object rotating around the circle is rotating at 720 degrees per second.

If we have a formula “y = cos (360 * 0.1t)”, then the cycles-per-second frequency is 0.1 cycles per second. The angular frequency is 360 * 0.1 = 36 degrees per second. The object revolving around the circle is travelling at 36 degrees per second.

Angular frequency without circles

Angular frequency, and how it represents degrees per second, is a reasonably simple concept if we think of the circle. However, if we had never thought of the circle from which the Sine and Cosine waves are derived, or could have been derived, and only thought of the waves themselves, the concept is a lot harder to visualise. If we look at the formula or the graph for “y = sin 360t”, there is no mention of angles in it. We have to know what Sine and Cosine really are, so that we can understand the meaning behind angular frequency properly.

Angular speed

Measuring frequency in terms of degrees per second can be thought of as “angular frequency” because we are counting the angles completed per second. Another term for exactly the same thing is “angular speed”.

If we think of an object moving around the circle as completing so many 360th portions of a cycle every second, then “angular frequency” is apt; if we think of a point travelling “an angle *distance*” per second then “angular speed” is apt. Of the two terms, “angular frequency” is used more often. The fact there are two terms for the same thing, and they mention frequency and speed, reinforces the idea how frequency in waves and around the circle is analogous to the speed of objects moving in straight lines.

[Note that there is another term in maths, which is not relevant to this book, called “angular velocity”. This is a slightly different concept, and I only mention it here to avoid confusion].

The symbol for angular frequency

For situations where the frequency in a formula is unknown, unimportant, or could be a range of values, the cycles-per-second frequency can be portrayed by the letter “f”. Therefore, a generic formula for a Sine wave that takes into account frequency is:
“y = sin (360 * f * t)”
... and its Cosine equivalent is:
“y = cos (360 * f * t)”.

These would more typically be written as:
“y = sin (360 * ft)”
... and:
“y = cos (360 * ft)”
... or even:
“y = sin 360ft”
... and:
“y = cos 360ft”.

There is also a symbol that groups the “360” and the “f” together to represent them as a single value of *angular* frequency. That symbol is the lower-case Greek letter Omega, “ω”, which is one of the Greek equivalents to the lower-case Latin letter “o” (as in “oscar”). Confusingly, it looks like the Latin letter “w” (as in “whiskey”). Many people write it by hand in the same way as they would write the letter “w”. [If you want to handwrite “ω” and “w” in a way that distinguishes them, you can write “ω” in a more bulbous way than “w”, and with the lines on each side curving upwards towards the centre.] The symbol “ω” represents the “360” *and* the cycles-per-second frequency in one. In other words, the formula, “y = sin (360 * ft)” can be rephrased as “y = sin ωt”.

If “f” in a Sine or Cosine wave formula is 10 cycles per second, then “ω” will be 360 * 10 = 3600 degrees per second. Such a formula without symbols could be given as either “y = sin (360 * 10t)” or “y = sin (3600 t). They mean exactly the same thing, but some people prefer one way of writing a formula to another.

The purpose behind using the “ω” symbol is that it saves having to write out 360 in every formula. It really is nothing more than shorthand for writing out “360 multiplied by f”. Generally, the more academic the person writing the formulas, the more likely it is that they will use angular frequency “ω” instead of regular frequency “f”.

One thing to note is that angular frequency is usually measured not in degrees per second, but in radians per second. [I explain radians in Chapter 22]. Radians are another way of dividing a circle, but instead of there being 360 divisions, there are 2π divisions. Most of the formulas in books that mention "ω" will not be using it with the units of degrees per second, but instead radians per second. This means that "ω" will not represent "360 * f", but instead "2π * f".

Some people are confused by angular frequency, so do not be surprised to see "ω" occasionally *misused* to represent *normal* cycles-per-second frequency. Once you become used to waves, it is easy to spot when someone has made this mistake.

Notation in this book

In formulas in this book, I will generally keep the 360 and the frequency (and later the 2π and the frequency) as separate entities to make things clearer. In other words, if there is a formula "y = sin (360 * 2 * t)", I will write it in that way, or more probably, as "y = sin (360 * 2t)", instead of writing it as "y = sin 720t". These are ultimately the same thing, but in the latter formula, it is harder to know the cycles-per-second frequency.

Potential sources of confusion

What the circle represents

One important fact relating to time and frequency is that the circle still represents *angles*. It represents the angles of a rotating object at any particular moment in time. If we know the frequency of a rotating object, we can measure the angle and calculate the time when the object would be at that angle. It is easy to become mistaken or confused and think that the circle specifically represents 1 second of movement, or some other time value, but the circle still represents 360 degrees of rotation. [Of course, if the object rotates at 1 cycle per second, then one cycle and one second will be the same, but even in that situation, it is always better to think of the circle as 360 degrees].

This important distinction is clearer when thinking about the helix. One rotation around the helix is 360 degrees, no matter how far the object moves down the time axis.

Normal frequency and angular frequency

One possible source of confusion is deciding whether people are talking about normal "cycles-per-second" frequency or angular frequency. Angular frequency, when using degrees, is always 360 times more than normal frequency, as I have explained before. Therefore, if we see a formula that multiplies the time by 360, or a multiple of 360, then we can assume that it is referring to angular frequency. When we introduce the concept of radians in Chapter 22, we will divide the circle up into portions that are not based on 360 divisions, and therefore angular frequency will require a different multiplication – it will be 2π times the normal frequency. However, it will still be obvious as to whether we are talking about angular frequency or not. Note that very occasionally, you will read things written by people who use the word "frequency" as shorthand for "angular frequency", which most people would agree is confusing.

Frequency without time

Another source of confusion might seem a matter of pedantry. Frequency refers to the number of repetitions in an amount of *time*. Therefore, we cannot have frequency without mentioning time. It would be as if we spoke about the speed of a vehicle without mentioning time.

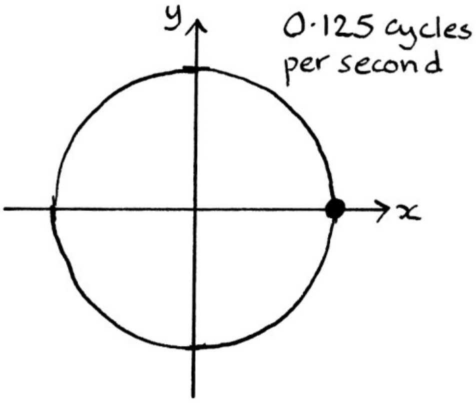
Occasionally, you might read things by people who are talking about frequency, but neglect to mention time in their formulas. For example, they will give a formula such as this: "y = sin (360 * θ)" with the implication that "θ" refers to degrees and not time. With no context, this formula is mathematically valid. However, the mention of 360 (if we are thinking in degrees, or the mention of 2π if we are thinking in radians) implies that the formula refers to the wave's angular frequency with respect to time. If it did not refer to time, then there would be no point in the multiplication by 360 (or 2π). If the "θ" in the formula refers to angle, as by convention it usually does, then the formula is not saying what the author of the formula thinks it does. The Sine of an angle does not require correcting to make it repeat every 360 degrees, and any graphs based on this will repeat once every degree, not every 360 degrees. The formula should either have "t" in it as so: "y = sin (360 * t)", or there should be an explanation before or after the formula to indicate that, despite convention, "θ" in this case is being used to refer to time in seconds.

If we have a formula that is based around "θ" or "x", instead of "t", and there is not an explicit comment that the "θ" or the "x" refers to time, then technically, any value scaling the "θ" or the "x" is not frequency. The results of formulas will be the same as if they were related to time, but strictly speaking, it is not *frequency*. When dealing with "discrete waves", as discussed in a later chapter, waves are treated as being sequences of individual y-axis values, without necessarily having an explicit mention of time. However, there is still an *implied* time attribute, so formulas can still relate to frequency in those situations, even if time is not directly mentioned.

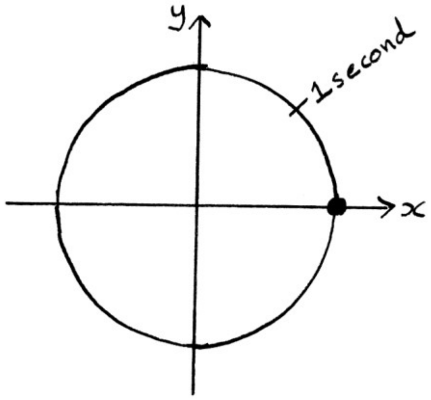
Frequency on the circle

Whereas a circle of a particular radius implies the existence of a Sine wave and a Cosine wave of a particular *amplitude*, a circle on its own cannot indicate the *frequency* of those waves. There are several solutions to this:

1. We can just write the frequency on the circle. This is probably the best solution.



2. We can mark lines on the circle to indicate the angular distance completed in one second. This does not look particularly tidy, and becomes confusing or meaningless if the frequency is particularly low or high.



3. We can think of the helix that the circle, the Sine wave and the Cosine wave come from. The helix indicates the properties of the circle, the Sine wave, the Cosine wave and all their attributes. We can see differences in frequencies by looking at how closely the lines in the helix are together. However, as I have said before, the helix is difficult to draw and very difficult to read meaningful values off. It is useful to think of the helix as a concept, but at this stage in learning, it is probably better just to use the circle and write the frequency on it in words.

Formulas for frequency

The general formula for a Sine wave that takes into account frequency, and uses degrees is:
y = sin (360 * f * t)
... where “f” stands for “frequency” measured in cycles per second or hertz.

For a Cosine wave, the formula is:
y = cos (360 * f * t)

Both of these made more concise are:
y = sin 360ft
y = cos 360ft

If we were dealing in radians, the formulas would be:
y = sin 2πft
y = cos 2πft

The general formula for a Sine wave that treats angular frequency as one variable is:
y = sin ωt
... where “ω” represents angular frequency and is, when using degrees, an abbreviation for “360 * f”. When using radians, it is an abbreviation for “2π * f”.

The corresponding formula for a Cosine wave is:
y = cos ωt

Chapter 7: Phase

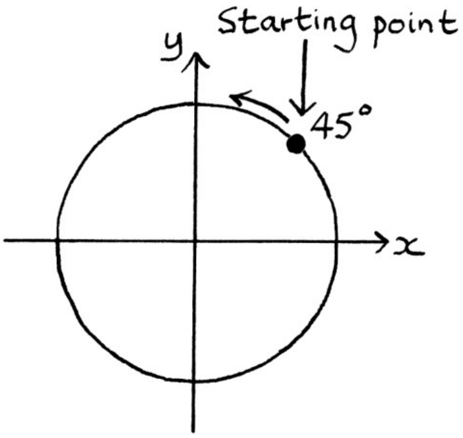
Phase

The concept of phase is not difficult to understand, but there are numerous ways in which it can be confusing. In this chapter, I will explain what phase is, while trying to remove any possible sources of confusion.

The term "phase" is really just another word for "angle", but with a more nuanced meaning. Specifically, phase can be thought of as meaning "starting angle".

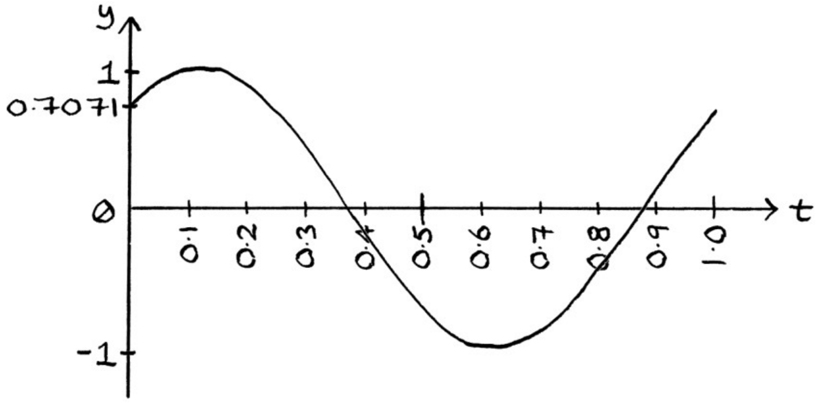
When thinking about circles, "phase" is the angle that indicates the starting point of an object about to rotate around the circle. So far in this book, all objects rotating around the circle have started at 0 degrees, but it does not have to be this way.

For example, an object could start at 45 degrees. If an object starts at 45 degrees on the circle, then we can say, "Its phase is 45 degrees":

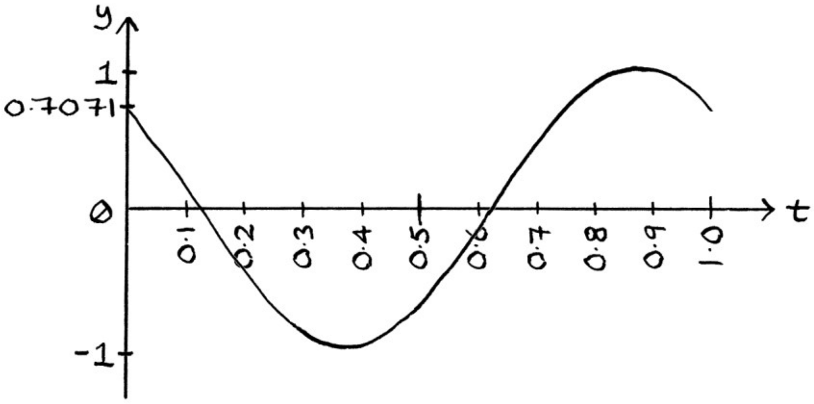


What is more, the Sine and Cosine waves derived from the movement of the object around the circle will reflect the object's new starting place.

The Sine wave will start with the object's y-axis value on the circle when the object is at 45 degrees (0.7071 units) and then continue as normal from that point.

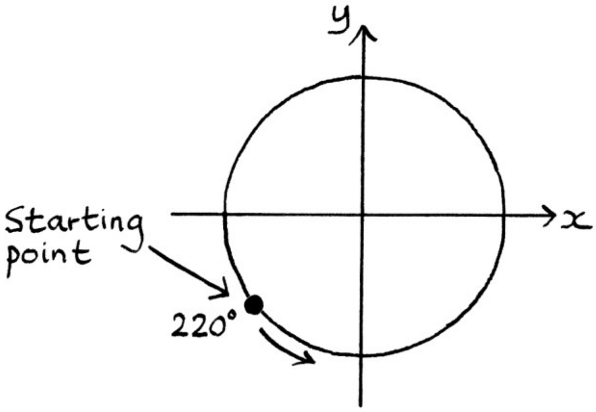


The Cosine wave will start with the object's x-axis value on the circle when the object is at 45 degrees (also 0.7071 units in this particular case), and then continue as normal from that point.

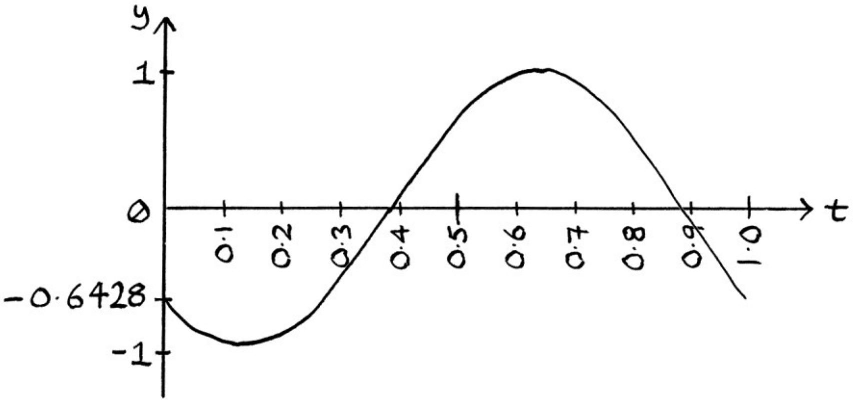


In other words, the Sine and Cosine waves derived from a rotating object start based on where the object starts on the circle. Strictly speaking, they have always done this in previous examples (when the object started at 0 degrees), but now we can see that they do this wherever the object starts.

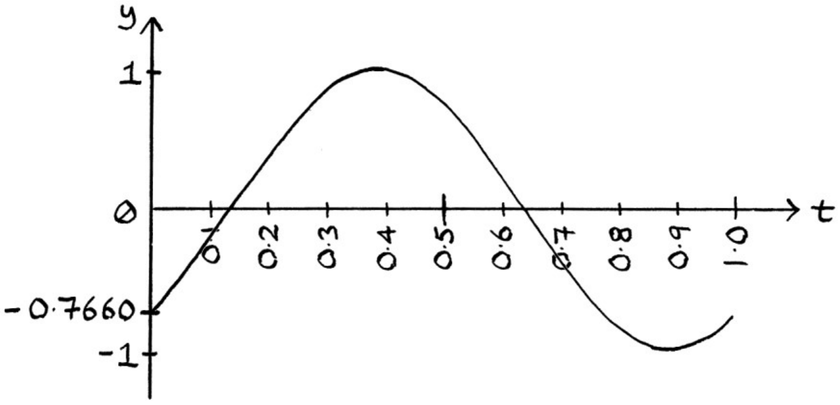
If an object starts at 220 degrees, then its phase is 220 degrees:



The above circle's Sine and Cosine waves will also reflect the object's starting place. The Sine wave starts by showing the object's y-axis value on the circle at 220 degrees (−0.6428 units), and then continues normally from that point:

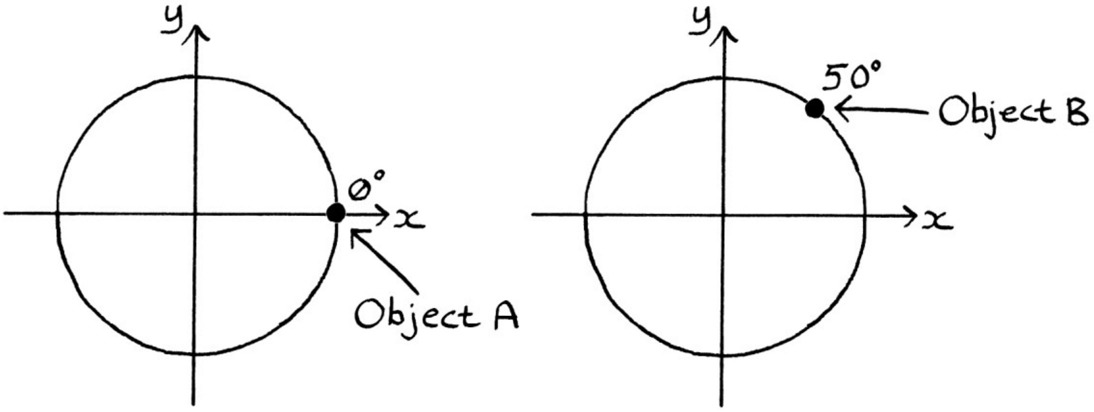


The Cosine wave starts by showing the object's x-axis value on the circle at 220 degrees (−0.7660 units), and then continues normally from that point:



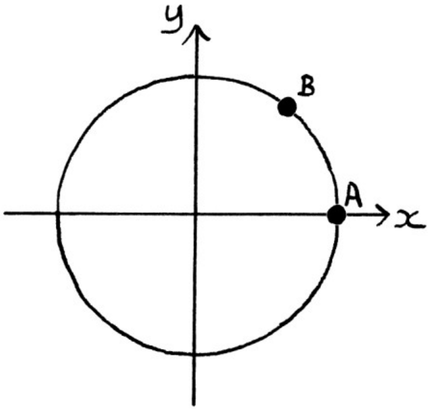
Imagine we have two objects rotating around identical circles at the same frequency of 1 cycle per second. We will call the objects “Object A” and “Object B”. We will say that Object A starts at 0 degrees, and Object B starts at 50 degrees.

At t = 0 seconds, the two circles and the two objects will look like this:

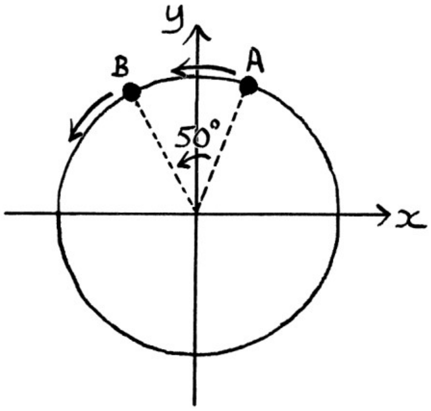


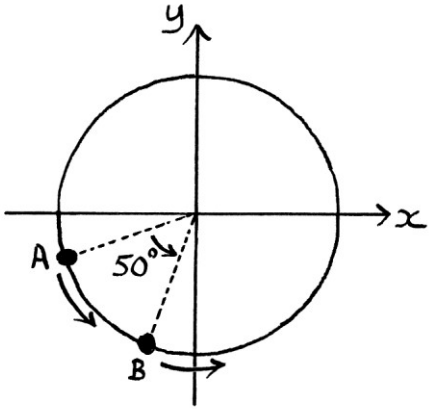
Object A has a phase of zero degrees, in that it starts at 0 degrees, and Object B has a phase of 50 degrees, in that it starts at 50 degrees.

When the two objects start moving, Object B will have a 50-degree head start on Object A. This is most obvious if we draw the objects on the same circle:



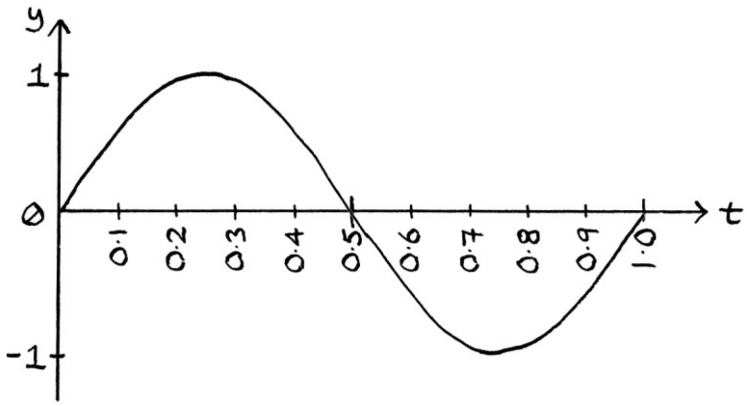
For the whole duration of their rotations, Object B will always be 50 degrees ahead of Object A. The reason for this is that they are both moving around at the same rate of 1 cycle per second.



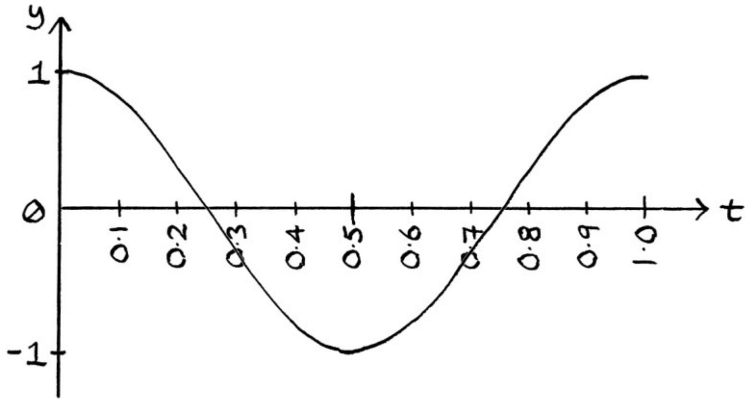


When looking at the wave graphs derived from each circle, the Sine and Cosine waves derived from Object A's movement will look like this:

The Sine wave:



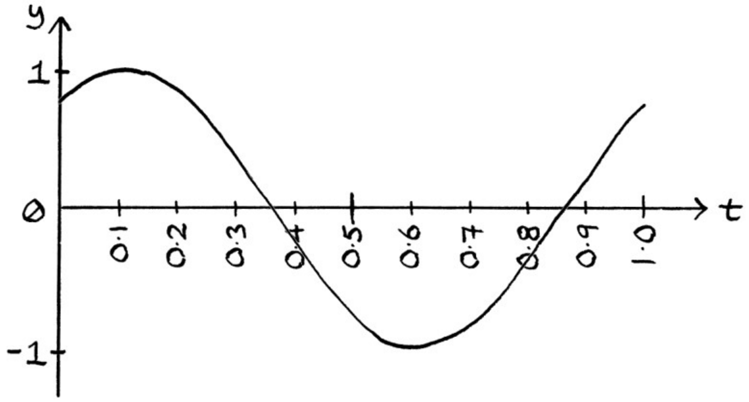
The Cosine wave:



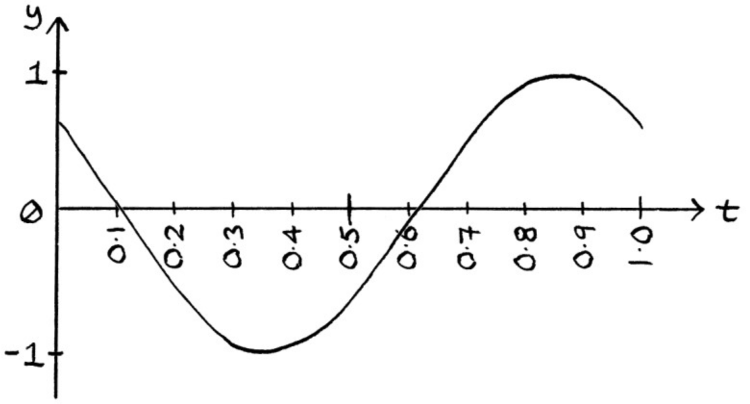
... and the Sine and Cosine waves derived from Object B's movements will look like this:

The Sine wave:

The Sine wave:

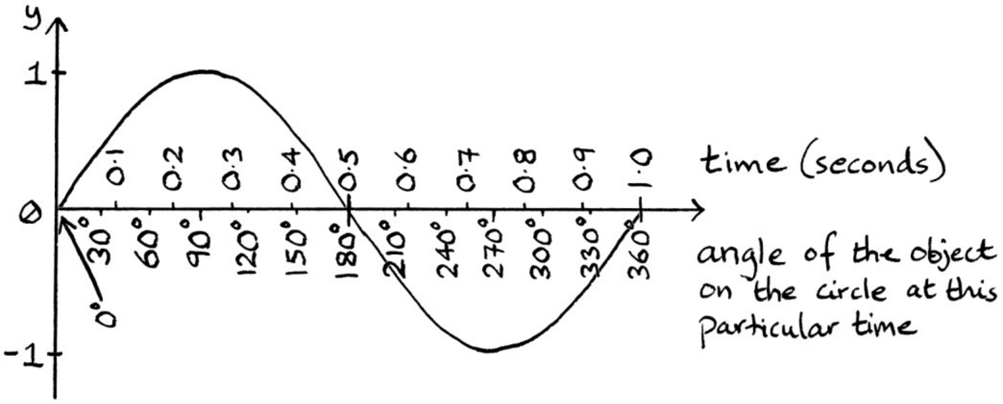


The Cosine wave:

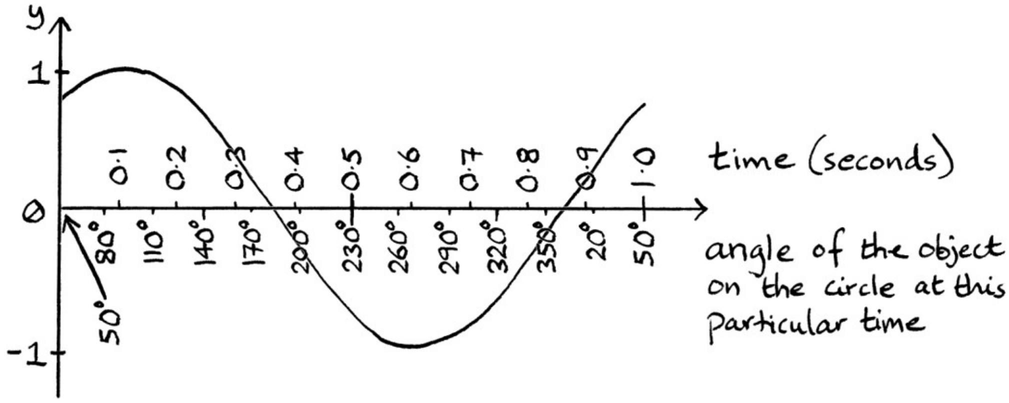


We can see that although the shapes of Object A and Object B's waves are the same, Object B's Sine and Cosine waves reach the peaks and dips sooner than those of Object A. This can be seen more clearly if we write the object's angle on the circle at that particular moment in time. In these graphs, I have written the time passed in seconds above the axis, and the angle of the object on the circle at that particular time beneath the axis.

Object A's Sine wave looks like this:

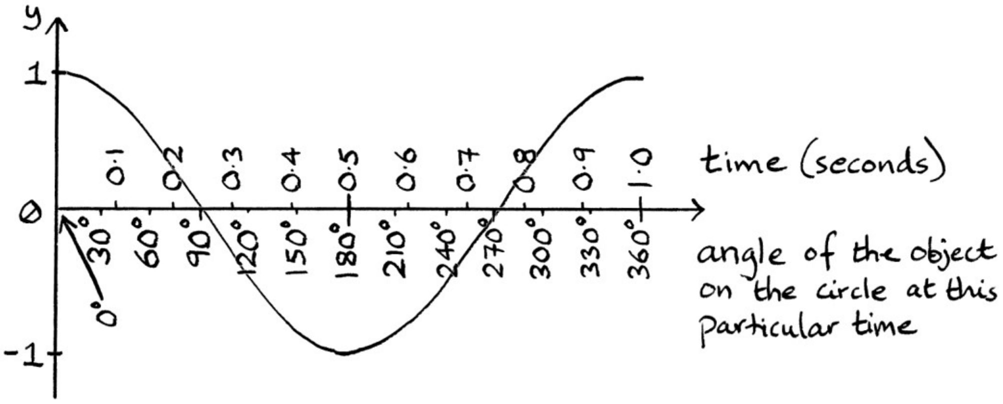


Object B's Sine wave looks like this:

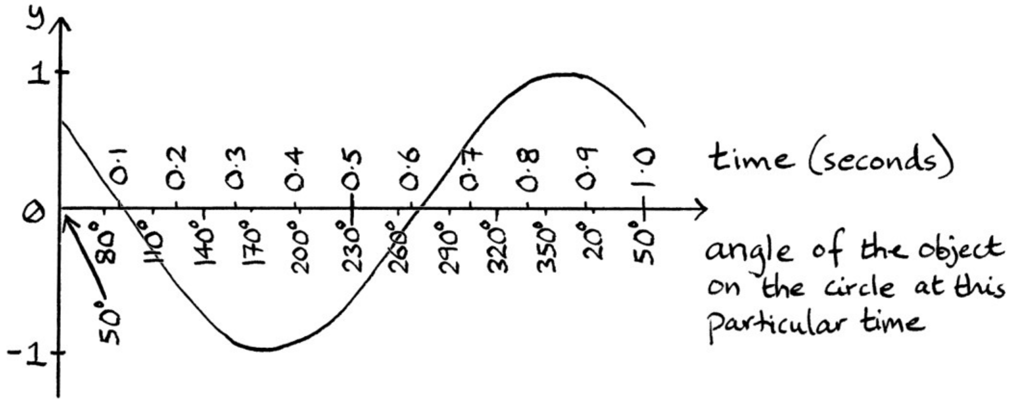


Object A's Sine wave starts when Object A is at 0 degrees on the circle. Object B's Sine wave starts when Object B is at 50 degrees on the circle.

Object A’s *Cosine* wave looks like this:



Object B’s Cosine wave looks like this:



As we can see, the wave curves for Object B are shifted 50 degrees to the left along the time axis in comparison to those of Object A. The wave curves from Object B have a 50-degree “head start”. They reach the peaks and dips 50 degrees earlier. The technical way to describe this shifting is to say that the Sine and Cosine waves derived from the movement of Object B have a “phase of 50 degrees”.

The formulas for the Sine and Cosine waves for Object A are:
“y = sin 360t”
“y = cos 360t”
... but we could also give these formulas as:
“y = sin (360t + 0)”
“y = cos (360t + 0)”
... which mean exactly the same thing. The value 0 is the phase in the formula measured in degrees.

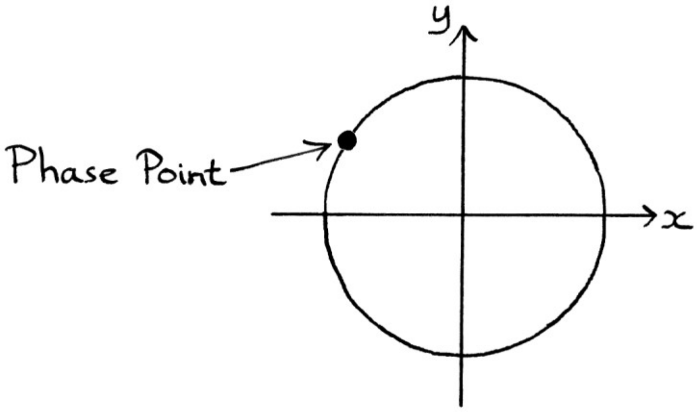
The formulas for the Sine and Cosine waves for Object B are:
“y = sin (360t + 50)”
“y = cos (360t + 50)”
... where the value 50 is the phase in the formula as measured in degrees.

I will explain more about the formulas shortly.

It is important to note that if a Sine wave has a particular phase in its formula, then the corresponding Cosine wave will have the same phase in *its* formula, and vice versa. This is because they both derive from the same object on the circle, and one object only has one starting point.

The waves derived from an object moving around a circle with a non-zero phase still have the same amplitude and frequency as they would have done if they had been derived from an object and circle with zero phase. The only difference is that the phase changes the starting point of the object about to rotate around the circle. Phase is independent of amplitude and frequency. This might be obvious when you think about how the starting point of an object rotating around the perimeter of a circle will not affect the radius of the circle or the speed at which the object rotates around the circle.

In this book, I will refer to the starting point of an object rotating around a circle as “the phase point”. We could just as easily call it the “starting point” or other terms.



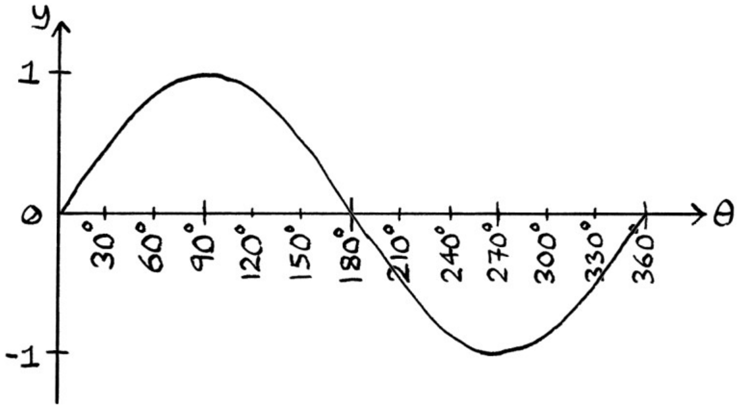
Waves

Phase is generally seen and thought about more on waves than it is on circles. Therefore, this section will look at just waves.

Angles

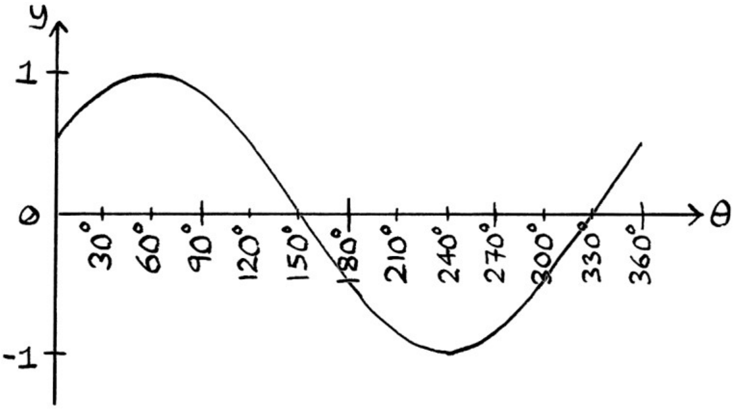
To simplify the explanation, we will go back to using waves with just angles. In other words, we will forget about time for now.

If we plot the points for one cycle of the wave "y = sin θ", we end up with this graph:



The graph starts when "θ" is zero, and its first y-axis value is also zero.

If we plot the points for one cycle of the wave "y = sin (θ + 30)", we end up with this graph:



We can see that when "θ" is zero, the y-axis value is greater than zero. In fact, when "θ" is zero, its y-axis value is 0.5, which is "sin (0 + 30)", or "sin 30". On the "y = sin θ" graph, the y-axis value at 180 degrees was zero. On the "y = sin (θ + 30)" graph, the corresponding zero point is at 150 degrees. All the y-axis values on the "y = sin θ" graph are reached 30 degrees earlier on the "y = sin (θ + 30)" graph. The whole of the "y = sin (θ + 30)" wave is identical to the "y = sin θ" wave, except it has been slid to the left by 30 degrees.

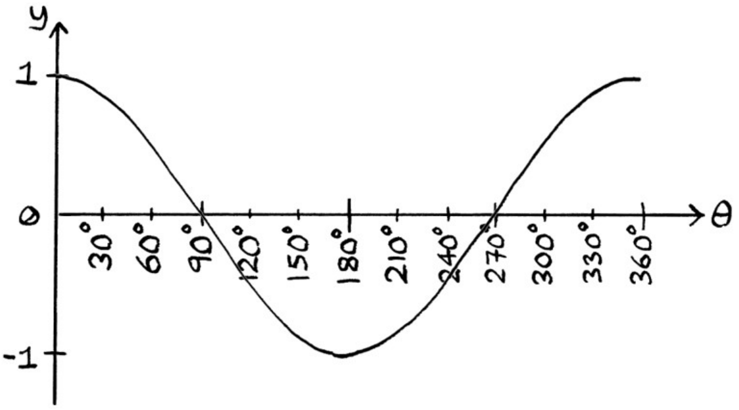
If we look at the *formula* of the "y = sin (θ + 30)" graph, we can see that every value of "θ" has 30 degrees added to it before it is Sined. This means that we are Sining a consistently larger number for every point on the graph. This wave will always have a 30 degree "head start" on the "y = sin θ" wave. It will reach the y-axis points 30 degrees sooner than the "y = sin θ" wave. The 30 degrees we have added on to the Sine wave in the equation is the "phase".

Among the observations we can make about the waves using wave-related terminology are:

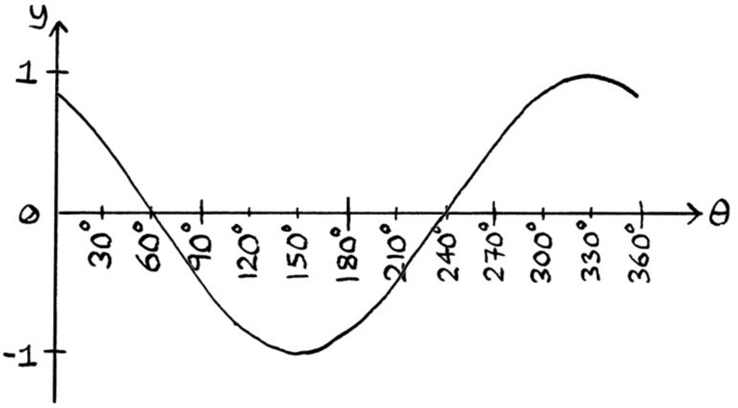
- The "y = sin (θ + 30)" wave has a phase of 30 degrees.
- The "y = sin θ" wave has a phase of zero degrees. This wave could also be described as "y = sin (θ + 0)".
- There is a "phase difference" between the "y = sin θ" and the "y = sin (θ + 30)" waves of 30 degrees.
- The "y = sin (θ + 30)" wave is a "phase shifted" version of the "y = sin θ" wave – it has been "shifted" along the θ-axis to the left.

If we look at the corresponding Cosine graphs, "y = cos θ" and "y = cos (θ + 30)", we see the same behaviour.

This is the "y = cos θ" graph:



This is the “y = cos (θ + 30)” graph:



Whereas the “y = cos θ” wave crosses the x-axis at 90 degrees and 270 degrees, the “y = cos (θ + 30)” wave crosses the x-axis at 60 degrees and 240 degrees. The “y = cos (θ + 30)” graph is identical to the “y = cos θ” graph, except it has been shifted to the left by 30 degrees.

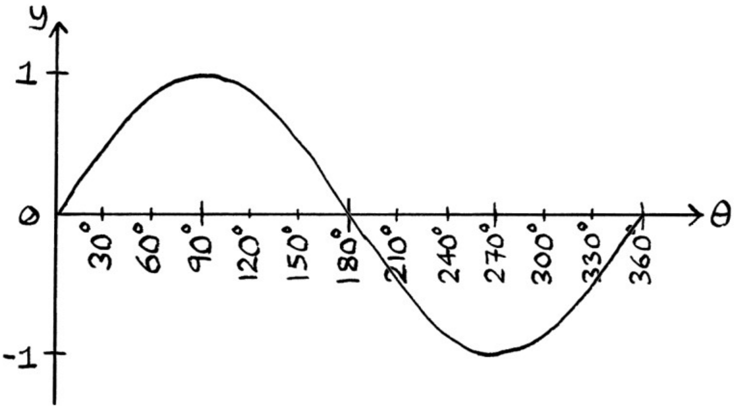
Two useful rules to know are that:

- A *positive* phase in the formula for a wave slides the wave to the *left* by that number of degrees in comparison to a wave with no phase.
- A *negative* phase slides the wave to the *right* by that number of degrees in comparison to a wave with no phase.

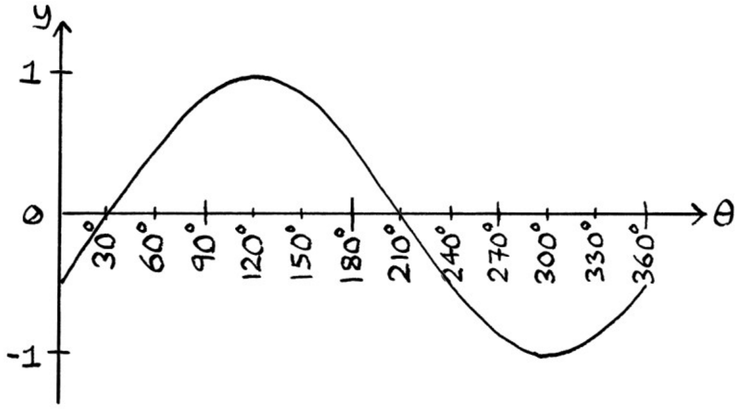
[Note that these are slightly simplified rules, as I will explain later].

If we compare a “y = sin θ” graph and a “y = sin (θ − 30)” graph, the “y = sin (θ − 30)” graph is the same but slid to the *right* by 30 degrees. This is because every value of “θ” is having 30 degrees taken off it before it is Sined. We are always Sining a lower number. This means the graph is delayed – it only reaches the same y-axis values as the “y = sin θ” graph 30 degrees later.

Here is the “y = sin θ” graph:



Here is the “y = sin (θ – 30)” graph:

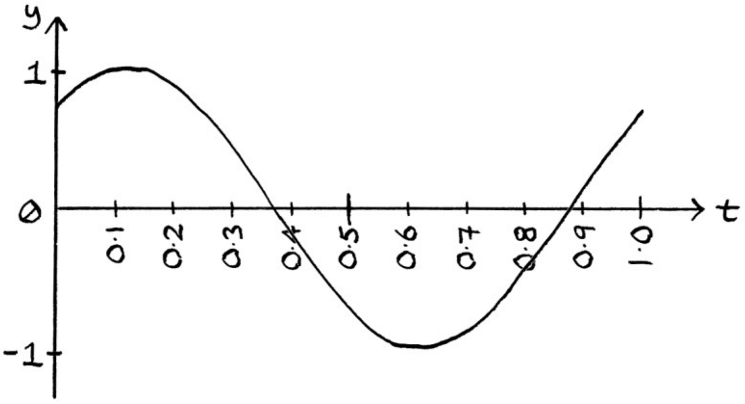


The “y = sin θ” graph crosses the x-axis at 0 degrees and 180 degrees (and 360 degrees). The “y = sin (θ – 30)” crosses the x-axis at 30 degrees and 210 degrees. If we drew the curve continuing for more cycles, it would also cross the x-axis at 390 degrees. The “y = sin (θ – 30)” graph is delayed by 30 degrees.

Time

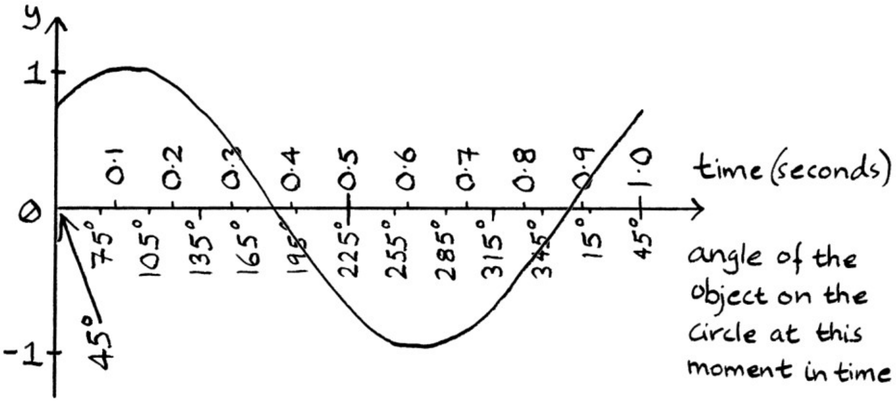
The concept of adding to the angle, or subtracting from the angle, before it is Sined or Cosined applies equally to waves that involve *time*. In such cases, the phase in the formula is still an angle, as we can tell by thinking about the circle – the phase in the formulas represents the starting angle of the object about to rotate around the circle. Referring to phase as an *angle* can be mildly confusing to newcomers if the wave graph is labelled with *time* as the x-axis. Therefore, as always, it pays to consider the circle from which the Sine and Cosine waves are, or could have been, derived. [I say “could have been” because in the real world not all waves literally or obviously derive from circles, but it makes things a lot easier to think about them as if they do.]

In this formula for a wave: “y = sin (360t + 45)”, the “45” is 45 degrees, even though the “t” refers to time in seconds. The graph looks like this:

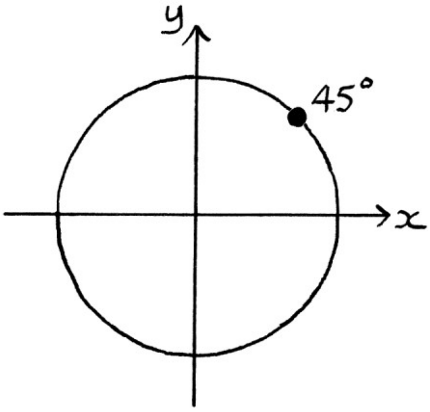


This graph has a head start of 45 degrees, which in this case amounts to a head start of 45 ÷ 360 = 0.125 seconds. It reaches all the peaks and dips that a “y = sin 360t” graph would reach, but always 0.125 seconds earlier.

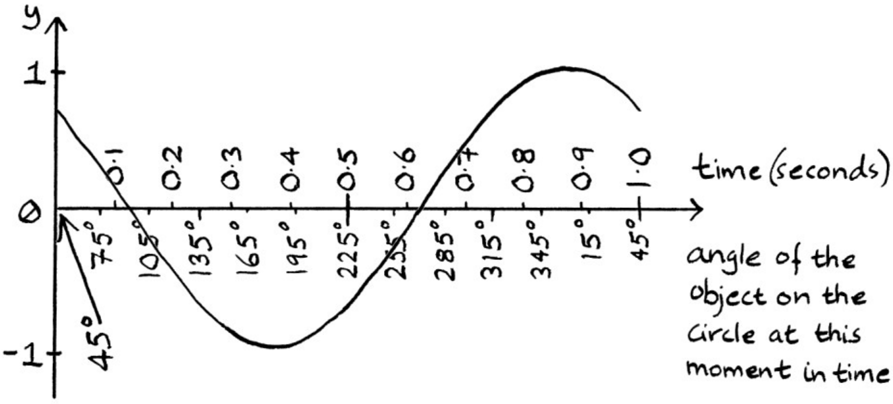
If we labelled the time axis with both time and the angle of the object on the circle at that particular moment in time, the graph would appear as in the following picture. Note, how after one full cycle, the angle of the object on the circle is back at 45 degrees.



The circle for this graph looks like this, with the starting point (the phase point) marked:

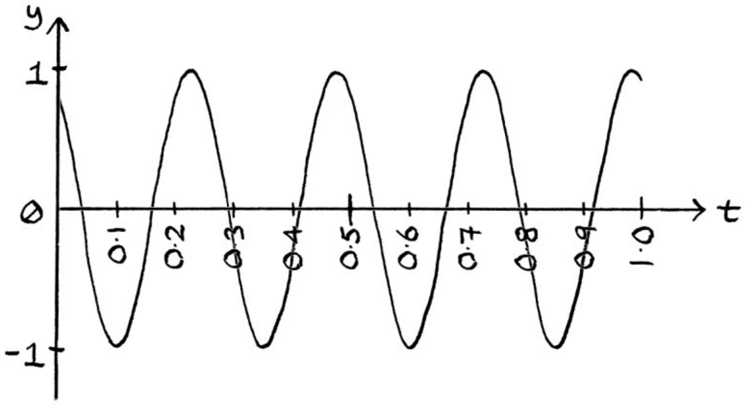


The wave's corresponding Cosine wave looks like this when drawn with a t-axis that shows both the time and the angle of the object on the circle at that particular time:

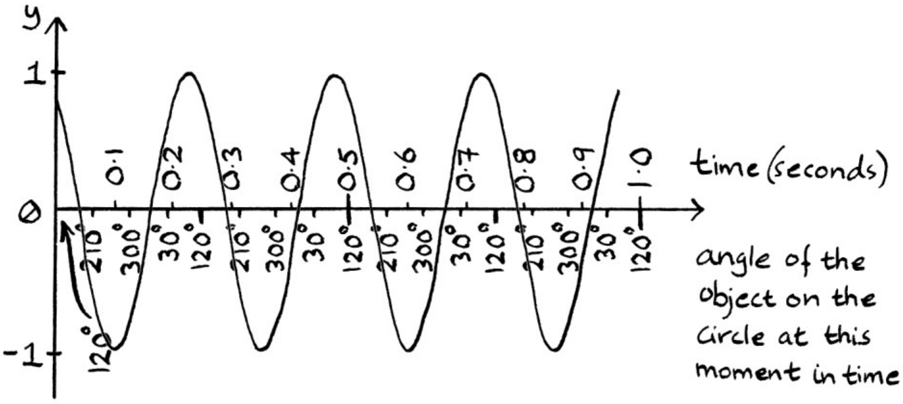


Another example

As another example, we will look at the wave “y = sin ((360 * 4t) + 120)”. This wave has a frequency of 4 cycles per second, and it has a phase of 120 degrees. The wave looks like this:



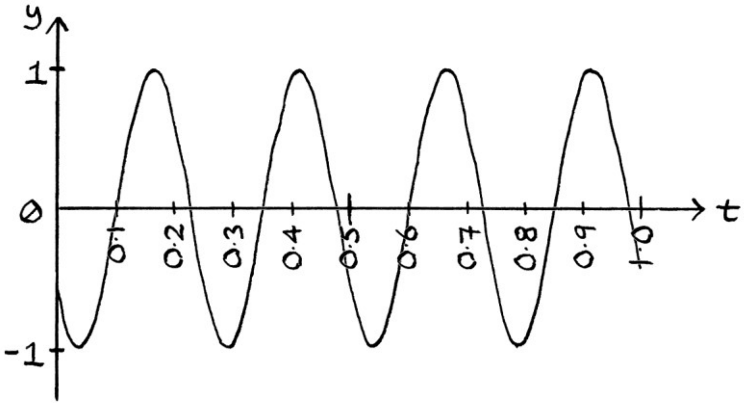
Note how it is harder to tell the effect of the 120-degree phase on this graph as the t-axis is time in seconds, and the frequency is not 1 cycle per second as it was in the previous example. Here is the same wave with the angle of the object on the circle at particular times written on it too. Now it is slightly easier to see how the phase of 120 degrees has affected the wave:



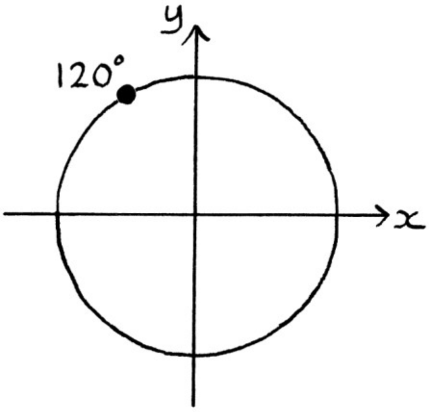
Note how after every cycle, the angle of the object on the circle is back at 120 degrees.

We can work out how the phase in degrees relates to a time in seconds by looking at the frequency in the formula. The angle of 120 degrees is 120 ÷ 360 = 0.3333 (a third) of the way around a circle. The frequency is 4 cycles per second. Therefore, a cycle is completed in 1 ÷ 4 = 0.25 seconds (in other words, the period is 0.25 seconds). A third of 0.25 seconds is 0.25 ÷ 3 = 0.08333 seconds (or in other words a twelfth of a second). Therefore, the phase of 120 degrees in this formula means that the wave is shifted 120 degrees to the left (relative to a wave with zero phase), which is also equivalent to a shift of twelfth of a second to the left.

The corresponding Cosine wave to this wave is “y = cos ((360 * 4t) + 120)”. As I said before, if a Sine wave has a particular phase in its formula, then its corresponding Cosine wave will have exactly the same phase.



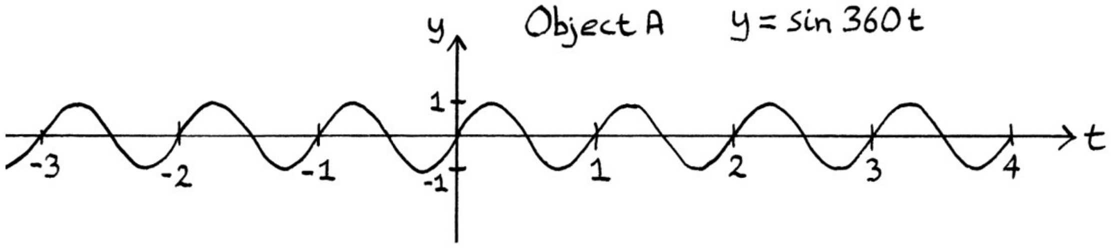
The circle from which these waves are derived looks like this, with the starting point (the phase point) marked:

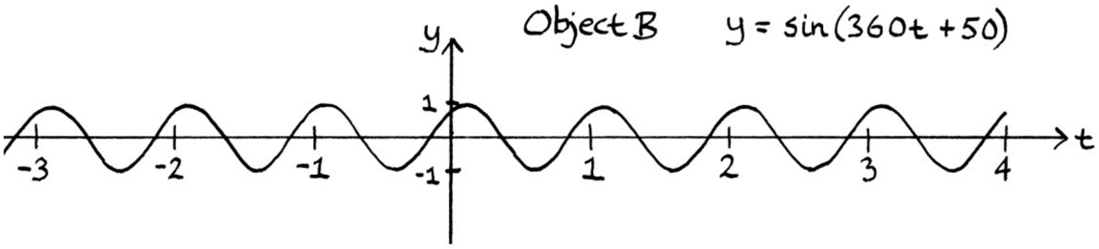


Forever waves

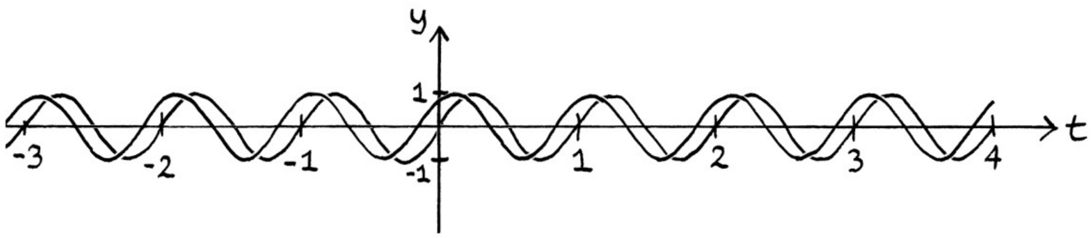
We will look again at the example of Objects A and B from earlier in this chapter. The formulas for the waves for Object A were “y = sin 360t” and “y = cos 360t”. The formulas for the waves for Object B were “y = sin (360t + 50)” and “y = cos (360t + 50).

If we extend the Sine waves of Objects A and B to represent the objects moving around the circles forever, the wave graphs would look like this: (These are drawn to a different scale so they fit on the page):





Every peak and dip for Object B’s graph occurs 50 degrees *earlier* than the peaks and dips for wave graph A. We can plot the two graphs over each other for comparison:



Phase phrases

There are multiple ways we can describe the phase aspect of Object A and Object B's Sine waves. We can say that:

- "Object B's Sine wave has a phase of +50 degrees."
- "Object B's Sine wave *formula* has a phase of +50 degrees".
- "The phase difference between Object A and Object B's waves is 50 degrees".
- "The phase difference between Object B and Object A's waves is 50 degrees".
- "Object B's Sine wave is the same as Object A's if Object A's had had a phase shift of 50 degrees to the *left*".
- "Object A's Sine wave is the same as Object B's, if Object B's had had a phase shift of 50 degrees to the *right*".
- "The waves are the same, but Object A's is a time lagged version of Object B's."

The same things can be said about Object A and Object B's *Cosine* waves.

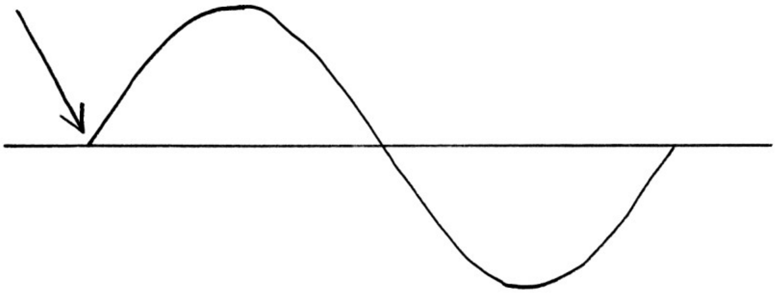
There are other phase-related comments we can make:

- "Object A's Sine wave formula has no phase."
- "Object A's Sine wave formula has zero phase"
- "Object A has zero phase."
- "Object B has a phase of +50 degrees."

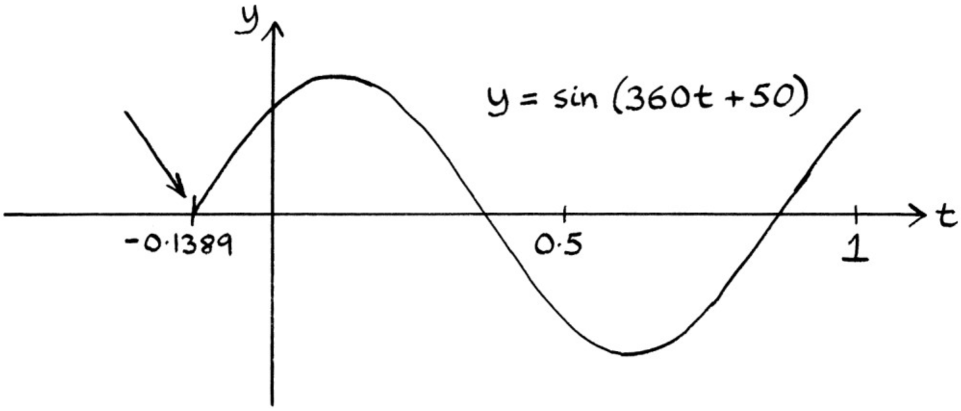
You might not generally hear people refer to an object rotating around a circle as having a phase, but in my view, it makes sense when discussing waves. An object's phase is its starting point on the circle.

Confusion about starting points

As we know, if we treat Object B as moving around before and after t = 0, we could extend its “y = sin (360t + 50)” Sine wave graph backwards and forwards as we did before. Doing this is a good way of demonstrating a mistake that some people make when thinking about phase shifts with waves. Some people consider the “start” of a wave as being the equivalent point on the curve as when a “y = sin 360t” wave is at t = 0. In other words, for them, the start of a Sine wave is here:

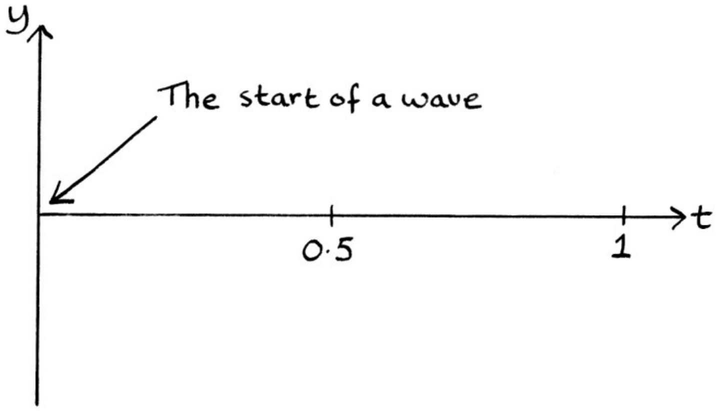


For them, the “y = sin (360t + 50)” wave *starts* when the time is −50 ÷ 360 = −0.1389 seconds. They would consider it “starting” earlier than the “y = sin 360t” wave.

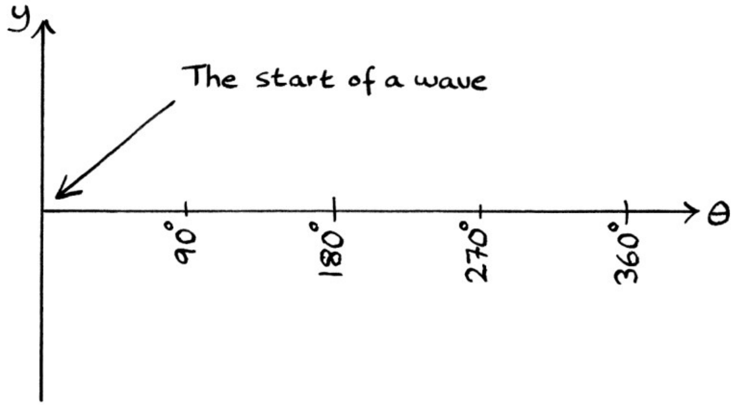


Thinking of the “starting” point in this way might help in contemplating the difference in similar waves, but it is a bad way to think about waves if you never want to become confused. In my opinion, the starting point of a wave relating to time is *always* at t = 0.

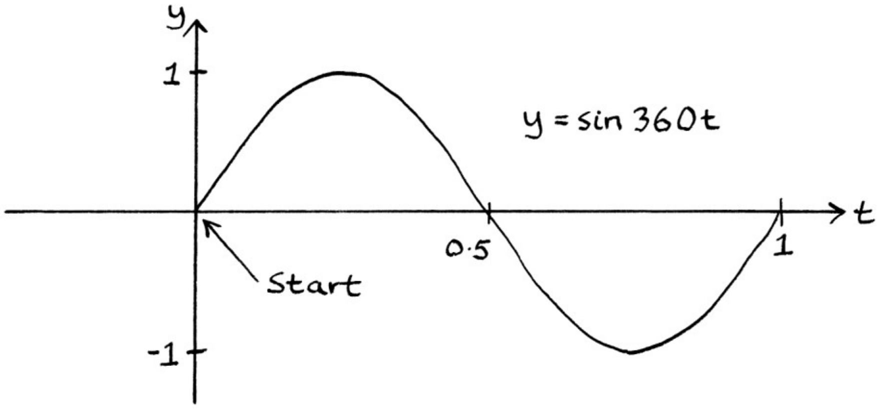
It does not matter what the shape of the wave is, a wave relating to time always starts here:

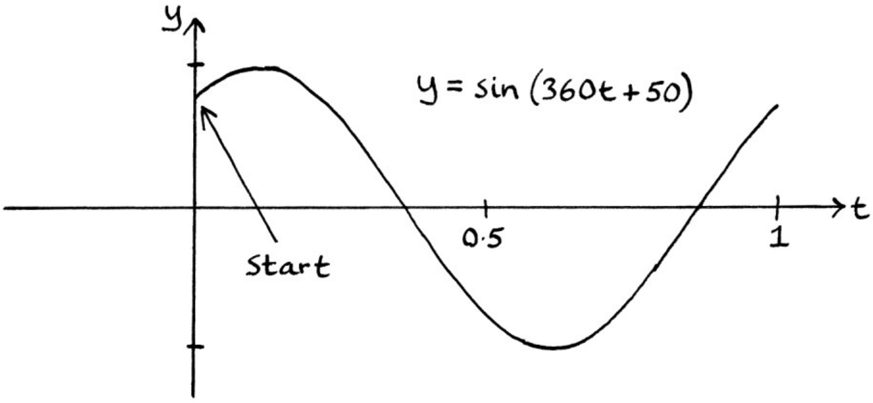


Similarly, the starting point of a wave relating to angles is *always* at the point when "θ" is zero. In the formula, "θ" might have values added to it or subtracted from it, but the graph starts when "θ" is zero. Therefore, the wave should be considered starting when "θ" is zero.



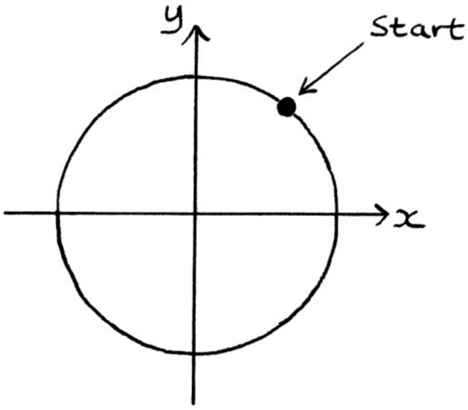
For our "y = sin 360t" and "y = sin (360t + 50)" waves, their starts are indicated as so:



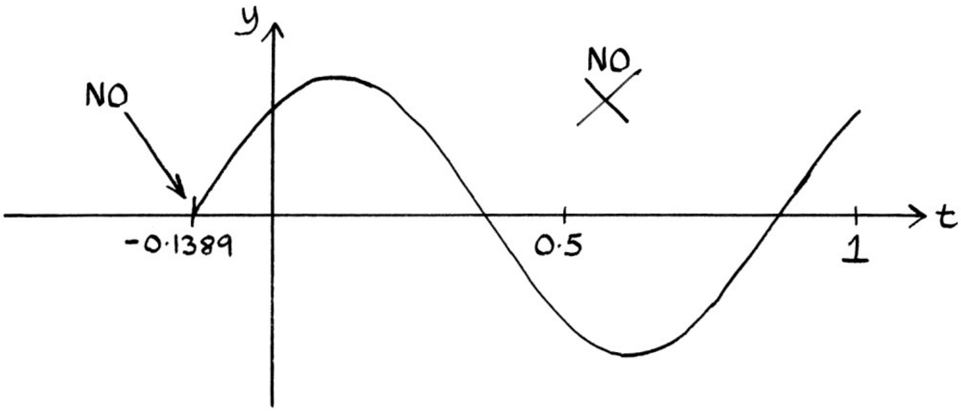


We can always draw the waves as they appear before t = 0 if we want, but the waves' *starts* are at t = 0, no matter how we draw them.

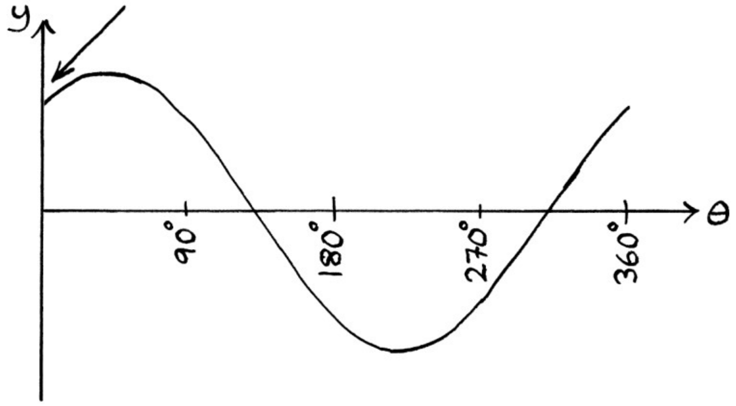
The proper way of thinking about the starting point is made clearer by thinking about the circle:



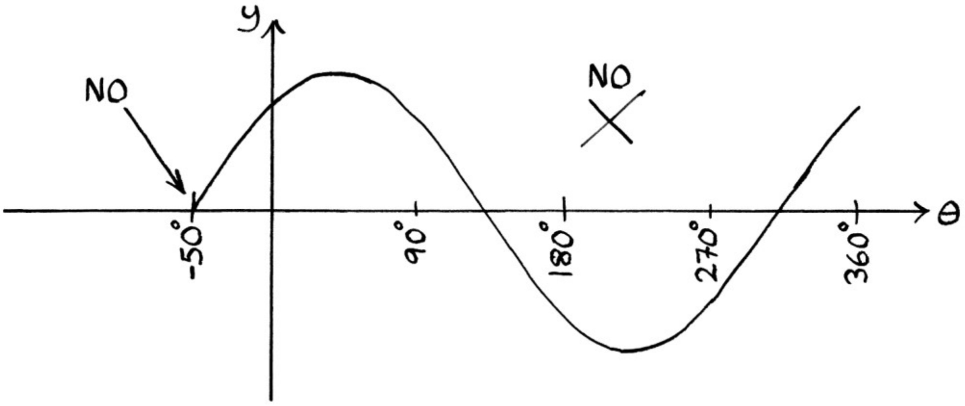
An object rotating around the circle that has as its formula, "y = sin (360t + 50)", *starts* at 50 degrees. When the time is zero seconds, it will be at 50 degrees on the circle. The wrong way of thinking about "starts" would imply it was starting at 0 degrees (i.e. −0.1389 seconds before it actually starts).



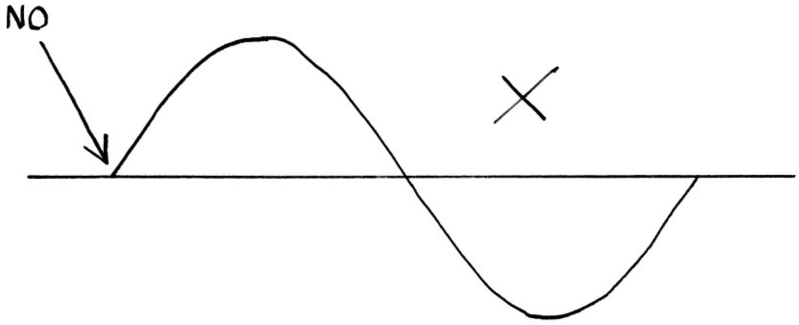
Similarly, if we were dealing with just angles and not time, “y = sin (θ + 50)” should start with the y-axis height of the point on the circle’s edge when the angle is 50 degrees, which is when “θ” is zero. When “θ” is zero, it has 50 added on to it, and the y-axis value on the wave graph is the Sine of θ+50.



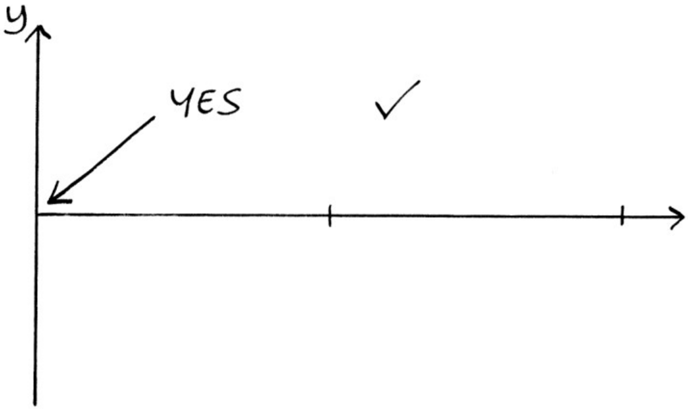
The wrong way of thinking about starts has the wave starting when “θ” is −50 degrees.



To reiterate all of the above, do not think of this as the start of the wave:



Instead, think of t = 0 (or θ = 0) as the start of the wave:



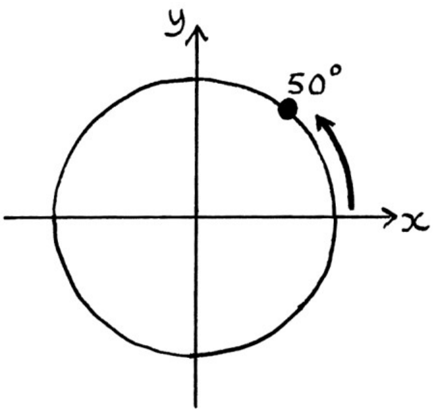
For most of the time that you deal with waves, this distinction might not matter. However, if you want to understand phase and not be confused by phase, it makes sense to think of the start of waves in this way.

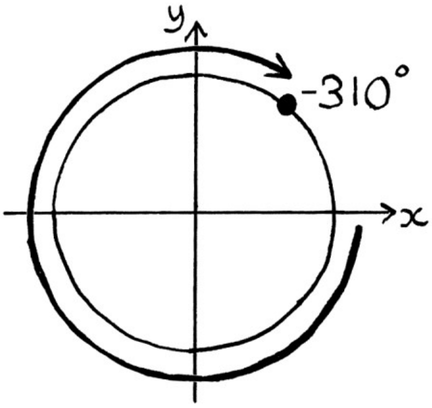
Negative phase in formulas

When dealing with objects moving around circles, it might be apparent that a positive phase in a formula could just as easily be treated as a negative phase equal to that value subtracted from 360, and vice versa. This is because on a circle we can measure the starting position of the object either way around.

Using our example of the object starting at 50 degrees, the formulas for the two derived waves could be:
“y = sin (360t + 50)”
“y = cos (360t + 50)”
... or they could be:
“y = sin (360t − 310)”
“y = cos (360t − 310)”

This is because the point at an angle of +50 degrees on the circle could also be identified by the angle −310 degrees. +50 and −310 refer to the same place on the circle.





The Sine and Cosine waves derived from the object’s movement around the circle will be identical no matter which way is used to identify that point. The waves show the object’s x-axis and y-axis position over time as it rotates from that point, and it does not matter how we describe that point.

As another example, a phase of +220 degrees is the same as a phase of −140 degrees. [360 − 220 = 140]. Therefore, the formulas:
“y = sin (360t + 220)” and “y = cos (360t + 220)”
... are identical in meaning to the formulas:
“y = sin (360t − 140)” and “y = cos (360t − 140)”

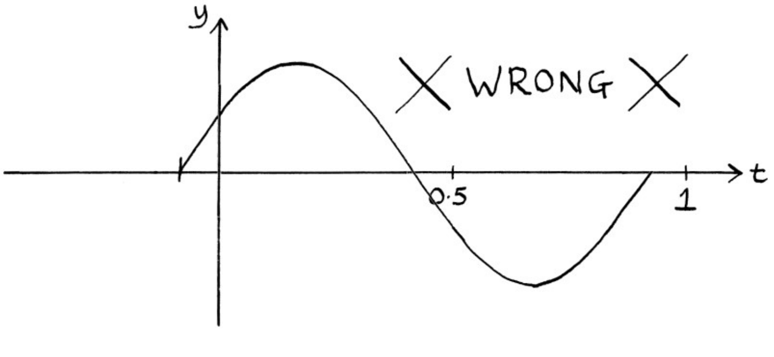
A phase of +170 degrees is the same as a phase of −190 degrees. [360 – 170 = 190]. Therefore, the formulas:
“y = sin (360t + 170)” and “y = cos (360t + 170)”
... are identical in meaning to the formulas:
“y = sin (360t – 190)” and “y = cos (360t – 190)”

A phase of +30 degrees is the same as a phase of −330 degrees. Therefore, the formulas:
“y = sin (360t + 30)” and “y = cos (360t + 30)”
... are identical in meaning to the formulas:
“y = sin (360t – 330)” and “y = cos (360t – 330)”

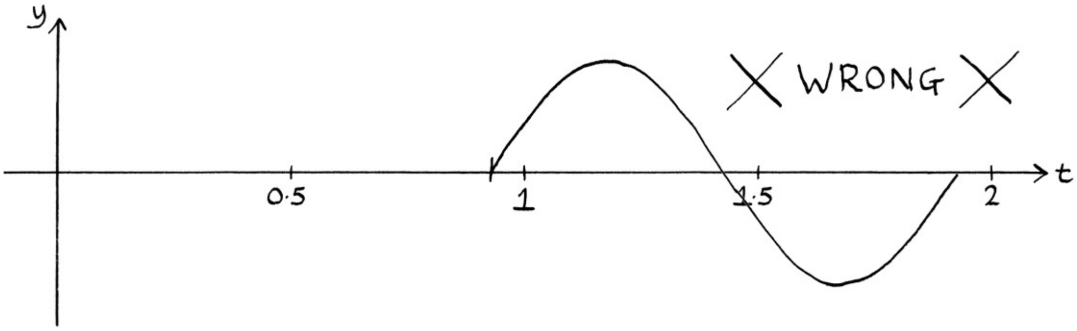
Although you might think that there is some subtle difference between these positive phase and negative phase wave pairs, in reality, they are identical. No matter what the value of “t”, it does not matter whether we consider the phase as being +30 or −330. They both result in the same values for each Sine wave or each Cosine wave. On the circle, +30 degrees and −330 degrees refer to *exactly* the same point. It does not matter whether we refer to that point as being at +30 degrees or −330 degrees, it is still the same point. Similarly, we could also identify exactly the same point as being +390 degrees, or +750 degrees. Although these measure more than once around the circle, they still refer to exactly the same point on the circle, and so the derived graphs will be identical.

For the Sine and Cosine waves with +30 degree phase, when t = 0, the y-axis values will be 0.5 and 0.8660 respectively. For the Sine and Cosine waves with −330 degree phase, when t = 0, the y-axis values will be exactly the same. The same is true if the wave have a +390 degree phase.

One reason you might think that there is a difference is if you are incorrectly thinking about where the start of a wave is. As I said before, a wave should only be thought of as starting when t = 0, or when θ = 0. If you think of the start in the wrong way, then you might imagine the first cycle of a +30 degree phase Sine wave looking like this, which would be wrong:



... and you might think of the −330 degree phase resulting in a Sine wave that looks like this for the first cycle, which is obviously wrong:



These graphs are wrong, and they show the wrong way to think about the waves. The waves start when t = 0. The waves do not start when t = −0.08333 seconds or when t = +0.9167 seconds. This is an example of how careless thought about starts can cause confusion with phase.

If we were dealing with just angles, and not time, then it would still be the case that a wave with a positive phase would be identical to a wave with that same phase described in a different way.

Side note: head starts

Strictly speaking, the use of the term “head start” is slightly ambiguous when thinking of circles and waves. An object’s position on the circle could be said to be both ahead and behind a different object’s position depending on how we want to think about the situation. A circle has no beginning or end, so it is a matter of debate whether we say that an object at, say, 270 degrees is ahead or behind an object at, say, 90 degrees. However, to make things simpler, I think it is good to consider an object at a higher angle on the circle as being “in front” or “ahead” of an object at a lower angle, with the rule that we treat all angles as being between 0 degrees and just under 360 degrees. For example, an object at an angle of 200 degrees has a head start on an object at an angle of 100 degrees. An object at an angle of 359 degrees has a head start on an object at an angle of 1 degree.

Similarly, the angle 730 degrees will be treated as 10 degrees; the angle −10 degrees will be treated as 350 degrees. Therefore, an object at an angle of −10 degrees is ahead of an object at an angle of +730 degrees.

All of this is arbitrary, but it will help avoid confusion when thinking about phase.

Positions of objects

It can sometimes seem counterintuitive that a wave with a higher phase than another has a head start, yet a wave that has a negative phase does not mean it starts later. In reality, it is the *position* of a point around the circle that dictates whether it has a head start or not. Whether that position is indicated with a positive or negative angle makes no difference to the literal position of that point, and therefore it makes no difference as to whether that particular point is ahead or behind any other point. A Sine or a Cosine wave derived from a circle will have a phase based on the position of that point on the circle, and it is irrelevant to the waves how the position of that point is described.

An apparent contradiction

Some points about phase when viewed solely on waves that you may have gathered so far are:

- A positive phase in the wave formula results in the wave being shifted *left* along the time or θ-axis by that number of degrees in relation to a wave with zero phase.

- A negative phase in the wave formula results in the wave being shifted *right* along the time or θ-axis by that number of degrees in relation to a wave with zero phase.

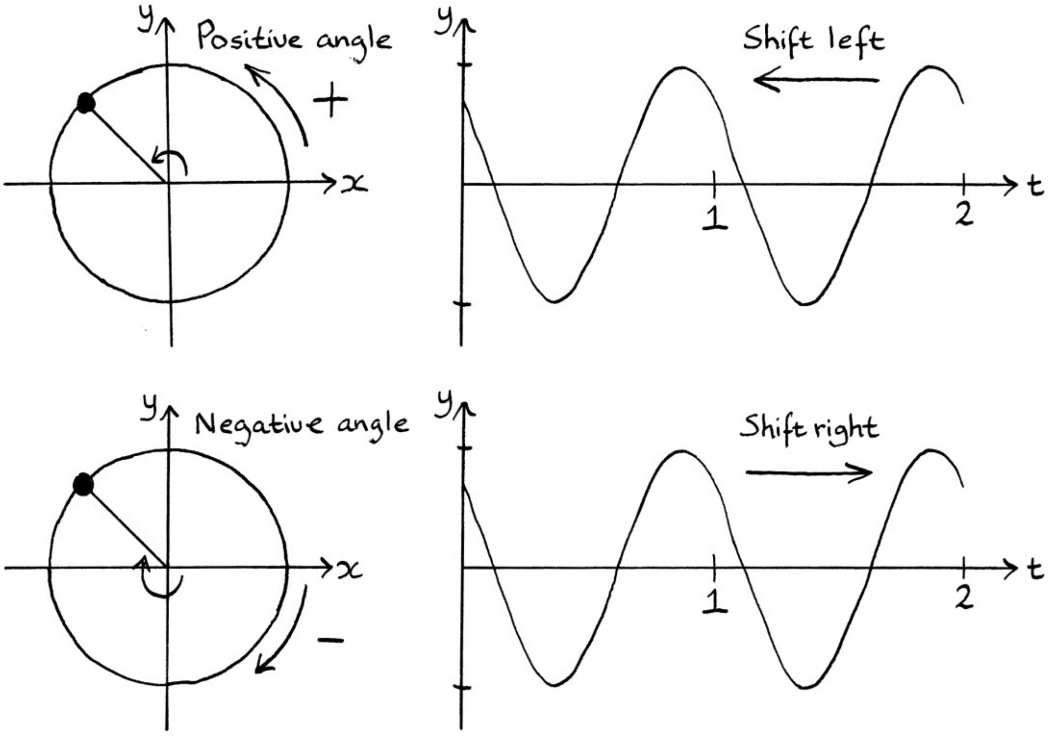
- A positive phase in the wave formula is equivalent to a negative phase of that value subtracted from 360.

- A negative phase in the wave formula is equivalent to a positive phase of that value subtracted from 360.

From these four points, there is seemingly a contradiction. If a positive phase shifts the wave left, and a negative phase shifts the phase right, then it cannot be the case that a positive phase can be equivalent to any negative phase. If that were true, then a shift left would be the same as a shift right.

However, because we are dealing with waves that constantly repeat their shapes (because they are derived from circles, which do not really have a beginning or an end), this is not a contradiction. If the waves did not repeat their shape, then it *would* be a contradiction to say that a shift left of so many degrees was equivalent to a shift right of that many degrees subtracted from 360. However, with periodic waves (in other words, waves that repeat their shape), this is not a contradiction – it is just something that might be confusing at first glance. The waves derived from circles repeat in an identical fashion for their entire existence.

As I have said before, the wave's starting position is related to the starting point of the object rotating around a circle. That point can be identified in countless ways. For example, −350 degrees, +10 degrees, +370 degrees and so on all refer to the same point. The circle helps with visualising the shifting. A positive phase, which appears as an anticlockwise rotation on the circle, appears as a leftwards shift for the waves derived from the circle. A negative phase, which appears as a clockwise rotation on the circle, appears as a rightwards shift for the waves derived from the circle. Obviously, we can measure around the circle in either direction to reach the same point, and similarly, we can shift the wave in either direction to achieve the same phase.



Ideally, there should be more comprehensive rules for describing shifting:

- A positive phase in the wave formula results in the wave being shifted *left* along the time or θ-axis by that number of degrees, or it results in the wave being shifted *right* along the axis by that number of degrees subtracted from 360. They amount to the same thing.

- A negative phase in the wave formula results in the wave being shifted *right* along the time or θ-axis by that number of degrees, or it results in the wave being shifted *left* along the axis by that number of degrees subtracted from 360. They amount to the same thing.

However, it is much easier just to simplify these and have the two basic rules with which we started. They do not give the full picture, but they are more useful as a consistent way of visualising what is going on.

- A positive phase in the wave formula results in the wave being shifted *left* along the time or θ-axis by that number of degrees, in relation to a wave with zero phase.

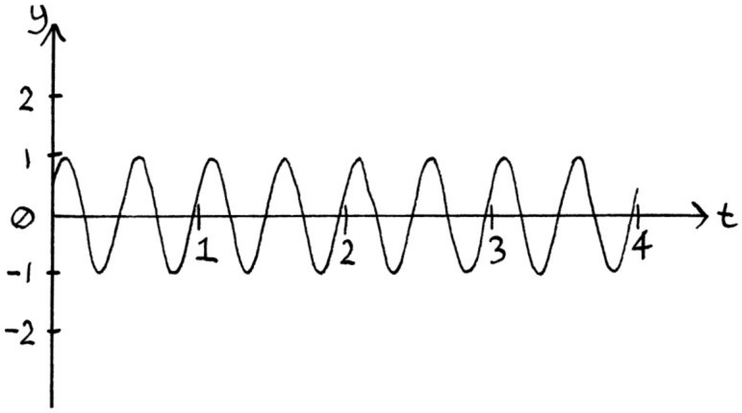
- A negative phase in the wave formula results in the wave being shifted *right* along the time or θ-axis by that number of degrees, in relation to a wave with zero phase.

Phase, frequency and amplitude

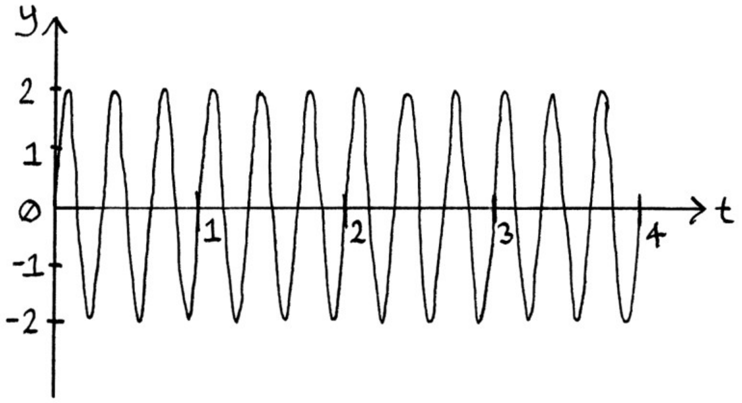
A wave formula with a non-zero phase can also mention frequency and amplitude. All three properties are independent of each other.

Examples

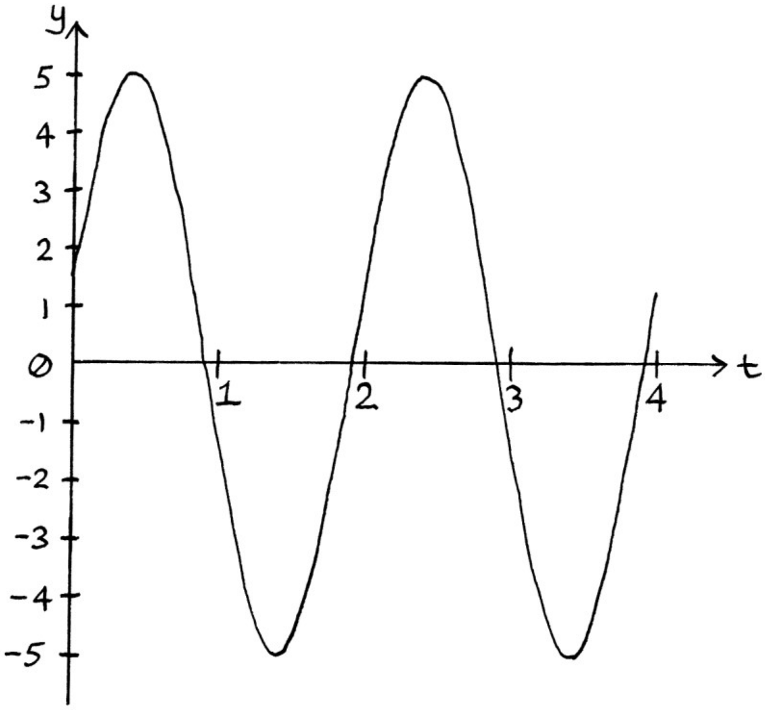
The wave, “y = sin ((360 * 2t) + 25)” has an amplitude of 1 unit, a frequency of 2 cycles per second, and a phase of 25 degrees.



The wave “y = 2 cos ((360 * 3t) + 270)” has an amplitude of 2 units, a frequency of 3 cycles per second, and a phase of 270 degrees.

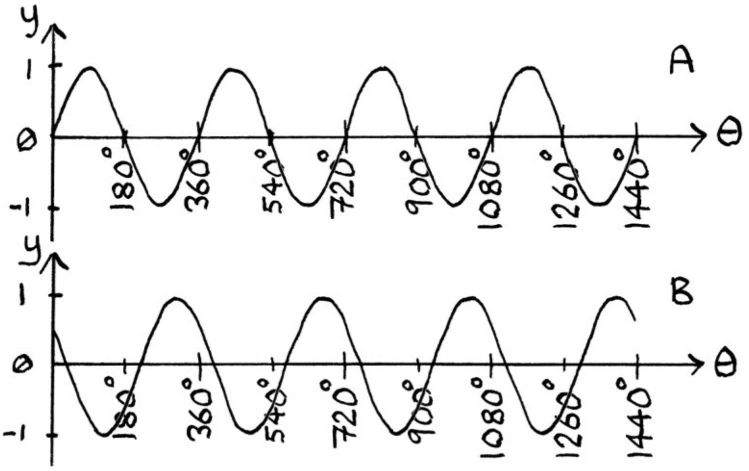


The wave “y = 5 sin ((360 * 0.5t) + 15)” has an amplitude of 5 units, a frequency of 0.5 cycles per second, and a phase of 15 degrees.

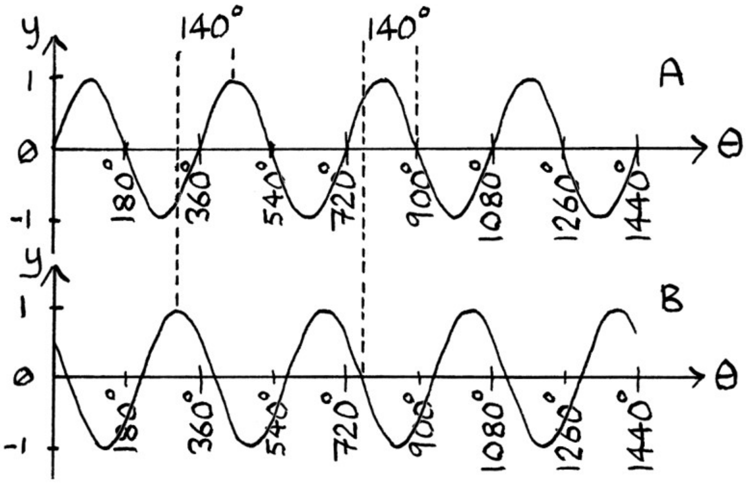


Relative phase

The difference in angle between the *formulas* of two waves of the same frequency is identical to the difference in angle between any two corresponding points on the cycles on the wave *graphs*. If we ignore time for a moment, and consider these two waves:
Wave A: y = sin θ
Wave B: y = sin (θ + 140)
... then their graphs look like this:

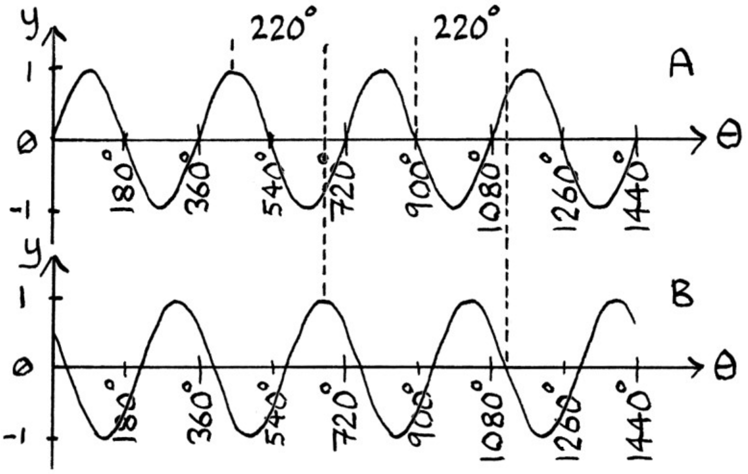


From the formula, we know that Wave B has a phase +140 degrees greater than Wave A's. On the graphs, one point on one wave will always be 140 degrees apart from the corresponding point on the other wave. This is most obvious when looking at the peaks, the dips, or the moments when the y-axis value is zero.

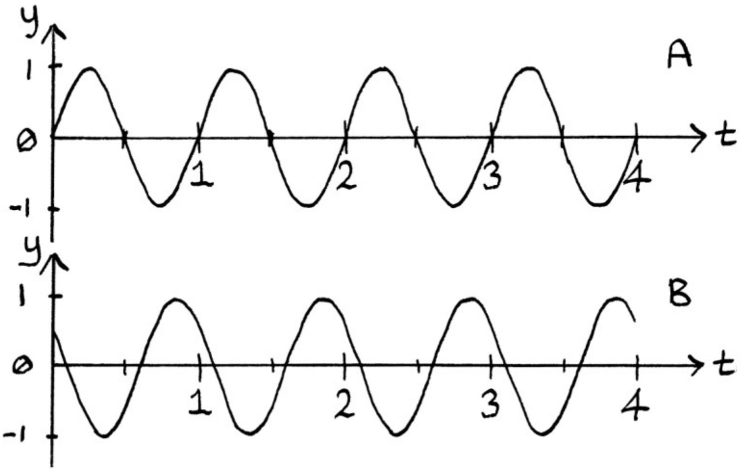


However, it is true for any corresponding point.

Another thing to notice is that we could also say that the points are 220 degrees apart too:



When dealing with time, the same principle applies, but it is slightly harder to know the phase difference. If we have these two waves:
Wave A: y = sin (360t)
Wave B: y = sin (360t + 140)
... then the wave graphs have the same shape as the non-time ones did, and they look like this:

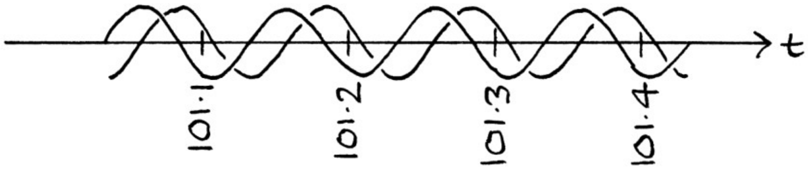


From the formula, we know that Wave B has a phase that is 140 degrees more than Wave A's. However, looking at the graphs, it is harder to see the phase difference because the x-axis is time. If we had very accurately drawn waves that were easy to read, we would be able to see that the time difference between the waves is 0.3889 seconds. Knowing the angle difference requires slightly more work. We do this by first working out what the frequency is: we can tell what the frequency is by seeing how often the cycles repeat – the frequency is 1 cycle per second. From that, we know that the waves repeat 360 degrees every second. Therefore, if we know that the time difference is 0.3889 seconds, and we know that the frequency is 360 degrees per second, we can calculate that the phase difference will be 0.3889 * 360 = 140.0004 degrees, which we will round to 140 degrees.

Unknown starts

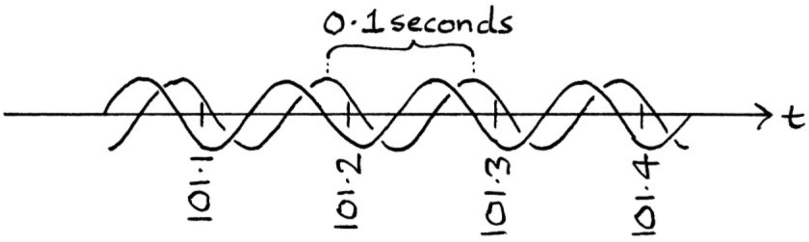
When dealing with real-world waves, we might have two waves of the same frequency for which we do not know the formulas, and for which we do not know when they started. We cannot say what their absolute phases are – we can only say what their relative phases are.

Consider these two waves drawn on the same graph. They have exactly the same frequency, but their phase is different.



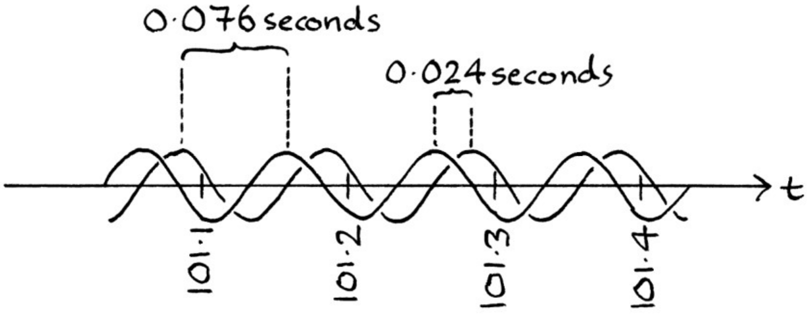
The t-axis here is "time since we started observing the waves", which, although almost the same as "time since the wave actually started", means we cannot tell what the actual phases are in the formulas for the two waves. However, we can see that they have different phases, and it is possible to calculate the difference in phase. Although the x-axis is time, we can still work out the phase difference between the waves as angles.

To do this, we first work out the frequency by seeing how often a cycle repeats every second. For these two waves, their cycles last 0.1 seconds (the periods are 0.1 seconds). Therefore, their frequencies are both 1 ÷ 0.1 = 10 cycles per second.



Then we measure the time difference between the two waves by picking corresponding points on the curves.

There are two time differences between the waves, depending on which wave we consider the “first” one.



The two time differences are 0.076 seconds and 0.024 seconds.

We will look at the 0.076 seconds measurement first. Although we just have waves on their own, we will imagine that they both derive from circles with objects rotating around them. It does not matter if they literally do or not – thinking about all waves in this way is always helpful. If the frequency of the objects rotating around the circles from which the waves can be said to derive is 10 cycles per second, then in 0.076 seconds, the objects would have travelled 0.076 * 10 = 0.76 cycles. As a full cycle is 360 degrees, 0.76 cycles is the same as 0.76 * 360 = 273.6 degrees.

Now we will look at the 0.024 seconds measurement. At 10 cycles per second, an object would have rotated 0.024 * 10 = 0.24 cycles of the circle. The value of 0.24 cycles is the same as 0.24 * 360 = 86.4 degrees.

From all this, either we can say that one wave precedes the other wave by 0.076 seconds or 273.6 degrees, or we can say that one wave precedes the other wave by 0.024 seconds or 86.4 degrees. As we do not know which wave started first, these are adequate deductions.

We cannot know the actual formulas of the waves, but we can still create formulas based on what we do know. We could give the formulas as:
“y = sin ((360 * 10t) + q)”
... and:
“y = sin ((360 * 10t) + 273.6 + q)”
... where “q” is an unknown phase, which might be any angle.

... or as:
"y = sin ((360 * 10t) + q)"
... and:
"y = sin ((360 * 10t) + 86.4 + q)"
... where "q" is an unknown phase, which might be any angle.

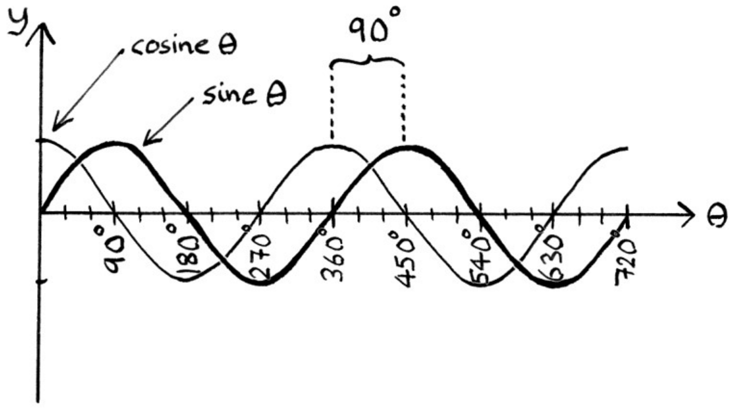
Analogies to phase

Phase, when thought about on the *circle*, is equivalent to someone starting a race and having a head start.

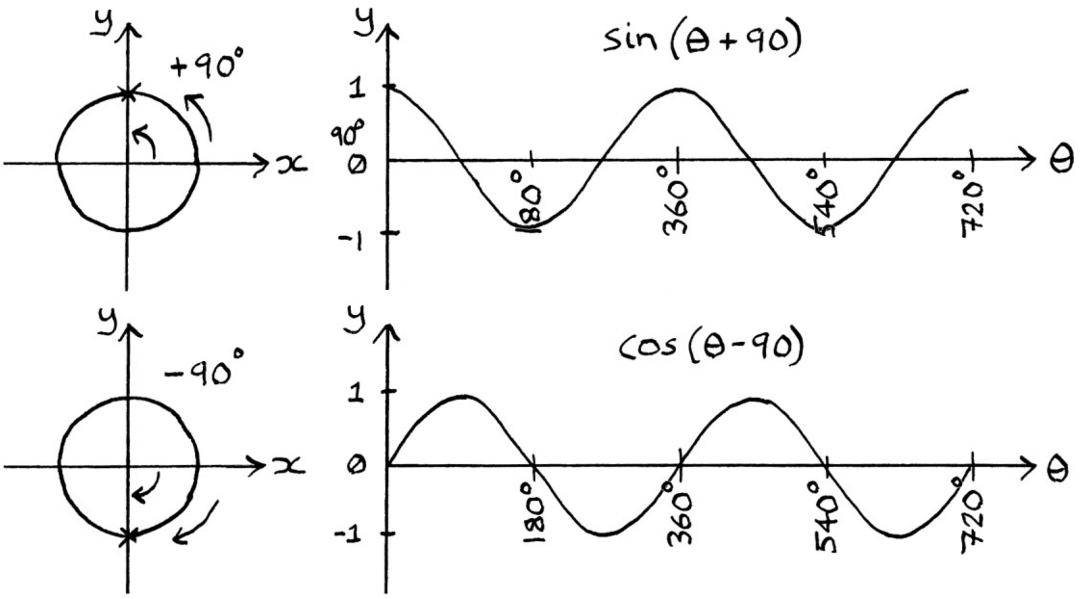
However, if you think of phase on the *wave* graph in this way, it can lead to confusion as you might think a head start would mean the wave "starts" further to the right on the time axis. Therefore, a better analogy for phase when thinking about *wave graphs* is the planting of a tree. If a tree had been planted further back in time, it would be more developed compared with one planted later.

Cosine is Sine with a non-zero phase

The Cosine of a value is the same as the Sine of that value plus 90 degrees. To put it another way, the Sine of "θ + 90" is the same as the Cosine of "θ". If we look at the "y = sin θ" wave and the "y = cos θ" wave, we will see that they are the same wave but with a 90 degree phase difference.



Although the Cosine wave is derived from the x-axis values of the circle, and the Sine wave is derived from the y-axis values of the circle, it is the case that their characteristics are identical except for the 90-degree phase difference. If for some reason, we could not be bothered to read the x-axis values from the circle to find the Cosine wave, we could just as easily read the y-axis values of a circle from 90 degrees further around to produce a Sine wave with a phase in its formula of +90 degrees. It would be exactly the same as the Cosine wave graph. Or, if we could not be bothered to read the y-axis values from the circle, we could just as easily read the x-axis values of a circle from 90 degrees earlier to produce a Cosine wave that had a phase of −90 degrees, and it would be identical to the Sine wave graph.



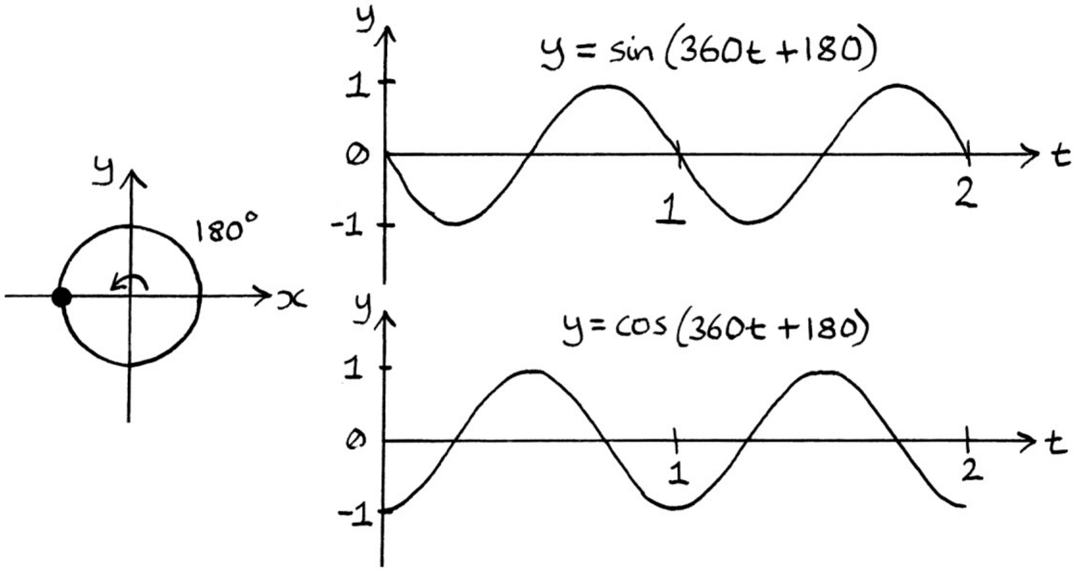
Concerning angle-based waves, a “y = cos θ” wave is identical to a “y = sin (θ + 90)” wave, and a “y = sin θ” wave is identical to a “y = cos − 90)” wave. Similarly, for time-based waves, a “y = cos 360t” wave is identical to a “y = sin (360t + 90)” wave, and a “y = sin 360t” wave is identical to a “y = cos (360t − 90)” wave. These can be summarised by saying that there is a 90-degree phase difference between Sine and Cosine. This idea was mentioned in Chapter 3, but now that we know the word “phase”, we can describe it differently.

On the “y = sin θ” and “y = cos θ” graphs, the Cosine wave reaches the peaks and dips 90 degrees earlier than the Sine wave. Some people describe this situation by saying that “the Cosine wave precedes the Sine wave by 90 degrees” or that “it leads the Sine wave by 90 degrees”. Of course, we could also say that a “y = cos θ” wave is identical to a “y = sin (θ − 270)” wave, and a “y = sin θ” wave is identical to a “y = cos (θ + 270)” wave. We could also say, “The Sine wave precedes the Cosine wave by 270 degrees,” because it means exactly the same thing as, “The Cosine wave precedes the Sine wave by 90 degrees.”

Other phase examples

180-degree phase shift

An object rotating around a circle that started at 180 degrees, would produce a Sine wave with the formula "y = sin (360t + 180)", and a Cosine wave with the formula "y = cos (360t + 180). The resulting Sine wave graph is the same as an upside down "y = sin 360t" graph, and the resulting Cosine wave graph is the same as an upside down "y = cos 360t" graph. Shifting the wave 180 degrees is the same as inverting it.



When looking at graphs of just angles, with no mention of time, a "y = sin (θ + 180)" wave is the same as an upside down "y = sin θ" wave, and a "y = cos (θ + 180)" wave is the same as an upside down "y = cos θ" wave.

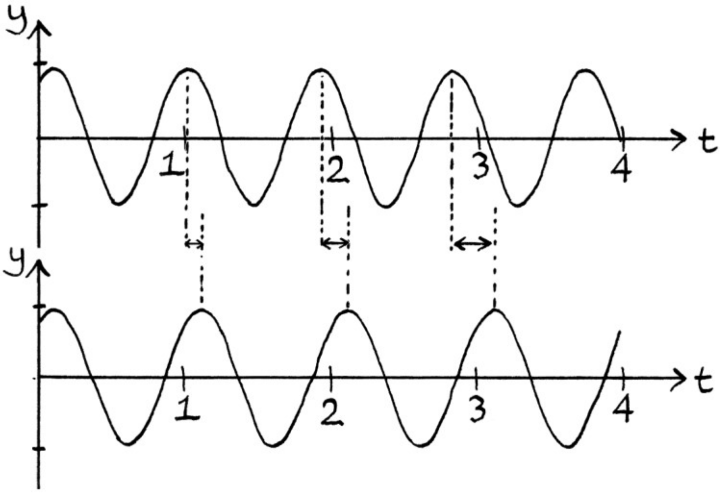
360-degree phase shift

An object that started rotating at 360 degrees would have identical graphs to if it had started rotating at 0 degrees. At t = 0, it would be in exactly the same position as if the angle had been given as 0 degrees. The angles of 0 degrees and 360 degrees refer to exactly the same place on the circle's edge.

Phase, frequency, and amplitude

Phase and frequency

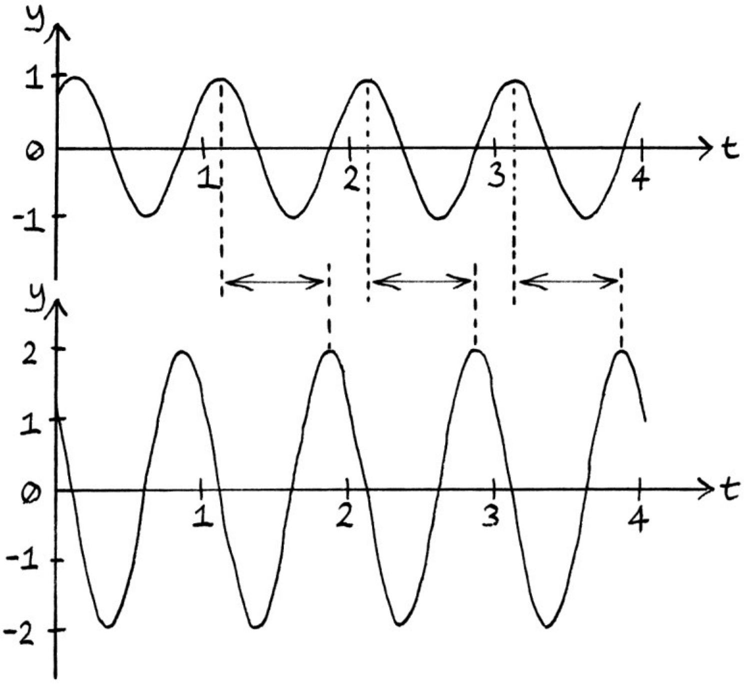
Comparing two waves of the *same* frequency, but with a different phase, can be useful. However, if the waves have different frequencies, the initial phase difference will only be apparent at the very beginning. Afterwards, the cycles of the waves will no longer be synchronised with each other. In the following picture, the second wave has a slightly slower frequency, so the corresponding points on each cycle move further apart as time goes on.



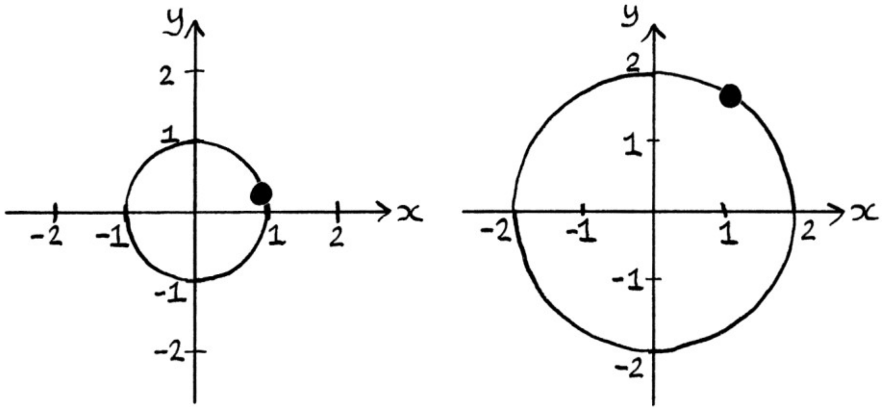
One way of testing if two consistently repeating waves have a different frequency is by seeing if the distance between corresponding cycles remains the same throughout their existence. If the phase difference is not the same for the existence of the wave, then they must have a different frequency.

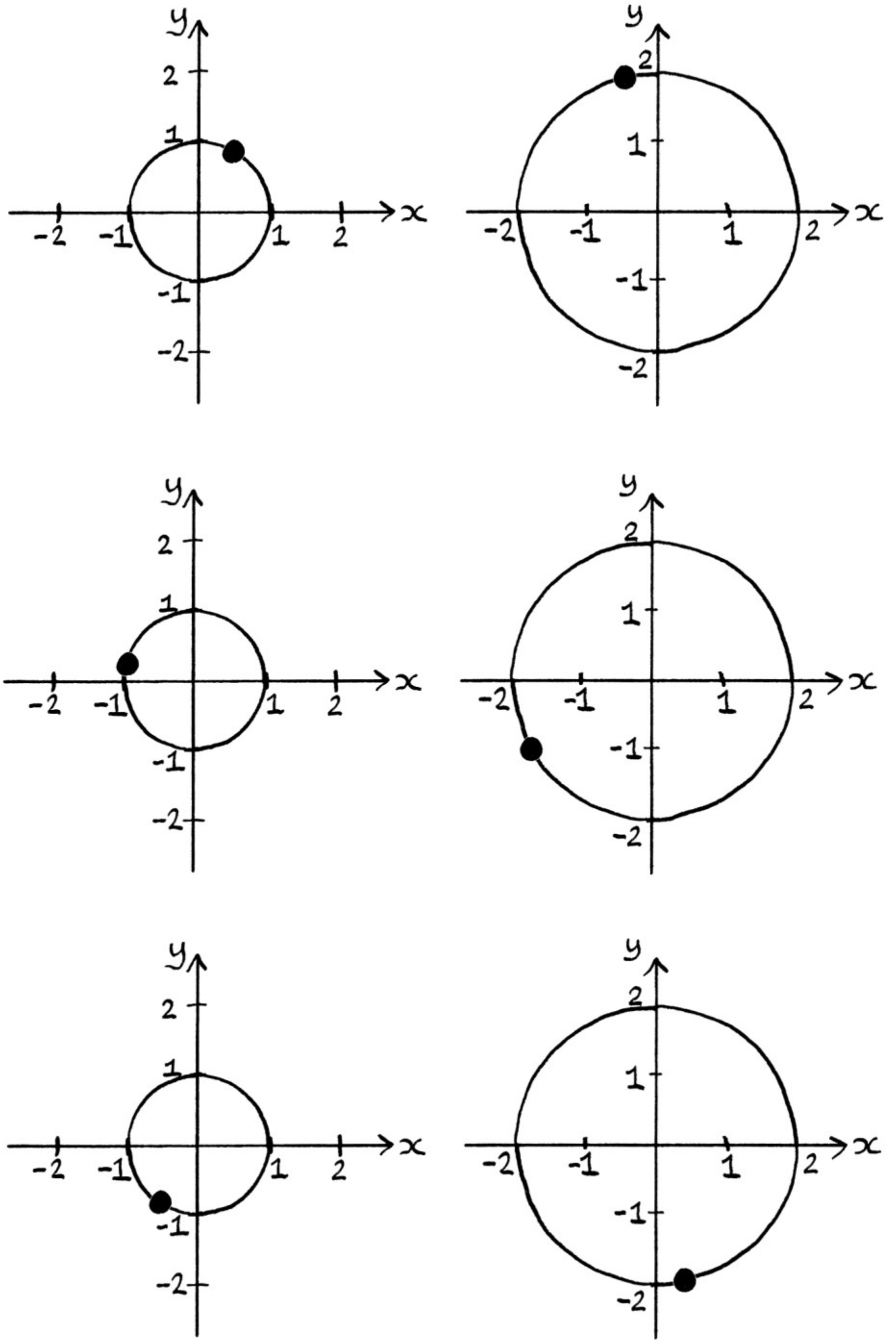
Phase and amplitude

If two waves have the same frequency but differ in amplitude and phase, the different phases will still be apparent for the duration of the waves.



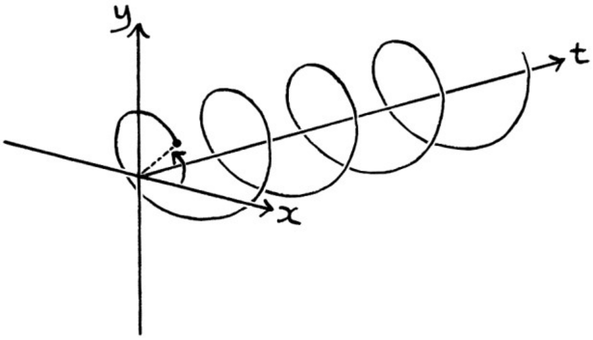
We can tell this would be the case by thinking of the circles from which the waves are derived. If we have two objects, each rotating around a circle at the same number of revolutions per second, then no matter how large the circles are, the objects will always be the same number of degrees apart.




y
x
-2
-1
1
2

Phase on the helix

Phase can be portrayed on the helix. If we had a Sine and Cosine wave pair with phases of 45 degrees, we would end up with the following helix. The object starts at 45 degrees.



Phase terminology

These are some of the words relating to phase that you might see in explanations or discussions of waves:

Phase angle

“Phase angle” is another way of describing the phase that appears in a formula. It has the same meaning as “phase”.

In phase

If two waves are described as being “in phase”, it means they have the same phase – they are “in phase with each other”. The shapes of the waves coincide with each other with respect to time (or angle if the wave does not relate to time). The waves might have different amplitudes, but the peaks and dips will occur at the same place. Any two waves of the same frequency that have the same phase value in their formulas will be “in phase” with each other. For time-based waves derived from objects rotating around circles, if two waves are in phase with each other, it means that the objects started at exactly the same angle on the circle. The waves “y = sin (360t + 24)” and “y = 2 sin (360t + 24)” are in phase with each other.

Out of phase

If two waves are described as being “out of phase”, it means that they have different phases. For example, the waves “y = sin (360t + 10)” and “y = sin (360t + 11)” are out of phase with each other.

In-phase

The hyphenated word “in-phase” is the adjective form of the phrase “in phase”. We might say, “I really like in-phase waves,” or “these two waves are in-phase waves”.

The term “in-phase” is commonly used as another way of referring to the x-axis of a circle chart on which waves of various phases are being portrayed. In such a case, it can refer to a phase of zero degrees, or it can refer to a phase with which all the other phases are being compared. I explain what this means later in this chapter.

Quadrature

In the sense used with waves, the word “quadrature” refers to two items that are at 90 degrees to each other, or 90 degrees apart. The Sine and Cosine waves derived from a circle are quadrature waves in that their corresponding points occur 90 degrees apart. If we have two waves with phases that are 90 degrees apart, we could describe them by saying, “These two waves are in quadrature.”

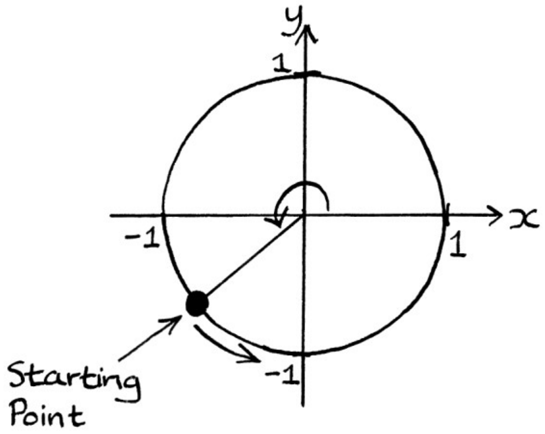
The term can also be used to say that one wave has a phase 90 degrees higher than another. In such a case, we might say, “A Cosine wave is the quadrature wave to a Sine wave,” or, “This wave is the quadrature wave to this reference wave.”

The word “quadrature” ultimately derives from the Latin word “quadrare” meaning “to make something square”. The word’s relevance here is due to the way that two lines that are 90 degrees to each other form a square’s corner. In the circle chart, the axes are at right angles to each other. An object starting at 90 degrees (on the y-axis) will be in quadrature to one starting at 0 degrees (on the x-axis).

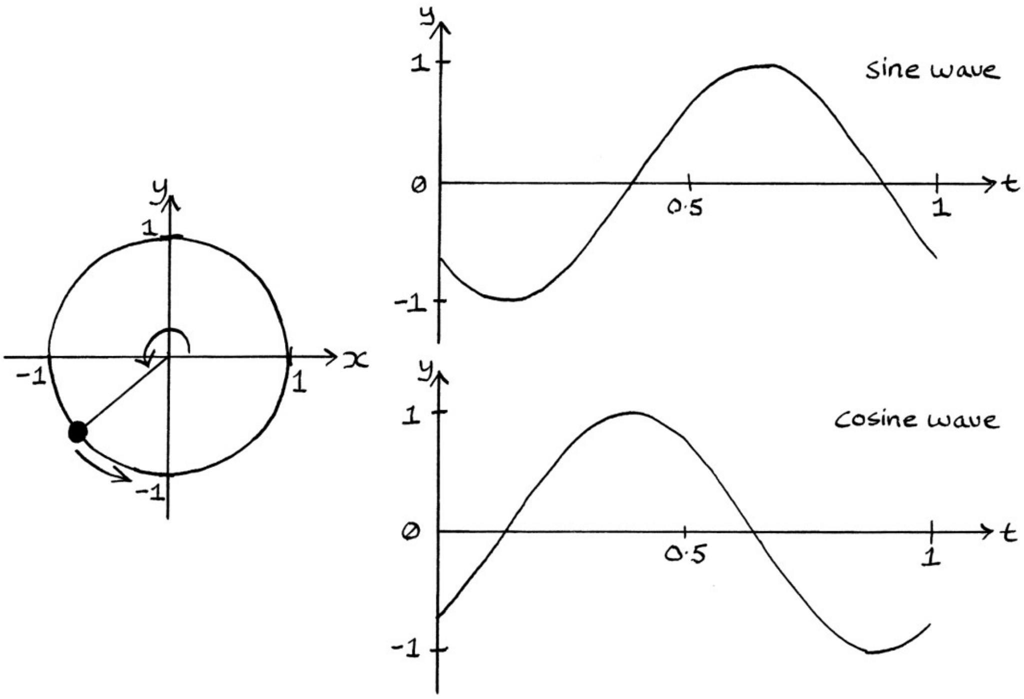
The term “quadrature” is commonly used to refer to the y-axis of a circle chart on which various phases are being portrayed, in which case, it can refer to a phase of +90 degrees, or to a phase that is +90 degrees higher than whatever phase the In-Phase axis (x-axis) is referencing. This will make more sense later in this chapter.

Distinguishing phases on the circle

So far, to mark the phase on the circle, we have placed a dot at the starting point of an object about to move around the circle. In other words, we have marked the angle of the object at t = 0:



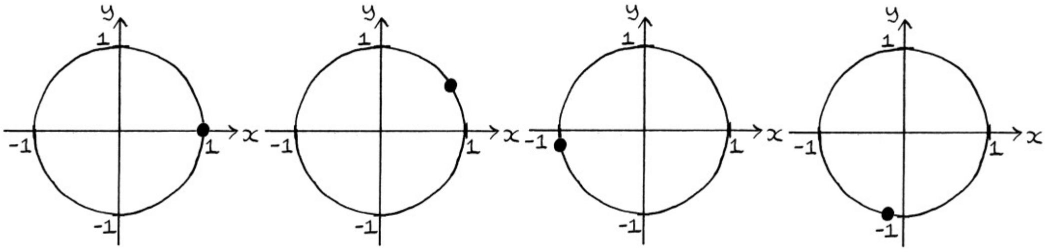
This is as if we froze the circle when t = 0. Plotting the phase means that the circle has all the information of both a Sine and Cosine wave for the moving object (except frequency), as it has always done, but it also has the phase of both too. The Sine wave is based on the y-axis values of the object as it moves around the circle, starting at the marked phase point, and the Cosine wave is based on the x-axis values of the object as it moves around the circle, starting at the marked phase point.



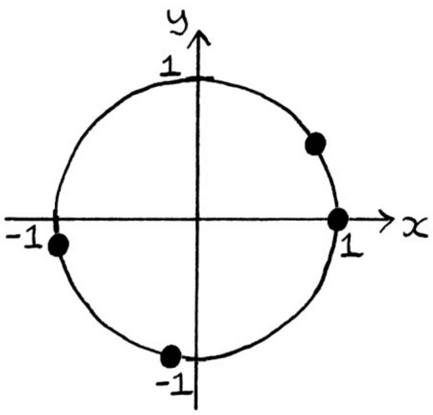
When dealing with several waves of the same frequency and amplitude but with different phases, an easy way to visualise and distinguish them is to mark the angles of their phases around the same circle all at the same time. As before, we can imagine that the plotted positions are the “starting points” of objects about to go around the circle – some objects have head starts on others by being further around at the beginning. In this sense, we can think of this drawing as being frozen in time. The circle is frozen at exactly t = 0 just before the objects start moving. For now, I will call this diagram, *as I am going to draw it,* a “combined phase circle” because it is a circle that combines all the phases of different waves.

As an example, we will consider four wave pairs with different phases. These represent the movement of four different objects moving around four different circles.
“y = sin 360t” and “y = cos 360t”
“y = sin (360t + 33) and “y = cos (360t + 33)”
“y = sin (360t + 190) and “y = cos (360t + 190)”
“y = sin (360t + 260) and “y = cos (360t + 260)”

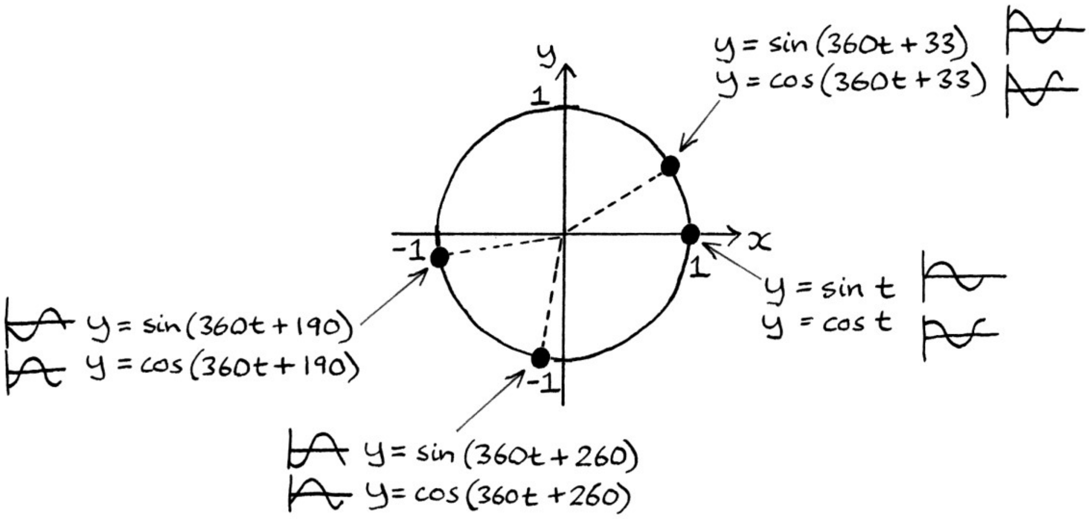
The four circles from which these waves are derived look like this:



We can put all the objects on to one single circle, and we end up with this:



To make the circle clearer, we can include the formulas of the waves that derive from each of those objects, and also draw the first cycle of each wave:



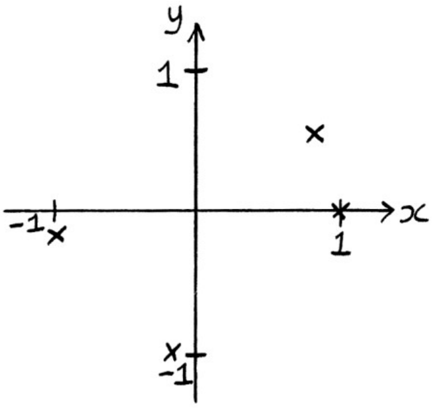
We now have the phases of each object marked around our completed combined phase circle. The reason for doing this is that we now have a better visual idea of how the waves are going to behave. We can see which waves start ahead of the others, and by how much. The little wave pictures help simplify everything too. The waves represented by this combined phase circle will maintain the same difference in phase throughout their existence. This is another way of saying that the objects rotating around the four initial circles will keep the same angular distance away from each other as they rotate. The objects will never catch up with each other.

The combined phase circle is a handy tool to compare the phases of different objects rotating around circles. Generally, such a diagram would not be used, but instead a much more basic diagram called a "constellation diagram".

Constellation diagram

What I am calling a "combined phase circle" is the underlying principle behind what is generally called a "constellation diagram" or a "constellation plot". A constellation diagram is often used in processing radio waves because it gives a way of visualising waves with different phases (but the same frequency) at the same time.

If we were to draw our combined phase circle as a proper constellation diagram, it would cease being a circle, and would end up as just the phase points marked on the graph:



Although this diagram looks a lot more basic than the combined phase circle, the information it portrays is still there. We can see that there are 4 points – each point represents the phase point (in other words, the starting point) of an object with a different phase. All the points represent objects that have the same frequency, and for the sake of this example, we will presume that we know that the frequency is 1 cycle per second.

From just knowing the position of a particular phase point, we can work out the derived Sine and Cosine waves that it represents. The phase of a Sine and Cosine wave pair is calculated by measuring the angle of a phase point from the origin of the axes. The amplitude of the waves is calculated by measuring how far away the phase point is from the origin of the axes. In this case, all the amplitudes are 1 unit.

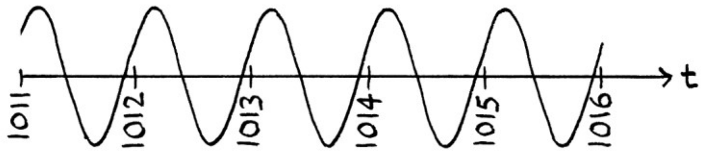
Whereas the combined phase circle spells out everything simply, a constellation diagram requires more thought to analyse. On the other hand, a constellation diagram is a lot quicker to draw. In everyday life, you might see a constellation diagram on a screen on a radio receiver, in signal processing software, or written out by hand to illustrate an idea. While my example shows the objects' positions at t = 0, a constellation diagram on a radio receiver will be showing the relative phase differences at a particular moment in time, which might vary depending on the situation.

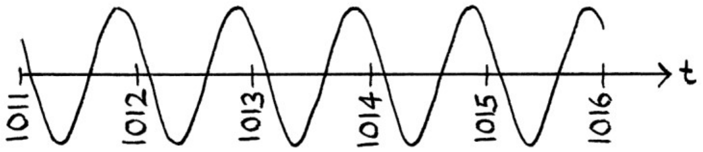
Constellation diagrams have several uses:

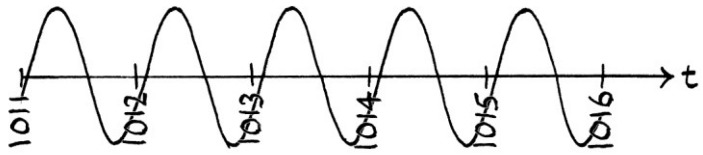
- They help visualise the phases of different waves.
- They can help in setting up radio transmitting and receiving equipment.
- They can help in matching up the phases of received radio waves – if we are receiving two identical signals and want to combine the power of each, we can move the antennas until the phases are identical. Viewing the received signals in real time on a constellation diagram display makes this a lot easier.
- Conversely, if we are receiving one signal on two different radio receivers, we can use a real-time constellation diagram display to watch how the phases change as the antennas are moved. This can be used to work out where the signal is coming from. Two receivers that are at slightly different distances from one transmitter will receive a signal with a different perceived phase.
- Constellation diagrams help indicate which possible phases are going to be received, and whether they are being received correctly, when receiving communications encoded with variations of phase. This is examined in the explanation on phase shift keying in a later chapter.

In real-world situations, a wave probably will not be observed at the very moment it starts, and therefore the initial phase in the formula for a wave will not be known. A constellation diagram can still be used in these situations, but the points around it will not be comparing with a Sine or Cosine wave with zero phase, but instead will compare with the phase of an arbitrarily chosen reference wave.

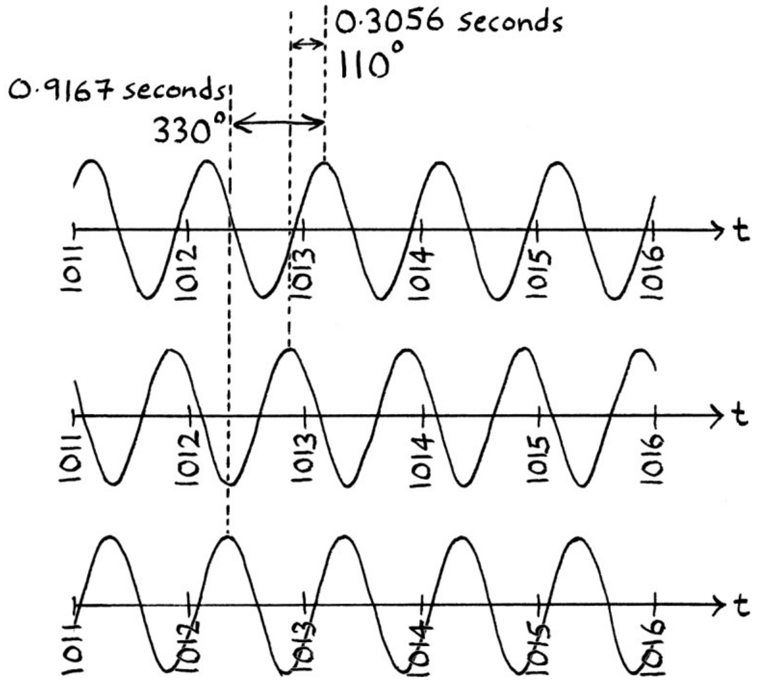
As an example, we will say that we have the following three waves, all with the same amplitude and frequency, plotted on a graph where the time axis refers to the time since we started observing them:



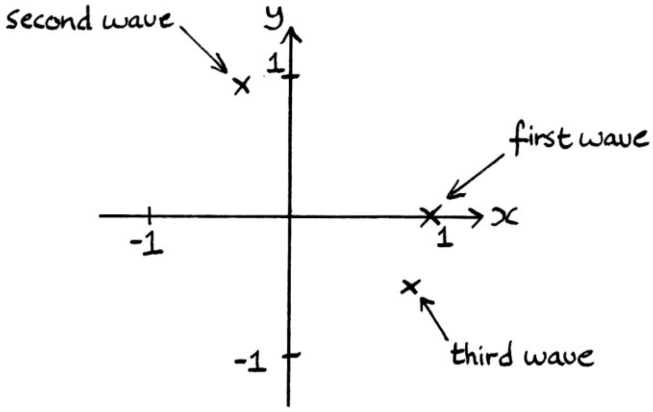




Although it is not marked, the amplitude of each wave is 1 unit. The frequency of each wave is 1 cycle per second. We will presume we can read the graphs with great accuracy. The second wave precedes the first wave by 0.3056 seconds, which means it has a phase that is 360 * 0.3056 ≈ 110 degrees greater than the first wave. The third wave precedes the first wave by 0.9167 seconds, which means it has a phase that is 360 * 0.9167 ≈ 330 degrees greater than the first wave.



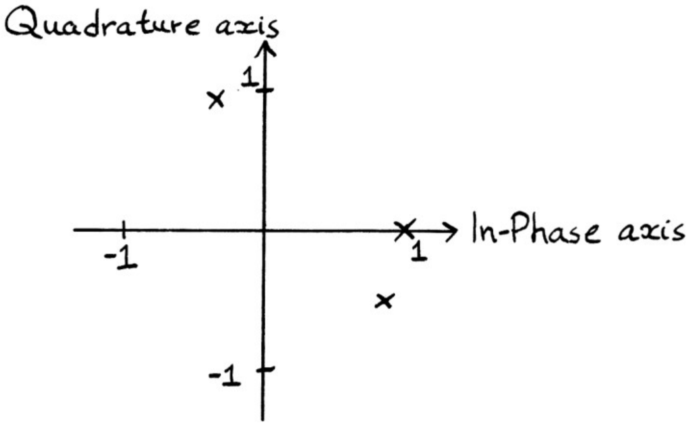
We can plot the phases all on a constellation diagram, but because we do not know the literal phase in their formulas, we will plot the phase angles relative to the first wave. In other words, we will pretend that the first wave has a phase of zero degrees, and then say that the phases of the second and third waves are equal to the phase difference between their waves and the first wave. Therefore, we plot the first wave at 0 degrees, the second wave at 110 degrees, and the third wave at 330 degrees.



Although we are plotting the phase points on the constellation diagram based on individual waves, the points really indicate both the Sine and Cosine waves for those phases. In this example, we do not know if the received waves are Sine waves or Cosine waves, but if we treat them as being all Sine waves or all Cosine waves, it does not make a difference when drawing the points. If we treat them as Sine waves, then we could infer the corresponding Cosine waves because they would have the same phase, amplitude and frequency, and if we treat them as Cosine waves, we could infer the corresponding Sine waves for the same reason.

Although we can label the axes "x" and "y", another way of labelling them refers to how the phases are thought of in relation to the reference wave: we label the x-axis, the "In-Phase" axis, and the y-axis, the "Quadrature" axis. The reference wave sits on the In-Phase axis – if there were any other waves on this axis, it would mean that they were "in phase" with the reference wave – in other words, they would have exactly the same phase. If there were a wave that had a phase +90 degrees higher than the reference wave, it would sit on the Quadrature axis. It would be in quadrature with the reference wave.

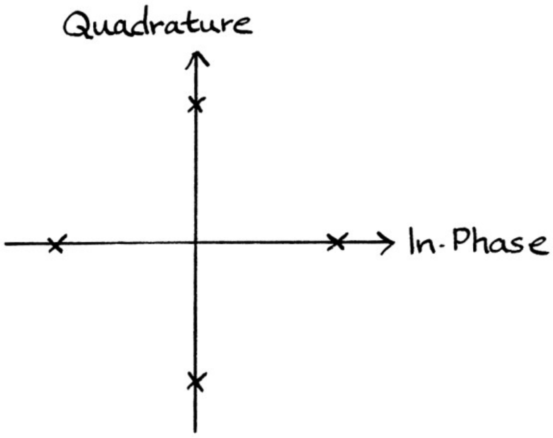
The previous constellation diagram redrawn with the axes relabelled looks the same but has different axis labels:



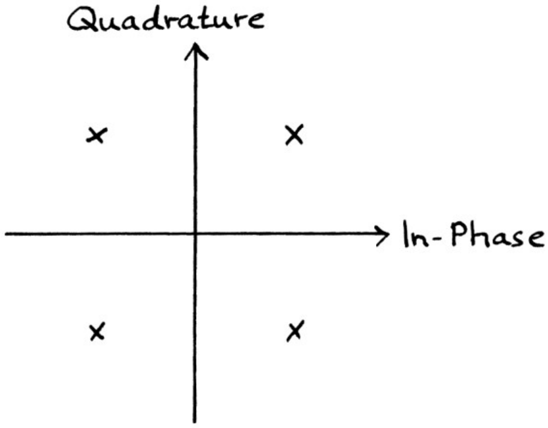
Given that we do not know the actual phases of the waves, and that we are treating the phases in reference to a particular wave, using the names In-Phase and Quadrature for the axes is appropriate. The actual phase of the reference wave probably is not zero, but we are treating it as if it were. When receiving real-world waves such as radio and sound waves, it is unlikely that we will know the actual phase of the waves, and therefore, all we can do is deal with the relative phases.

Labelling the axes "In-Phase" and "Quadrature" is not necessarily better than labelling them "x" and "y", but it reinforces what the purpose of the graph is, and how the phases are being portrayed in relation to a particular reference phase. The axis labels "In-Phase" and "Quadrature" are often abbreviated to "I" and "Q" to save space.

The reference wave does not actually have to be one of the plotted waves – it can be any chosen reference that is useful for plotting the waves. For the purposes of dealing with waves for which we do not know the formulas, it is the *difference* in phase that is important. In such cases, a constellation diagram such as this...



... could also be drawn in this way:



The difference in the phases is the same in each chart. Of course, if the *actual* phase values are important in whatever we are doing, for example, if we were using the constellation diagram as a reference to draw actual Sine and Cosine waves for which we know the formulas, then moving the points around changes the meaning.

Phase and amplitude on the constellation diagram

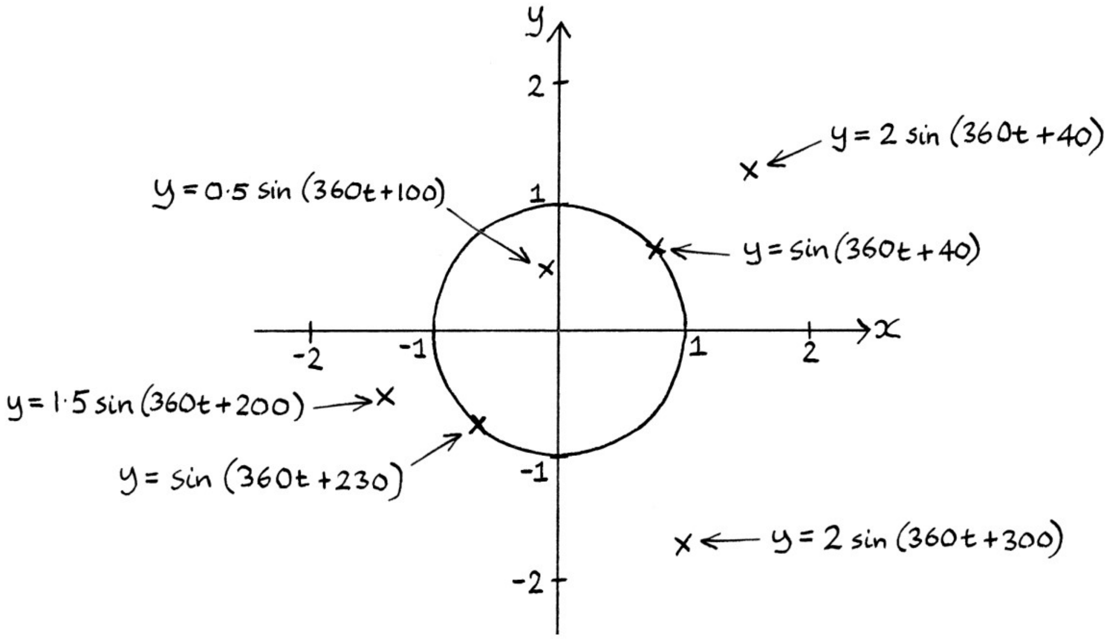
The constellation diagram and the combined phase circle only really make sense when dealing with waves of the same frequency. If the waves have a different frequency, the diagrams become less useful – we could still mark the starting points of the waves, but at any time after t = 0, the phase differences would have changed. If waves all have the same frequency, the waves on the constellation diagram and combined phase circle will keep the same phase difference throughout their existence.

The constellation diagram and the combined phase circle, however, do work with waves of different *amplitudes*.

For example, if we have six waves with differing phases and amplitudes as so:

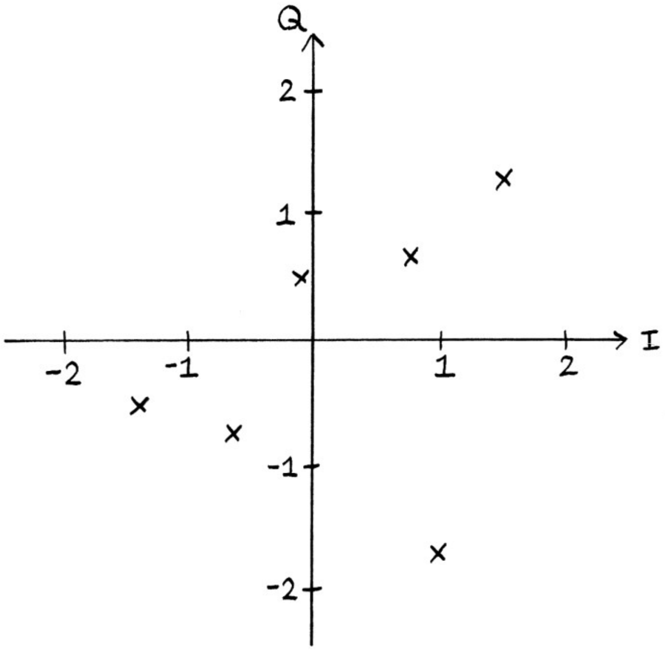
y = sin (360t + 40)
y = 2 sin (360t + 40)
y = 0.5 sin (360t + 100)
y = 1.5 sin (360t + 200)
y = sin (360t + 230)
y = 2 sin (360t + 300)

... then our combined phase circle would look like this (with the unit circle drawn in place).



The first thing to note is that drawing the unit circle becomes less useful when the waves do not have amplitudes of one unit. The second thing to note is that we are plotting phase points using just Sine waves, but each point also indicates the corresponding Cosine wave. This is because the phase, amplitude and frequency will be the same for both the Sine and Cosine wave derived from a circle.

The constellation diagram for these waves looks like this:

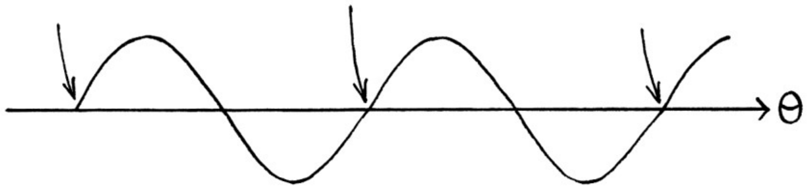


Calculating the phase of a wave

If we are given a wave, it is reasonably straightforward to calculate its phase. Here we will look at some methods to find the phase of a *Sine* wave that is centred at y = 0 on the graph.

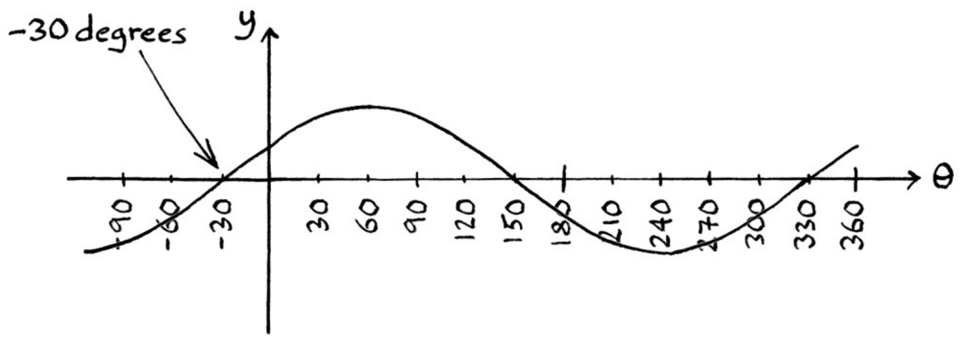
Sine waves based on angles

For a Sine wave based on angles and not time, we find a place where the curve's y-axis value is zero and about to rise, as illustrated in this picture:



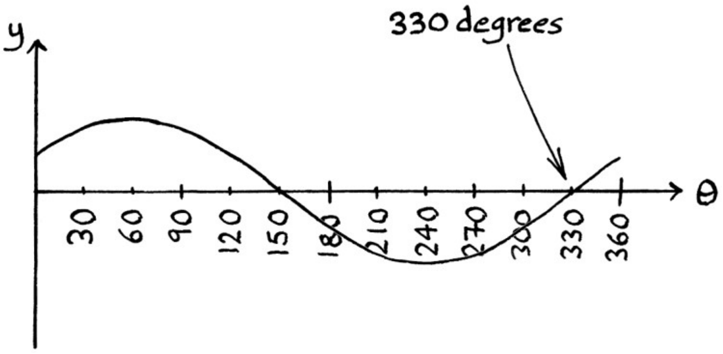
The phase of the Sine wave will always be the negative of the angle at one of those points. However, we might need to adjust that angle to make it into an angle between zero and 360 by adding or subtracting 360 degrees one or more times. To reduce the amount of adjusting, ideally, we take a reading from where the y-axis is zero and rising closest to θ = 0. It does not matter if we read from the negative or positive side of θ = 0.

If the graph is drawn with negative values of θ, then we can look in the negative half of the axis for the angle where the y-axis value of the curve is 0 and about to rise. [Ideally, we pick the place that this happens that is closest to θ = 0]. The phase of the wave will be the negative of the angle at that point. For example, in the following graph, the curve is at y = 0 and about to rise when "θ" is −30 degrees. The phase of the wave is, therefore, the negative of that, which is +30 degrees.



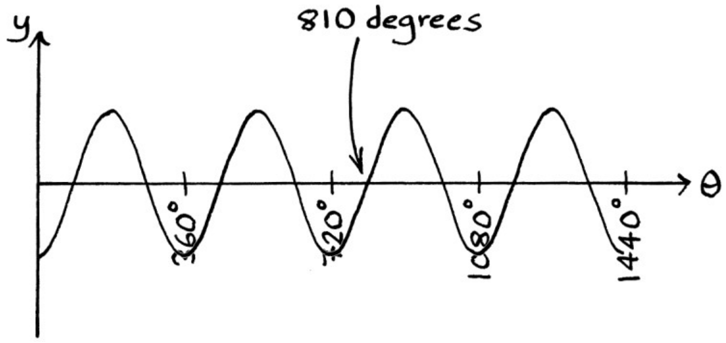
[Note how I have not written any numbering along the y-axis. This is because the details of the y-axis are irrelevant to the phase of the wave, except for where y = 0.]

If the graph is only drawn with positive values of θ, then we look at where the curve's y-axis value is 0 and about to rise in the positive half of the axes. [Again, ideally, we pick the place where this happens that is closest to θ = 0.] The phase of the wave will be the negative of that angle. For example, in the following picture, the relevant point is at θ = 330 degrees. Therefore, the phase of the wave is −330 degrees. We could leave the phase as a negative angle if we wanted, but here, we will say that we want a positive phase. Therefore, we convert −330 into a positive angle that means the same thing, by adding 360 to it, and we end up with:
–330 + 360 = 30 degrees.



The two graphs above actually refer to the same wave with the same phase. We could also have calculated the phase from the first wave by reading the angle at 330 degrees, but it is simpler to negate −30 degrees, than it is to negate +330 and then have to turn it into a positive phase by adding 360 degrees to it.

If we do not take the first place where the y-axis value is zero and about to rise (on either side of θ = 0), then our calculated phase will require more work to turn it into a value between 0 and 360 degrees. For example, in the following angle-based wave, we will read the angle where the curve is zero and rising at 810 degrees.



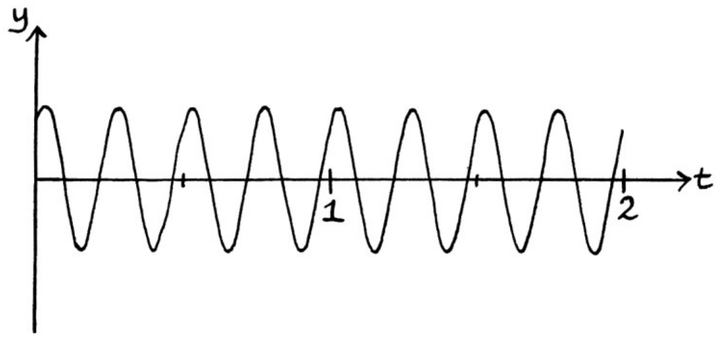
The negative of this is −810 degrees. Therefore, the phase of this wave is −810 degrees. This is an adequate answer, but we will say that we want a positive angle between zero and 360. Therefore, we keep adding 360 degrees to −810 until we have an angle between 0 and just under 360 degrees:
−810 + 360 = −450
Then: −450 + 360 = −90
Then: −90 + 360 = 270.

Therefore, the phase of this wave is 270 degrees.

Sine waves based on time: method 1

If we have a wave based on time, then we have to calculate the frequency first, then we can work out the angles of the curve, and then we can work out the phase of the wave.

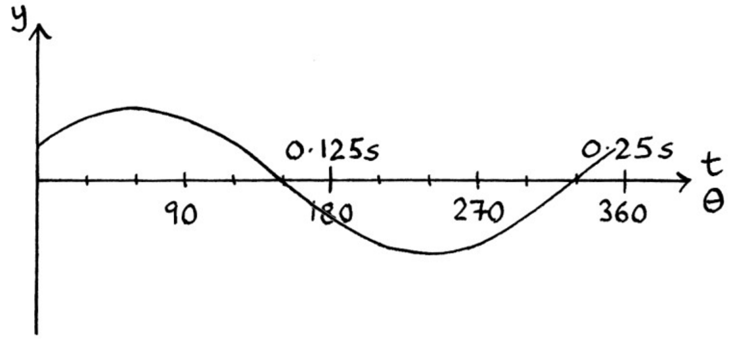
As an example, we will look at the following wave:



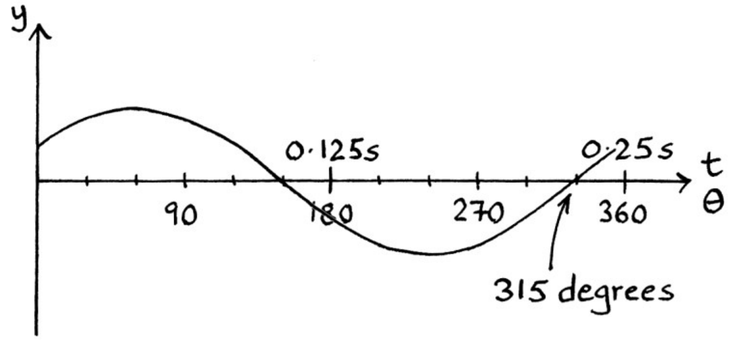
By counting how many cycles there are within a second, we can tell that the frequency of this wave is 4 cycles per second. As an easy way of visualising what this means, if we think of an object rotating around a circle at this rate, it would complete 4 revolutions in one second, which is the same as 1 revolution in a quarter of a second. In other words, it would complete 360 degrees in quarter of a second. We can also say that the *wave* completes 360 degrees in quarter of a second, but without thinking of a circle, this can be slightly harder to visualise.

If the wave completes 360 degrees in quarter of a second, then we just need to divide a quarter of a second into 360 pieces, and then we can read the angle of the wave for moments of time. [In practice, this might require a large graph to read the values accurately].

To do this, we will zoom into one quarter second of the graph, and divide it up into 360 pieces. I have drawn the zoomed-in graph with the x-axis showing both the time and the angle of the object rotating around the circle from which the wave is derived at each moment in time:



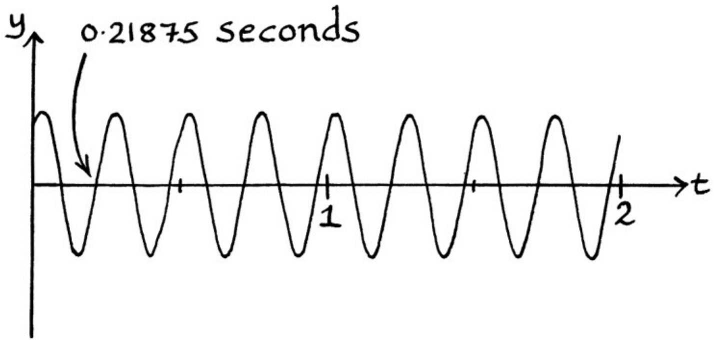
As we now know the angles of the wave, we can make a reading from it. The curve has a y-axis value of 0 and is rising when the angle is 315 degrees:



Therefore, the phase of the wave is −315 degrees, which is also −315 + 360 = 45 degrees.

Sine waves based on time: method 2

Another way we could have done the above is to see at what *time* the wave has a y-axis value of zero and is rising, and then calculate the angle for that time. The first moment that the wave has a y-axis value of zero and is rising is at 0.21875 seconds. (We will pretend that we can read the graph very accurately.)



If the time of interest is 0.21875 seconds, and the wave completes 4 cycles per second, then we know that 0.21875 seconds is 0.21875 * 4 = 0.8750 of the way through one cycle. If it is 0.8750 of the way through one cycle, then it is 0.8750 * 360 = 315 degrees of the way through 360 degrees. The angle for this time is, therefore, 315 degrees, and the phase of the wave is, therefore, −315 degrees, which is also 45 degrees.

A summary of this procedure is:

- Read the first time (in seconds) when the wave has a y-axis value of zero and the curve is rising.
- Calculate the frequency of the wave by counting how many cycles there are in one second.
- Multiply the time by the frequency, and the result of that by 360 degrees.
- The phase will be the negative of that result.
- (If needed, add or subtract 360 degrees until that number is between 0 and just under 360 degrees).

To put this slightly more mathematically:
“phase = – (time reading * frequency * 360)”

Cosine waves

Calculating the phase of a Cosine wave is similar to calculating the phase of a Sine wave. Instead of looking at the place where the curve is at y = 0 and rising, we look for where the curve is at its maximum and about to fall. We then negate the angle at that place. Alternatively, we can use the methods for Sine waves and subtract 90 degrees afterwards.

Potential sources of confusion

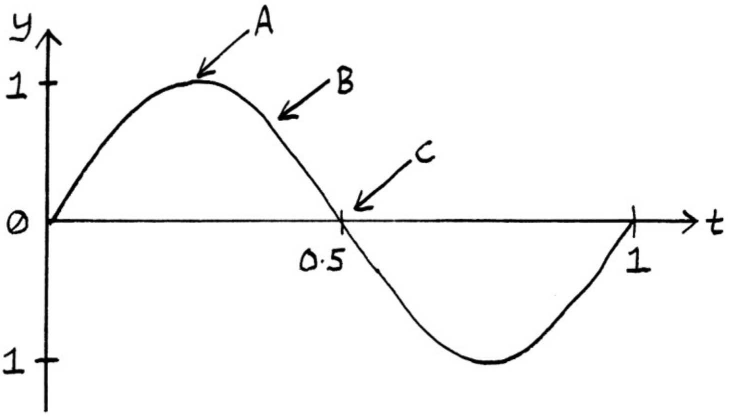
The two meanings of phase

As with amplitude, the term “phase” is often used to mean two slightly different things.

In this chapter, when I refer to “phase”, if I am talking about Sine waves, I mean the “phase difference” in relation to a Sine wave with the formula “y = sin 360t”, and if I am talking about Cosine waves, I mean the “phase difference” in relation to a Cosine wave with the formula “y = cos 360t”. Another way of thinking about this is that I am referring to the starting angle of an object about to rotate around a circle in comparison to if it were going to start at 0 degrees.

However, some people will use the word “phase” as another word for “angle”. When they say “phase”, they are not referring to the overall phase difference (the starting point of an object about to rotate around a circle), but to the specific angle at a particular time. In other words, they are referring to the angle of the object on the circle at a particular moment in time. What makes this slightly more complicated is that most of the time the term used in this sense will be directed at a point on a wave graph, with no mention of the circle from which the wave was, or could have been, derived.

In the following graph, the object rotating around the circle would be at 90 degrees at point A, it would be at 135 degrees at point B, and it would be at 180 degrees at point C:



Another way of saying all this is that the "instantaneous angle" of the *wave* is 90 degrees at point A, 135 degrees at point B, and 180 degrees at point C. Where things can become confusing is when people re-use the term "phase" to mean the "instantaneous angle" as well. Such people will say that the *phase* at point A is 90 degrees, the *phase* at point B is 135 degrees, and the *phase* at point C is 180 degrees. For them the terms "phase" and "angle" mean the same thing. This is obviously very confusing if you are not aware of this dual use of the word "phase".

Some people who use the term "phase" to mean the angle on the wave at any moment in time, or the angle of an object on the circle at any moment in time, will use the term "initial phase" to refer to the starting point of an object. In this way, what *I* call "phase", *they* call "initial phase", and what *they* call "phase", *I* would generally call "the angle at a particular moment in time".

If you insist on the word "phase" having two meanings, then you can remove any ambiguity by using language that is more precise. "Phase", in the sense that I have been using it in this chapter, could be called "initial phase", "starting phase", "overall phase", "overall phase difference" or "overall phase shift". "Phase", in the sense of "the angle of the object on the circle at a particular moment in time as represented on the wave graph" could be called "instantaneous phase", "instantaneous angle", or "the angle at this time". Personally, I think it is confusing to use "phase" on its own to mean anything other than the starting angle of an object. [I will not be using the term "instantaneous phase" until much later in this book.]

Different people have different ways of describing things, and all the rules for naming are socially constructed, so there is no true right or wrong use of the word "phase". Usually, there is a social consensus as to the best term to describe something, but when it comes to phase, there are really two opposing viewpoints.

As you become more used to waves and other people talking about waves, the context will usually indicate which meaning of the word "phase" is intended.

The ambiguous term "phase shift"

One of the most confusing things about phase is the way that people use ambiguous language when discussing "phase shifts". This careless word use will be the main cause of anyone becoming muddled by phase.

If *I* use the term "phase shift" when talking about the *formula* for a Sine or Cosine wave, it will refer to the phase element of that wave's formula. In other words, in this formula, "y = sin (360t + 45)", the phase shift is +45. The *positive* phase means that the wave is shifted to the *left* in relation to the same wave with zero phase ("y = sin 360t"). If the formula were "y = sin (360t − 30)", then the *negative* phase would mean that the wave is shifted to the *right* in comparison to the same wave with zero phase (also "y = sin 360t").

The confusing moment comes when some people use the term "phase shift" to refer to the sliding of waves on the *graph* without reference to the formula. It is common to hear someone say something such as, "This wave has had a positive phase shift," when what they mean is that the wave on the wave graph has been shifted to the *right* along the axis in relation to some other wave. It has been shifted up in the *positive* direction of the θ-axis or time axis. This is, of course, the opposite of how a positive phase shift works in a wave formula. Such people will also say, "This wave has had a negative phase shift," and what they mean is that the wave has been slid to the *left* on the wave graph. It has been shifted down in the *negative* direction of the θ-axis or time axis. Again, this is the opposite of how a phase shift works in a formula. You can see why people might think of phase in this way, though, as they are not thinking of the formula or the circle. They are thinking of the positive and negative ends of the θ-axis or time axis.

To summarise the way they see things: a *positive* phase shift for them means that the wave has been slid to the right; a *negative* phase shift for them means that the wave has been slid to the left. This is the *opposite* way round to positive and negative phases used in the formula. It also ignores the idea of phase as portrayed on the circle.

This can be a huge source of confusion if you are not aware of it. Even if you are aware of it, it can still be sometimes impossible to know exactly what someone means. This ambiguity means that it is imperative to be specific about what you mean when you talk about phase.

In my opinion, you should never use the term "phase shift" without clarifying what you mean:

- If the phase shift is in the context of the formula, you should say so, and say whether it is positive or negative. [If the angle is 180 degrees, it does not matter as +180 degrees is the same as −180 degrees.]
- If the phase shift is not in the context of the formula, you should indicate the direction along the θ-axis or time axis that the phase shift has been in by using the words "left" or "right". [Again, if the angle is 180 degrees, this does not matter.]

Therefore, instead of saying something vague such as:
"this wave has a phase shift of 45 degrees"
... you should say:
"this wave's *formula* has a phase of +45 degrees"
... or:
"this wave's *formula* has a phase of −45 degrees"
... or:
"this wave has been shifted to the *right* by 45 degrees along the time axis"
... or:
"this wave has been shifted to the *left* by 45 degrees along the time axis"
... depending on what you mean.

Ideally, whenever you use the terms "phase" or "phase shift" you should remember that it is easy to say something ambiguous.

The term "phase difference", which is another way of saying "phase shift", suffers from the same types of ambiguity.

Summary of phase terminology confusions

- We can have “phase” in the formula for a wave. For example, “y = sin (360t + 12)”. In a formula, a positive phase means the wave is shifted to the left; a negative phase means the wave is shifted to the right. When thinking about circles, “phase” in this sense refers to the starting point of an object rotating around a circle.

- We can have “phase” meaning “instantaneous phase” meaning the angle on a wave at a particular moment in time, which, if thinking about circles, means the angle of the object on the circle at that particular moment in time.
- We can have a “phase shift”. Outside of speaking about a formula, this is often, but not always, *intended* to mean that a positive phase shift moves the wave to the right on the graph, while a negative phase shift moves it to the left on the graph. However, the term is ambiguous without clarification.

When you have more experience of waves, you will become better at understanding what people mean when they refer to phase, phase shift and phase difference. However, it is still sometimes impossible to know what someone means.

It pays to notice that careless use can be confusing. When I use any of these terms in this book, I will try to use them unambiguously.

Phase thoughts

Often in explanations of waves and signal processing, it seems that people are slightly afraid of phase. As you read more about waves, you will notice how some people needlessly try to avoid non-zero phases, and thus make things much more complicated for themselves. Often you will see people use Sine and Cosine waves as if they were really completely different entities, and not just the same thing with different phases. I am sure that one reason for the “fear of phase” is their being confused by the ambiguities of the language surrounding phase.

Related to the above paragraph, it pays to be aware that, frequently, people use the terms “Sine wave” and “Cosine wave” when what they really mean is “Sine wave with zero phase” and “Cosine wave with zero phase”. Sometimes, this is to save time; sometimes, it is because they have forgotten about phase altogether.

Formulas for phase

The general formula for a Sine wave relating to time that takes into account phase is:
$y = \sin(360t + \phi)$
... where "ϕ" is the lower-case Greek letter "Phi" used to represent phase. The letter "ϕ" is the Greek equivalent of the lower-case Latin letter "f" (as in "foxtrot") or a "ph" sound.

The Cosine equivalent is:
$y = \cos(360t + \phi)$

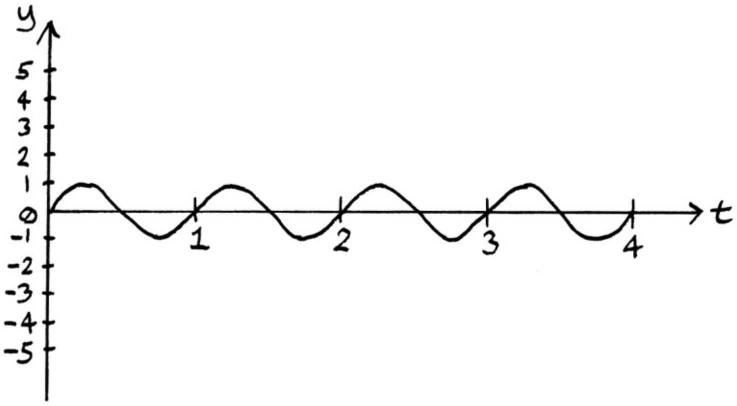
The same formulas without reference to time would be:
$y = \sin(\theta + \phi)$
$y = \cos(\theta + \phi)$

Chapter 8: Mean levels

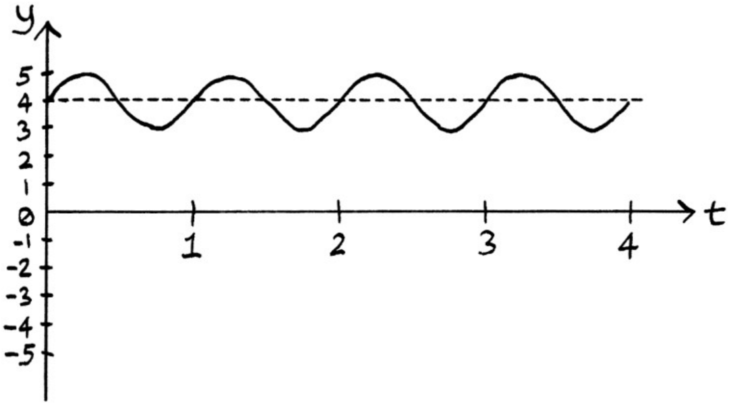
Mean levels

The mean level of a wave is less commonly thought about than the amplitude, frequency or phase. It is often the neglected property of a wave. However, it is still important to know about mean level to have a full understanding of waves.

Mean level is the y-axis value of the centre of the wave. It is literally the mean, as in the average, y-axis value of a wave. In a Sine wave or a Cosine wave derived from a circle that is centred on the origin of the x and y-axes, the mean level will be zero – the points on the waves fluctuate around zero, as shown in this "y = sin 360t" Sine wave:

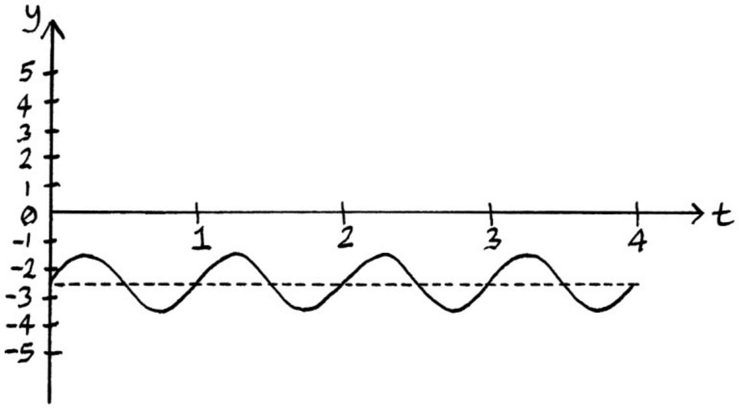


The mean level can be altered, thus shifting the whole wave up or down the y-axis. If we change the "y = sin 360t" formula to be "y = 4 + sin 360t", it will raise the Sine wave upwards by 4 units. Whatever the result of "sin 360t", it will have 4 units added to it afterwards, so it will always be 4 units higher up the y-axis than before. The average y-axis value and, therefore, the mean level of the wave will be 4 units.

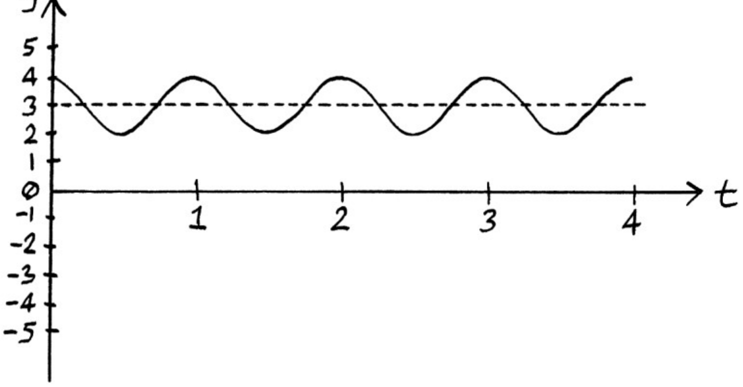


[Note how the formula:
"y = 4 + sin 360t"
... could also be written as:
"y = 4 + (sin 360t)"
... but usually the brackets would be left off.]

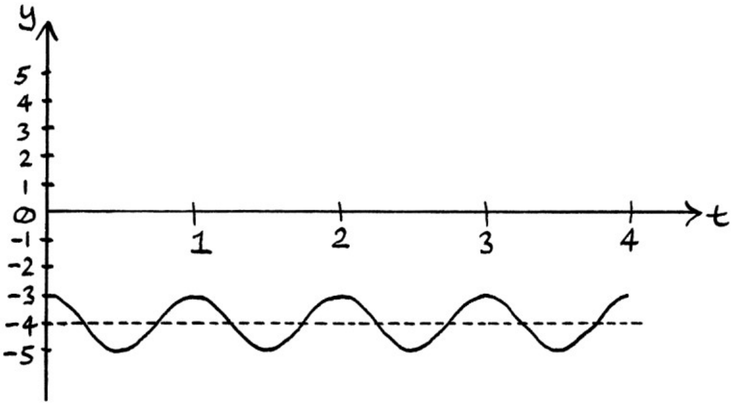
If we change our formula to be "y = −2.5 + (sin 360t)", which would more commonly be written as "y = −2.5 + sin 360t", it will lower the Sine wave. Whatever the result of "sin 360t", it will have 2.5 units removed from it afterwards, so it will always be 2.5 units lower than before. The mean level of the wave will be −2.5 units.



For Cosine, mean level has the same meaning as for Sine. If we change the "y = cos 360t" formula to be "y = 3 + cos 360t", it will raise the Cosine wave upwards by 3 units. Whatever the result of "cos 360t", it will have 3 added to it afterwards, so it will always be 3 units higher than before. The average y-axis value of the wave will be 3 units. The mean level of the wave will be 3 units.



If we change the formula to “y = −4 + cos 360t”, it will lower the Cosine wave downwards by 4 units compared with “y = cos 360t”. Whatever the result of “cos 360t”, it will have 4 units subtracted from it afterwards, so it will always be 4 units lower than before. The mean level of the wave is −4 units.



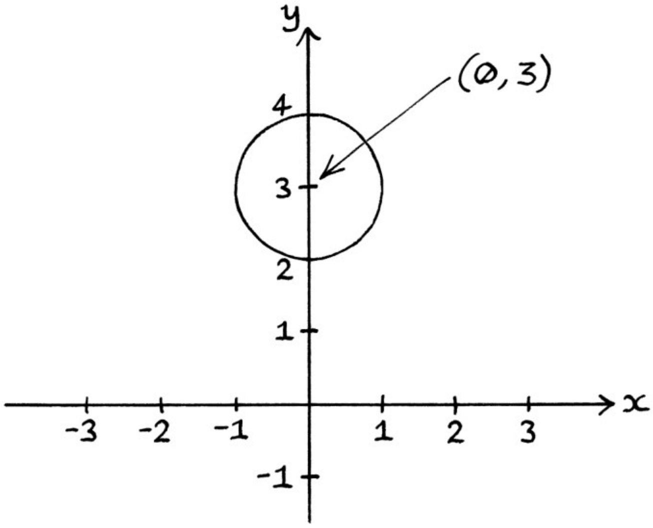
Mean level on the circle

The mean level of a wave is directly related to the position of the centre of the circle from which that wave is derived. When it comes to the circle, it might be apparent that there are really two mean levels. The circle can be shifted both vertically and horizontally around the circle chart. The vertical (y-axis) position of the centre of the circle will be the same as the mean level of the derived Sine wave, and the horizontal (x-axis) position of the centre of the circle will be the same as the mean level of the derived Cosine wave.

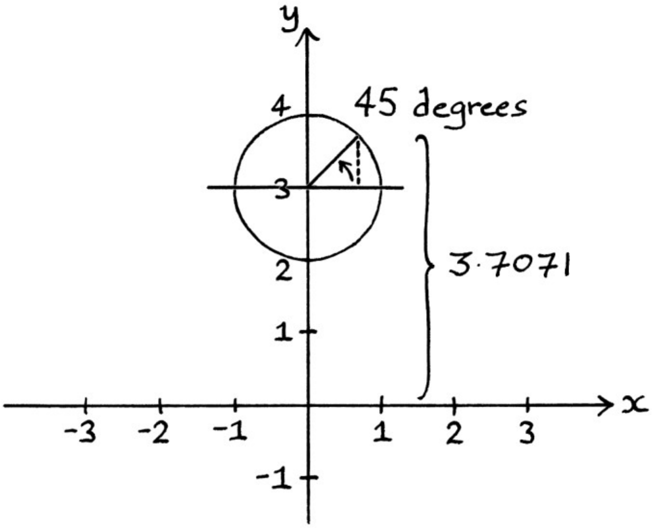
Vertically shifted circles

Normally, when we create an *angle*-based Sine wave from a circle, we measure the y-axis values of points around the circle’s edge at evenly spaced angles from its centre. When a circle is not centred on the origin of the axes, we still do exactly the same thing. The difference being that the resulting Sine wave graph will be raised up its own y-axis by the same amount that the centre of the circle is raised up *its* y-axis.

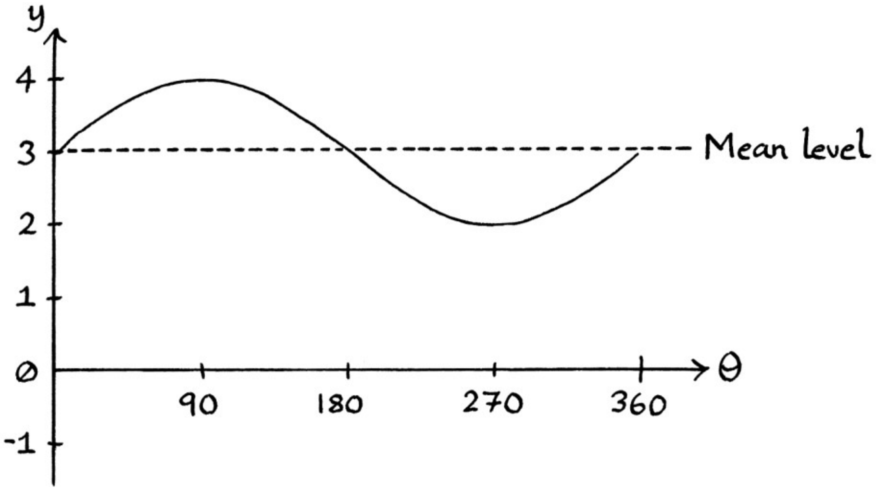
As an example, we will look at the following circle, which has a radius of 1 unit, and is centred at the coordinates (0, 3):



To obtain the derived Sine wave, we measure the y-axis values of points on the edge of this circle at evenly spaced angles from *the circle's* centre. For example, the y-axis value of the point on the circle's edge at 45 degrees is 3.7071 units:



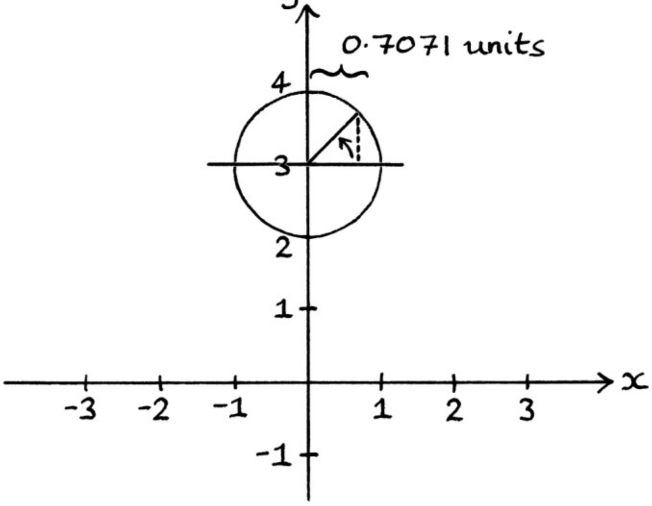
If we measure all the points around the circle, we will end up with this Sine wave graph, which has a mean level of 3 units:



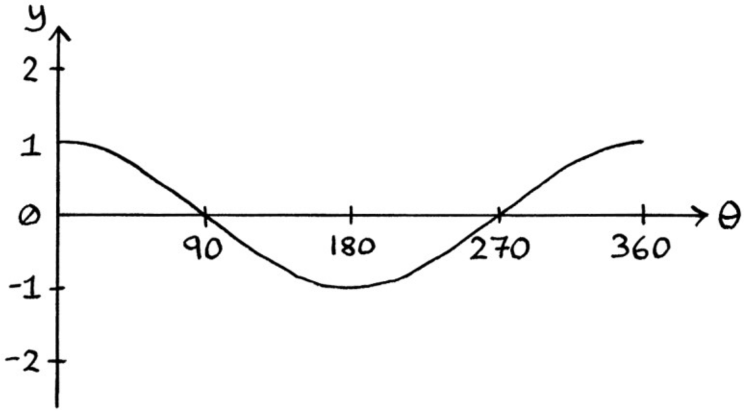
It is important to note that we measure the y-axis values of the points around the circle's edge at evenly spaced angles from *the circle's centre*, and not at evenly spaced angles from the *origin of the axes*. [The angles are measured from the centre of the circle, but the y-axis measurements are measured from the origin of the axes.]

The resulting Sine wave has the formula: "y = 3 + sin θ".

We can create an angle-based Cosine wave too. To do this we measure the x-axis values of points around the circle at evenly spaced angles from the circle's centre. For example, the point on the circle's edge at an angle of 45 degrees has an x-axis value of 0.7071 units:



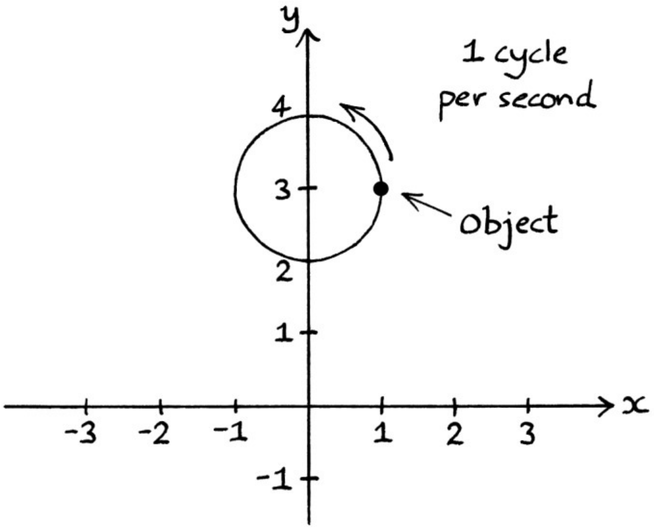
The angle-based Cosine wave derived from the circle looks like this:



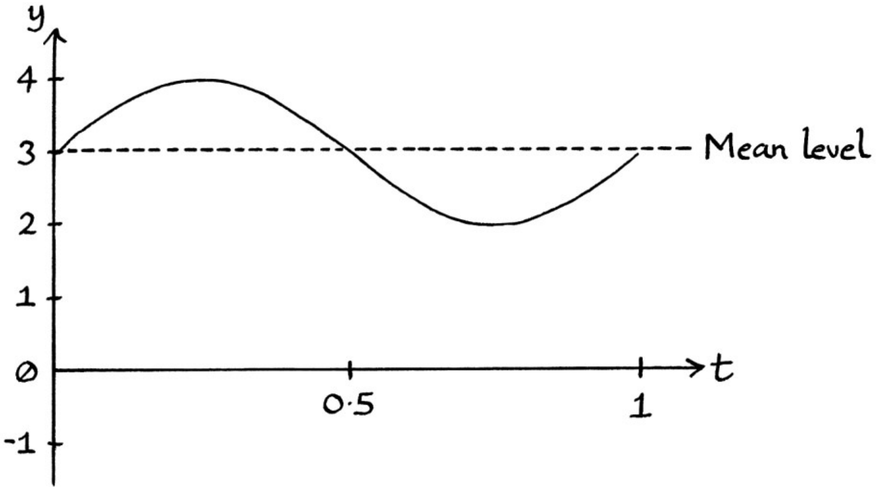
Its formula is "y = cos θ". Note how this wave has a mean level of zero units – it fluctuates around y = 0. This is because on the circle graph, the circle was still centred around x = 0. The position of the circle in this particular example has no effect on the mean level of the derived Cosine wave.

Time

We can also use circles that are not centred on the origin of the axes to create *time*-based Sine and Cosine waves. Normally, when we create a time-based Sine wave from an object rotating around a circle, we measure the y-axis value of the object at evenly spaced moments in time. For a circle that is not centred on the origin of the axes, we do exactly the same thing. Similarly, the time-based Cosine wave is based on the x-axis values of the object at evenly spaced moments in time. As an example, we will imagine an object rotating around the previous circle at a rate of 1 cycle per second.

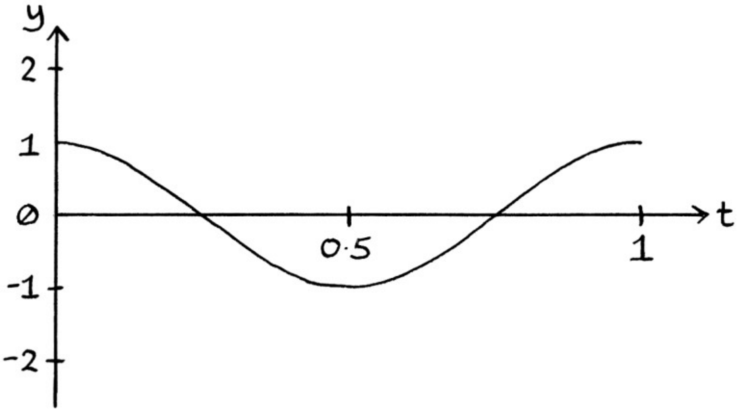


The Sine wave derived from its movement has the formula “y = 3 + sin 360t” and looks like this:



[It has exactly the same shape as the angle-based Sine wave for that circle.] The wave is centred around y = 3, which is the same as the y-axis value of the centre of the circle.

The Cosine wave has the formula “y = cos 360t” and looks like this:

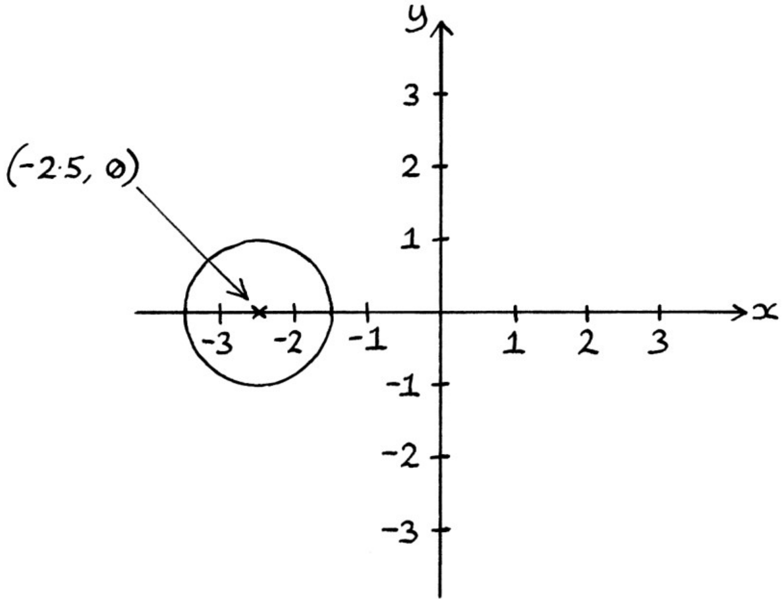


[It, too, has exactly the same shape as the angle-based Cosine wave.] The Cosine wave is centred around y = 0, which is the same as the *x-axis* value of the centre of the circle.

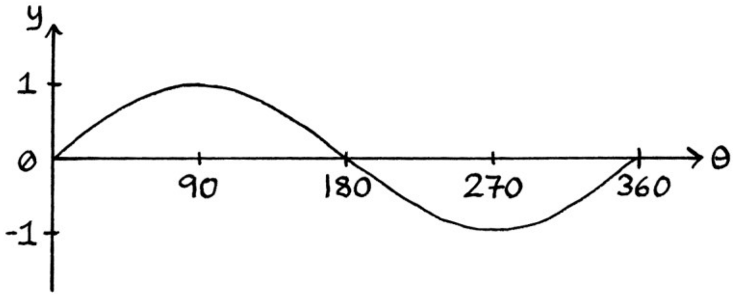
Horizontally shifted circles

If the circle has its centre to the right or left of x = 0, the derived Sine and Cosine waves are still created in the same way.

As an example, we will look at this circle, which has its centre at (−2.5, 0):

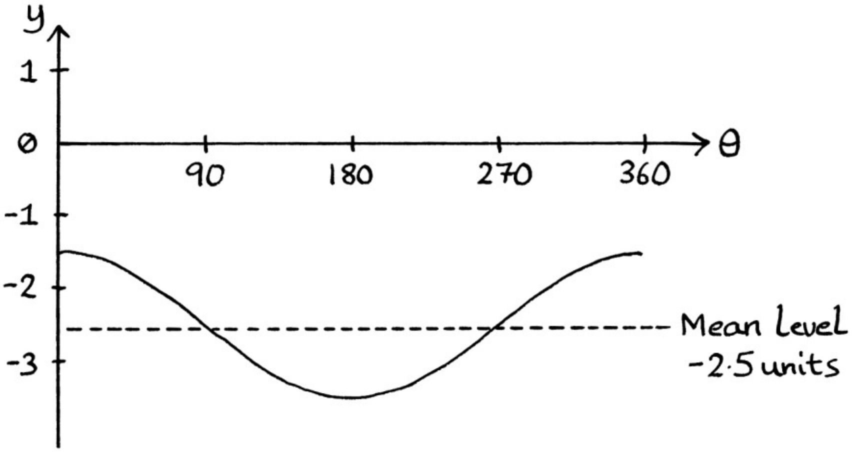


The angle-based Sine wave for this circle shows the *y-axis* values of points on the circle's edge at evenly spaced angles from its centre. It looks like this:



It has a mean level of zero units because the circle's centre has a y-axis of zero units. Its formula is "y = sin θ".

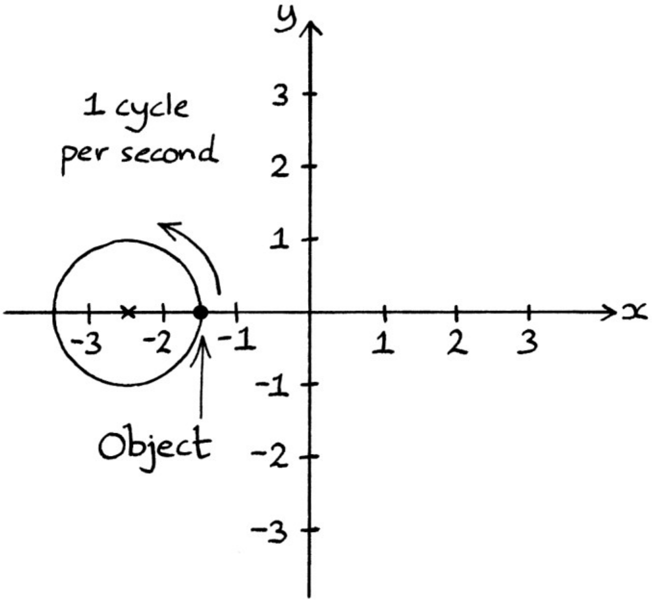
The angle-based Cosine wave shows the *x-axis* values of points around the circle's edge at evenly spaced angles from its centre. It looks like this:



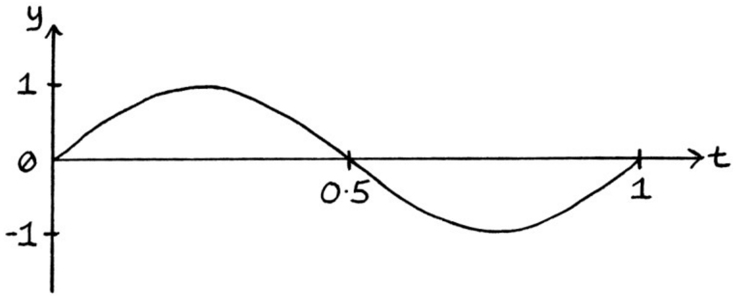
It has a mean level of −2.5 units because the circle's centre has an *x-axis* value of −2.5 units. The wave's formula is "y = −2.5 + cos θ".

Time

If we imagine at object rotating around the previous circle at 1 cycle per second, we can have time-based Sine and Cosine waves that portray its position over time.

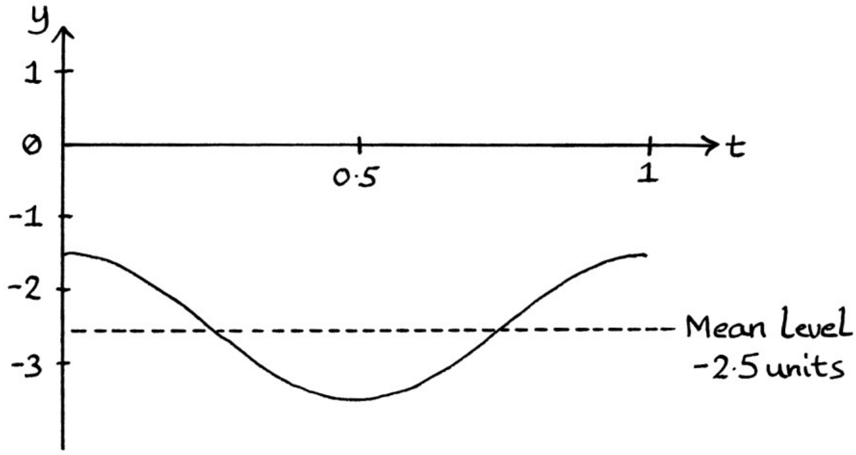


The time-based Sine wave shows the object's *y-axis* position at equally spaced moments in time. It looks like this:



It has a mean level of zero units. Its formula is "y = sin 360t".

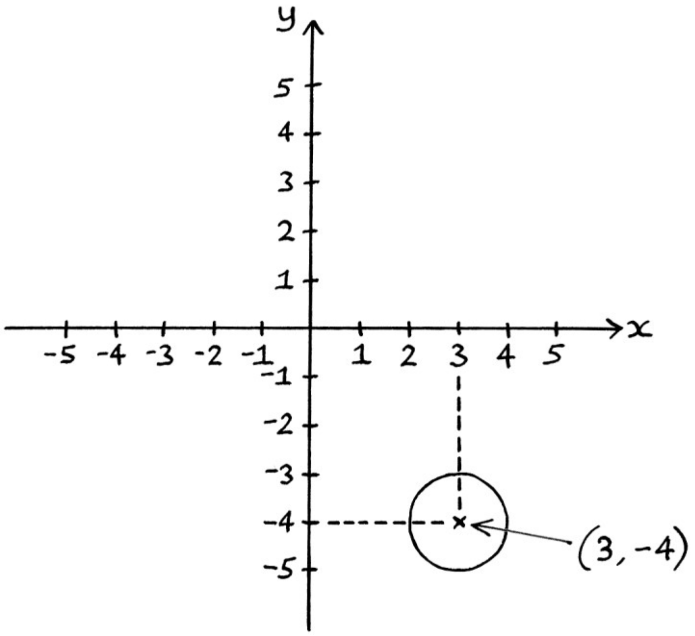
The time-based Cosine wave shows the object's *x-axis* position at equally spaced moments in time. It looks like this:



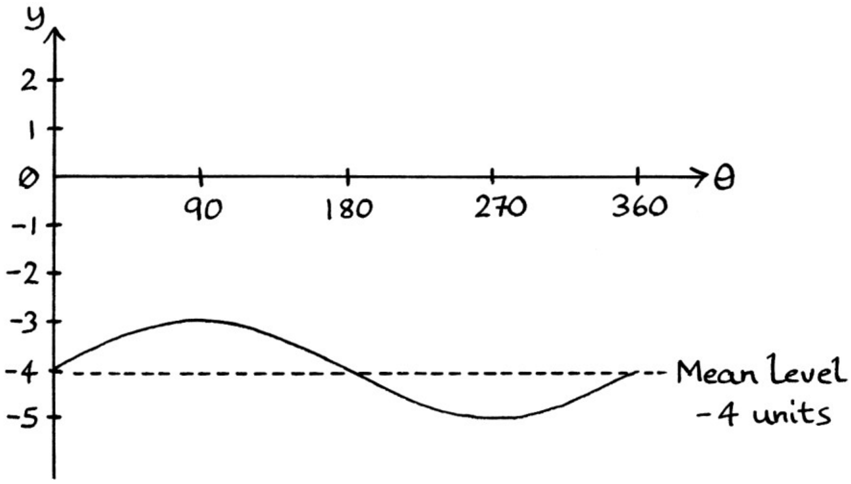
It has a mean level of −2.5 units. Its formula is "y = −2.5 + cos 360t".

Circles anywhere on the axes

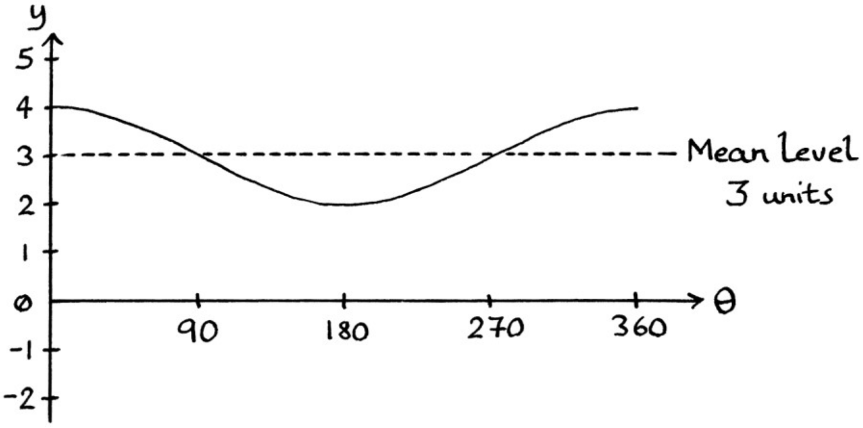
As we have seen, if the circle is centred over the y-axis, the derived Cosine wave will have a zero mean level. If the circle is centred over the x-axis, the derived Sine wave will have a zero mean level. It is, of course, possible to have a circle centred away from each axis. As an example, we will look at the following circle, which is centred at the coordinates (3, −4):



The angle-based Sine wave will show the y-axis values of points around the circle's edge that are at evenly spaced angles from its centre. Its formula is "y = −4 + sin θ", and it looks like this:



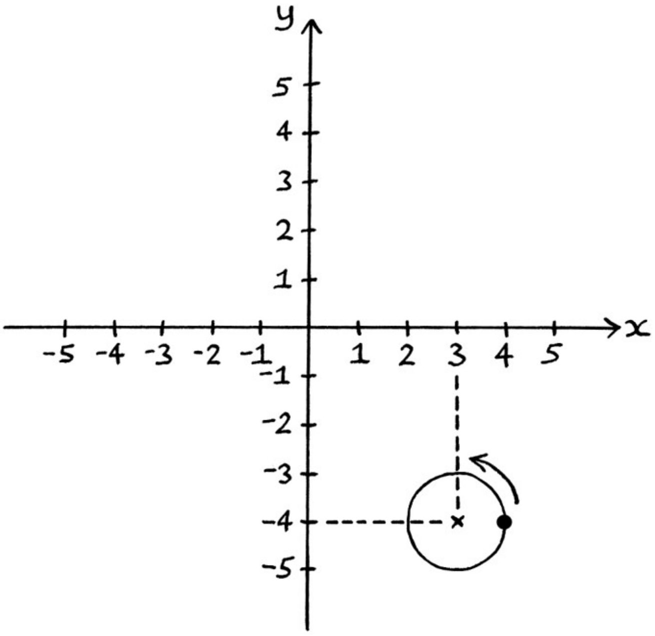
The angle-based Cosine wave shows the x-axis values of points around the circle's edge that are at evenly spaced angles from its centre. Its formula is "y = 3 + cos θ", and it looks like this:



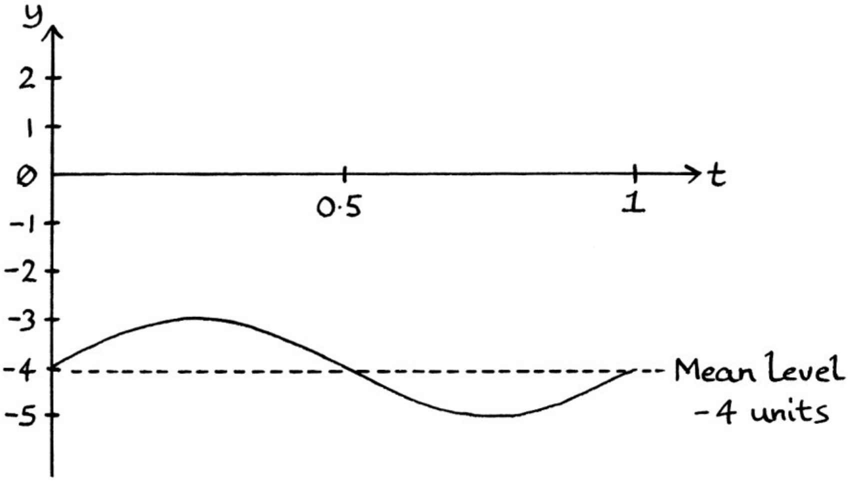
Note how the two derived angle-based waves both have non-zero mean levels. The mean level of the Sine wave is the same as the y-axis coordinate of the centre of the circle. The mean level of the Cosine wave is the same as the x-axis coordinate of the centre of the circle.

Time

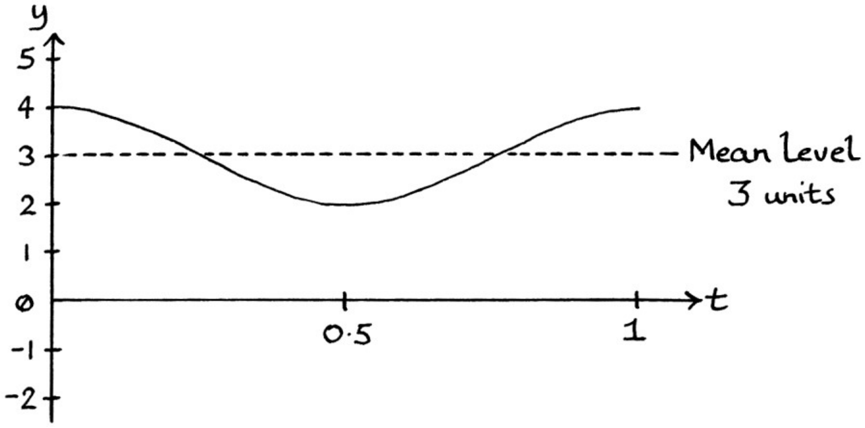
If we draw an object rotating around the above circle at a rate of 1 cycle per second, it will look like this:



The time-based Sine wave that portrays the object's y-axis values over time will have the formula "y = −4 + sin 360t":

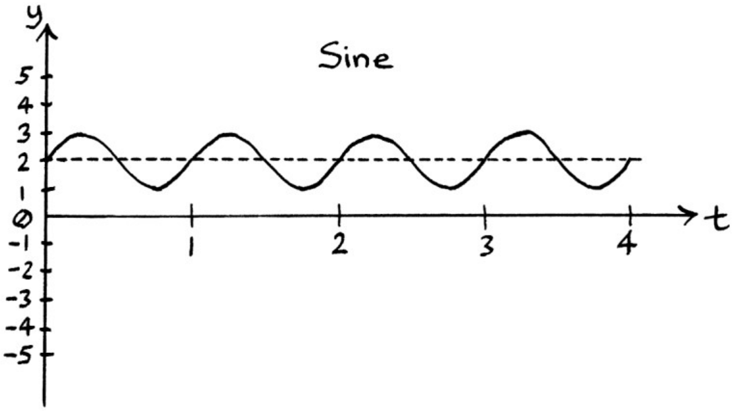


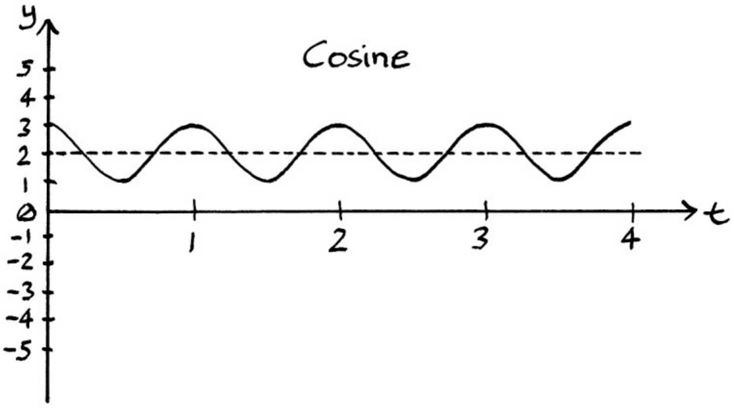
The time-based Cosine wave that portrays the object's x-axis values over time will have the formula "y = 3 + cos 360t":



From all of the above, we can see that from one circle, there are two independent mean levels: one mean level for the Sine wave, and one mean level for the Cosine wave. Although we can know that any Sine wave has a corresponding Cosine wave with the same amplitude, frequency and phase and vice versa, if we only have one wave, we cannot know the corresponding wave's mean level. Therefore, if we have one wave, and we want to recreate the other wave and the circle, the best we can do is guess, perhaps incorrectly, that the other wave has either the same mean level or no mean level at all.

If the circle were centred at a point where the "x" and "y" values are the same, then the Sine and Cosine waves derived from the circle would have the same mean level. For example, if the circle were centred at (2, 2) on the circle chart, the derived Sine wave and the derived Cosine wave would each have a mean level of 2 units.





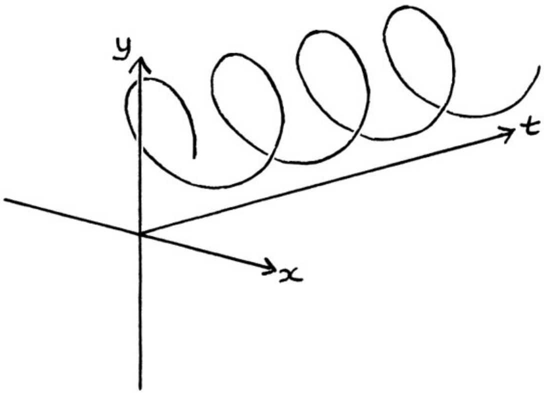
The two mean levels in practice

Having two mean levels is more of a theoretical concept for most situations where waves are used. A radio wave or a sound wave, as it is received, has no mean level – it is centred around zero. If calculations are performed on such a wave, there might become one mean level. Therefore, with radio and sound waves, mean level is only treated as existing in one dimension on the circle, that is to say for one wave.

If a radio or sound signal is treated as a Sine wave or a sum of Sine waves, then the mean level will only apply to the Sine waves – the mean level for the corresponding Cosine waves is undefined. Similarly, for Cosine waves, the mean level for the corresponding Sine waves is undefined. More usually, you will find that the mean level for the other wave is not just undefined, but not even considered as a concept. Due to the way that waves are often taught at a basic level without mentioning circles, and given that most maths on waves is done with radio or sound waves, it is rare to see two mean levels mentioned at all. As always, it pays to remember that radio and sound waves are not the only types of waves.

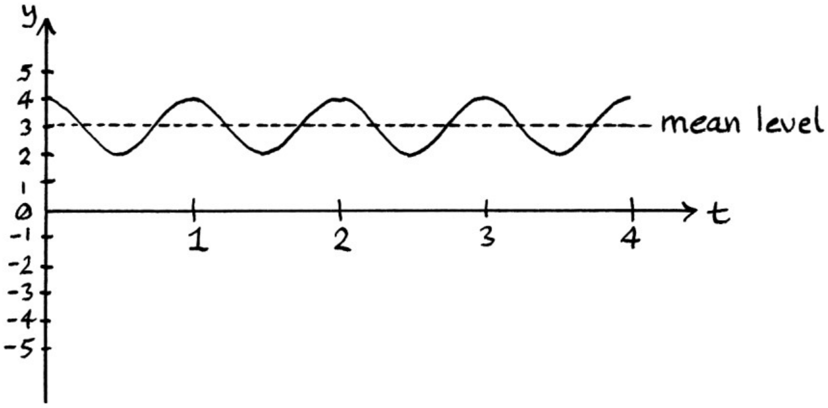
Mean level on the helix

Mean level can be portrayed on the helix. Here is a helix constructed with a Sine wave with a positive, non-zero mean level, and a Cosine wave with a zero mean level. It could also be said to be based on a circle centred on a positive, non-zero y-axis value and a zero x-axis value.

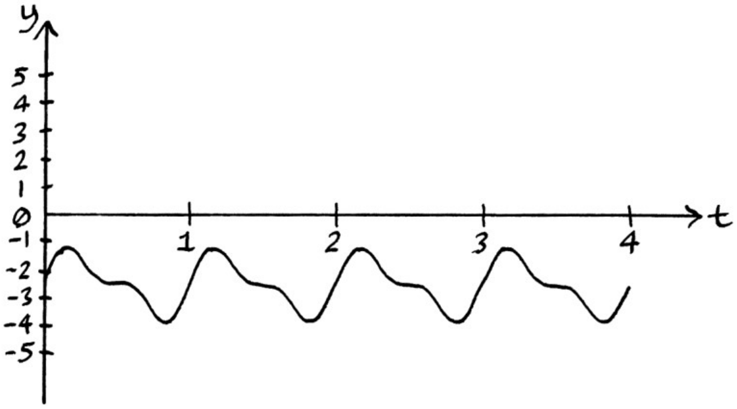


Calculating the mean level

The mean level of a wave is the y-axis value around which the wave fluctuates:



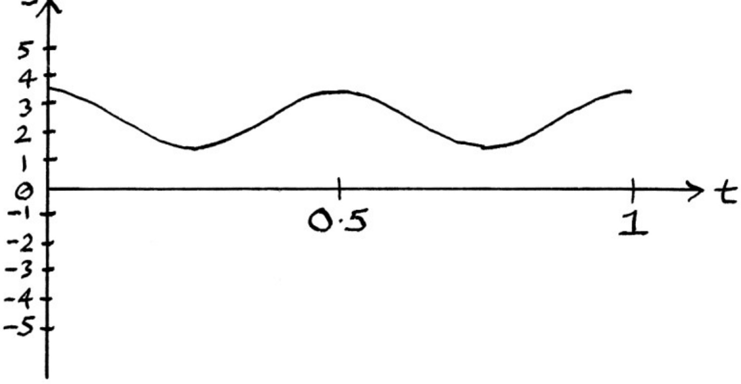
Although it is obvious where the centre of a simple wave is, and therefore what its mean level is, it can be good to know other ways of calculating it. This is because such methods also apply to calculating the mean level of more complicated signals where the mean level is less obvious, such as with the following graph, which is made from adding two waves together:



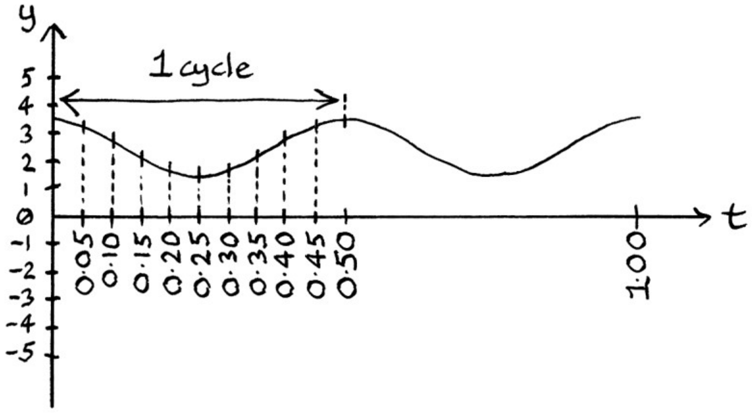
Mean level is the mean level

The mean level is, as I have said before, the average level of a wave. Therefore, one way to calculate it is to measure every y-axis value along a wave and take the average. Of course, this would be impossible as a wave might continue forever, and there are really an infinite number of points in any one section of a wave. However, if a wave repeats its shape in an identical way, we can take a series of evenly spaced y-axis values for one cycle and calculate the average of those. The more of these evenly spaced values we read off the graph, the more accurate the result will be.

As an easy example, we will calculate the mean level for the following wave, which we will pretend is drawn very accurately. [Note that the y-axis value at t = 0 is not necessarily the mean level for a Sine wave as the wave formula might have a non-zero phase].



This wave repeats every half a second, so its period is 0.5 seconds, and its frequency is 2 cycles per second. As we only need to measure from one cycle, we only need to consider times from 0 seconds up to 0.5 seconds. For the purposes of this example, we will measure ten evenly spaced y-axis values for one cycle. We will read the y-axis values at every 0.05 seconds. We will pretend that we can read the values off the graph to 2 decimal places.



The y-axis value at t = 0.00 is 3.50
The y-axis value at t = 0.05 is 3.31
The y-axis value at t = 0.10 is 2.81
The y-axis value at t = 0.15 is 2.19
The y-axis value at t = 0.20 is 1.69
The y-axis value at t = 0.25 is 1.50
The y-axis value at t = 0.30 is 1.69
The y-axis value at t = 0.35 is 2.19
The y-axis value at t = 0.40 is 2.81
The y-axis value at t = 0.45 is 3.31

We do not read the y-axis value at t = 0.5 seconds as that is the point where the next cycle starts. If we included that, our results would not be correct.

We add the values together and divide by the number of values, and we have the average: 25.0 ÷ 10 = 2.5

Therefore, we can say that the mean level of this wave is approximately 2.5 units. In this case, the result is completely correct, which we can tell just by looking at the graph. The actual wave formula was “y = 2.5 + sin ((360 * 2t) + 90)”. Ideally, we would need to calculate our average with more values to be sure that our result is reasonably accurate.

If the total of our y-axis values had been zero, we would not have needed to do the step of finding the average because we would know that the average would be zero too. This is because zero divided into any number of pieces is still zero. However, if the total is not zero, then we must calculate the average.

Thoughts

The above method of calculating the mean level requires an accurately drawn graph of the wave. In practice, such a thing is unlikely to be available. Later on in this book, when we look at discrete waves (where waves are stored as a sequence of y-axis values at evenly spaced moments in time), calculating the mean level becomes extremely easy and straightforward as all the measuring is done for us. In such cases, we just add up the given y-axis values for one cycle and find the average. For now, the idea of working out the mean level from an accurately drawn graph helps in visualising what mean level really is.

The name of mean level

The term "mean level" is just one of many ways to describe the same thing. I am choosing "mean level" as it refers to the average point of the wave – in other words, the y-axis value around which the wave is centred. You will often see other terms used to describe the same thing. In signal processing, a common term is "DC component", where "DC" stands for "Direct Current" in the sense of electrical current. The direct current part of a wave representing alternating current results in the wave being higher or lower on the y-axis. The term is used even when the wave has nothing to do with electricity at all. Personally, I think for the first steps in learning about theoretical waves derived from circles, the term "DC" is slightly confusing and that is why I will use the term "mean level" in this book.

Instantaneous amplitude

In Chapter 5 on amplitude, we were introduced to the term "instantaneous amplitude". Instantaneous amplitude is the name for the y-axis value of a wave at a particular moment in time. The word "instantaneous" distinguishes the idea from normal "amplitude", which is the extent up and down the y-axis value that the points of a wave reach either side of its centre. As the term "instantaneous amplitude" refers to the y-axis value at any particular time, if there is a mean level, then that mean level will be included in the instantaneous amplitude. For example, if we have the wave:
"y = 2 + sin 360t"
... then the instantaneous amplitudes will vary from +1 units up to +3 units depending on where on the graph we make the reading.

If we have the wave:
"$y = \sin 360t$"
... then the instantaneous amplitudes will vary from −1 units to +1 units.

Formulas for mean level

In this book, I will give the general formula for a time-based Sine wave that takes into account mean level as:
$\mathbf{y = h_s + \sin 360t}$
... where "h_s" represents the height of the mean level for the Sine wave.

The corresponding formula for a Cosine wave is:
$\mathbf{y = h_c + \cos 360t}$
... where "h_c" represents the height of the mean level for the Cosine wave.

Note that I am labelling each "h" with a subscript ("s" or "c") to indicate that they are different for each wave. For future examples in this book, if the distinction is not relevant or is obvious from the context, I might use just the letter "h" with no subscript.

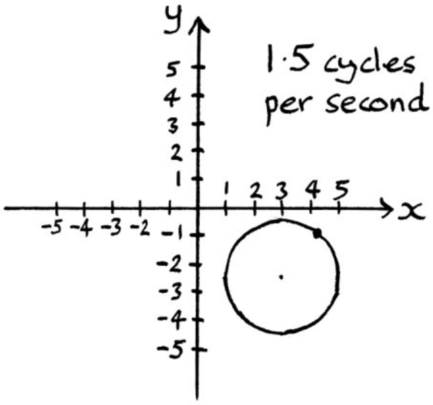
For waves relating to angles and not time, the formulas will be:
$\mathbf{y = h_s + \sin \theta}$
$\mathbf{y = h_c + \cos \theta}$

Chapter 9: All the wave attributes

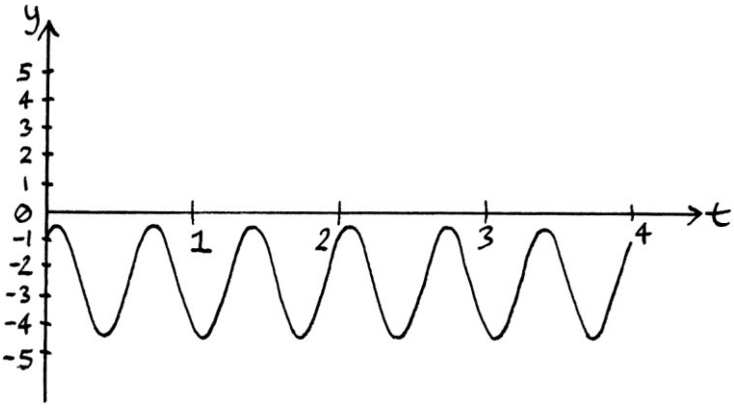
Independent properties

Each of the four properties of a wave can be altered independently without having any effect on the others. We can change one or more of them to produce different effects. For example, we can raise the frequency, increase the amplitude, increase the phase in the formula, and change the mean level all at the same time, and each attribute will remain distinguishable.

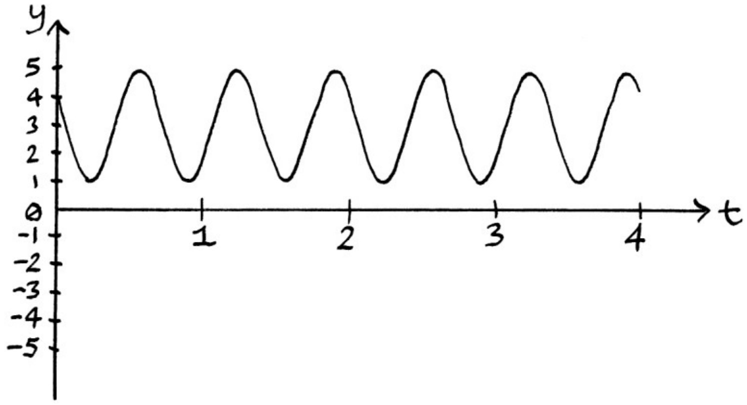
If we have this circle...



... then the Sine wave graph will look like this:



... and the Cosine wave graph will look like this:



In both the Sine wave and the Cosine wave graphs, it is possible to distinguish the amplitude, frequency, phase, and mean level. All four attributes are independent of each other. A Sine wave and its corresponding Cosine wave will always share the same amplitude, frequency and phase. However, their mean levels might or might not be different.

Angle-based formulas

The formulas for angle-based waves that have the attributes of amplitude, phase and mean level are as follows:

Amplitude:
“$y = A \sin \theta$”
“$y = A \cos \theta$”
... where “A” is the overall amplitude, and “θ” is the angle in degrees.

Phase:
“$y = \sin (\theta + \phi)$”
“$y = \cos (\theta + \phi)$”
... where “ϕ” is the overall phase of the wave measured in degrees.

Mean level:
“$y = h_s + \sin \theta$”
“$y = h_c + \cos \theta$”
... where “h_s” is the average height of the Sine wave, and “h_c” is the average height of the Cosine wave.

The two formulas that take into account all three attributes at the same time are:
"$y = h_s + A \sin (\theta + \phi)$"
"$y = h_c + A \cos (\theta + \phi)$"

[There is no frequency attribute here because the concept of frequency is meaningless in an angle-based wave. Angle-based waves just show the position of static points around a circle's edge].

Time-based formulas

The formulas for time-based waves showing amplitude, frequency, phase and mean level are as follows:

Amplitude:
"$y = A \sin 360t$"
"$y = A \cos 360t$"
... where "A" represents the overall amplitude, and "t" is the time in seconds.

Frequency:
"$y = \sin 360ft$"
"$y = \cos 360ft$"
... where "f" represents the overall frequency.
or:
"$y = \sin \omega t$"
"$y = \cos \omega t$"
... where "ω" represents angular frequency and is shorthand for "360 * f". [If we were working in radians, it would be shorthand for "2π * f".]

Phase:
"$y = \sin (360t + \phi)$"
"$y = \cos (360t + \phi)$"
... where "ϕ" represents the overall phase of the wave as measured in degrees.

Mean level:
"$y = h_s + \sin 360t$"
"$y = h_c + \cos 360t$"
... where "h_s" represents the average height of the Sine wave, and "h_c" represents the average height of the Cosine wave.

Formulas that take into account all these possible attributes at the same time are:
“$y = h_s + A \sin (360ft + \phi)$”
“$y = h_c + A \cos (360ft + \phi)$”

... or, if we group the “360 * f” together as one entity:

“$y = h_s + A \sin (\omega t + \phi)$”
“$y = h_c + A \cos (\omega t + \phi)$”

Unneeded ones and zeroes

Generally, when using formulas for waves, if there is a multiplication by 1 (for example, in the amplitude or frequency), it will be left out, and if there is an addition of 0 (for example, in the mean level or phase), it will be left out. The meaning is the same, but doing this keeps the formulas shorter and more succinct.

For example, if the mean level is zero, then there is not much point in mentioning it in the formula. Therefore, instead of giving a formula as:
“$y = 0 + 3 \sin ((360 * 2t) + 180)$”
... it would usually be given as:
“$y = 3 \sin ((360 * 2t) + 180)$”.

If the phase is zero too, then that will be left out. Therefore, the formula:
“$y = 3 \sin ((360 * 2t) + 0)$”
... would usually be given as:
“$y = 3 \sin (360 * 2t)$”.

If the amplitude is 1, then that does not need to be stated. Therefore, the formula:
“$y = 1 \sin (360 * 2t)$”
... would usually be given as:
“$y = \sin (360 * 2t)$”.

If the frequency is 1, then too does not need to be stated. Therefore, the formula:
“$y = \sin (360 * 1t)$”
... would usually be given as:
“$y = \sin (360 * t)$”
... or more commonly:
“$y = \sin (360t)$”
... or:
“$y = \sin 360t$”.

Formulas tend to be given with the minimum information necessary.

In this book, I will sometimes include unneeded ones and zeroes to make an explanation clearer. For example, if I am comparing two frequencies and one of them is 1 cycle per second, then I might leave the 1 in the formula to make it easier to compare. For example, I might list two waves as:
"y = sin (360 * 2t)"
... and:
"y = sin (360 * 1t)".

If the fact that a mean level or phase is zero is relevant or might help with understanding, then I might leave the zero in the formula.

Other people's symbols

It is important to know that you will often see different symbols to those being used here to describe waves, and even different formulas, depending on the context and the whim of the person using the formulas.

Other people's use of different symbols can be confusing at first, but in the context of a wave formula, we can tell what everything is by its position or what is being done to it. For example, with the meaningless formula:
"y = ? + ? sin ((360 * ? * ?) + ?)"
... we can tell that:

- The first "?" must represent mean level, because it is being added on to the rest of the equation.
- The second "?" must be amplitude, because it is being multiplied by the Sine part of the equation. This "?" is scaling the result of the Sine function.
- The third "?" must represent either frequency or time.
- The fourth "?" must represent whichever of frequency or time was not represented by the third "?".
- The fifth "?" must be phase because it is being added to the rest of the part of the equation being Sined.

Once you know how wave formulas work, you can figure out any unknown symbols. Among the alternatives for the symbols that you might see in wave formulas are:

For amplitude, you might see: “A”, “a”, “A_0”, “A_1” etc, “v” among other things.

- The letter “A” stands for “amplitude”.
- The letter “A” with a number next to it, e.g. “A_0”, is usually used if there are going to be other mentioned waves with amplitudes of other values. Using numbered letter “A”s avoids confusing the reader with even more symbols. Such numbered letters can appear muddling when you first see them, but you will become used to them quickly.
- The letter “v” stands for volts, and is used if the y-axis is referring to volts. Do not confuse this “v” with “v” or “ν” being used to represent frequency. As well as “v”, you might see any abbreviation for the unit in which the amplitude is being measured.

For phase, you might see: “ϕ”, “φ” and “θ” among other things.

- “ϕ” is the lower-case Greek letter “phi”.
- “φ” is also the lower-case Greek letter “phi”, but shown in a font that more closely resembles handwriting. Depending on the font, this can, confusingly, look a bit like the lower-case Greek letter “psi” (“ψ”), which is not relevant to this book.
- “θ” is the lower-case Greek letter “theta”. The letter “θ” is sometimes used for phase in time-based waves (in which case, “θ” will not be in the formula to represent the angle, so there will not be two thetas). The letter “θ” is usually used to represent an angle, so it is consistent to use it for phase, which is also an angle. However, using a symbol other than “θ” makes it more obvious that it represents phase.

For frequency, you might see “f”, “F”, “f_0”, “f_1”, “ν”, “v”, “ω” among other things.

- “f” and its variations obviously stand for “frequency”. Do not confuse this “f” with the “f” that stands for “function”.
- “ν” is the lower-case Greek letter “nu” (pronounced “new”), and is the Greek equivalent to the lower-case Latin letter “n” (as in “number”). Confusingly, it looks just like the lower-case Latin letter “v” (as in “victor”). This is often used to represent frequency in such academic subjects as optics. One advantage of using a symbol other than “f” is that it is not confused with the letter “f” being used to mean “function”. Often, you will see “ν” used to represent the equivalent of frequency in distance-based waves, or in other words, waves where the graphs have the x-axis as distance instead of time. This “ν” does not represent time-based frequency as measured in cycles per second, but distance-based frequency as measured in cycles per *metre*. It is the inverse of wavelength. It is a different concept, and I will explain more about it in Chapter 31.

- "v" is the lower-case Latin letter "v" (as in "victor"). Sometimes, people use the letter "v" because it is quicker than typing a "ν" (nu).
- "ω" represents *angular* frequency as described earlier in Chapter 6, but it is usually only used when dealing with circles that are divided into portions based on radians instead of degrees. You might occasionally see this misused to represent normal cycles-per-second frequency. If someone cannot type "ω" on their keyboard, they might use a "w" instead.

For mean level, you might see: "h", "DC" or other symbols depending on whether the y-axis on the wave graph is representing a theoretical value or a real-world measurement. One important thing to note is that usually an author will present waves as either being all Cosine waves or all Sine waves, and therefore they will not distinguish between the mean levels for Sine and Cosine. For them, there is only a mean level for whichever wave type is being discussed, and the other type is either ignored or considered as having no mean level.

- "h" stands for height. This is not a common symbol for mean level, but it is as good as anything else. In this book, when it is relevant, I distinguish between the mean level for Sine and the mean level for Cosine by giving the "h" a suffix: "h_s" and "h_c".
- "DC" stands for "Direct Current". This is often used even when the waves have nothing to do with electricity.

For the actual angle in a wave, you might see: "θ", "Ω".

- "θ" is the generally used symbol.
- "Ω" is rarer.

For time, you might see: "t", "n".

- "t" is the most common symbol for time.
- "n" is used with *discrete* waves, as discussed in Chapter 39. It does not represent time as a continuous entity, but instead it essentially represents individual moments in time. You do not need to understand what this means yet.

You will also see other symbols used in formulas. Sometimes, you will see symbols swapped around because they relate to earlier parts of a calculation, and it is consistent to continue using them in a later wave formula. For example, imagine that we are using "ϕ" to refer to the phase of one wave, and we have a second wave's amplitude that is dependent on the phase of that wave. In such a case, it would be simpler to use that "ϕ" symbol as the amplitude of the second wave, than to confuse matters by redefining our symbols.

Although it can be temporarily confusing to see new symbols used to represent concepts, or even see old symbols used in different ways, it is just social convention that has led to certain symbols being used to represent particular ideas. There is no true “right” or “wrong” symbol, as all of this is just a social construction. However, in my view, while learning or teaching, it pays to be consistent with the general accepted symbols just to avoid confusion and potential ambiguity. [Note that I have not done this with the letter “h” for mean level]. I also think it pays to use symbols that are distinctive and cannot be misread.

A good way of thinking about symbols and terminology in general is to follow Jon Postel’s ideology for computer networking, as mentioned in RFC1122: “Be liberal in what you accept, and conservative in what you send.” Although this is intended for networking, you can use the idea with waves and signals: when you write formulas, use terms, or describe things, try to be as unambiguous as possible, and use commonly used terms and symbols. When dealing with what other people have written or said, do not be surprised if they are ambiguous and confusing, but try to understand them anyway. As you learn more about waves, the confusion you might have about different symbols will diminish.

There are often disputes as to what is the best symbol for a particular concept (but less so for the more basic things described in this chapter), and there exists a minority of people who adamantly insist that their arbitrary opinion on what a symbol should be is somehow more “correct” than the opinions of other people.

Historically, mathematicians in Western Europe picked random Greek and Latin letters to represent concepts for the piece of work that they were doing at a particular time. They used different symbols at different times to refer to the same thing, depending on how they felt on the day, in much the same way as one might use the symbol “x” nowadays. Over time, a few of these symbols became set in stone to represent certain concepts, and now we are stuck with slightly unhelpful and similar symbols for things. Even today, when we have easy access to characters from countless non-western European alphabets, mathematicians prefer to stick to Latin and Greek symbols.

Other people's formulas

Generally, in more advanced maths, wave formulas are given with the angles in radians instead of degrees, and therefore the waves relating to time will mention angular frequency in radians too. This means that in situations where we would use "360t" when working in degrees, we would use "2πt" instead (because radians divide the circle into 2π divisions, instead of 360 divisions). This means that we will often see waves with formulas such as this: "y = sin 2πft". I explain π and radians in Chapters 21 and 22.

Apart from the different formulas for degrees and radians, sometimes you will see formulas for waves that are written in a different way to how I have described them here. A few of these differences are as follows:

Negative phase

A common difference you might see is a formula where the phase is negative instead of positive. In other words, instead of this: "y = A sin (360ft + ϕ)", you will see this: "y = A sin (360ft − ϕ)". This means that where I might say a phase of +100 degrees, they will instead give a phase of −260 degrees. As I explained in Chapter 7 on phase, it does not make any difference whether the phase is given as a positive angle or a negative angle, as long as the angles identify the same point on the circle's edge. The phase indicates a particular point on the edge of the circle – it is the same point whether we measure it with a positive angle or a negative angle. The Sine and Cosine wave derived from that circle will be identical in shape and meaning because their phases are based on the position of that point, and not how that point is described.

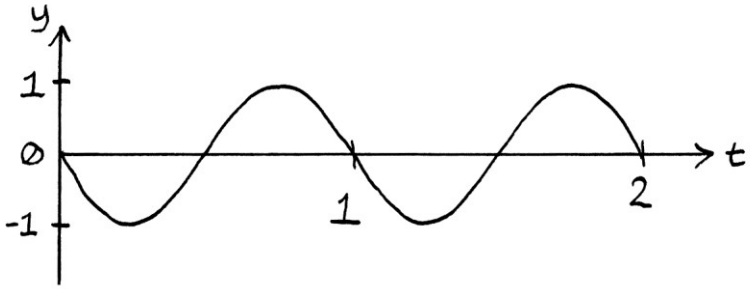
It is a matter of choice whether you decide to use positive or negative phases, but personally, I think a positive phase looks better.

Negative amplitude

Sometimes, you might see a formula that has a negative overall amplitude. For example:
“y = −1 sin 360t”

The wave drawn from this formula would be upside down in comparison to the same wave with a positive amplitude. Every result of “sin 360t” would be made negative and so appear in the other half of the y-axis. In fact, the wave curve would be identical to a wave formula with a *positive* amplitude and a phase of 180 degrees:
“y = 1 sin (360t + 180)”

Both “y = −1 sin 360t” and “y = 1 sin (360t + 180)” look like this:



The formula “y = −1 sin 360t” is a perfectly valid formula, but if you want to be consistent and think of the circle from which a wave is, or could have been, derived, then a negative amplitude does not really make much sense – you cannot have a circle’s radius being a negative length.

When you are first learning about waves, it is easiest not to use negative amplitudes unless it is a temporary measure as part of a calculation. Using only positive amplitudes will help reinforce the idea that Sine and Cosine waves are derived from circles. In my opinion, even if a Sine or Cosine wave has no apparent connection with a circle at all, it makes sense to treat it as if it had been derived from a circle – it makes things easier to understand and easier to deal with.

When you are first learning, if you ever see a negative-amplitude formula, you can turn it into a positive-amplitude formula by adding 180 degrees to the phase and making the amplitude positive. [Or subtract 180 degrees as that will have the same effect]. The result will refer to an identical curve on a graph, and will be consistent with the idea of amplitude being the same as the radius of a circle.

For example, if we are presented with the formula:
“y = −2 sin ((360 * 5t) + 10)”
... then we add 180 degrees to the phase, and make the amplitude positive:
“y = +2 sin ((360 * 5t) + 10 + 180)”
... which ends up as this:
“y = 2 sin ((360 * 5t) + 190)”.
This formula refers to exactly the same wave curve as the negative-amplitude formula.

Instead of adding 180 degrees, we could have subtracted 180 degrees (and again made the amplitude positive). This is because adding 180 degrees to an angle produces the same result as subtracting 180 degrees.
“y = 2 sin ((360 * 5t) + 10 − 180)”
... which is the same as this:
“y = 2 sin ((360 * 5t) − 170)”.
The angle of −170 degrees is the same as the angle −170 + 360 = 190, so we can rephrase the wave as:
“y = 2 sin ((360 * 5t) + 190)”.
... which is the same result.

If we see the wave “y = −56.7 sin ((360 * 41t) + 300)”, then we can rewrite it as, or think of it as:
“y = +56.7 sin ((360 * 41t) + 300 + 180)”
... which is this:
“y = 56.7 sin ((360 * 41t) + 480)”
... and, as the angle of 480 degrees is the same as the angle 480 − 360 = 120 degrees, we can give the formula as:
“y = 56.7 sin ((360 * 41t) + 120)”

Negative amplitudes are only confusing when you are learning because they contradict how you should visualise Sine and Cosine. Negative amplitudes will not stop formulas from working – they are just “conceptually wrong”. As you become more used to waves, you will find the concept of negative amplitudes easier to visualise, and you will come to see that negative amplitudes can be a useful idea. Frequently, a maths problem is more straightforward if you use negative amplitudes. Accepting the concept of negative amplitudes will also make your knowledge of waves more thorough.

When first learning about waves, I think it is good to have a rule that amplitudes can only be positive, but to break that rule now and then, but only briefly, and only if it makes things simpler for a particular calculation or observation. As you become more adept with waves, you can relax the rule or ignore it completely.

"f(t) =" instead of "y ="

This is not really a difference in the formula, but a different way of thinking of the formula. Instead of "y = A sin (360ft + ϕ)", you might see this:
"f(t) = A sin (360ft + ϕ)"

In this case, "f(t)" refers to "the function's effect on t" – the first "f" stands for function. In this sense, the equals sign is not saying that "f(t)" is equal to the *result* of the formula, but really that "f(t)" is equivalent to the *whole* formula as an entity. We can think of the formula as saying, "The function that is having an effect on the variable 't' is 'A sin (360ft + ϕ)'".

Another way of thinking about this is that "f(...)" is a way of not having to write out the whole formula again in the future. If we say "f(t) = A sin (360ft + ϕ)", then we can simply say "f(t)" in future sentences instead of spelling out "A sin (360ft + ϕ)" each time.

Using the symbol "f" to mean function is slightly confusing here because "f" is also being used to represent frequency.

Period instead of frequency

Sometimes people give the formula for a wave in terms of its period instead of its frequency. The period of a wave is the reciprocal of the frequency (1 ÷ frequency), and the frequency of the wave is the reciprocal of the period (1 ÷ period). One common symbol for period is the upper-case letter "T" (as in "Tango"), which means that 1 ÷ f = T, and 1 ÷ T = f. [Another common symbol for period is the upper-case letter "L" (as in "Lima").]

What I would write as:
“y = sin 360ft”
... some people might write as:
“y = sin (360t ÷ T)” or “y = sin (360 * $\frac{1}{T}$ * t)”
... where, in this case, the period is being represented by the upper-case Latin letter “T”, which is slightly confusing because we are using the lower-case Latin letter “t” to represent time.

The formula phrased to include the period of the wave is just another way of saying exactly the same thing as a formula that includes frequency. However, it is longer and requires a bit more thought to analyse. Personally, I think it only makes sense to mention period in the formula if period is of particular importance to the subject being discussed at the time. Otherwise, frequency is better, especially when we consider how it is good to treat Sine and Cosine waves as derived from circles – if someone tells us how many cycles per second an object completes, we instantly know what this means. If someone tells us the period, we have to do a calculation before it makes sense. Similarly, waves are generally categorised by their frequency (or sometimes their wavelength if they are real-world waves where wavelength is relevant), and not their period, so frequency is better. When analysing signals created by adding waves, it can be useful to consider the period, but there are some people who will give every wave formula in terms of the period.

Phase given as a time instead of an angle

In a formula such as “y = cos (360t + 45)”, the phase is given as an angle in degrees. It is also the case that the time, after having been multiplied by 360, is being treated as if it were really an angle in degrees too. Sometimes, you will read formulas where an author has decided to put the phase as a time in seconds. For example:

“y = cos (360 * (t + 0.125))”

The 0.125 in this case refers to the phase given as a time in seconds. To give the time in seconds, it has been necessary to factor out the 360 from both the time and the phase, and end up with a formula that is twice as difficult to read.

Another way of writing the phase as a time in seconds is just to convert the phase part to seconds, and indicate in the formula that it indicates seconds. This makes the formula slightly odd:

“y = cos (360t + 0.125 seconds)

There is not really much point in giving the phase as a time in seconds, unless the phase as a time in seconds is immediately important to the subject we are discussing. Otherwise, it is just an unnecessary complication. We cannot immediately tell what the frequency or the phase is when the formula has been factored – instead we have to do the multiplication first. It is confusing to people trying to learn. Fortunately, it is rare to see the phase given as a time in seconds, especially in anything remotely academic.

Unnecessary factorisation

A similar idea to giving the phase as a time instead of an angle, is to factor the part of the formula being Sined or Cosined, but without any purpose behind it whatsoever apart from some psychological need to do it. You will often see this in tasks set for schoolchildren.

For example, where I would give this formula:
y = sin (360t + 180)
... someone else might give this formula instead:
y = sin (180 * (2t + 1)).

Where I would write:
y = cos ((360 * 12t) + 45)
... they would write:
y = cos (45 * (96t + 1))

First, there is absolutely no use in doing this – it does not have any advantages whatsoever. It is wasted effort. Second, it makes the formula much harder to understand – we have to do a calculation before we can know the phase or the frequency. We cannot imagine how these look on the circle, on the helix, or even on a wave graph by looking at the formula. Third, this makes the whole concept of waves much harder for anyone trying to learn. It is a good example as to why so many people grow up to dislike maths.

Summary

The point of this section is just to make you understand the variety of symbols and formulas you might see. There is a lot of variety, and in every new explanation you read, you might have to adapt slightly. Seeing a variety of symbols and formulas can give you a better understanding of everything. Generally, it is straightforward to know what someone means, but occasionally you will read something written by someone who is just being difficult. Occasionally, but more often than you would expect, you will read something by someone who has made a mistake. In any book over a certain size, it is probably impossible not to make some mistakes. Similarly, when a book is published, the publisher will often introduce more mistakes during the conversion to their chosen layout. This is especially true if someone at the publisher has to retype all the formulas.

When you are first learning about waves, the different symbols and formulas can seem complicated and unintuitive. However, as you continue to learn, you will become so used to them that you will forget them ever being a problem.

Wave pairs

Circles centred on the origin of the axes

For a circle that is centred on the origin of the x and y-axes, its Sine and Cosine waves will have a mean level of zero – in other words they will be centred around y = 0 on each of their graphs.

For the waves derived from such a circle, it will always be the case that if we have one of the waves, the other corresponding wave will have the same basic formula.

In other words, if the Sine wave derived from a circle is:
“$y = A \sin (360ft + \phi)$”
... then the Cosine wave will be:
“$y = A \cos (360ft + \phi)$”
... where A, f, and ϕ have the same value in each equation.

If the Cosine wave is:
“$y = A \cos (360ft + \phi)$”
... then the Sine wave will be:
“$y = A \sin (360ft + \phi)$”.

For the pair of waves derived from the same circle, they cannot be anything else.

Similarly, a particular Sine wave could only ever have come from one particular circle, and a particular Cosine wave could only ever have come from one particular circle. If we have a Sine wave, then it implies a particular circle, and that circle implies a particular Cosine wave. If we have a Cosine wave, then it implies a particular circle, and that circle implies a particular Sine wave.

For example, if we have an object rotating at 4 cycles per second around a circle (centred on the origin of the axes) with a radius of 5 units, and the object started at 45 degrees, then the formulas for the Sine and Cosine waves would be:
“$y = 5 \sin ((360 * 4t) + 45)$”
... and:
“$y = 5 \cos ((360 * 4t) + 45)$”.

They cannot be anything else. The Sine wave “$y = 5 \sin ((360 * 4t) + 45)$” has as its corresponding Cosine wave, “$y = 5 \cos ((360 * 4t) + 45)$”. These two waves are a pair. The circle produces the pair, and the pair produces the circle.

If a Sine wave and Cosine wave are derived from the same circle, then they will be the same wave but with a 90 degree phase difference between them. Therefore, if we have the Sine wave as:
“$y = A \sin (360ft + \phi)$”
... the corresponding Cosine wave’s curve will *always* be the same as:
“$y = A \sin ((360ft) + (\phi + 90))$”
... for circles centred on the origin of the axes.

Similarly, if we have the Cosine wave as:
“$y = A \cos (360ft + \phi)$”
... then the corresponding Sine wave’s curve will *always* be the same as:
“$y = A \cos ((360ft) + (\phi - 90))$
... for circles centred on the origin of the axes.

The only real difference is that Sine waves are read from the y-axis and Cosine waves are read from the x-axis.

Rules

For circles *that are centred on the origin of the x and y-axes*, some important rules are:

- Any particular circle implies the characteristics of its Sine and Cosine waves.

- Any particular Sine wave implies the characteristics of the circle it is derived from.

- Any particular Cosine wave implies the characteristics of the circle it is derived from.

- Any particular Sine wave implies the characteristics of its Cosine wave twin.

- Any particular Cosine wave implies the characteristics of its Sine wave twin.

- Any particular circle, Sine wave or Cosine wave implies the characteristics of the helix it is derived from.

- If we have the details of a circle, a helix, a Sine wave or a Cosine wave, we can calculate the details of the other parts.

- Any of the four parts implies the attributes of the others.

Note that these rules are only true for Sine and Cosine waves with no mean level. In other words, they are only true for waves that can be described by these equations:

“$y = A \sin (360ft + \phi)$” and “$y = A \cos (360ft + \phi)$”
... or:
“$y = A \sin (\theta + \phi)$” and “$y = A \cos (\theta + \phi)$”.

To put this another way, they are only true for waves derived from circles that are centred on the origin of the axes.

All circles

If we consider circles that are not centred on the origin of the axes, then the above rules are still true, but it will not be possible to deduce the mean level of a wave from its twin. We would have to have the original circle to calculate the mean levels of each wave. A standard rule would be:

"For any Sine wave derived from a circle, there will be a corresponding Cosine wave that has the identical amplitude, frequency and phase. However, it will not necessarily have the same mean level. The same is true for a Cosine wave and its corresponding Sine wave."

It pays to think of every wave as having come from a circle, even if it has not obviously been derived from a circle. Similarly, it pays to think of every wave as having a corresponding twin wave.

Chapter 10: Some terminology

In this chapter, we will look at some terms to do with waves.

Pure wave

Other people do not use the term "pure wave" particularly often, but it is a useful term. A "pure wave", as I am defining it here, is a wave that is, or could have been, derived from a single circle. In other words, it is a wave derived either from the points around a circle (if we are dealing with just angles), or from the position of an object rotating around a circle over time (if we are dealing with time). Other people might call a pure wave, a "sinusoid", but see later in this chapter for why I do not do that.

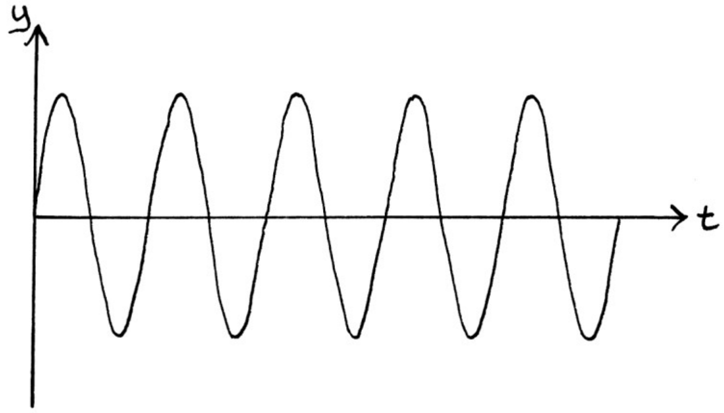
The significant attribute of a *pure* wave is that it can be portrayed using either of the following formulas, depending on whether it is angle-based or time-based:
$y = h_s + A \sin (\theta + \phi)$
$y = h_s + A \sin (360ft + \phi)$

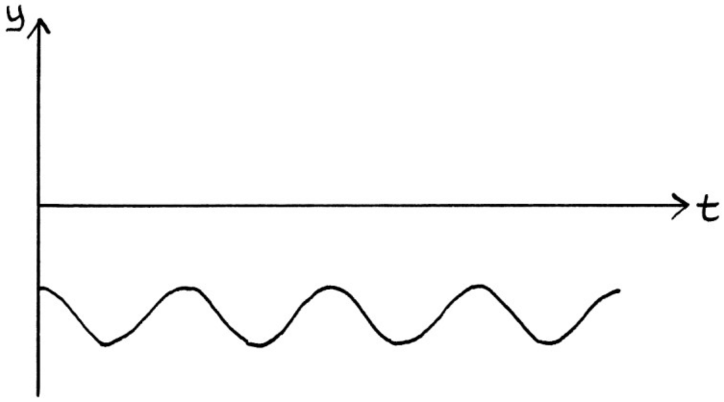
Because a Cosine wave is the same as a Sine wave with a +90 degree phase in its formula, the above formulas include these formulas too:
$y = h_c + A \cos (\theta + \phi)$
$y = h_c + A \cos (360ft + \phi)$

Therefore, we could, if we wanted, say that a pure wave is a wave that can be portrayed using one of the above four formulas. Strictly speaking, all time-based formulas are just special cases of angle-based formulas where the angle is scaled by a particular amount. The Sine and Cosine functions operate on values that are treated as angles, regardless of whether we want to think of the values as times or not. Therefore, we could just say that a pure wave is any wave of the form:
$y = h_s + A \sin (\theta + \phi)$

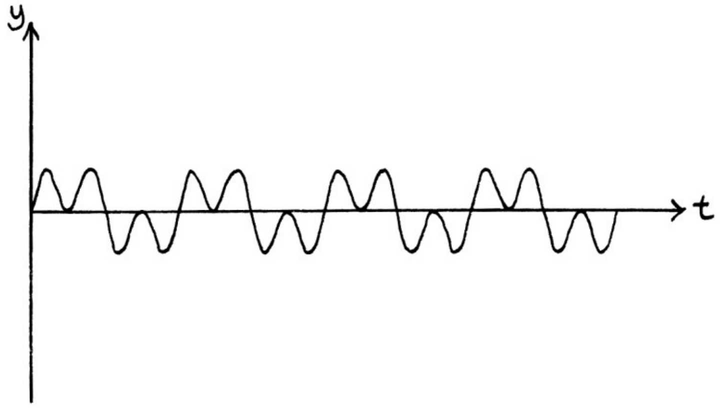
If a wave cannot be portrayed using one of the above formulas (in which case, it is also true that it would not be possible to derive it from one circle) then it is *not* a pure wave. In basic terms, a pure wave is just a typical Sine wave or a Cosine wave.

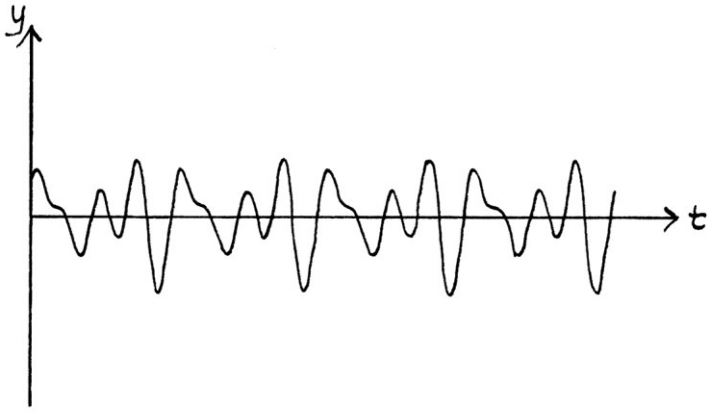
These are pure waves:





The following are *not* pure waves because they cannot be described using any one of the above formulas, which is equivalent to saying that they could not be derived from a single circle:



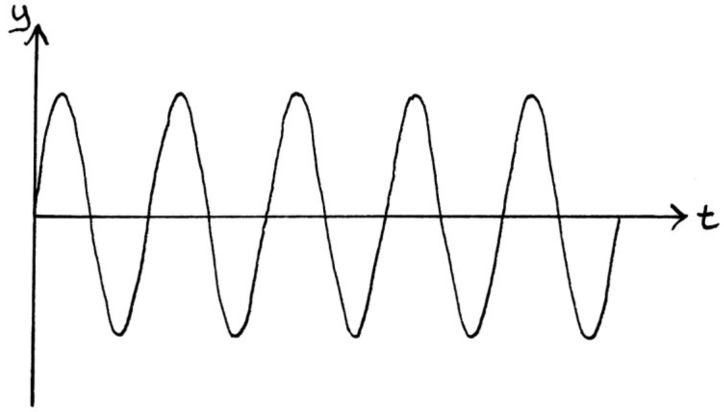


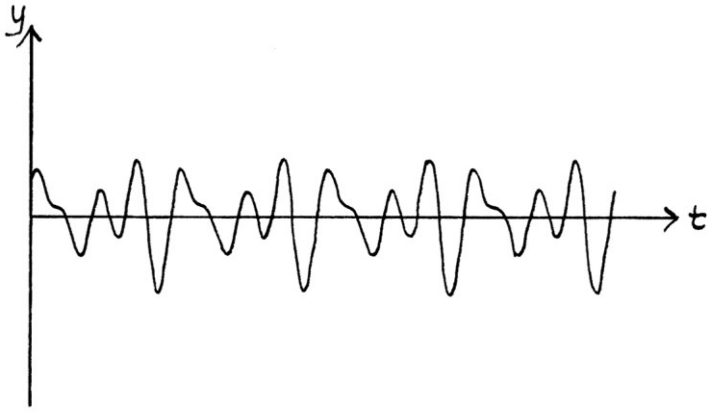
It is usually obvious if a wave is a pure wave or not, although occasionally we might see a wave that resembles a pure wave, but which has just a very slightly different shape.

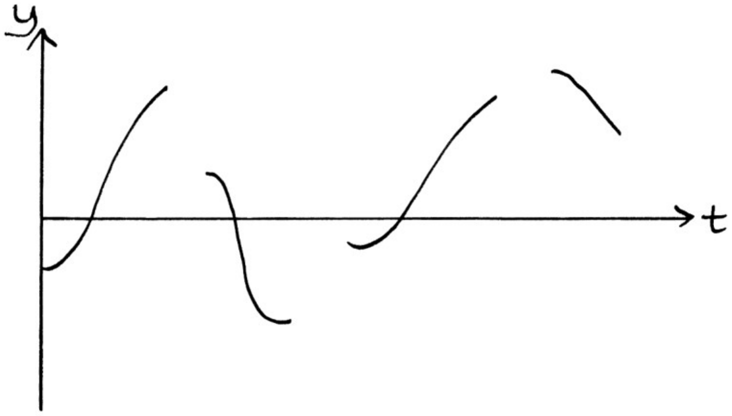
Signal

The word “signal”, in its broadest sense, could be said to be any alteration in the state of an entity over time that conveys, or could be interpreted as conveying, information. For the purposes of this book, we will simplify that definition to say that a signal is any ongoing set of values that can be plotted on a graph. Note that this simple definition is probably not correct for some situations outside those described in this book. It is easiest to explain the definition with examples. A pure wave is a signal. The combined sum of lots of pure waves is a signal, but not necessarily a pure wave. A signal does not have to have anything to do with circles, Sine waves or Cosine waves.

These are all signals:







It is impossible to show a graph of something that is not a signal, because if it cannot be drawn on a graph, then, by my definition, it is not a signal.

The word "signal" can be useful to identify things that are wave-like but are not necessarily pure waves. Sometimes, I will use the term "impure signal" to distinguish between signals that are pure waves and signals that are not pure waves.

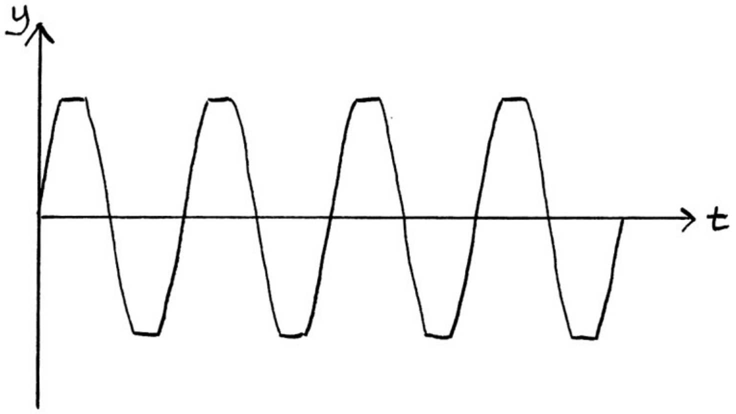
Periodic signal

A periodic signal is a signal that constantly repeats its shape. Sine waves and Cosine waves are periodic signals. It is possible to add or multiply some waves to create signals with shapes that never repeat. Such signals are not periodic signals, but "non-periodic" or "aperiodic" signals. As such signals never repeat, they have infinitely long periods.

Sinusoid

I try to avoid using the term “sinusoid” in this book because other people use the term in a variety of ways, so it can be ambiguous. In this book, I want terms to be unique in meaning and as unambiguous as possible. The word “sinusoid” is *usually* used to refer to what I am calling a “pure wave”. In other words, a sinusoid is a typical Sine or Cosine wave, with any frequency, phase, amplitude, or mean level. Nearly everyone who uses the term “sinusoid” will use it in this way.

The suffix “-oid” means “of that form” or “resembling that type”. Strictly speaking, from a linguistic point of view, a sinusoid could be anything *resembling* a Sine wave. Depending on how pedantic we want to be, this could mean something that is not a pure wave, but just vaguely resembles one. Therefore, some people might think of this as a sinusoid, but I definitely would not:



Cosinusoid

Where the term “sinusoid” becomes completely meaningless is when some people (admittedly very few people) use the term “cosinusoid”, which is not a term I will ever use. The possibly-not-a-word word “cosinusoid” is used to mean a wave that resembles a Cosine wave. If you use the word “cosinusoid”, then you have a word that specifically means something that resembles a Cosine wave *with zero phase* – if it did not refer to a Cosine wave with zero phase, then there would be no need to distinguish between a Cosine wave and any other wave – you could have just used the word “sinusoid” instead. This means that the use of the word “cosinusoid” implies that you only use the word “sinusoid” to mean something that only resembles a Sine wave *with zero phase*. Therefore, the use of the word “cosinusoid” immediately implies that you have no word for a Sine or Cosine wave with a non-zero phase.

It makes sense never to use the word “cosinusoid”, and instead let “sinusoid” refer to any pure wave with or without a phase. In this book, I would use the word “sinusoid” instead of “pure wave” if it were not for the minority of people who use the word to mean something else.

Wave form

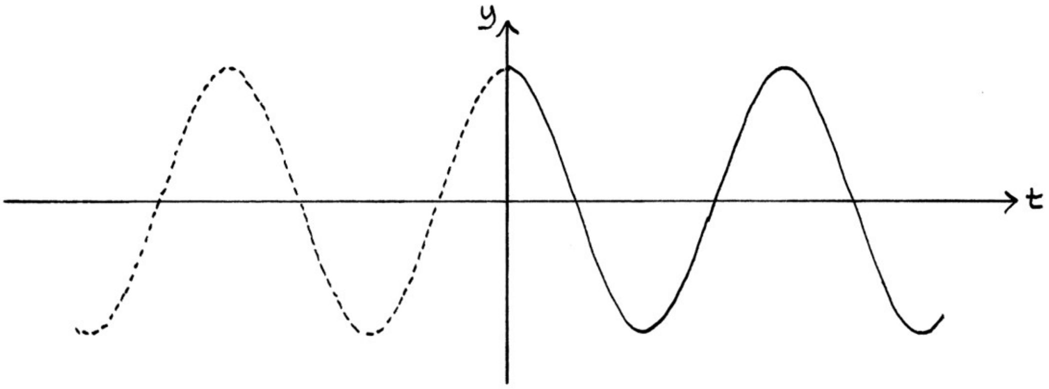
The term “wave form” or “waveform” is used in different ways. It is generally accepted to mean any signal, but some people use it to refer to a pure wave or any vaguely wave-like signal. Given its different uses and the way the name implies a wave, despite how it can be used to describe things that are not waves, I do not use the term in this book.

Functions

For the purposes of this book, we can think of a “function” as being another name for a mathematical process or a mathematical procedure. The effect of Sine on a value is a “function”, as is the effect of Cosine on a value. We can say that Sine and Cosine are “functions”.

Odd and even functions

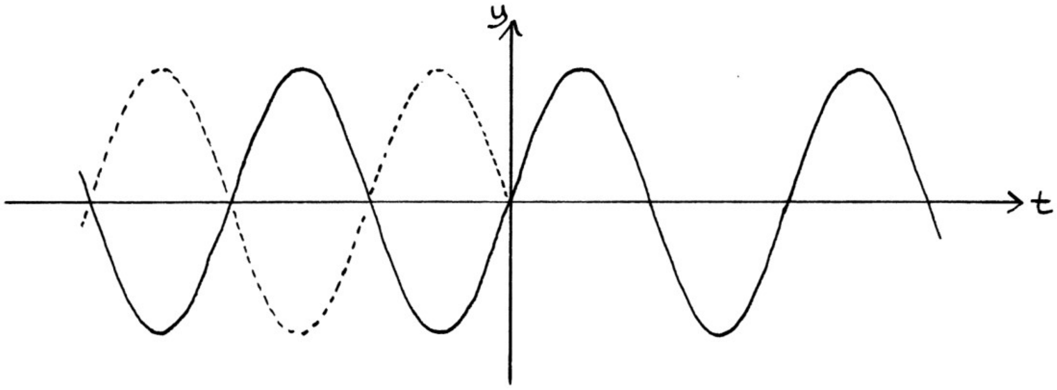
The terms “odd function” and “even function” refer to the symmetry of the graphs of functions across the y-axis. If the graph of a function can be reflected over the y-axis and keep the same shape, then it is called an “even” function; if the graph of the function does not retain its shape, then it is called an “odd” function. If a zero-phase Cosine wave is reflected over the y-axis, it will still look the same:



Another way of saying this is that on a *zero phase* Cosine wave, the y-axis values for a particular time or angle are the same as the y-axis values for the negative of that time or angle. A more mathematical way of expressing this is:
cos θ = cos −θ
... and:
cos 360t = cos −360t.

This property of a Cosine wave with zero phase leads people to call it an "even function".

If a Sine wave *with zero phase* is reflected in the y-axis, it will not maintain the same shape.



Instead, the reflected half is an upside down version of what the Sine wave should look like. Another way of saying this is that on a Sine wave, the y-axis values for a particular time or angle are the negative of the y-axis values for the negative of that time or angle. A more mathematical way of expressing this is:
sin θ = −sin −θ
... and:
sin 360t = −sin −360t.

This property of a zero-phase Sine wave leads people to call it an "odd function".

For the positive half of a zero-phase Sine wave to be mapped on to the negative half of a zero-phase Sine wave, it would need to be mirrored across the y-axis, and then mirrored across the x-axis. A zero-phase Cosine wave would only need to be mirrored across the y-axis.

You will sometimes see Sine referred to as an "odd *wave*" and Cosine referred to as an "even *wave*", which mean the same thing.

You will often see people use the terms "odd function" and "even function", and the terms can sometimes be useful depending on the situation. When it comes to Sine and Cosine waves, the terms are only relevant or useful *if the waves have zero phase*. When reading about odd and even functions, you will frequently see the terms "Sine wave" and "Cosine wave" used without explicitly stating "with zero phase", even though that is what is meant.

Some people attach much more significance to whether a function is an even one or an odd one than I think is worthwhile.

Circle

To save time and make everything easier to read, I will often use the term "circle" to mean an "object rotating around a circle". I may say that a circle has a frequency, but really, I mean that the object rotating around the circle has a frequency. It makes sentences a lot shorter and succinct to use the term "circle", so I am stretching the definition of a circle to include an object rotating around a circle.

Always remember that one circle represents both a Sine wave and a Cosine wave. The circle created from a Sine wave is the same as the circle created from the Sine wave's corresponding Cosine wave. There is no such thing as a "Sine circle" or "Cosine circle" – there are only circles that represent both at the same time.

Circle chart

I will call the graph that shows the circle, "the circle chart". Sometimes, the shape being drawn will not be a circle, but I will still call the chart, the "circle chart". Later on in this book, this name might vary depending on how we think about the circle. For now, because it is a chart showing a circle, the name "circle chart" is the most appropriate name.

Helix chart

I will call the three-dimensional graph that shows the helix, "the helix chart". Sometimes, the shape being drawn will not be a helix, but I will still call the chart, the "helix chart".

Magnitude

In this book, I try to treat amplitude as something that is best kept positive, so that the idea of amplitude is consistent with the idea of the radius of a circle. If a radius can only be positive, then an amplitude should only be positive. I am doing this to make learning easier. When learning, this is a good thing to do – it keeps things simple, and means that when you see a negative amplitude, you will remember how to convert it into a positive amplitude, and so have a better understanding of waves. Most people do not worry about using negative amplitudes because the way that they learnt about waves did not emphasise the connection between waves and circles to the extent that I am trying to do in this book. There are many situations in the study of waves where it helps to allow the idea of negative amplitudes, whether you are learning or not.

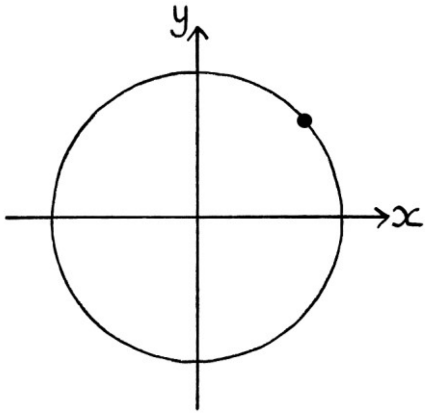
The word "magnitude" is essentially another word for "amplitude", but the difference being that magnitude is always positive – even for people who like to use negative amplitudes. Magnitude is the absolute value of the amplitude. For example, if the amplitude of a wave is −5 units, then the magnitude is +5 units. If the amplitude of a wave is +5 units, then the magnitude is also +5 units.

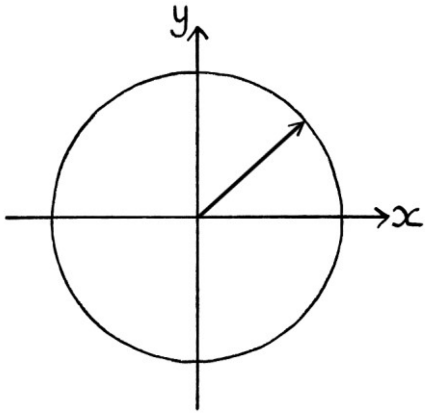
If a wave formula has a positive amplitude, then its magnitude will be the same value, as will be the radius of the circle from which it is, or could have been, derived. If a wave formula has a negative amplitude, its magnitude will be the positive of that amplitude, as will be the radius of its circle.

Vectors

A vector is the name for a line for which we pay attention to its length and angle. A vector is still "just a line", but we make a note of its characteristics. Since we are making a note of its angle, we also know its direction, in the sense of where it starts and where it ends. Vectors can be added to each other, in which case, they are placed end to end, and the result is the straight line from the start of the lines to the outermost point of the joined lines. This resulting straight line has its length and angle noted, and is another vector. Using vectors, or thinking in terms of vectors, can simplify some concepts, while making other concepts unnecessarily more complicated.

We can describe the position of the phase point on a circle by stating the length and angle of a line going from the circle's centre out to the phase point. In this sense, we would be using a vector. The length of the vector would be equal to the radius of the circle, and the angle of the vector would be the angle of the phase point. The picture on the left shows the phase point of a circle; the picture on the right shows a vector indicating exactly the same place:





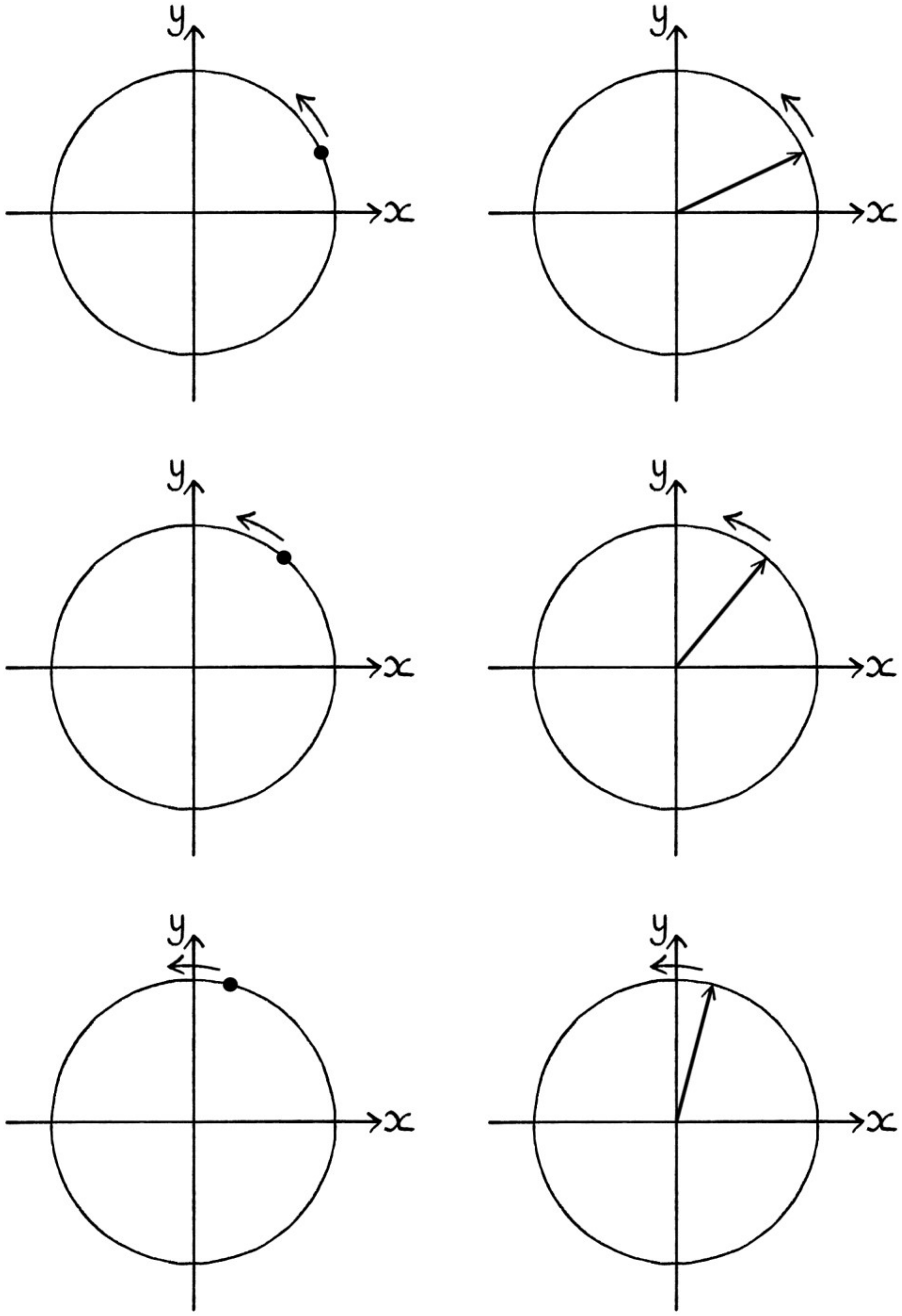
We can make the use of vectors slightly more complicated – given that we can use a vector to identify the position of the phase point of a circle, we can use the same vector to identify the amplitude and phase of the Sine wave or Cosine wave derived from that circle. In the same way that the amplitude of a wave is the same as the radius of the circle from which it is derived, and the phase of a wave is the same as the angle of the phase point, so is the amplitude of a wave equal to the length of the vector, and the phase of a wave equal to the angle of the vector. [The length of a vector is usually called its "magnitude".]

In most explanations of waves, vectors are used much more than they will be in this book. In this book, it will not matter if you do not understand what a vector is. When you have read this book, other people's use of vectors to describe wave-related matters will be much more straightforward.

Phasors

A phasor is a rotating vector. In other words, it is a line for which we keep a note of its length and angle, but its angle is always changing because the line is rotating. Usually, the idea of a phasor can be difficult to grasp because it is a complicated-sounding term, and it is slightly abstract. In our examples of an object rotating around a circle, we could draw a (rotating) line from the centre of the circle out to the *moving* object, and that line would be a phasor. Its angle at any one time would be the angle of the object at that moment in time. Its length would be the radius of the circle. In explanations that do not emphasise objects rotating around circles as much as we do in this book, a phasor seems a much more complicated idea. However, a phasor is really just a line that connects the centre of the circle to the object rotating around the circle. [Later on when we have shapes instead of circles, a phasor connects the centre of the shape (or the origin of the axes) to the object moving around the outline of the shape.]

In the following pictures, the circles on the left have an object rotating around them, and the circles on the right have a phasor indicating the position of that object.



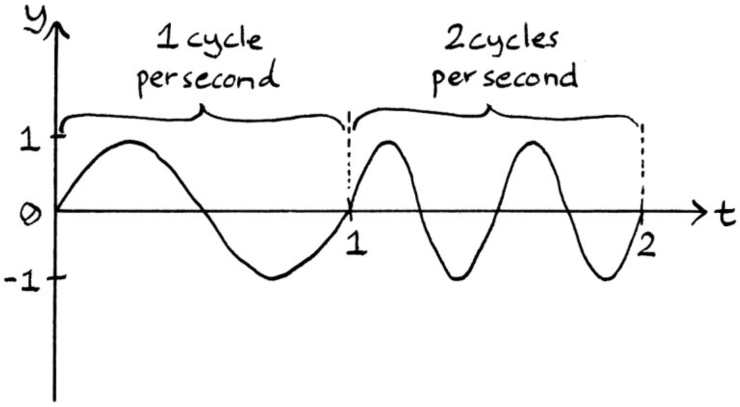
In this book, we will focus on objects rotating around circles, but many other explanations focus on phasors, which are essentially another way of describing exactly the same thing. When you have read this book, other people's use of phasors will be relatively straightforward if you remember that phasors just describe the line connecting the centre of the circle to the object rotating around the circle.

Chapter 11: More about frequency

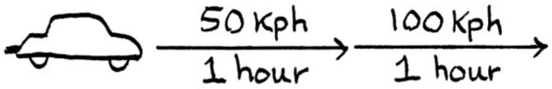
In this chapter, we will look at slightly more complicated concepts relating to frequency.

Varying frequency

In the same way that a vehicle does not have to travel an entire journey at the same speed, neither does an object rotating around a circle have to rotate at the same frequency all the time. For example, this Sine wave graph represents an object moving around a circle at 1 cycle per second for the first second, and then at 2 cycles per second for the next second:



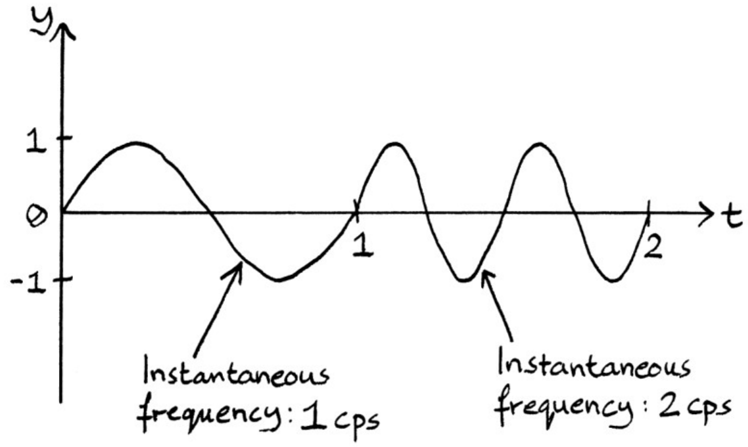
This is analogous to a vehicle travelling at, say, 50 kilometres per hour for the first hour, and then 100 kilometres per hour for the second hour.



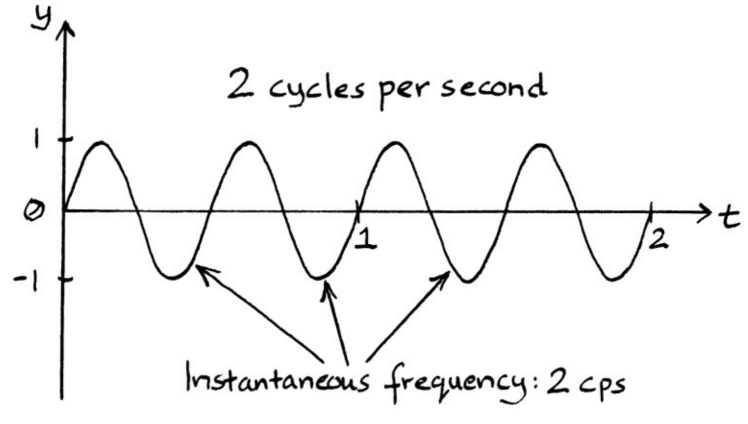
The concept of varying frequency is easy to understand, but it makes things much more complicated when analysing waves and signals. Signals that vary in frequency will be seen when we use waves to encode information much later in this book.

Overall and instantaneous frequency

Supposing we were looking at a Sine or Cosine wave based on an object that varied in speed as it travelled around a circle, it might be necessary to identify the frequency at a particular moment in time. In that case, we can use the term "instantaneous frequency". Instantaneous frequency is the frequency at a particular point. For example, in the graph from before, where the first second has a frequency of 1 cycle per second, and the next second has a frequency of 2 cycles per second, we can identify the instantaneous frequencies at particular moments using arrows:



For an object that rotates around a circle at a fixed frequency, the instantaneous frequency at any point will be the same as the overall frequency. As an example, if a wave has a constant frequency of 2 cycles per second, the instantaneous frequency at any point in time will also be 2 cycles per second.



If this idea is not clear, then imagine a vehicle travelling at 50 kilometres per hour for a length of time. At any moment in its journey, its instantaneous speed will be 50 kilometres per hour.

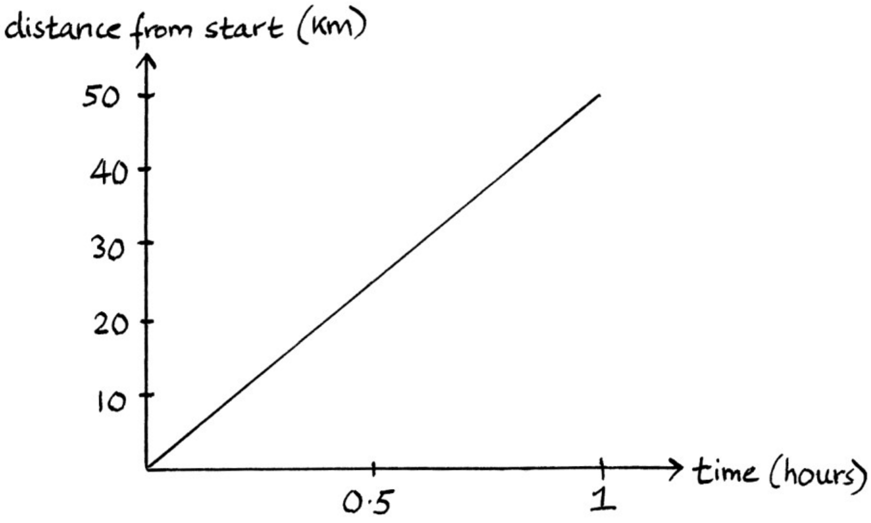
While the use of the term "instantaneous frequency" is not particularly useful in the above examples, if we have an object that rotates at ever changing frequencies, and, therefore, we have constantly changing frequencies in wave graphs, it becomes a convenient phrase.

More on instantaneous frequency

At first glance, the concept of instantaneous frequency seems obvious, but as with many things to do with waves, on closer inspection there is much more to it.

Instantaneous frequency is the frequency of an object at any particular moment in time. For an object that rotates around a circle at a fixed rate for its entire journey, its frequency at any moment in time will be the same as its overall frequency. This means that the frequency of the derived Sine and Cosine waves at that moment in time will also have the same frequency as the overall frequency of the waves.

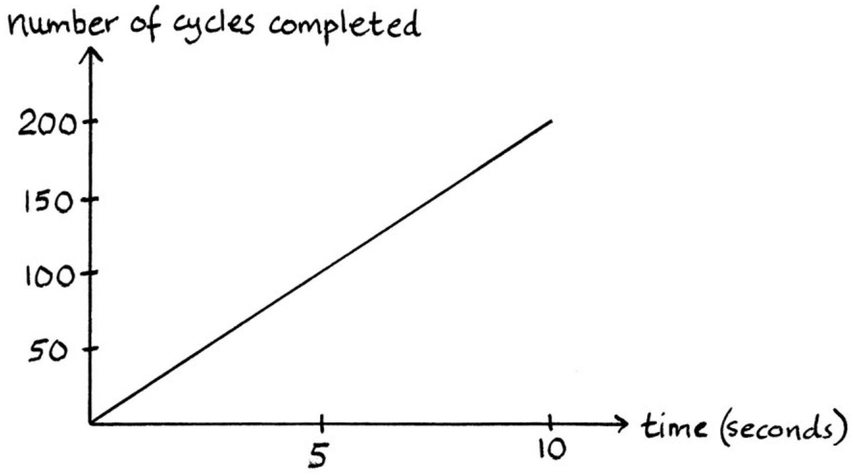
This and other ideas are best explained by imagining a vehicle moving at a fixed speed over its entire journey. We can draw a graph with the y-axis as "distance from the starting point" and the x-axis as time. If the vehicle travels for an hour, and it travels at 50 kilometres per hour, then the graph will look like this:



The overall speed of the vehicle is the complete distance it travels divided by the time it takes to do it. Therefore, its overall speed is 50 kilometres per hour.

The vehicle's speed at any one moment in time will also be 50 kilometres per hour because it maintains the same speed for the whole journey.

We can portray the equivalent ideas for an object rotating around a circle in the same way. We can draw a graph with the y-axis as "the number of cycles completed" and the x-axis as the time. If the object rotates at 20 cycles per second, and it travels for 10 seconds, the graph will look like this:



The overall frequency of the object is the total number of cycles it completes divided by the time it takes to complete them. Therefore, its overall frequency is 20 cycles per second. The object's frequency at any one moment in time will also be 20 cycles per second because it keeps the same frequency for the whole journey.

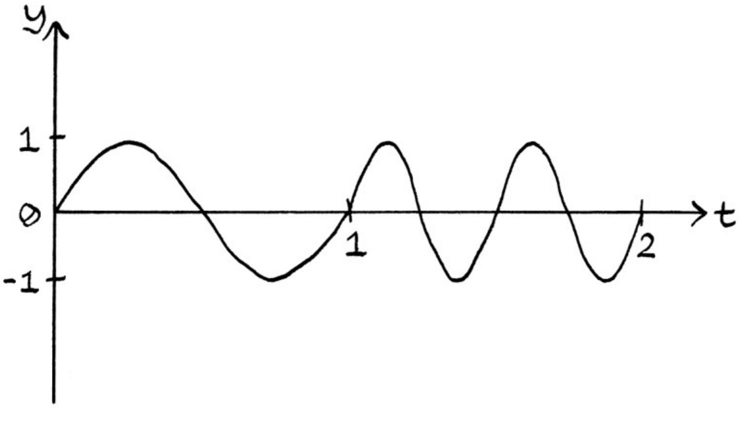
Going back to the speed of the vehicle, supposing we measured the time and distance at *one* particular point on the graph (for example at 25 kilometres and 30 minutes), we would not be able to calculate the speed of the vehicle because to calculate the speed we have to know the *change* in distance and *change* in time. One value is not enough to calculate a change. It would be equivalent to having a photograph of a vehicle and trying to use that to calculate the speed. In reality, we need at least two values to see the difference in time and distance. If we had the values "25 kilometres at 30 minutes" and "25.8333 kilometres at 31 minutes", then we could calculate the speed: the distance travelled in that time is 0.8333 kilometres, and the time taken is 1 minute. Therefore the speed is 0.8333 ÷ (1 ÷ 60) = 50 kilometres per hour.

If we define speed as the change in distance between two times, divided by the difference in those times, we have to have at least two values to calculate the speed of the vehicle. When looking at the graph itself, we can tell by how it is a straight line that the speed is constant over the journey, and therefore at any particular point, the speed should be the same. However, if we were actually *measuring* from the graph, we would need at least two points from the graph.

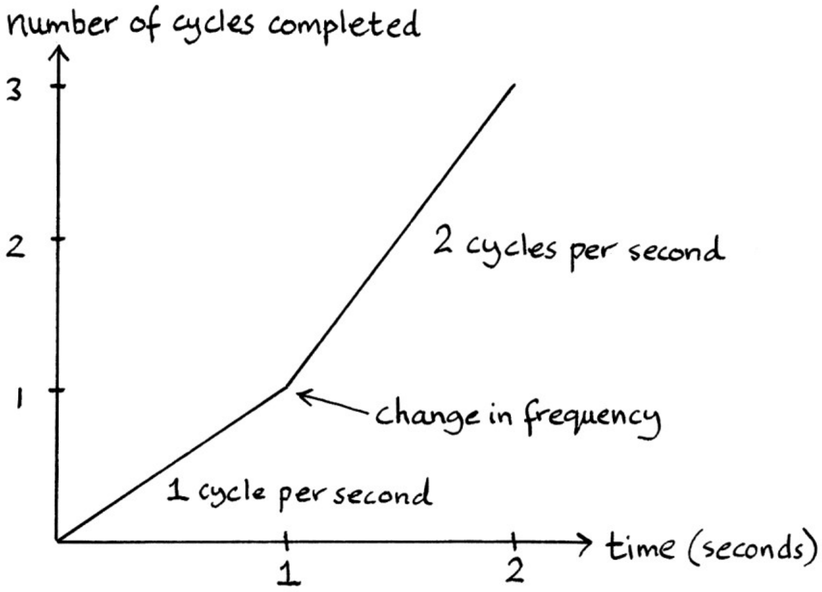
All of this leads to the idea that the concept of the speed of the vehicle at any particular *moment* in time could, if we were being philosophical and pedantic about it, be meaningless. Reading a single point from the graph is equivalent to freezing time itself, and therefore the concept of speed in such a situation does not make any sense, as there can be no movement or change in time if time is frozen. On a graph, a single point on its own could have come from any curve, and would not be sufficient to know the route of the curve through that point.

From all of the above, we could say that the idea of speed as a concept at one individual moment in time is meaningless. We can measure the speed by taking measurements at smaller and smaller intervals, but a calculation of speed still requires two times. In this way of thinking, “instantaneous speed” is really just “average speed” over an infinitesimally small time. In the real world, we might say that literal instantaneous speed is meaningless, but, as with so much in the real world that cannot exist, the concept is valid in the world of maths. In much the same way that people do not worry about negative numbers not existing in reality, it can make sense not to worry about instantaneous speed not existing in reality. [Many people would say that negative numbers *do* exist in reality, which shows how far maths has entered mainstream thinking.]

All of the above is also true for the frequency of an object rotating around a circle. Frequency is the number of cycles completed per second. We can put this more pedantically as, “frequency is the number of cycles completed between two moments in time, divided by the difference in those times.” Given that definition, “instantaneous frequency” would have to refer to the average frequency between two very close moments in time. In the world of maths, specifically in calculus, literally instantaneous frequency is accepted as a valid idea, and everything is much easier because of it. [I explain calculus in Chapter 30]. How we think of instantaneous frequency becomes more relevant when we have graphs such as the following one, where we might want to know the frequency at the time when the curve changes from 1 cycle per second to 2 cycles per second:



While the graph is in the 1 cycle per second section, the instantaneous frequency will always be 1 cycle per second, and while the graph is in the 2 cycle per second section, the instantaneous frequency will always be 2 cycles per second. However, the frequency at the point where the frequency changes from 1 to 2 is harder to discern. It is easiest to think about this by drawing a graph of cycles completed against the time in seconds:



If we read two points either side of the frequency change, we can calculate the average frequency for that duration. For example, at 0.9 seconds, 0.9 cycles have been completed, and at 1.1 seconds, 1.2 cycles have been completed. Therefore, the frequency between these times is (1.2 − 0.9) ÷ (1.1 − 0.9) = 0.3 ÷ 0.2 = 1.5 cycles per second.

We can measure closer times to see if we obtain a more accurate measurement. At 0.99 seconds, 0.99 cycles have been completed, and at 1.01 seconds, 1.02 cycles have been completed. Therefore, the frequency between these times is (1.02 – 0.99) ÷ (1.01 – 0.99) = 0.03 ÷ 0.02 = 1.5 cycles per second again. There is no advantage in reading closer times in this example. We can say that the instantaneous frequency at the moment of the frequency change is 1.5 cycles per second.

Technically, what we are doing here is calculating the gradient of the graph between two points. The gradient between two points gives us the frequency.

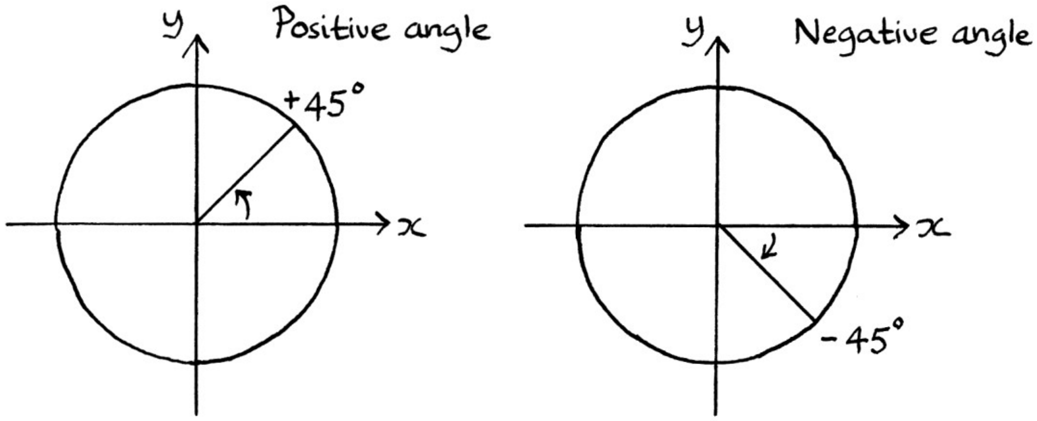
Confusions: frequency and instantaneous frequency

The term "instantaneous frequency" is used to mean the frequency at a particular moment in time. For a wave that has a constant frequency, the instantaneous frequency at any moment and the overall frequency will be the same. Sometimes, people can be unclear as to whether they are referring to overall frequency or instantaneous frequency, and use the term "frequency" for both ideas. The context should help us decide which is intended. Similarly, people might use words other than "instantaneous frequency", which mean the same thing.

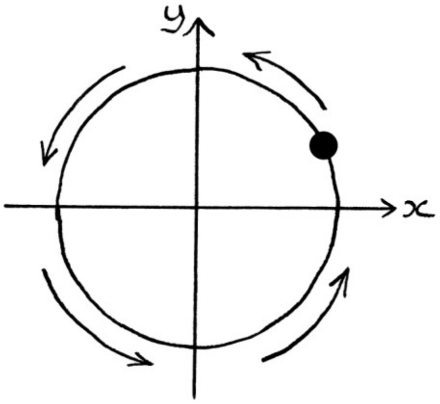
Negative frequency

Frequency, as I have explained before, refers to the number of cycles per second that an object rotates around a circle anticlockwise. *Negative* frequency refers to the number of cycles per second that an object rotates around a circle *clockwise* – in other words, in the "wrong" direction.

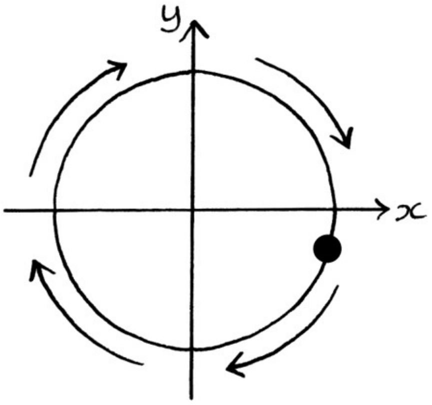
By convention, angles are treated as being positive if they increase when measured anticlockwise around the circle from zero degrees. They are treated as negative if they increase when measured clockwise around the circle.



Frequency relates to how fast an object moves around the circle. Given that we generally think about objects moving around circles as moving at ever-increasing angles, and that those angles increase anticlockwise, frequency relates to how fast an object moves around a circle anticlockwise.



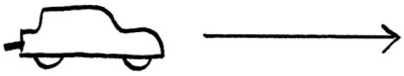
If an object moves around the circle *clockwise*, it will still have a frequency, in that it will still complete so many cycles per second, but the movement will be in the opposite direction to normal. It will move around at ever-*decreasing* angles, so can be said to have a *negative* frequency.



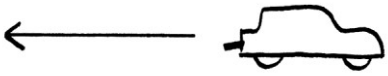
In this sense, the negative aspect is really an indicator to show it is moving in the opposite direction to the generally considered “proper direction”. The “proper direction” is an arbitrary social construction, but given that we live in a society where people generally treat angles as increasing anticlockwise, it is consistent to consider an object moving clockwise as moving in the “wrong direction” and thus having a negative frequency.

In the same way that frequency is analogous to the speed of a vehicle, so is negative frequency analogous to the speed of a vehicle moving backwards. To be more pedantic, it is analogous to the velocity. Negative frequency is the same basic idea as negative velocity. If we say that a vehicle has a positive velocity if it moves in one direction, then we could say that it has a negative velocity if it moves in the opposite direction.

Positive velocity:



Negative velocity:



Negative frequency is a much harder concept to understand if you are only introduced to it using Sine or Cosine waves. Given that Sine and Cosine waves are derived from circles, it makes sense to consider negative frequency as ultimately relating to circles.

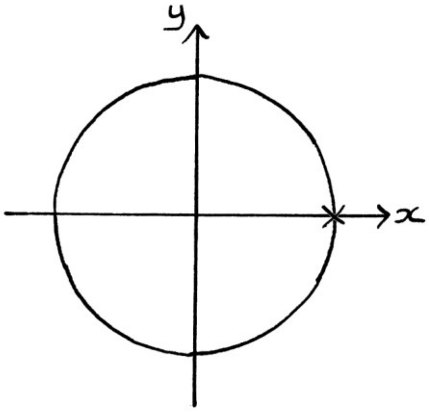
Now we will look at how negative frequency on the circle appears in the waves derived from that circle.

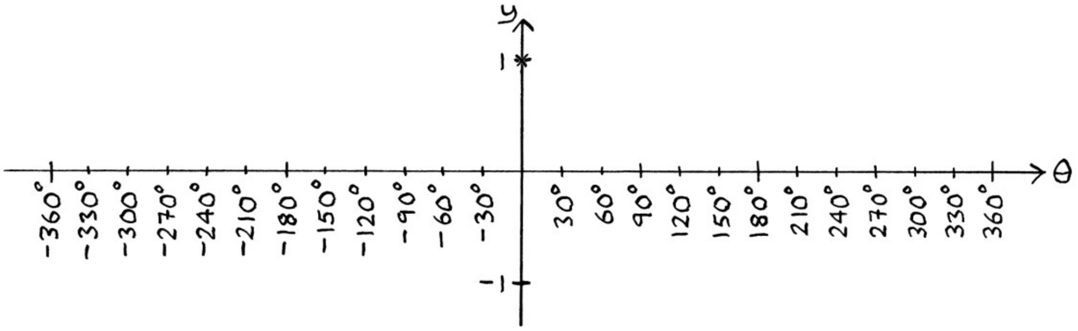
Negative-frequency Cosine

Negative-frequency angles

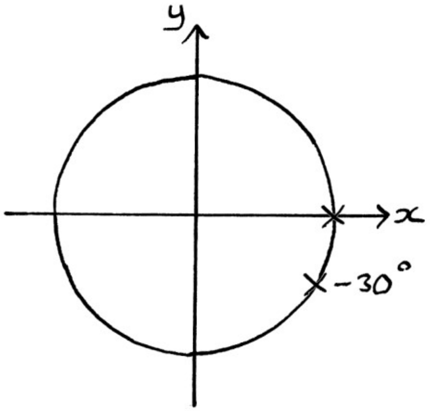
To show what a negative-frequency Cosine wave looks like, we will first look at an angle-based Cosine wave with zero phase. As there is no time component, we will really just be describing ever-decreasing angles around a circle.

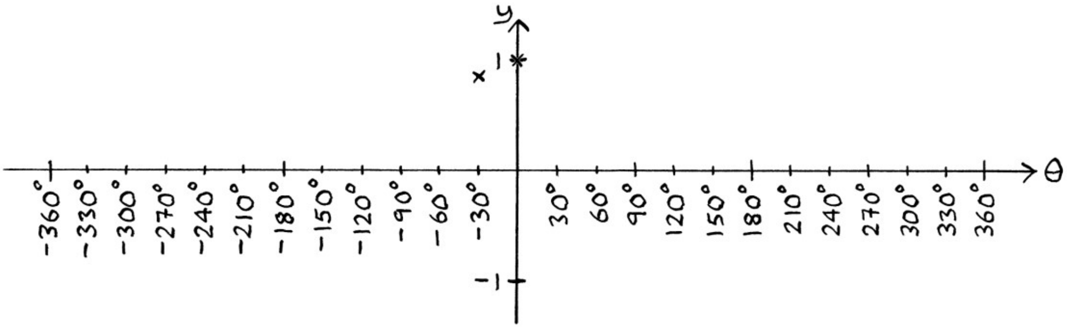
If we plot the values of a Cosine wave for all the decreasing angles of a unit-radius circle, we could start by plotting the x-axis value when the angle is 0 degrees:



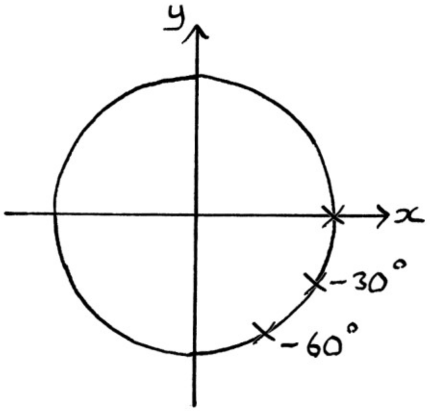


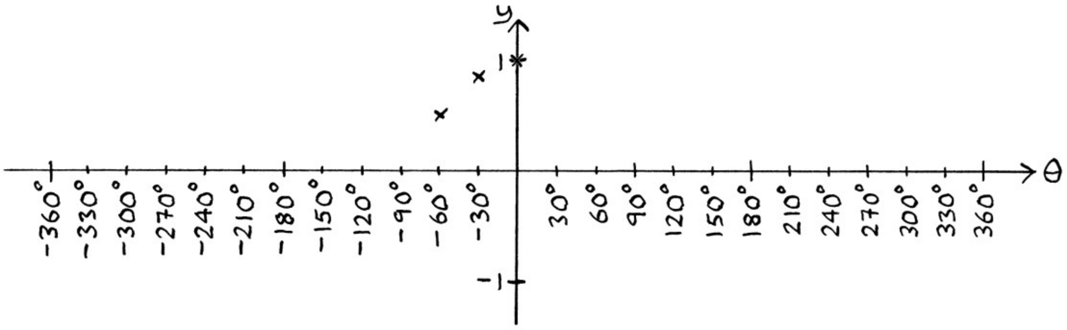
Then we plot the x-axis value for when the angle is −30 degrees:



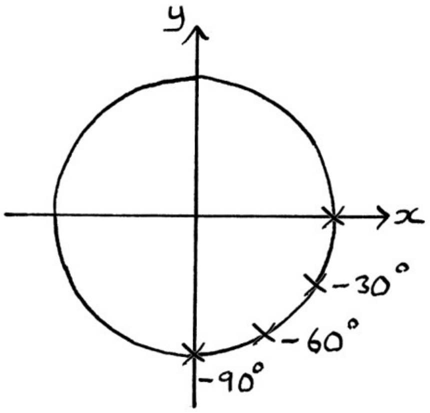


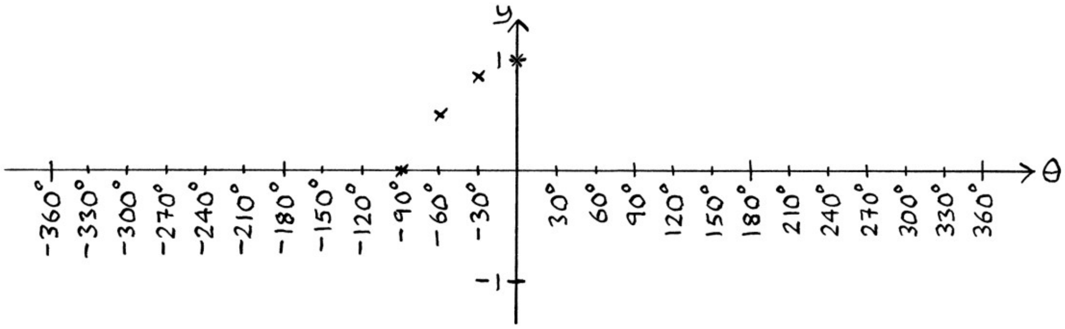
Then −60 degrees:



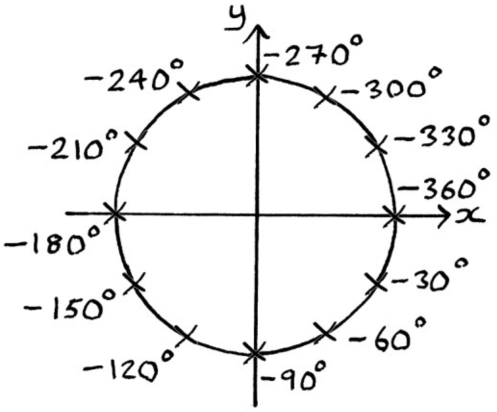


Then −90 degrees:

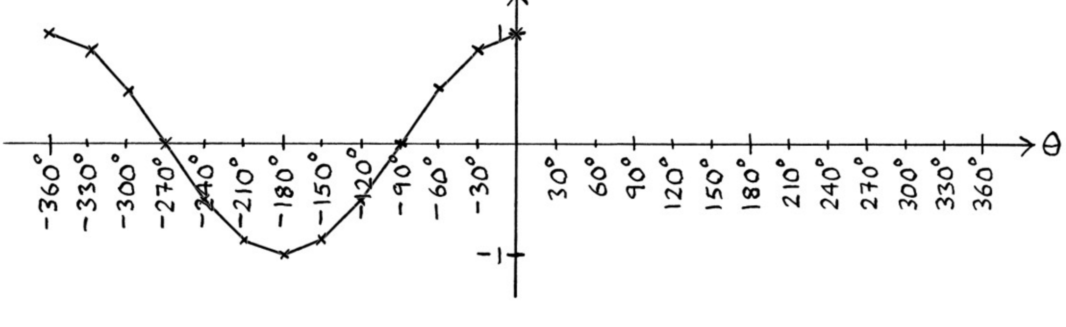




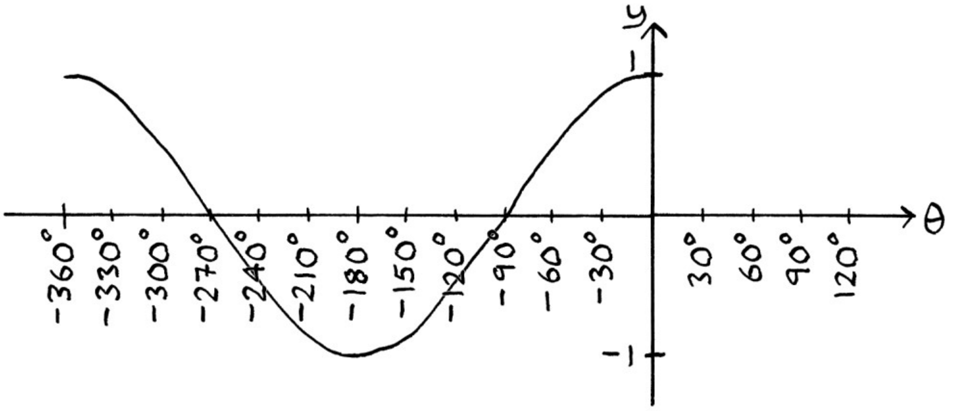
... and so on. Eventually, after reaching −360 degrees on the circle...



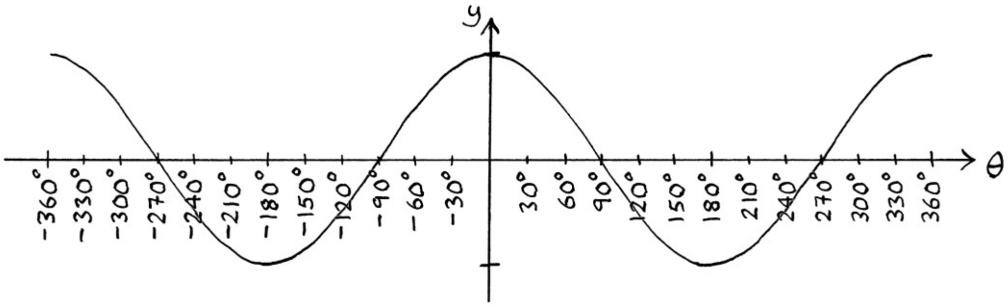
... we would have this graph:



If we had taken many more points from the circle, we would end up with the following smooth graph:



If we consider that −360 degrees on the circle is the same as 0 degrees, and that −259 degrees is the same as 1 degree and so on, we can see that we have just drawn a repetition of the wave, one cycle further to the left along the θ-axis. As we are dealing with the circle, we could have started our graph drawing at +360 degrees, and counted down from that. We could have continued counting at 0 and counted down to −360 degrees. We would have ended up with this graph:

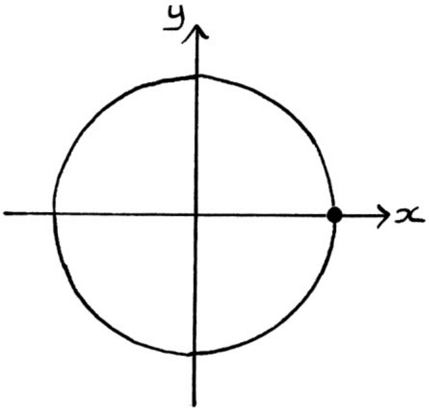


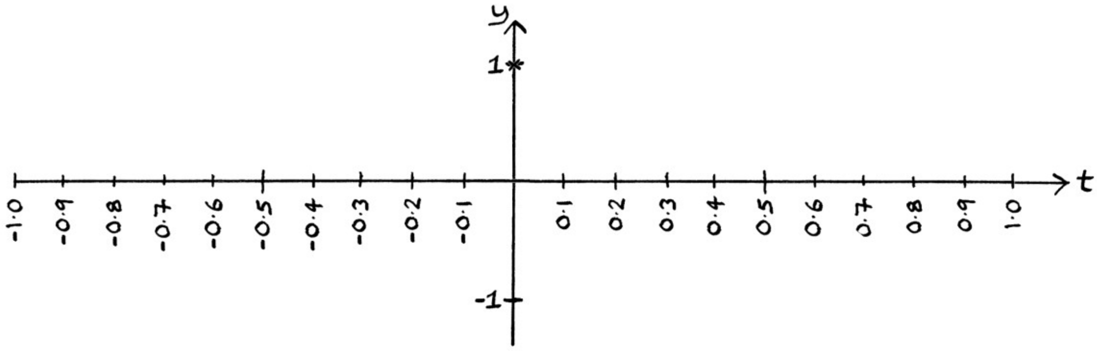
From this, it is clear that the Cosine wave graph for all negative angles is the same as the Cosine wave graph for all positive angles. They are the same thing. This should really be expected, given what we know about waves, but I am portraying it here to emphasise the idea.

Negative frequency time

Now we will redraw the Cosine wave graph with time. Therefore, we will consider an object moving around the unit-radius circle *clockwise*, completing one cycle every second. In every 1/360th of a second, it completes one degree of travel. We will start at t = 0 seconds and measure the position of the object moving around the circle at intervals of thirty 360ths of a second.

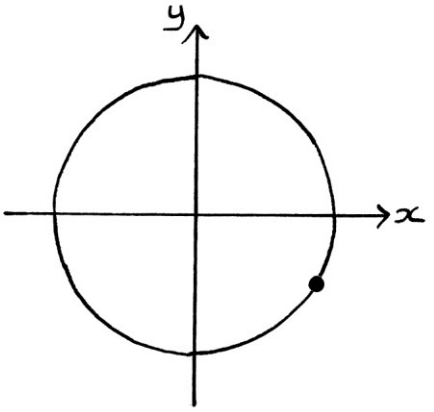
At t = 0, the angle of the object is 0, and the object's x-axis position on the circle's edge is at 1. The circle and the beginning of the graph look like this:

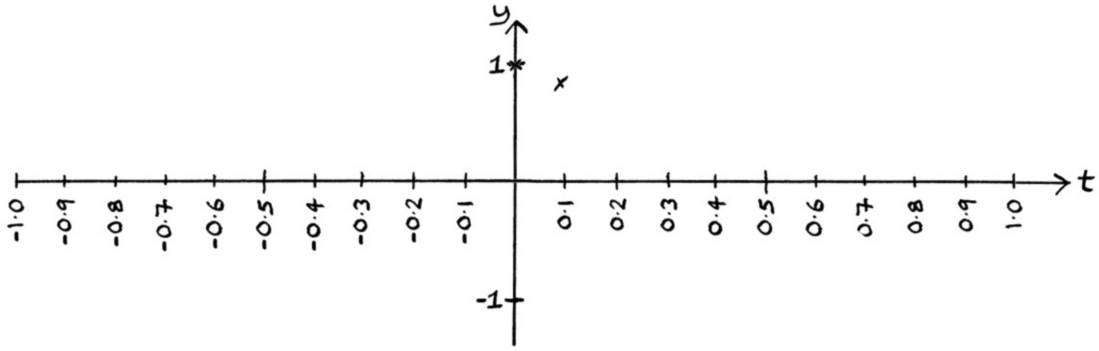




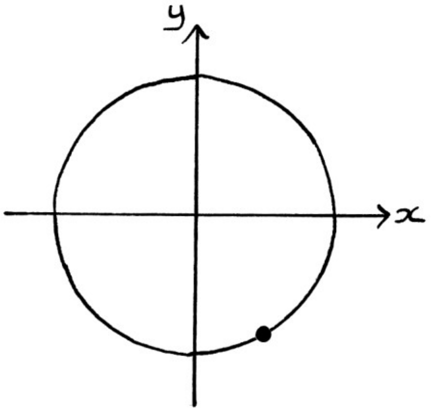
At thirty 360ths of a second, which is 0.08333 seconds, the object will be at −30 degrees. We read off the x-axis value of the object on the circle's edge at this time (cos −30), and plot it on the Cosine time graph for that *time*.

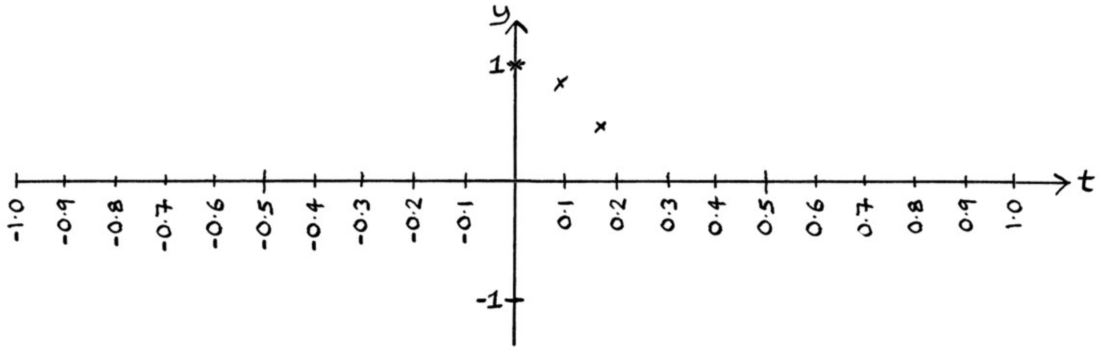
Therefore, we plot the x-axis value of the object at t = 0.08333 seconds.



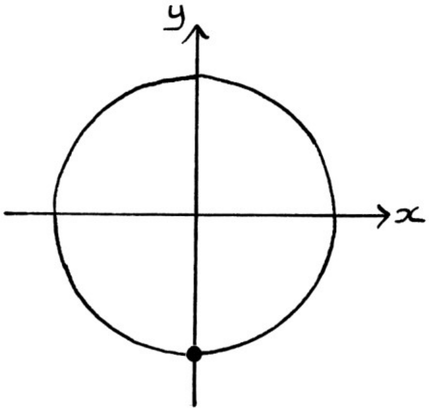


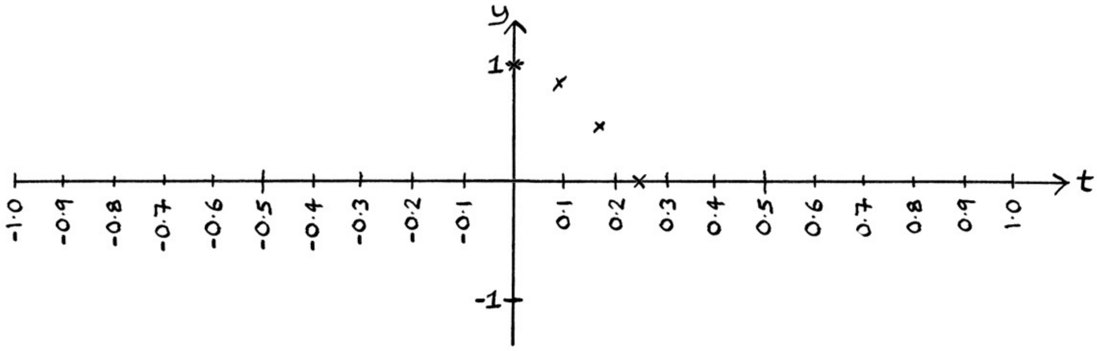
We then do the same for the next time, which is sixty 360^{ths} of a second, or 0.1667 seconds. The angle of the object is −60 degrees. The x-axis value will be "cos −60".



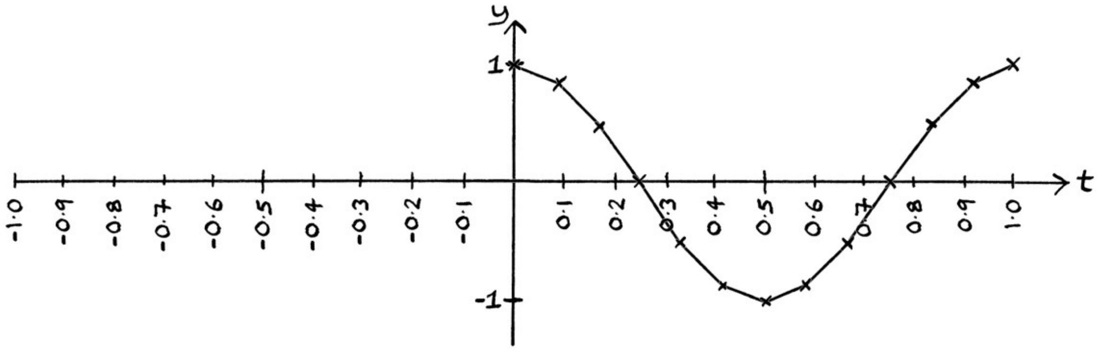


We then do the object's position at ninety 360$^{\text{ths}}$ of a second, which is 0.25 seconds:

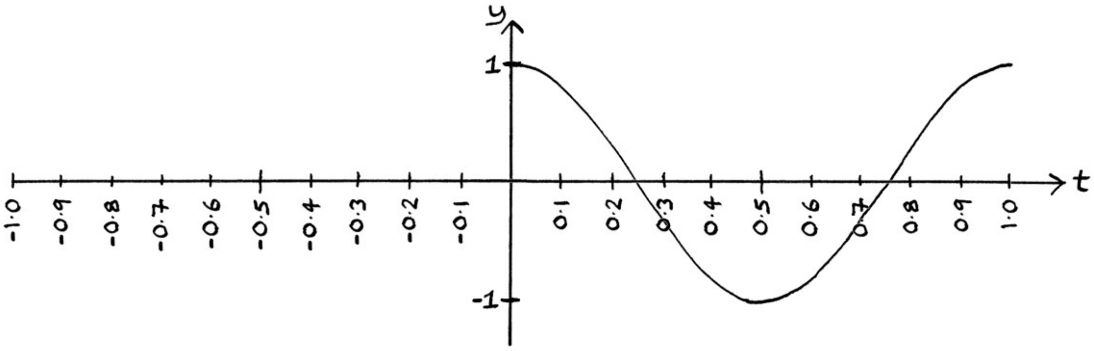




We continue doing this until the object has completed one revolution the wrong way around the circle, and we end up with this graph:

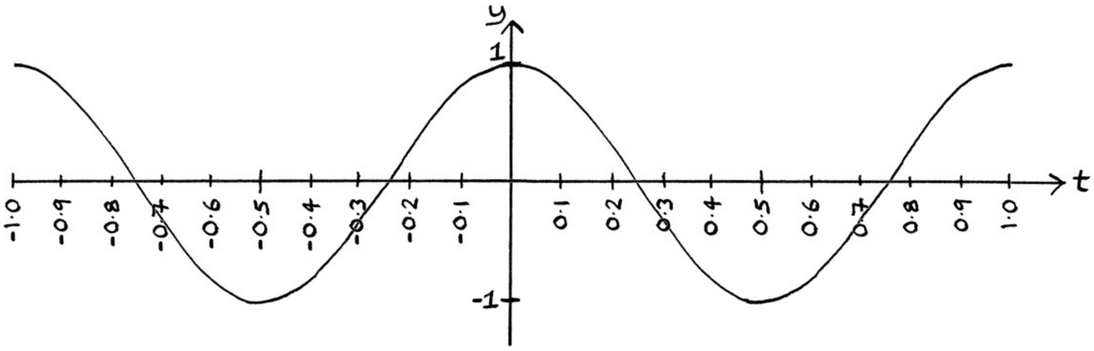


If we had taken more readings, we would have ended up with the following smooth curve:



The Cosine graph for *angles* went in the left-hand direction (down to ever lower negative angles) as it was being drawn, because we were reading off the x-axis value of points around the edge of the circle at angles below 0 degrees, and plotting those points at the corresponding angles on the graph. The Cosine graph for *time* goes in the normal direction (up to ever higher positive times), because we are reading off the x-axis values of the position of an object around the circle at certain points in time, and those times always increase upwards.

If we consider all moments in time, the time graph looks like this:



The angle-based Cosine wave created from reading ever-decreasing angles is identical to the angle-based Cosine wave created from reading ever-increasing angles. The time-based Cosine wave for a negative frequency is identical to the time-based Cosine wave for a positive frequency.

We could have guessed that the Cosine wave for negative frequency would be identical to the one for positive frequency by thinking about the formula for a Cosine wave with zero phase: "y = A cos (360ft)". A negative frequency would mean that the total value being Cosined would be negative. The Cosine of a negative value is the same as the Cosine of that value made positive. We can see this from the graph of a Cosine wave with zero phase – it is mirrored around x = 0. Therefore, the effects of a negative frequency on Cosine with no phase make no difference to the graph.

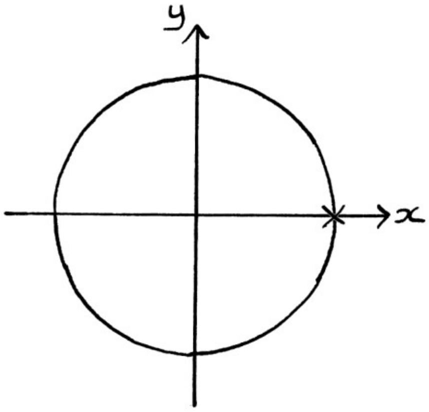
Negative-frequency Sine

Negative-frequency angles

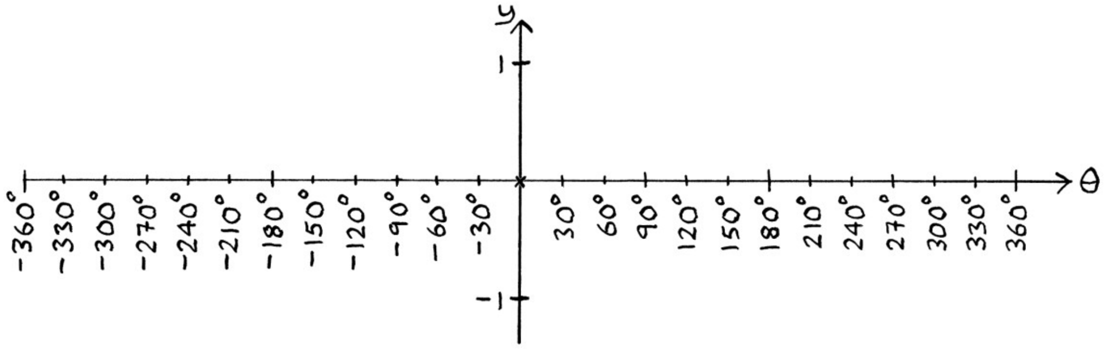
To show what a negative-frequency Sine wave looks like, we will begin by examining an angle-based Sine wave with zero phase.

We will read off the y-axis values for all angles of a unit-radius circle, starting at 0 degrees, moving down to −30 degrees, then −60 degrees, −90 degrees, all the way down to −360 degrees, and plot those on a graph.

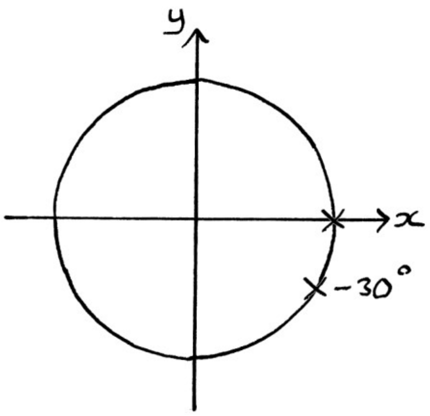
First, we measure the y-axis value on the circle's edge at 0 degrees:



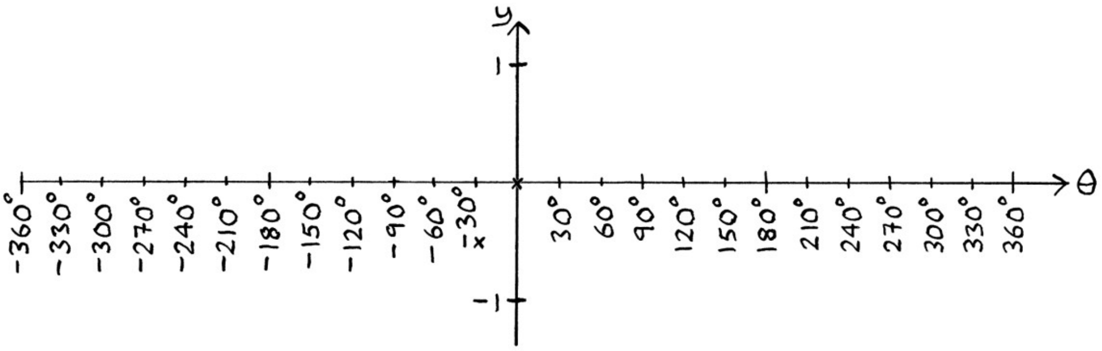
It is zero, and we mark that y-axis value on the graph at 0 degrees:



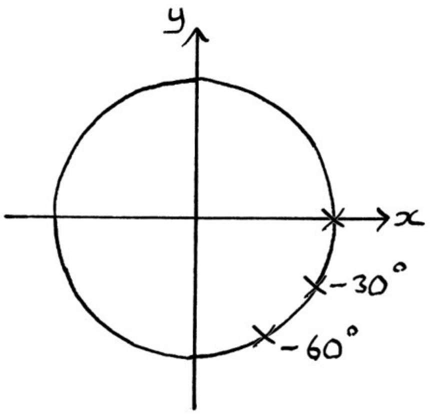
Then, we measure the y-axis value on the circle at −30 degrees:



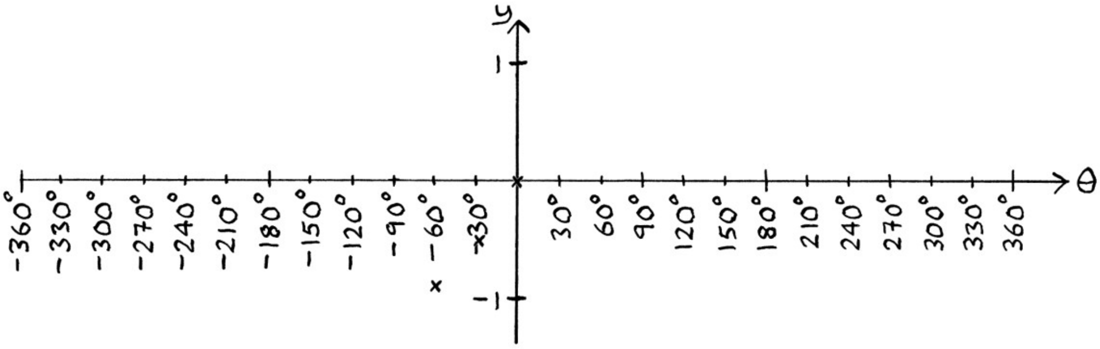
We mark that value on the graph as the y-axis value for when “θ” is −30 degrees:



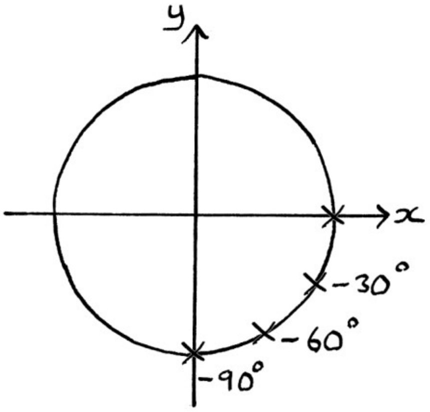
Then, we do −60 degrees:



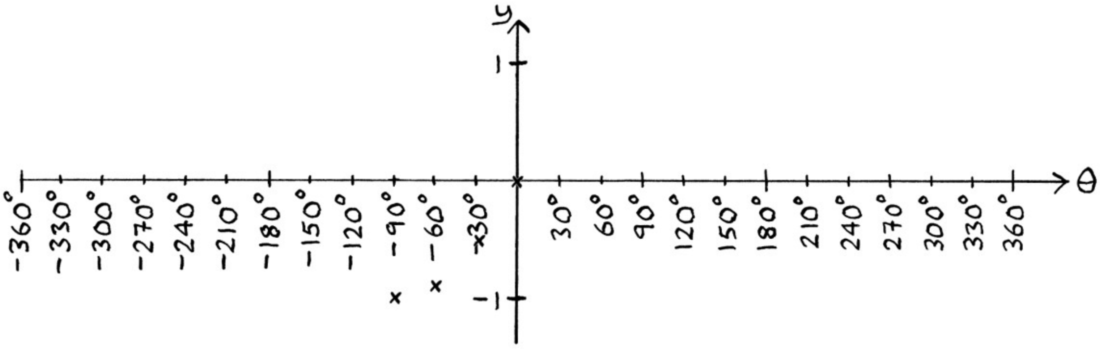
... and we mark that value on the graph:



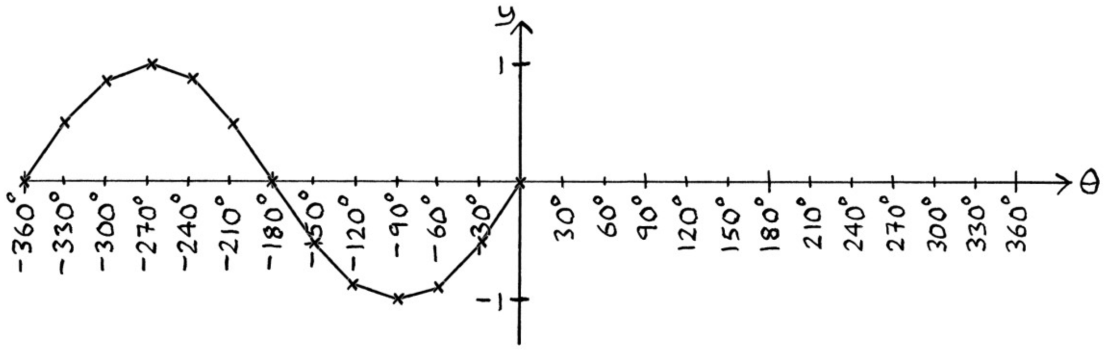
Then, we do −90 degrees:



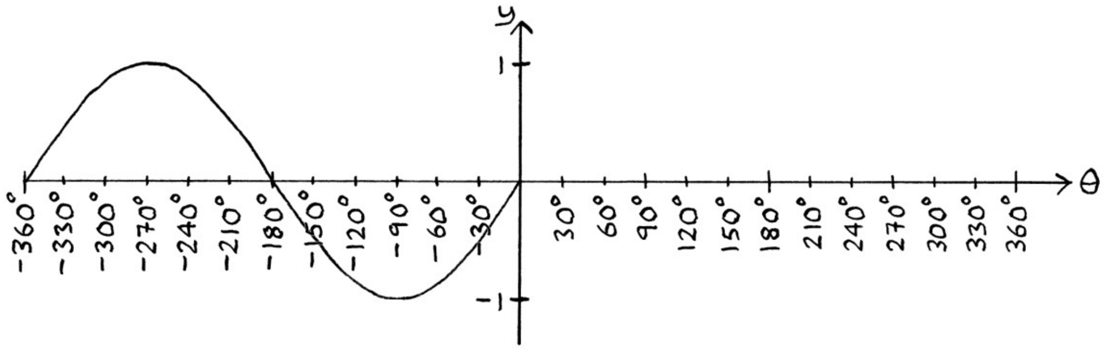
... and we mark that value on the graph:



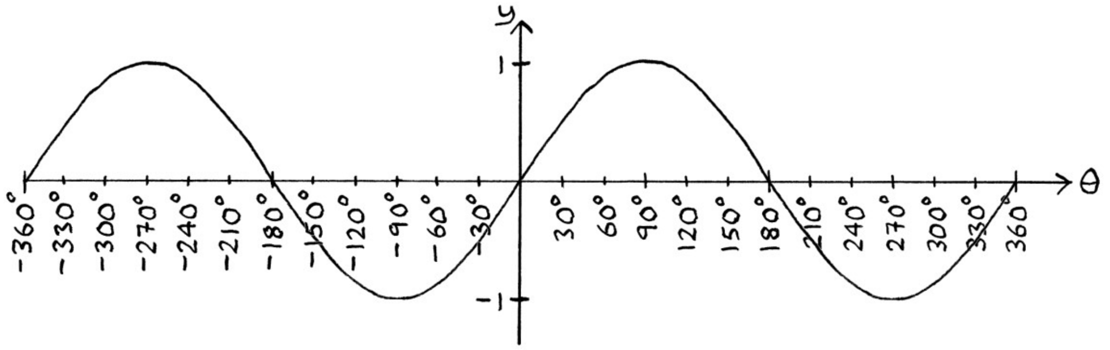
We continue in this way until, eventually, after reaching −360 degrees, we would have this graph:



If we had taken more measurements from around the circle, we would have ended up with the following smooth curve:



As we are dealing with the circle, we could have started our graph by measuring at +360 degrees on the circle, and counted down from that, all the way down to −360 degrees. We would have ended up with this graph:

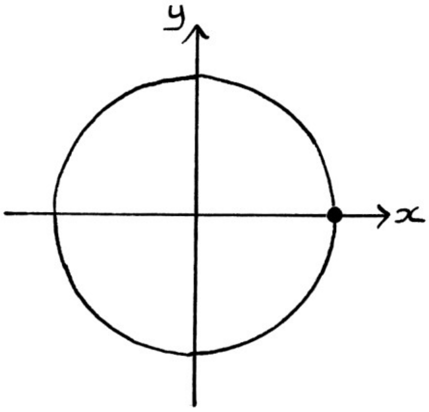


The shape of this graph is identical to the shape of a Sine wave graph on which we had marked ever-increasing angles. This is to be expected. The graph does not really show anything we did not know before. The moment when everything changes is when we draw a Sine wave graph that relates to time.

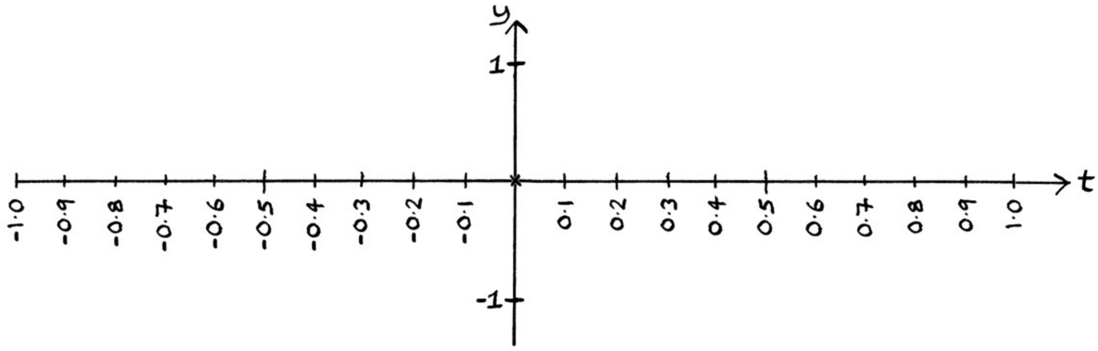
Negative frequency time

As with Cosine and time, for Sine and time, we will consider an object moving around the unit-radius circle clockwise, completing one cycle every second. In every 1/360th of a second, it moves one degree around the circle.

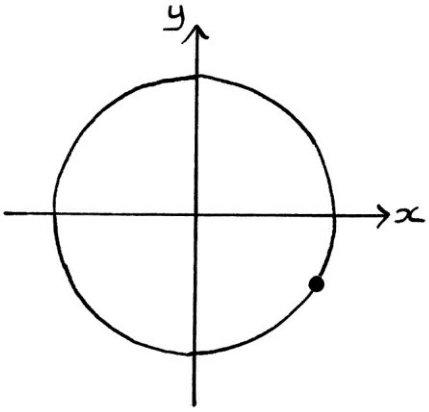
We will draw the graph in the same way. We start at t = 0. At this time, the angle of the object on the circle is 0 degrees, and the height of the object on the circle's edge is 0 units. In other words, the Sine of the angle is zero.



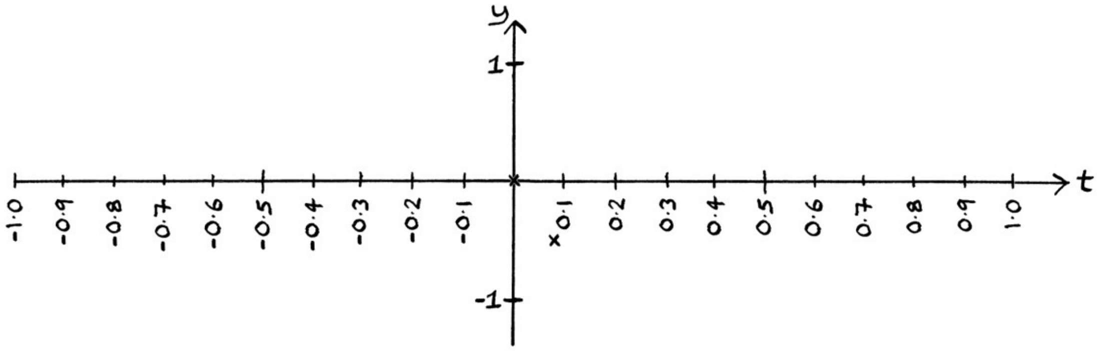
We plot this value at t = 0 on our Sine wave graph.



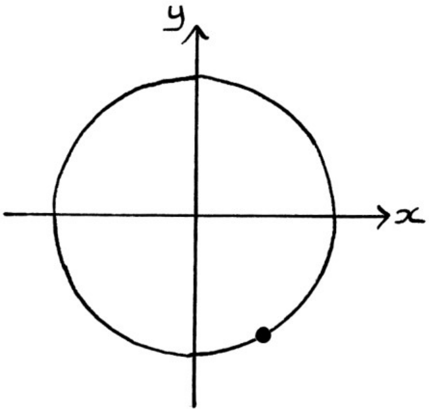
Next, we measure the object's y-axis position at thirty 360ths of a second, which is 0.08333 seconds. The object will be at −30 degrees on the circle, so its y-axis value will be Sine −30 = −0.5:



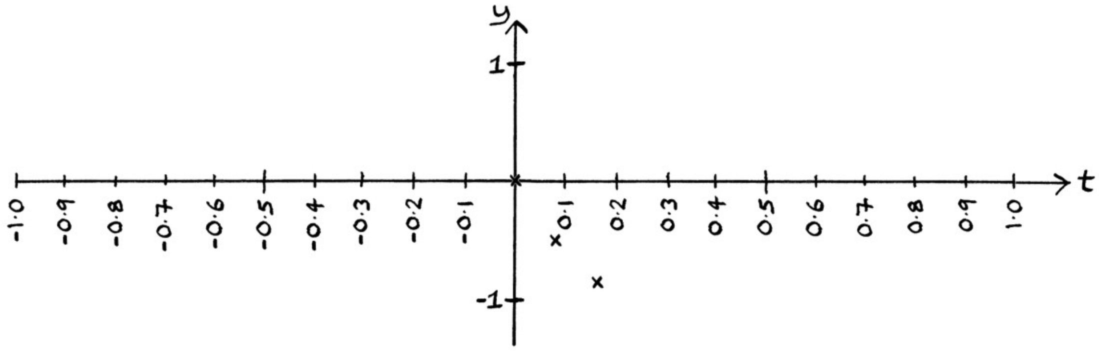
We plot that y-axis value on to the graph for the time of 0.08333 seconds:



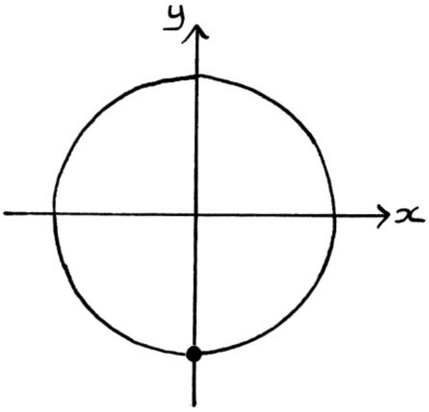
Next, we measure the y-axis position of the object at sixty 360ths of a second, which is 0.1667 seconds. The object is at −60 degrees on the circle, which means it has a y-axis height of −0.8660 units.



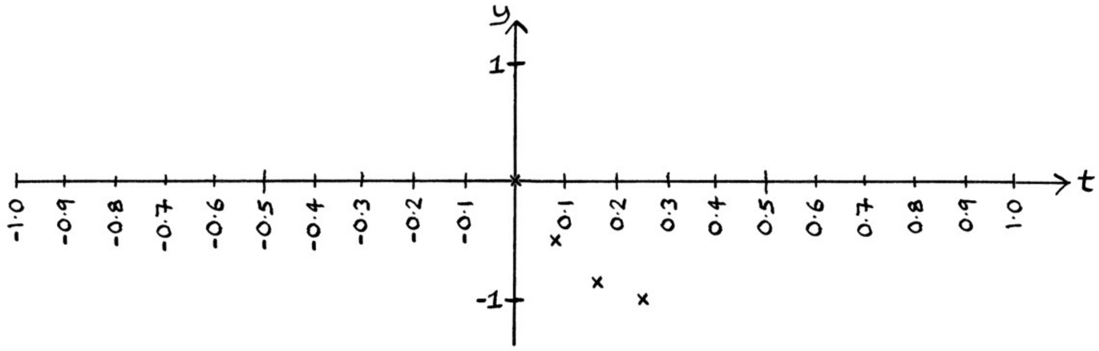
We mark that value on the graph at 0.1667 seconds:



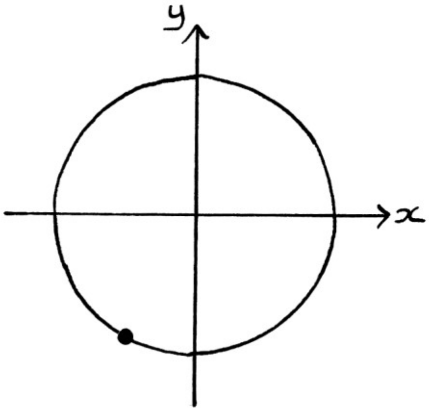
Then we measure the y-axis position of the object after 0.25 seconds. The object will be at −90 degrees on the circle, which means it has a y-axis value of −1.



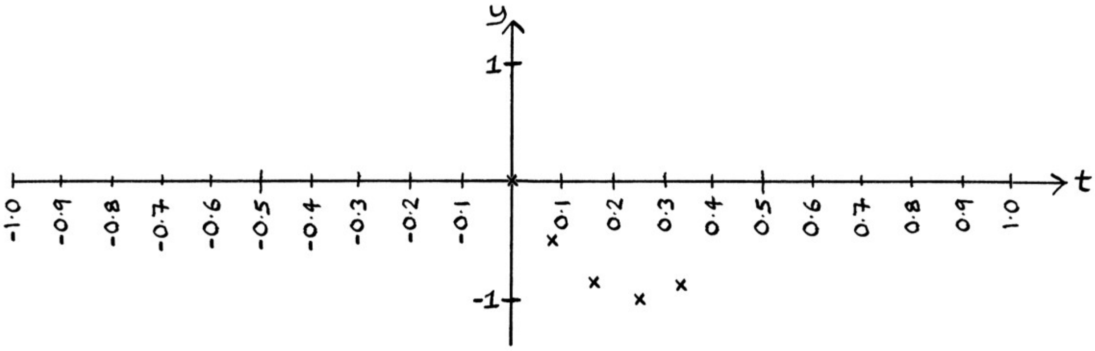
We mark that value on the graph at the time of 0.25 seconds:



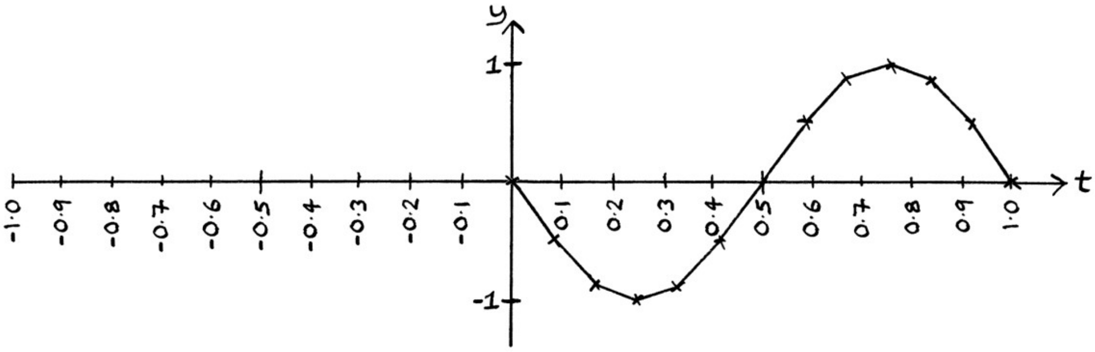
Next, we measure the y-axis position of the object at 120 360^{ths} of a second, which is 0.3333 seconds. Its angle on the circle is −120 degrees, so its y-axis position is −0.8660:



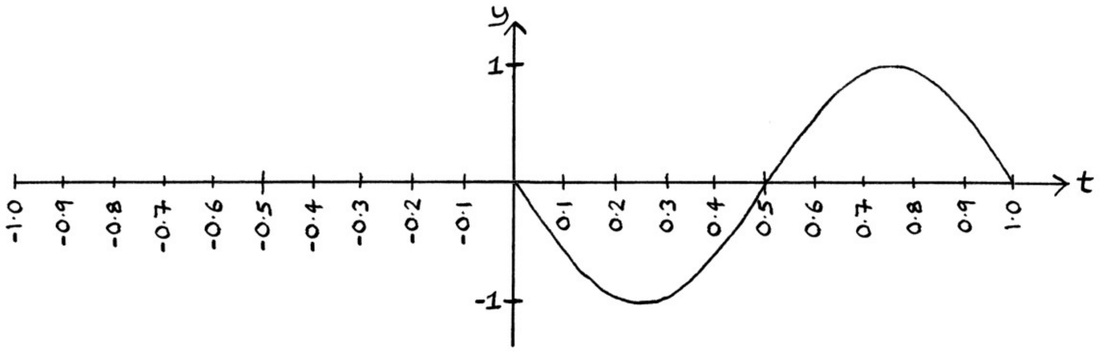
We mark that value on the graph at 0.3333 seconds:



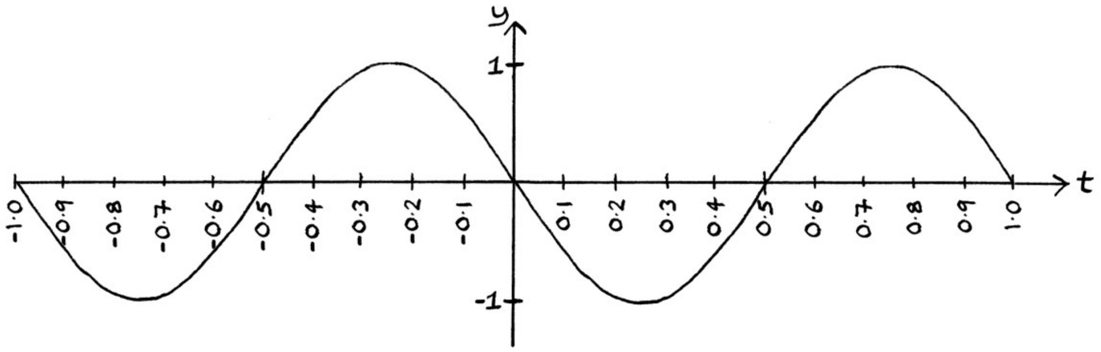
We continue observing the object move around the circle at *ever-decreasing* angles, and we plot the values on our graph at *ever-increasing* times. We continue until we reach 1 second, at which time the object is at −360 degrees. Our final graph will look like this:



If we had measured moments in time that were closer together, we would have had the following smoother curve:



If we extend the graph for all time, we end up with this graph:



As we can see, the negative-frequency Sine wave time graph is an inverted version of the positive-frequency Sine wave time graph. We might have been able to guess this from thinking about the formula for a Sine wave with no phase:
"y = A sin (360ft)"

A negative frequency would have caused the total value being Sined to be negative. The Sine of a negative number is the negative of the Sine of that number made positive, or to put this mathematically:

sin (−x) = −sin (x)

Therefore, the wave curve will be upside down. We could also say that the curve is the same as:

- a positive-frequency Sine wave with a +180 degree phase in its formula.
- a negative-*amplitude*, positive-frequency, zero-phase Sine wave.

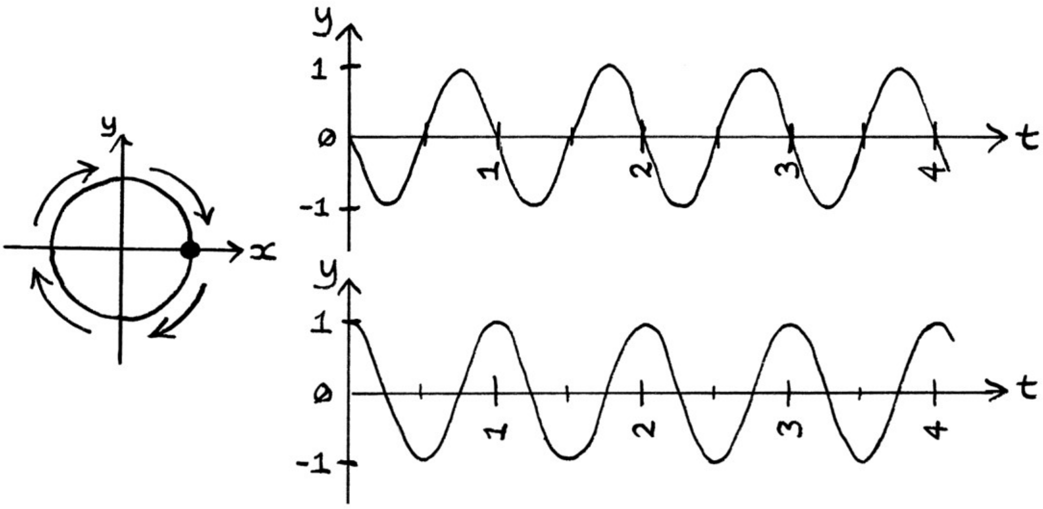
Each of these appears as an "upside-down" Sine wave.

Summary so far

From all of this, we can see that for a time-based Cosine wave with zero phase, negative frequency amounts to exactly the same thing as positive frequency. Thinking about this another way, the x-axis values of an object moving around a circle will be the same at particular moments in time whether that object is moving clockwise or anticlockwise. Given that a Sine wave with a 90-degree phase is the same as a Cosine wave with no phase, it is also true that a time-based Sine wave with a 90-degree phase and a negative frequency is identical to one with a positive frequency.

A zero phase, negative frequency, time-based Sine wave will appear as an inverted version of a zero phase, positive frequency time-based Sine wave. Each y-axis value on the negative-frequency wave will be the negative of the corresponding y-axis value on the positive-frequency wave. We could also say that a zero phase, negative-frequency Sine wave is the same as a positive-frequency Sine wave with a 180-degree phase. Another way of thinking about the idea is that the y-axis values of an object moving clockwise around a circle will be the negative of the y-axis values of an object moving anticlockwise around a circle for particular moments in time.

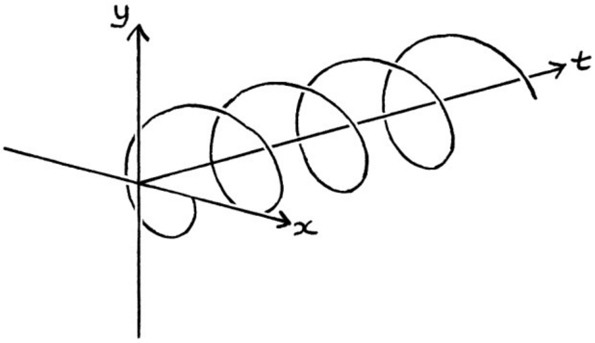
Our object moving around a unit-radius circle with a negative frequency produces the following negative-frequency waves, where the negative-frequency Sine wave is the same as an inverted positive-frequency Sine wave, and the negative-frequency Cosine wave is the same as a positive-frequency Cosine wave:



Negative frequency time helix

An object moving around a circle with a negative frequency, when viewed on the helix chart, would still be moving outwards down the time-axis, but it would be rotating *clockwise* as it did so. This might be obvious, as if an object rotates around the circle clockwise, then it will be rotating away down the helix clockwise too.

If we viewed the negative frequency time helix end on, it would appear as a circle; if we viewed it with the x-axis pointing directly at us, it would appear as an "upside-down" Sine wave graph; if we viewed it from underneath, that is with the y-axis pointing directly away from us, we would see the normal Cosine wave graph.

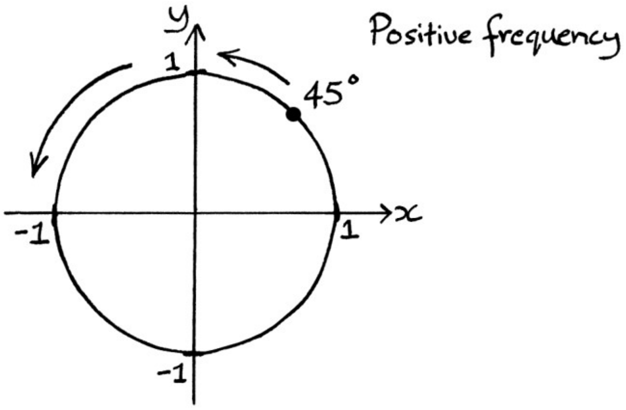


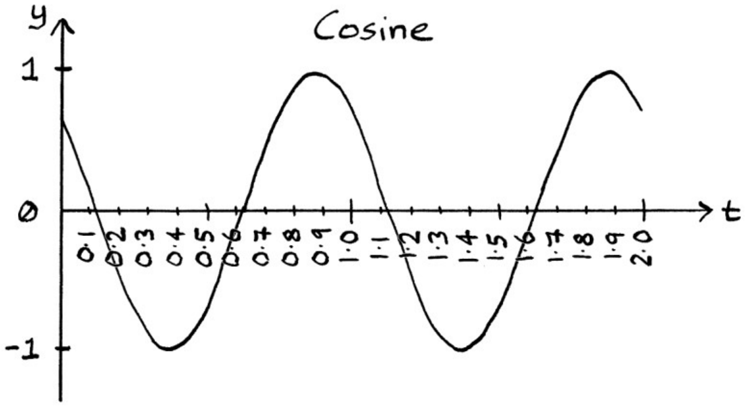
Phase with negative frequencies

So far, we have looked at negative-frequency Sine and Cosine waves with zero phase. When we are dealing with negative frequencies, the phase for an object rotating clockwise around the circle still indicates its starting position. However, because the object is moving in the other direction, a phase that would have given the object a head start if the object were moving anticlockwise, will now cause the object to be delayed by the same amount. Similarly, a phase that would have delayed an object moving anticlockwise, will give the object a head start when it is moving clockwise. This means that the derived Sine and Cosine waves from a negative frequency object with a positive phase will be slid to the *right* in comparison to a negative frequency object with zero phase. [With a positive-frequency Sine or Cosine wave, a positive phase means they have been slid to the *left*].

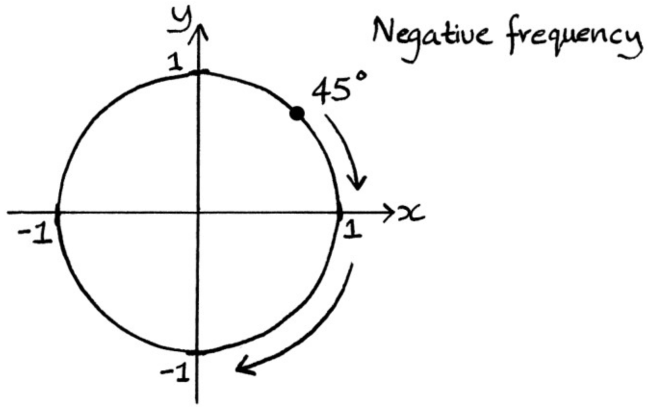
Cosine waves with negative frequency

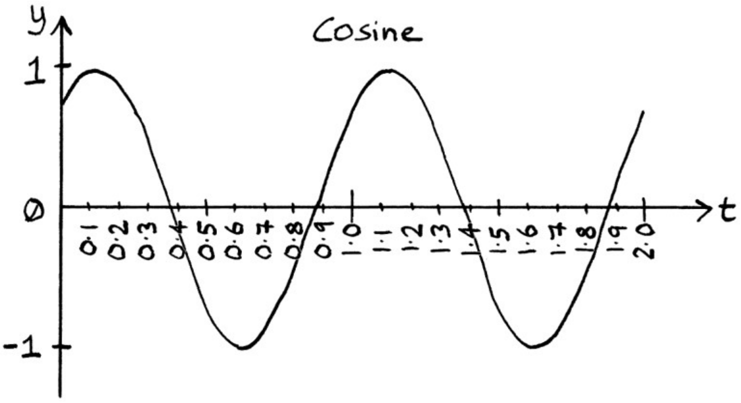
We will say that an object is rotating around a unit-radius circle at 1 cycle per second. The phase point (the starting point of the object) is at 45 degrees. For normal, *positive* frequency, this phase means that the object has a 45-degree (0.125 second) head start. The derived Cosine wave is shifted to the left by 45 degrees (0.125 seconds).





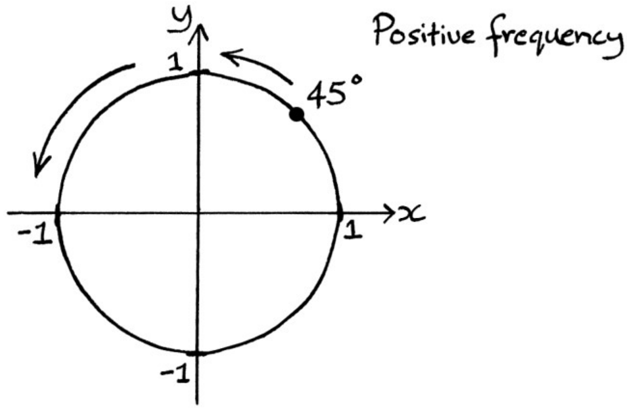
However, for *negative* frequency, this phase means that the wave has a 45-degree delay, or in other words, the object is delayed by 0.125 seconds. The reason for this is that the phase is in the other direction to the direction of movement. Therefore, it takes 0.125 seconds to reach the moment when previously it would have been at 0 degrees. The negative-frequency Cosine wave is shifted to the *right* by 45 degrees (0.125 seconds).

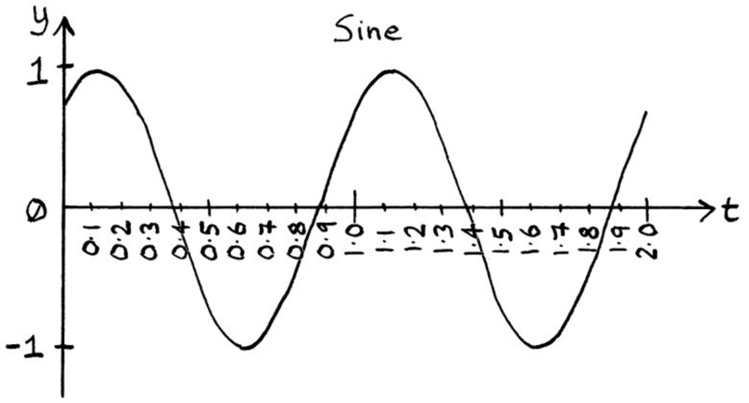




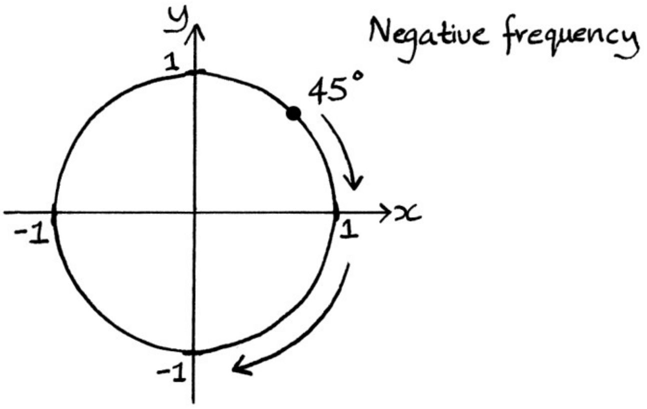
Sine waves with negative frequency

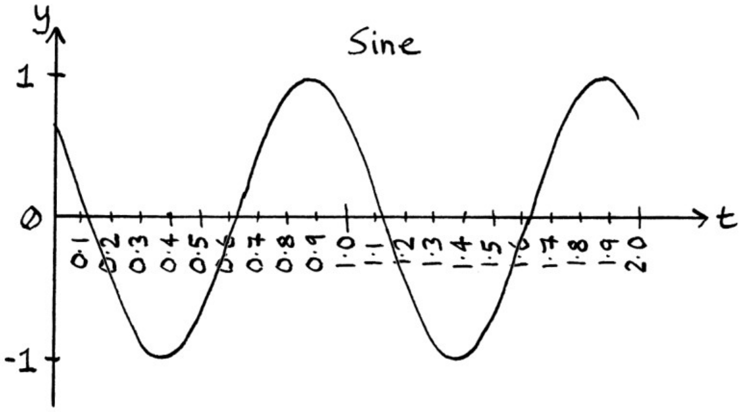
We will use the same circle for the Sine wave example. The phase point is still at 45 degrees. For a normal, *positive* frequency, this phase means that the object has a 45 degree (0.125 second) head start. The derived Sine wave is shifted to the left by 45 degrees (0.125 seconds):





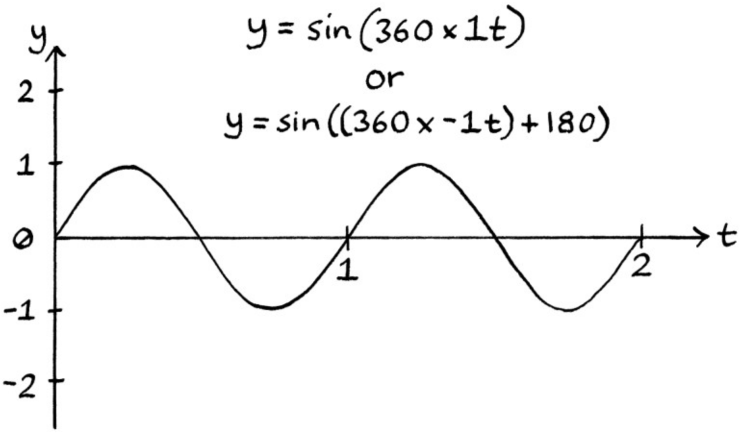
However, for *negative* frequency, this phase means that the object is delayed by 45 degrees (0.125 seconds). It takes 0.125 seconds to reach the point at 0 degrees. The wave is shifted to the right by 45 degrees or 0.125 seconds. Remember also that the negative frequency means that the Sine wave is upside down as well, so there are two factors in play here:



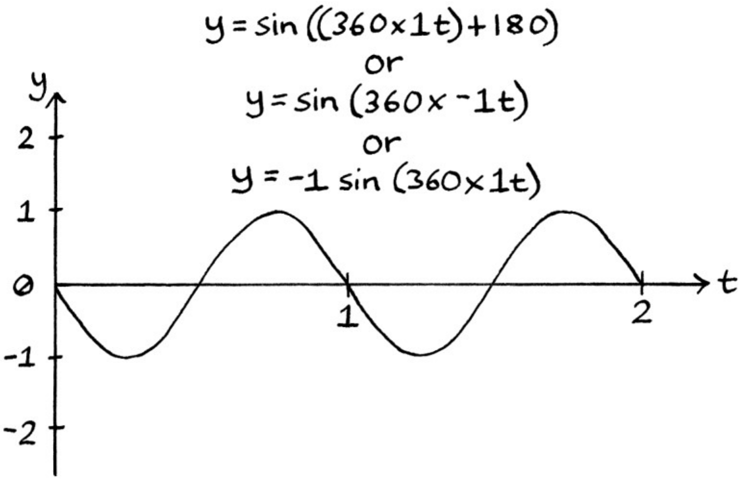


Equivalent formulas for Sine

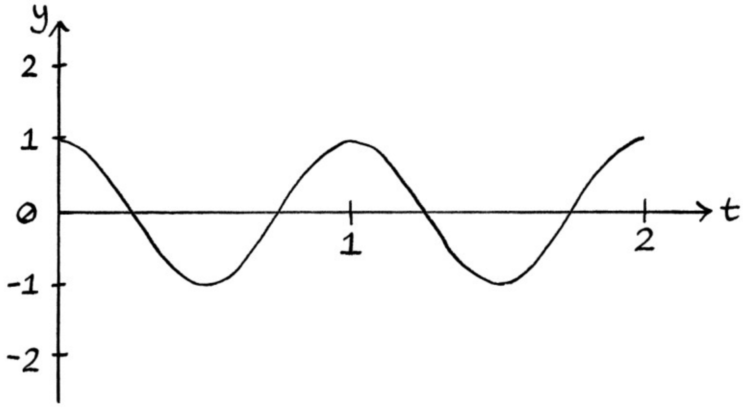
A Sine wave formula with a positive frequency and zero phase refers to exactly the same curve as that same wave formula with a negative frequency and a phase of 180 degrees. For example, the actual curves described by:
"y = sin (360 * 1t)"
... and:
"y = sin ((360 * −1t) + 180)"
... are identical.



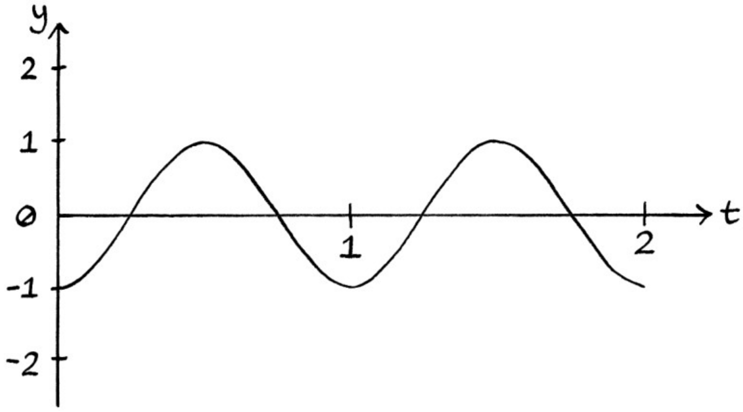
A Sine wave formula with a positive frequency and a 180-degree phase portrays exactly the same curve as that same formula with a negative frequency and a phase of 0 degrees. For example:
"y = sin ((360 * 1t) + 180)"
... is identical to:
"y = sin (360 * −1t)". It is also the same curve as that formula with a positive frequency, a phase of 0 degrees, and a negative amplitude – in other words, "y = −1 sin (360 * 1t)".



A Sine wave formula with a positive frequency and a phase of +90 degrees refers to exactly the same curve as that same formula with a negative frequency and a phase of +90 degrees. This is the same as saying that a Cosine wave formula with a positive frequency and zero phase refers to the same curve as that Cosine wave formula with a negative frequency and zero phase (because a Cosine wave with zero phase is a Sine wave with a 90-degree phase).



A Sine wave formula with a positive frequency and a phase of +270 degrees refers to an identical curve as that formula with a negative frequency and a phase of +270 degrees.



There is a rule for finding the equivalent positive-frequency and negative-frequency formulas:

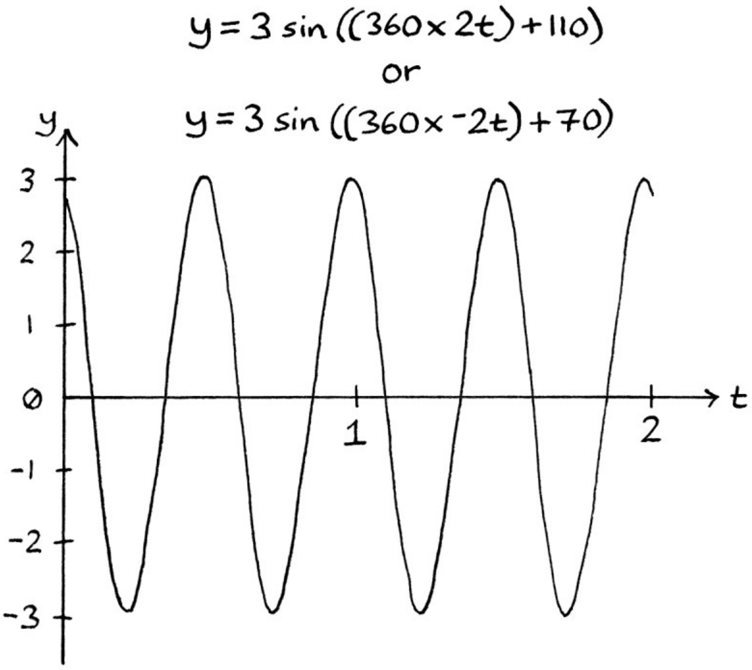
> "A Sine wave formula with a positive frequency and a phase so many degrees *above or below* 90 degrees will refer to the same curve as that formula with a negative frequency and a phase that many degrees *below or above* 90 degrees."

In other words, if we have a positive-frequency Sine wave formula, we see how many degrees away from 90 degrees the phase is, and the negative-frequency formula that refers to exactly the same curve will have a phase that same number of degrees away from 90 degrees in the other direction.

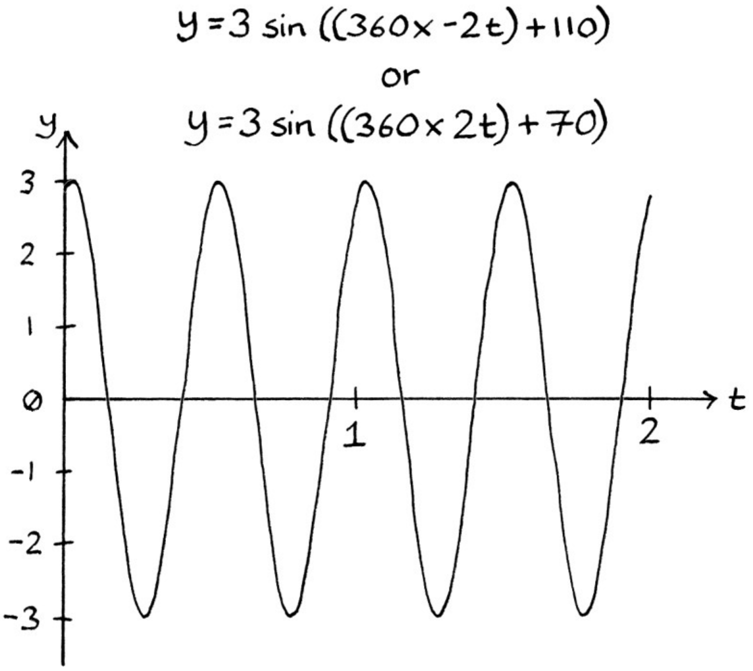
Conversely, if we have a negative-frequency Sine wave formula, we see how many degrees away from 90 degrees the phase is, and the positive-frequency formula that refers to exactly the same curve will have a phase that same number of degrees away from 90 degrees in the other direction.

To put this more mathematically:
"A sin ((360 * f * t) + (90 + x)) = A sin ((360 * −f * t) + (90 − x))"
... and:
"A sin ((360 * −f * t) + (90 + x)) = A sin ((360 * f * t) + (90 − x))"
... where (90 + x) is the phase of the formula for which we want to find the equivalent. [In other words, if the actual phase is 120 degrees, then "x" will be 30, because 90 + 30 = 120].

As an example, we will look at the formula "y = 3 sin ((360 * 2t) + 110)". The phase of 110 degrees is 20 degrees *more* than 90 degrees. Therefore, the phase of the equivalent negative-frequency formula will be 20 degrees *less* than 90 degrees, which is 70 degrees. Its formula will be "y = 3 sin ((360 * −2t) + 70)".



The rule works both ways, so if we have the formula "y = 3 sin ((360 * −2t) + 110)", its phase is 20 degrees *more* than 90 degrees, so the phase of the equivalent *positive*-frequency formula will be 20 degrees *less* than 90 degrees. The equivalent formula will be "y = 3 sin ((360 * 2t) + 70)".



We will look at the formula "y = 11 sin ((360 * 7t) + 330)". The angle of 330 degrees is 240 degrees *above* 90 degrees. Therefore, the equivalent negative-frequency formula will have a phase 240 degrees *below* 90 degrees. This is 90 – 240 = −150 degrees. We make this into a positive phase by adding 360 degrees to it, and we end up with 210 degrees. The equivalent negative-frequency formula will be "y = 11 sin ((360 * −7t) + 210)".

Now we will look at the formula "y = 2 sin ((360 * 4t) + 6)". The angle of 6 degrees is 84 degrees *below* 90 degrees. Therefore, the equivalent negative-frequency formula will have a phase 84 degrees *above* 90 degrees. This is 90 + 84 = 174 degrees. The equivalent formula will be "y = 2 sin ((360 * −4t) + 174)".

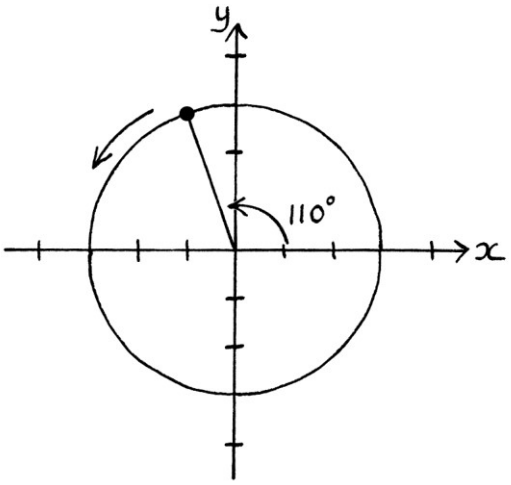
We can also do the calculation from 270 degrees, instead of 90 degrees. Therefore, a second rule that has the same meaning is this:

> "A Sine wave formula with a positive frequency and a phase so many degrees *above or below* 270 degrees will refer to the same curve as that formula with a negative frequency and a phase that many degrees *below or above* 270 degrees."

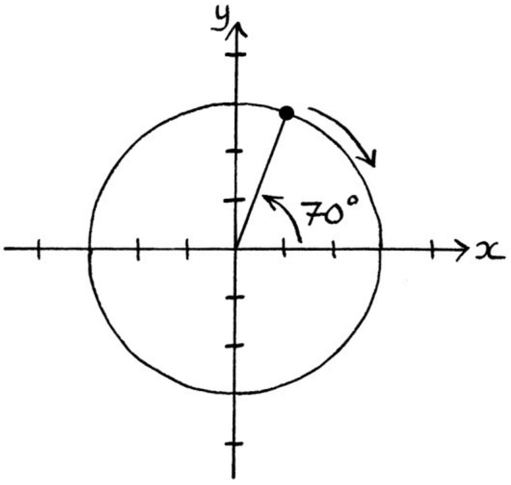
The rules written more mathematically are:
“A sin ((360 * f * t) + (270 + x)) = A sin ((360 * −f * t) + (270 − x))”
... and:
“A sin ((360 * −f * t) + (270 + x)) = A sin ((360 * f * t) + (270 − x))”
... where (270 + x) is the phase of the formula for which we want to find the equivalent.

Using the number 270 can make it easier to do the calculation mentally if the phases are close to 270 degrees; using the number 90 can make it easier if the phases are close to 90 degrees.

The reason why both these rules work makes much more sense when we look at circles. The circle for: “y = 3 sin ((360 * 2t) + 110)” looks like this:

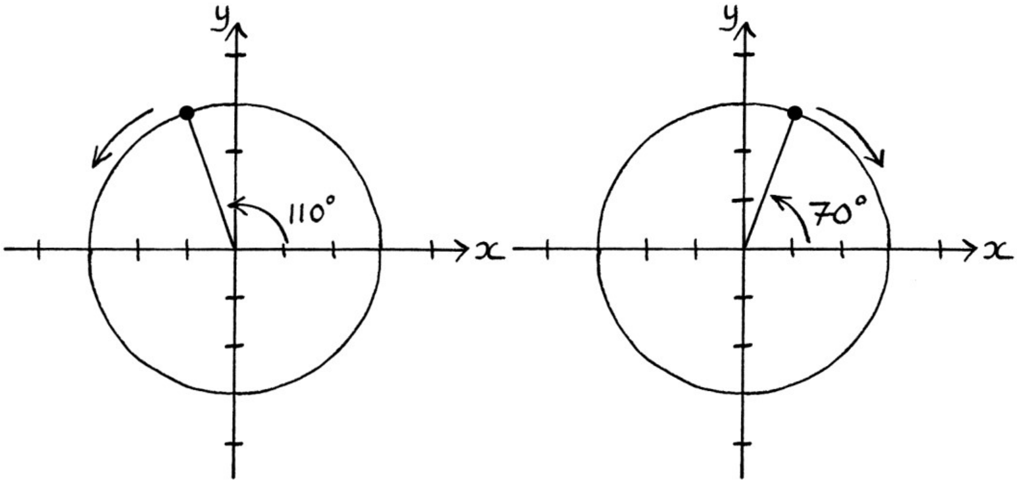


The circle for the equivalent negative-frequency formula:
“y = 3 sin ((360 * −2t) + 70)”
... looks like this:



These are the same circle but mirrored left or right across the y-axis. The phase points stay in the same place on each circle, but are mirrored too.

The two circles look like this, drawn side by side:

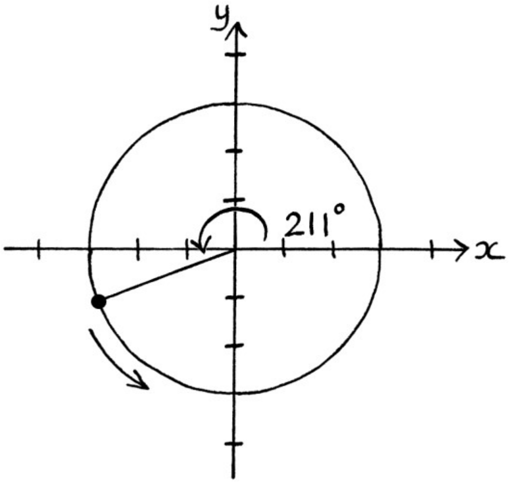


To find the circle that produces the equivalent Sine wave formula with the opposite frequency direction, we look at the mirror image of the circle.

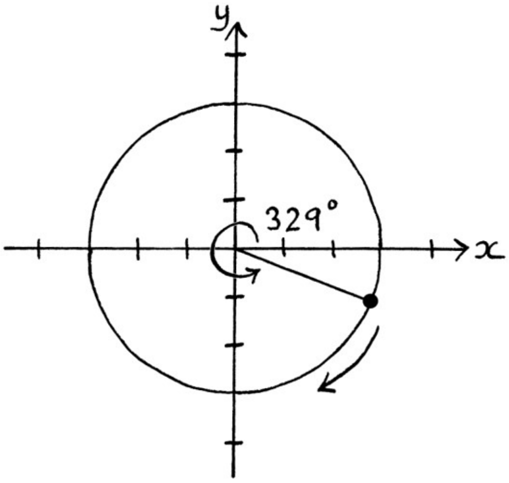
To find the equivalent wave curve with the opposite frequency direction, we look at the mirror image of the circle from which that Sine wave was, or could have been, derived.

Mirroring the circle left or right works for finding both the negative-frequency equivalent of a positive-frequency formula, and for finding the positive-frequency equivalent of a negative-frequency formula. It is much easier to understand the situation, and to perform the maths mentally if we imagine the circles.

As an example, we will look at the formula “y = 3 sin ((360 * 3t) + 211)”. The circle from which this wave is derived looks like this:

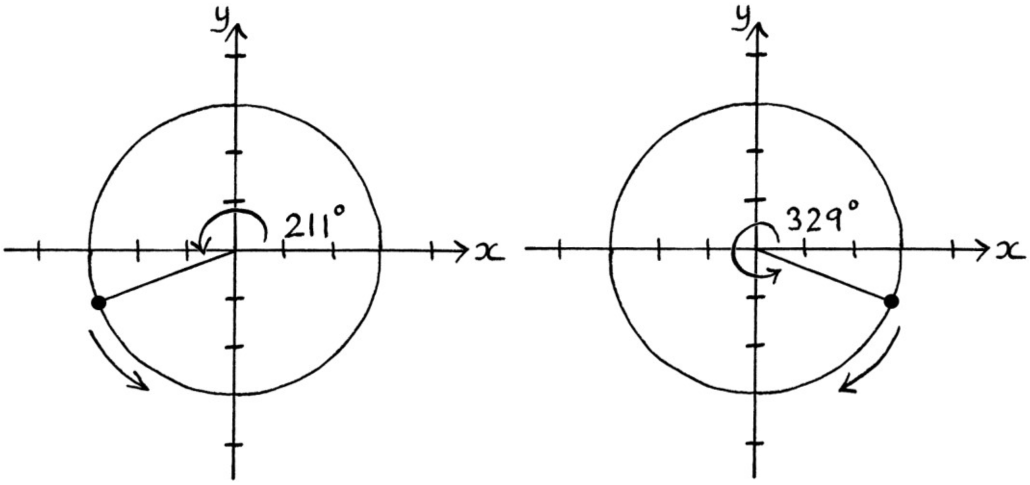


If we mirror this circle left or right, we obtain its negative frequency equivalent. It looks like this:

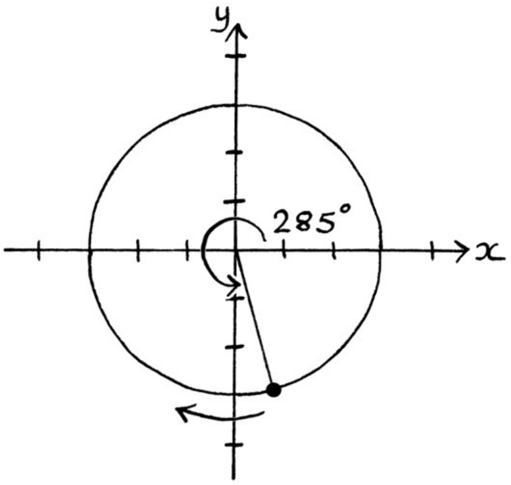


From this, we can see that the phase point of the mirrored circle is 329 degrees. Therefore, the phase of the negative-frequency Sine wave formula is 329 degrees, and the full formula for the wave will be “y = 3 sin ((360 * −3t) + 329)”. Note how both phase points are equidistant from 90 degrees. They are also equidistant from 270 degrees.

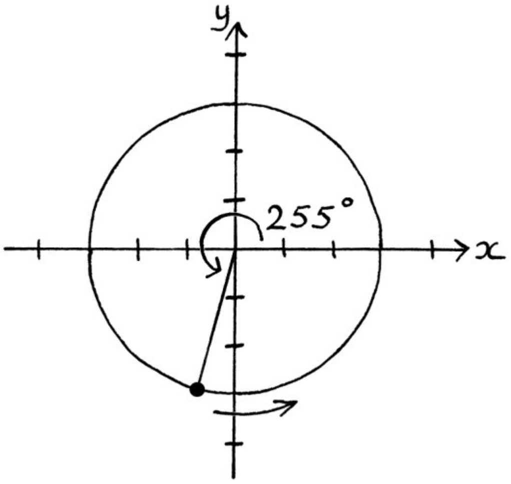
The two circles side by side look like this:



If we had the formula “y = 3 sin ((360 * −1.4t) + 285)”, the circle from which it is derived would look like this:

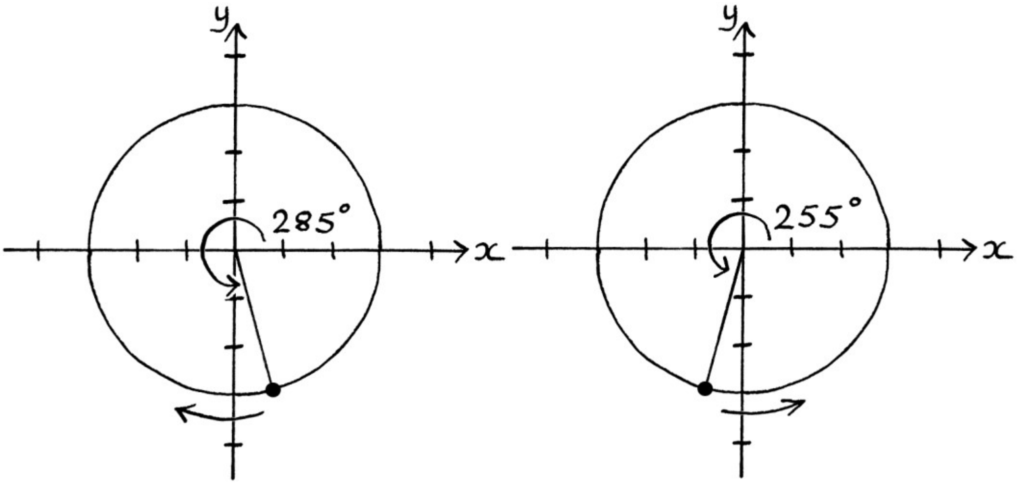


We will find the positive-frequency circle by mirroring the circle, and we will end up with this:

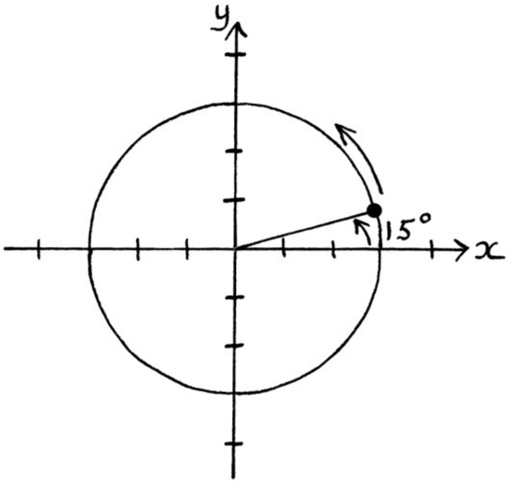


This mirrored circle has its phase point at 255 degrees. Therefore, it represents the Sine wave formula “y = 3 sin ((360 * 1.4t) + 255)”.

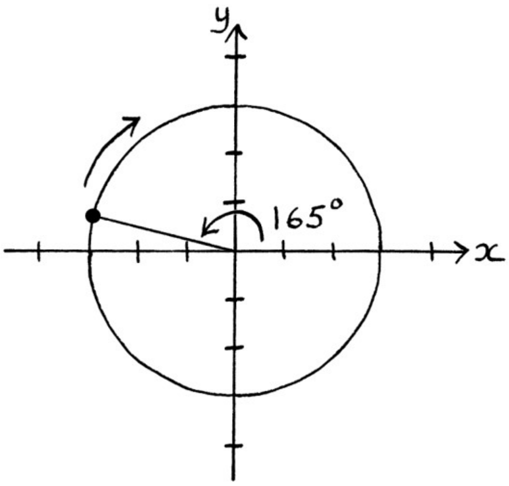
The two circles side by side look like this:



Supposing we had this wave, “y = 3 sin ((360 * 3t) + 15)”, then its circle looks like this:

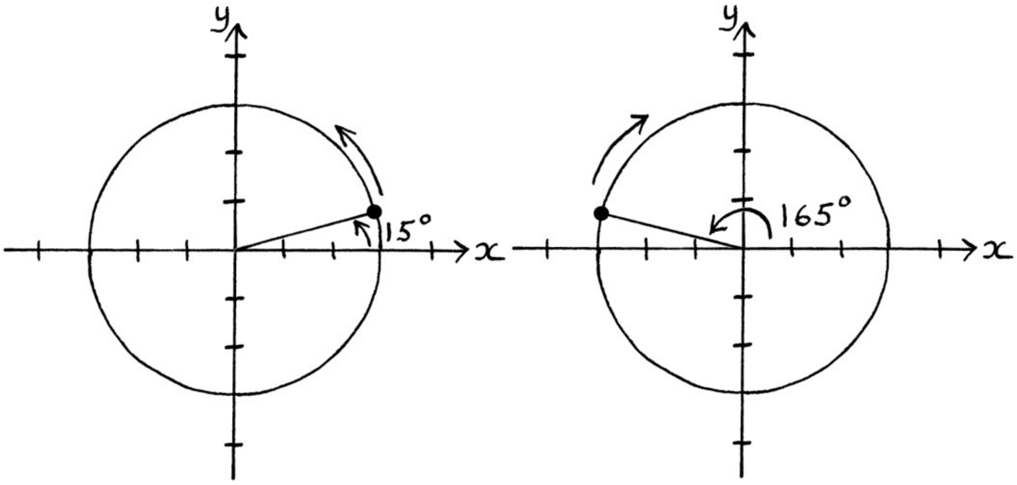


We mirror the circle left or right, and we have this:



The phase point is at 165 degrees. Therefore, the negative-frequency formula that is equivalent to the original wave is: “y = 3 sin ((360 * −3t) + 165)”.

The two circles side by side look like this:



There are two phases that are of interest when finding the corresponding negative-frequency Sine wave – these are when the Sine wave formula has a phase of 90 degrees or a phase of 270 degrees. If the phase is 90 degrees, then the equivalent phase will also be 90 degrees. This is because 90 is 0 degrees away from 90, so the equivalent phase will be 0 degrees away from 90. If the phase is 270 degrees, then the equivalent phase will also be 270 degrees.

We can phrase these equivalences more mathematically as:
“A sin ((360 * ft) + 90) = A sin ((360 * −ft) + 90)”
... and:
“A sin ((360 * ft) + 270) = A sin ((360 * −ft) + 270)”

Instead of seeing how many degrees above or below 90 degrees a phase is, we can subtract the phase from 180 degrees. This has exactly the same effect. If the original phase is 89 degrees, then subtracting it from 180 degrees will produce the angle of 91 degrees. If the phase is 10 degrees, we calculate 180 – 10 = 170 degrees. If the original phase is 359 degrees, then we calculate 180 – 359 = −179 degrees, which is 181 degrees. This method can be useful if we need to rephrase Sine wave formulas to solve equations.

To summarise everything so far, there are really four methods we can choose from to negate the frequency of a Sine wave formula while keeping the curve the same:

- We can see how many degrees above or below 90 degrees the phase is, and choose the angle that same number of degrees below or above 90 degrees.
- We can see how many degrees above or below 270 degrees the phase is, and choose the angle that same number of degrees below or above 270 degrees.
- We can imagine the phase point on the circle from which the Sine wave could have been derived, and mirror it left to right.
- We can subtract the phase from 180 degrees.

Equivalent formulas for Cosine

When it comes to Cosine waves, the rules are slightly different. For one thing, mirroring the circle left-to-right does not work. This is because a Cosine wave is really a Sine wave with a phase of 90 degrees. We are starting with a 90-degree phase already built in.

One way to find the equivalent formula is to convert the Cosine wave into a Sine wave, find the equivalent formula for that, and then convert the result back into a Cosine wave. For example, if we have the formula:
"y = 2 cos ((360 * 2t) + 20)"
... then it is the same as this Sine wave:
"y = 2 sin ((360 * 2t) + 20 + 90)"
... which is:
"y = 2 sin ((360 * 2t) + 110)".

The equivalent negative-frequency Sine wave is:
"y = 2 sin ((360 * −2t) + 70)"
... because 110 and 70 are both 20 degrees away from 90 degrees (and are also equidistant from 270 degrees).

Turning this back into a Cosine wave, we have:
"y = 2 cos ((360 * −2t) + 70 – 90)"
... which is:
"y = 2 cos ((360 * −2t) – 20)"
... which is:
"y = 2 cos ((360 * −2t) + 340)"

Therefore, the formulas:
“y = 2 cos ((360 * 2t) + 20)”
... and:
“y = 2 cos ((360 * −2t) + 340)”
... are equivalent and refer to the same wave.

If we think about the circle, the phases of these two Cosine waves are both equidistant from 0 degrees. They are also equidistant from 180 degrees. As might have been guessed, to find the equivalent phase for a positive-frequency or negative-frequency Cosine wave formula, it is necessary to find the value the other side of zero degrees. On the circle, we would flip the circle upwards or downwards.

The rule for equivalent Cosine waves is:

> “A Cosine wave formula with a positive frequency and a phase so many degrees *above or below* 0 degrees will refer to the same wave curve as that formula with a negative frequency and a phase that many degrees *below or above* 0 degrees.”

Another rule that also works is:

> “A Cosine wave formula with a positive frequency and a phase so many degrees *above or below* 180 degrees will refer to the same wave curve as that formula with a negative frequency and a phase that many degrees *below or above* 180 degrees.”

For example, if we have the formula:
“y = 4 cos ((360 * −7.7t) + 35”
... then +35 degrees is, obviously, 35 degrees more than 0, so the phase for the equivalent positive-frequency formula will be −35, which is −35 + 360 = 325 degrees. The full formula is:
“y = 4 cos ((360 * 7.7t) + 325”

If we have the formula:
“y = 2.2 cos ((360 * 2t) + 123”
... then +123 degrees is 123 degrees more than 0, so the phase for the equivalent negative-frequency formula will be −123, which is −123 + 360 = 237 degrees. The full formula is:
“y = 2.2 cos ((360 * −2t) + 237”

We could also do the calculation by seeing how many degrees the phase is from 180 degrees, in which case the equivalent phase will be that number of degrees the other side. Using the previous example, “y = 2.2 cos ((360 * 2t) + 123”, the phase of 123 is 57 degrees *below* 180, so the equivalent phase will be 57 degrees *above* 180. Therefore, it will be 180 + 57 = 237, which is the result we had before.

Having two methods makes it easier to do the calculation mentally. If the phase is near to 180, we can compare against 180; if the phase is near 0, we can compare against 0.

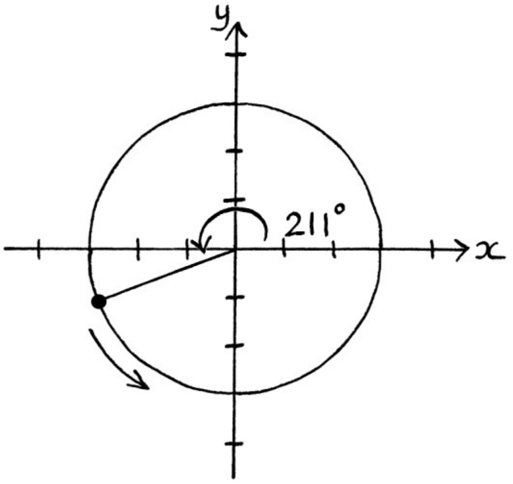
The rules given more mathematically are:
“A cos ((360 * f * t) + ϕ) = A cos ((360 * −f * t) − ϕ)”
... and:
“A cos ((360 * −f * t) + ϕ) = A cos ((360 * f * t) − ϕ)”
... where ϕ is the phase of the formula for which we want to find the equivalent.

The rules with 180 degrees are:
“A cos ((360 * f * t) + (180 + x)) = A cos ((360 * −f * t) + (180 − x))”
... and:
“A cos ((360 * −f * t) + (180 + x)) = A cos ((360 * f * t) + (180 − x))”
... where (180 + x) is the phase of the formula for which we want to find the equivalent.

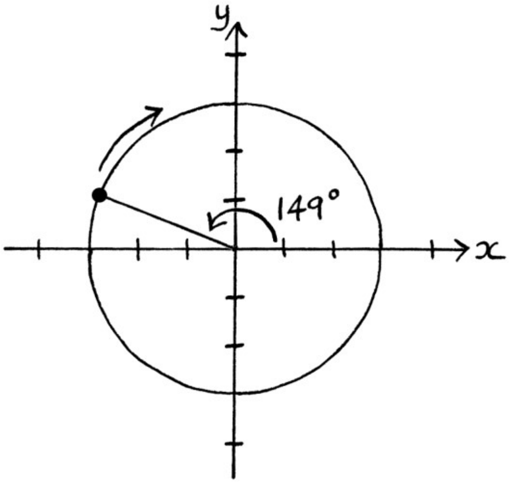
In many ways, finding an equivalent Cosine wave is easier than finding an equivalent Sine wave, as with a Cosine wave, we can just negate the sign of the phase. For example, a Cosine formula with a phase of 45 degrees has an equivalent with a phase of −45 degrees. A Cosine formula with a phase of 220.76 has an equivalent with a phase of −220.76 degrees. However, we still have to do some work if we want the resulting phase to be positive. [As always, the decision to have a positive phase or a negative phase in the formula is a matter of personal choice.]

When it comes to circles for Cosine, we flip the circle upwards or downwards.

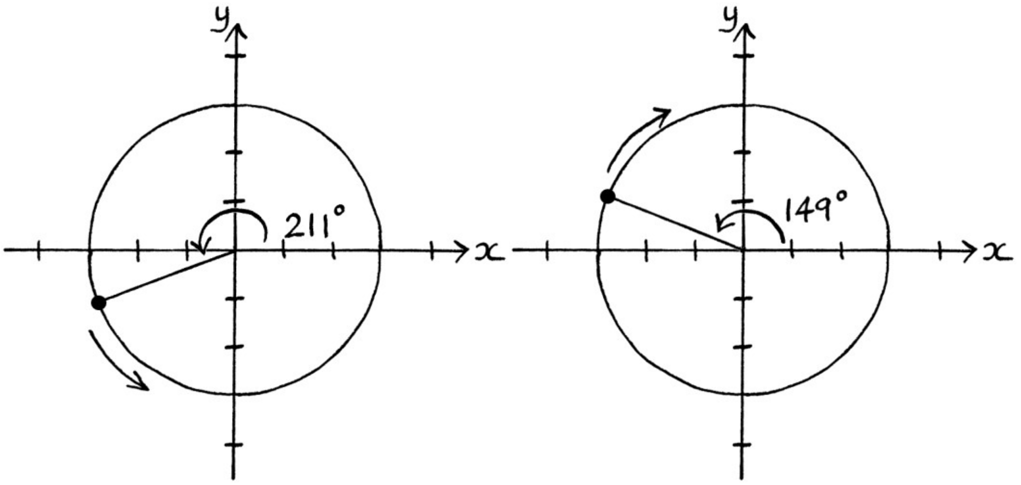
For example, if we have the formula “y = 3 cos ((360 * 3t) + 211)”, the circle from which it is derived looks like this:



Flipping the circle either upwards or downwards results in this:



The phase point ends up at 149 degrees, so the equivalent formula is “y = 3 cos ((360 * −3t) + 149)”. The two circles side by side look like this:



Two Cosine examples of note occur when the formula has a phase of 0 or a phase of 180 degrees. If we think of the circles being flipped up or down, the equivalents should be clear. Also, if we think of the rules, the equivalents should be clear. A phase of 0 degrees is 0 degrees away from 0, and therefore, the equivalent will still have a phase of 0 degrees. To put this a different way, a Cosine formula with a positive frequency and a phase of 0 degrees shows the same wave as that Cosine formula with a negative frequency and a phase of 0 degrees. We actually already knew this from knowing that a Cosine wave with no phase is the same whether the frequency is positive or negative.

When it comes to a phase of 180 degrees, the equivalent formula also has a phase of 180 degrees. This is because 180 degrees is equidistant from 0 degrees. It is also because 180 degrees is equidistant from 180 degrees.

We can phrase these equivalences more mathematically as:
"A cos ((360 * ft) + 0) = A cos ((360 * −ft) + 0)"
... and:
"A cos ((360 * ft) + 180) = A cos ((360 * −ft) + 180)"

It is important to realise that if we use a circle to find the equivalent negative-frequency Sine wave by mirroring it left or right, it will not find the equivalent negative-frequency Cosine wave. Flipping the circle left or right, or up or down, will only find the equivalent negative-frequency wave for one of the waves derived from that circle.

As a summary of this section, there are five methods we can choose from to negate the frequency of a Cosine wave formula while keeping the same curve:

- We can see how many degrees above or below 0 degrees the phase is, and choose the angle that same number of degrees below or above 0 degrees.
- We can see how many degrees above or below 180 degrees the phase is, and choose the angle that same number of degrees below or above 180 degrees.
- We can negate the phase. If it is positive, we make it negative; if it is positive, we make it negative.
- We can subtract the phase from 0 degrees. [This is identical to negating the phase].
- We can imagine the phase point on the circle from which the Cosine wave could have been derived, and flip it upwards or downwards.

A potential source of confusion

If we had a Sine or Cosine formula with a negative *amplitude*, then, no matter what the phase, we could add 180 degrees to the phase to find the equivalent positive-amplitude formula that referred to exactly the same curve. We could also subtract 180 degrees, as that ultimately means the same thing. To turn a positive-amplitude formula into a negative-amplitude formula, while keeping the curve it represents the same, we also add 180 degrees. [We could also subtract 180 degrees as that is the same thing]. For example:
"y = −2 sin ((360 * 4t) + 1)"
... represents the same curve as:
"y = +2 sin ((360 * 4t) + 181)"

Similarly, if we had a negative-*frequency* Sine wave formula with *zero phase*, then we could add (or subtract) 180 degrees to the phase to make it into a positive-frequency Sine wave formula that portrayed exactly the same wave. For example:
"y = 3 sin (360 * −5t)"
... represents the same curve as:
"y = 3 sin ((360 * 5t) + 180)"
This works in the other direction too. [As a Cosine wave formula with zero phase refers to the same curve whether it has a positive or negative frequency, this does not work for Cosine waves with zero phase].

It is very important to realise that adding 180 degrees to the phase only works for frequencies if the Sine wave formula has a phase of zero or 180 degrees. If the phase is any other value, then this will not work.

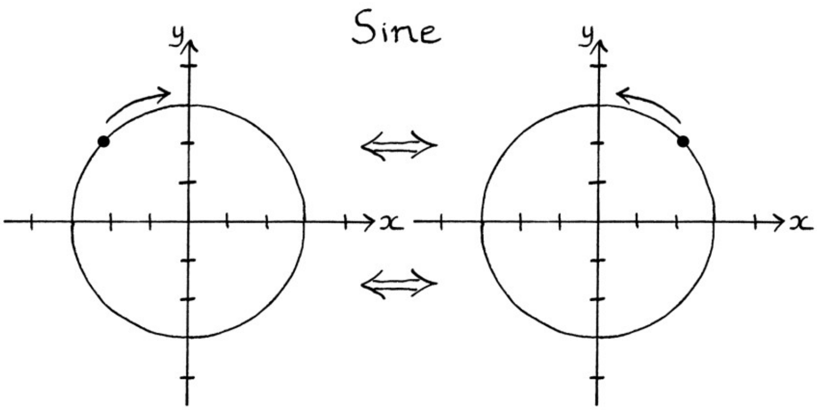
It is easy to become confused and think that converting a negative frequency into a positive frequency works in the same way as converting a negative amplitude into a positive amplitude. It does not. Converting negative frequencies to positive frequencies and back again by adding 180 degrees only works for:

- Sine waves with phases of 0 or 180 degrees, and that is because 0 and 180 degrees are equidistant from 90 degrees. [They are also equidistant from 270 degrees].
- Cosine waves with phases of 90 or 270 degrees, and that is because 90 and 270 are equidistant from 0 degrees. [They are also equidistant from 180 degrees].

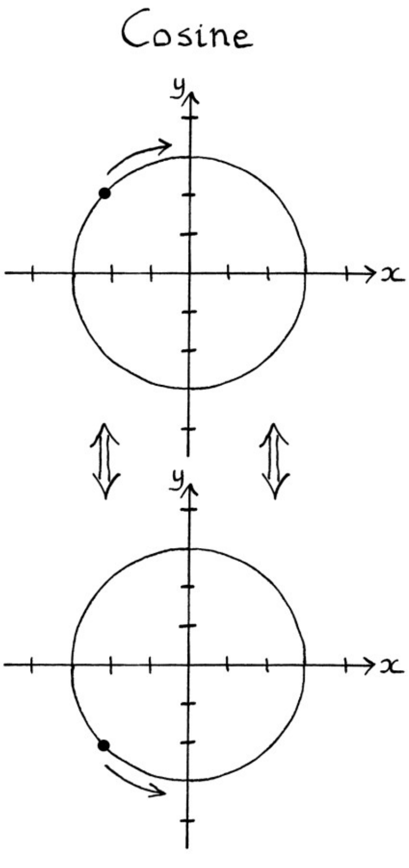
The two situations when adding 180 degrees works should really be thought of as coincidences, or else it can lead to mistakes. Now you have seen how flipping circles left or right, or up or down, works, you are less likely to make this mistake.

A summary of the best way of thinking about finding the equivalent positive-frequency and negative-frequency formulas is:

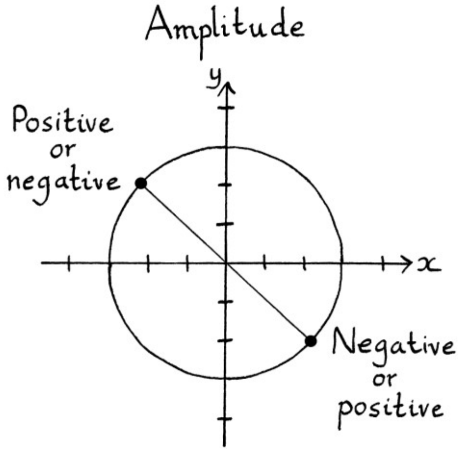
- To convert a Sine wave, or the circle that represents that Sine wave, with any phase to the opposite frequency, we mirror the circle left or right:



- To convert a Cosine wave, or the circle that represents that Cosine wave, with any phase to the opposite frequency, we flip the circle up or down:



Just so we can compare the methods for positive and negative frequency with the method for amplitude: when it comes to negative and positive *amplitudes*, to find the equivalent phase with the opposing amplitude, we find the point on the circle's edge on the other side of the origin of the axes. In other words, we draw a line from the phase point through the origin of the axes, and out the other side until it touches the circle again. Where it meets the circle's edge will be the phase of the equivalent formula. This will always be the original phase plus 180 degrees (or minus 180 degrees, as that is the same thing on a circle).



Negative frequency period

The period of a wave is the amount of time that one cycle takes to complete. With *positive* frequencies, the period is the reciprocal of the frequency. If frequency is represented by the letter "f" and period by the letter "T", we can say the following:
T = 1 ÷ f
... and:
f = 1 ÷ T
... when "f" is a positive value.

There are two ways in which we can think about the period of a *negative*-frequency wave:

- First, we could choose to say that the period of any wave is always positive, whether the frequency is positive or negative. The thinking behind this choice is based on how, when we have a negative frequency, the time it takes to complete a cycle will still be a positive value. It does not matter if a wave has a positive frequency or a negative frequency – the period will always be a positive time. Another reason for thinking in this way is that a negative-frequency wave has an identical curve to a positive-frequency wave with a possibly different phase. Therefore, in this way of thinking, the period of a negative-frequency wave should be the same as that wave given a positive frequency.

- Second, we could choose to say that the period of a negative-frequency wave is a negative value. This way of thinking is consistent with the "T = 1 ÷ f" and "f = 1 ÷ T" formulas. In this way of thinking, a wave with a frequency of −5 cycles per second would have a period of −0.2 seconds. A period being negative would indicate that the frequency of the wave to which it applies was also negative. Despite a negative time being harder to visualise, this choice is mathematically consistent.

Personally, I think the first choice is the best one when we are discussing *waves* (as opposed to circles) and it is what I would use in this book if the idea ever became relevant. The second choice is harder to visualise, but it could still be a useful idea in some situations when we are discussing circles.

Negative frequency summary

It is good to know about negative frequency for two reasons. First, it is another concept that is relevant to waves – the more you know about all the potential properties of waves, the better your understanding of waves will be. Second, the concept comes up later on when dealing with the frequency domain. However, when you see negative frequencies portrayed in the frequency domain, they will usually be a result of a different way of thinking about waves.

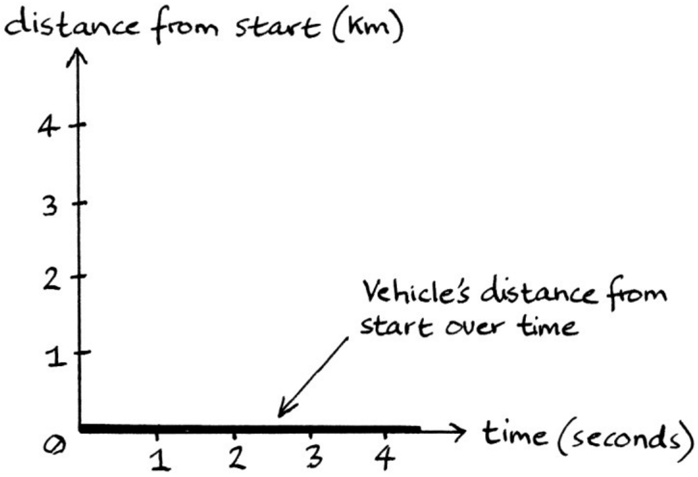
The idea of negative frequencies might be hard to visualise if you are presented with just a wave, so as always, it pays to consider the circle from which the wave is, or could have been, derived.

You will find a few people who think that negative frequency is not a valid concept. This is because they only think of waves, and not the circles that those waves are derived from, or could have been derived from. Given that a zero-phase Cosine wave formula refers to the same wave curve whether it has a positive or negative frequency, and given that a zero-phase negative-frequency Sine wave formula refers to the same wave as that formula made positive and given a phase of 180 degrees, it is easy to see why people would discount the idea of negative frequency if they had never thought about circles. If you only think of waves, and not circles, then there is little to show that negative frequency is much different from positive frequency. Negative frequency really only makes sense if you think about circles.

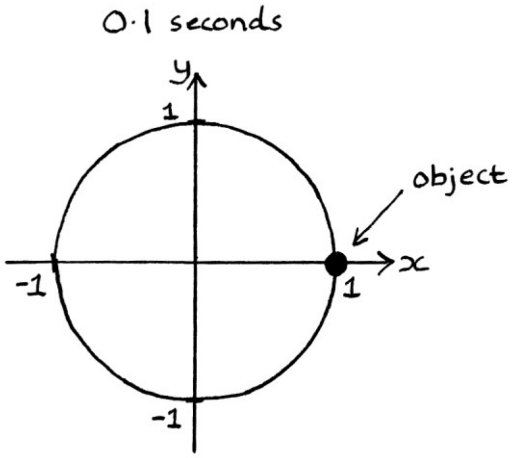
Zero frequency

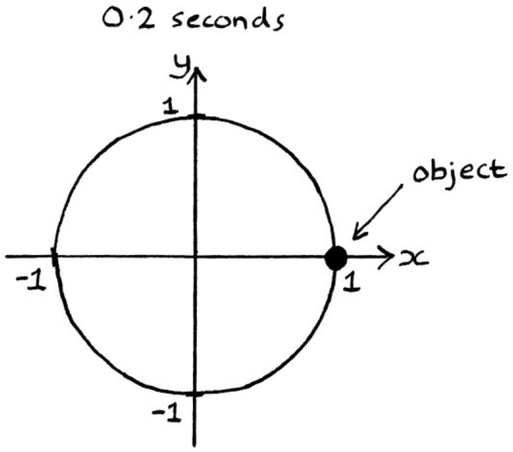
A concept that is slightly related to negative frequency is that of zero frequency – in other words, a frequency of zero cycles per second. This idea can seem puzzling, but it is straightforward if we remember that frequency is analogous to speed.

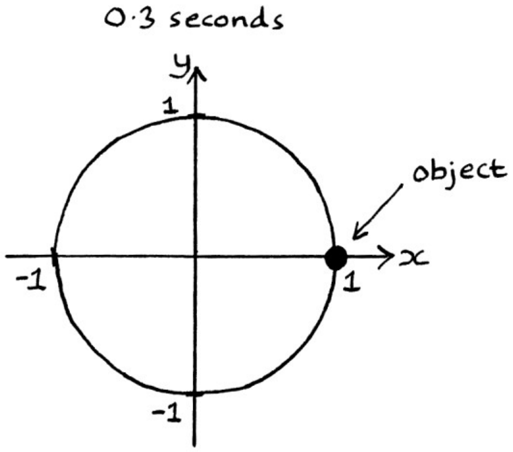
If we imagine a vehicle with zero speed – in other words a vehicle that is not moving – after 0.1 seconds, it will be in exactly the same place as it was to start with. After 0.2 seconds, it will be in the same place. After 0.3 seconds, it will be in the same place, and so on. For every moment in time, it will be in the same place. It does not move. We could draw a graph of the vehicle's distance from its starting place, and the result would just be a straight line along y = 0.



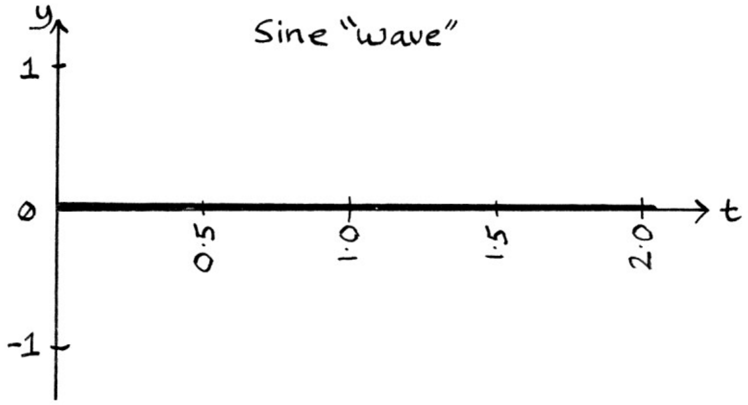
If we imagine an object “rotating” around a unit-radius circle at 0 cycles per second, it, similarly, is not moving. After 0.1 seconds, it will be at its starting point. After 0.2 seconds, it will still be at its starting point. After 0.3 seconds, it will still be at its starting point. For any moment in time, it will always be at its starting point.



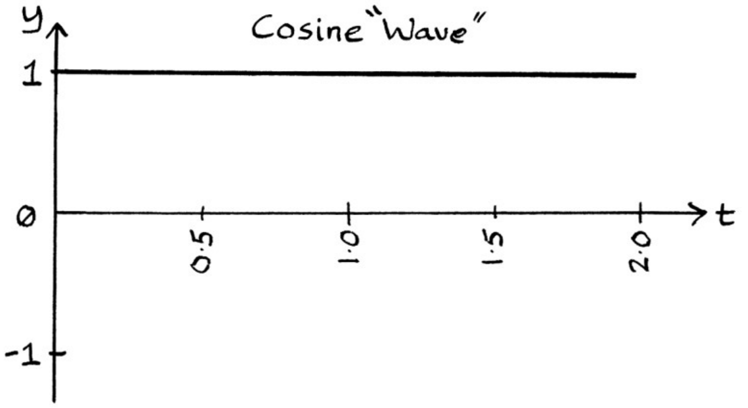




As always, the Sine wave derived from such a circle will be the object's y-axis position over time. As the object is not moving, its y-axis position will be the same for every moment in time. For the above circle, it will always be zero. The resulting graph will not actually be a wave, but just a horizontal line along y = 0:

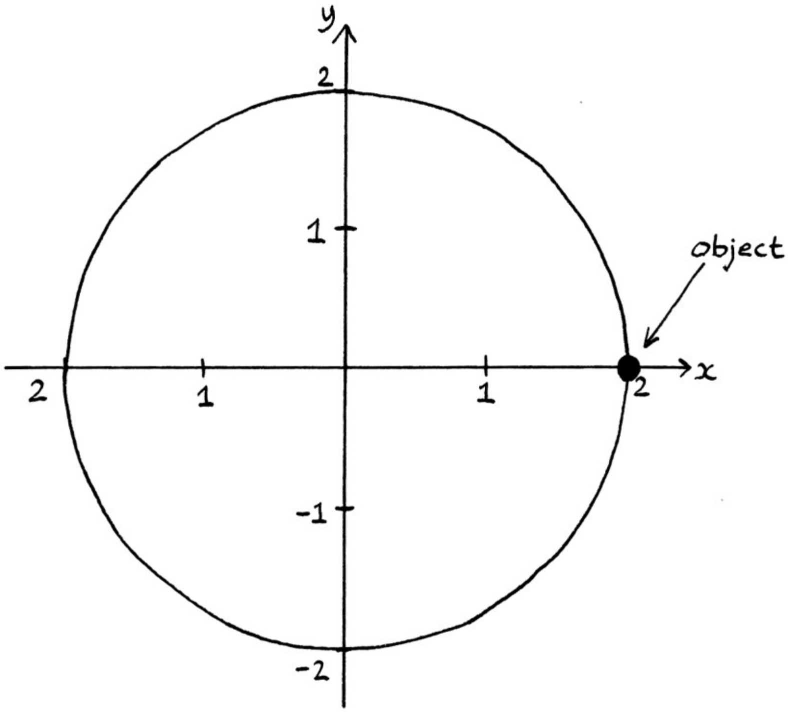


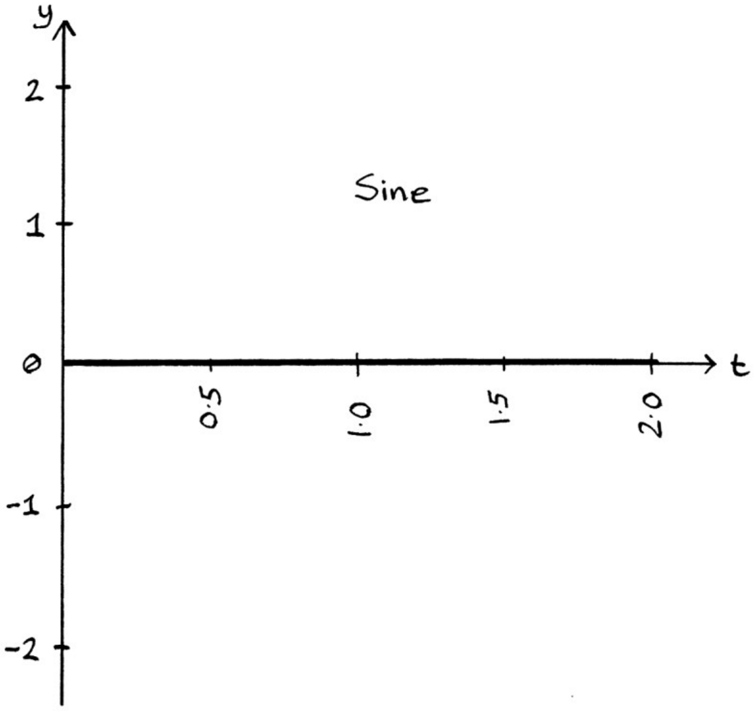
As always, the Cosine wave derived from such a circle will be the object's x-axis position over time. As the object is not moving, its x-axis position will be the same for every moment in time. On the above unit-radius circle, the object's x-axis position will be +1 at 0 seconds, +1 at 0.1 seconds, +1 at 0.2 seconds, +1 at 0.3 seconds and so on. It will always be +1. Therefore, the Cosine graph derived from the above circle will be a straight line where "y" is always +1.

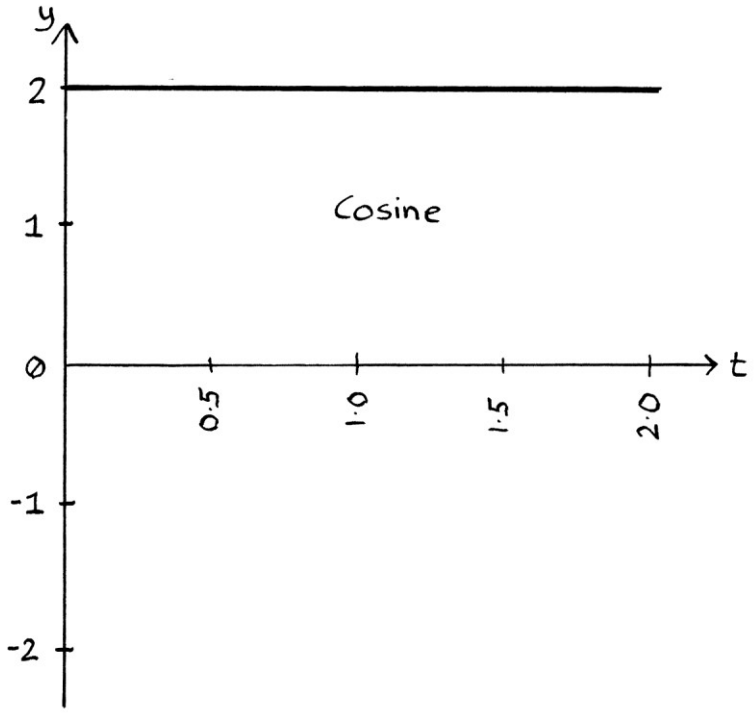


As I have said, the Sine and Cosine graphs with zero frequency indicate the starting point of the object for all time. Therefore, if the circle has a radius of 1 and a phase of zero, the Sine graph will be a horizontal line going through y = 0, and the Cosine graph will be a horizontal line going through y = 1. If there is zero phase and an amplitude other than 1, then the Sine graph will still be a line going through y = 0, but the Cosine graph will be a line at the height of the amplitude.

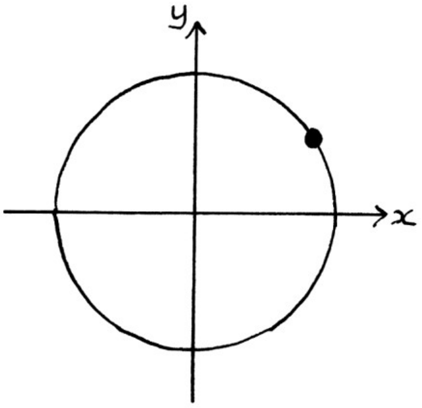
For example, if the circle has a radius of 2 units, then the Sine graph will still be a straight line going through y = 0, but the Cosine graph will be a straight line going through y = 2.

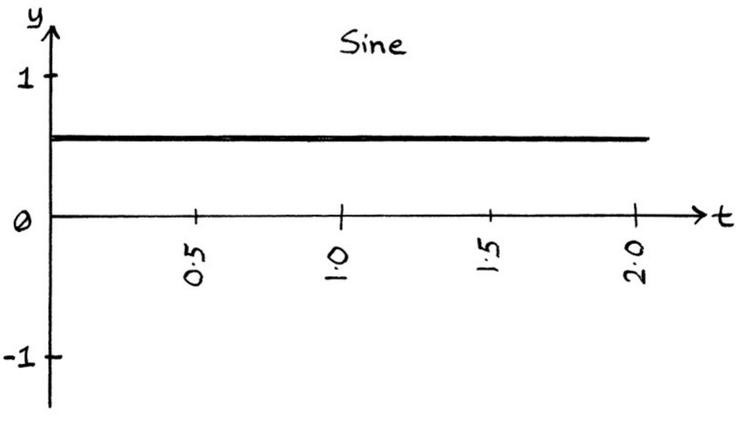


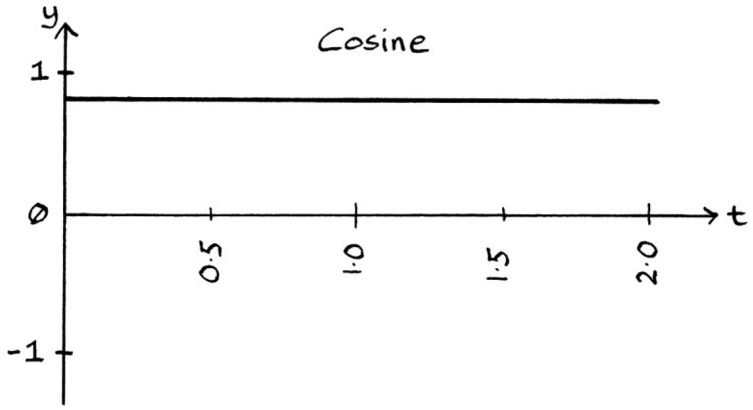




If there is a non-zero phase on the circle, the Sine graph will be a straight line at the y-axis value of the phase point on the circle, and the Cosine graph will be a straight line at the x-axis value of the phase point on the circle.







Zero frequency in the formulas

The nature of the Sine and Cosine graphs is also reflected in the formulas of the two “waves”.

A zero frequency in the Sine wave formula looks like this:
“$y = h_s + A \sin ((360 * 0 * t) + \phi)$”
... which ends up as:
“$y = h_s + A \sin (0 + \phi)$”
... or
“$y = h_s + A \sin \phi$”

This means that if there is no phase or mean level on the circle, the result of the formula will be zero, because sin 0 = 0, and the amplitude multiplied by zero will still be zero. The graph for the formula with zero phase and zero mean level would be a horizontal line at y = 0.

If the mean level is zero, but the phase is not zero, then the graph is the amplitude multiplied by the Sine of the phase for all time. If the mean level is not zero, then the line will be raised or lowered by the amount of the mean level.

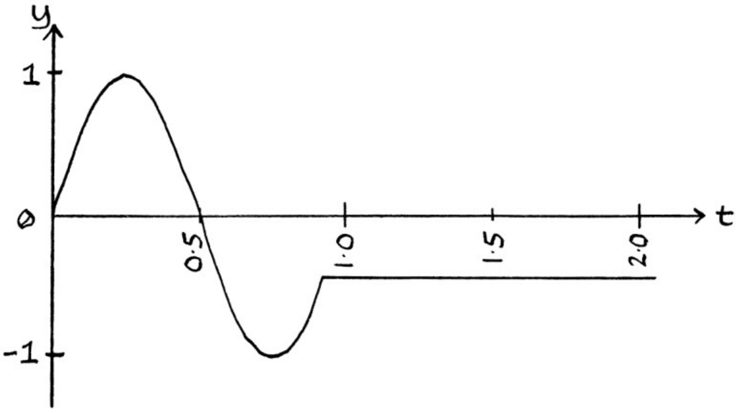
A zero frequency in the Cosine wave formula looks like this:
“$y = h_c + A \cos ((360 * 0 * t) + \phi)$”
...which ends up as:
“$y = h_c + A \cos (0 + \phi)$”
... or:
“$y = h_c + A \cos \phi$”

This means that if there is no phase and no mean level, the formula will be “A * cos 0”, which will always be “A * 1”, which is the amplitude. Therefore, if there is no phase or mean level, and the amplitude is 1, the graph will always be a horizontal straight line at y = 1.

If the mean level is zero, and the phase is non-zero, the Cosine graph’s y-axis points will be the amplitude multiplied by the Cosine of the phase, for all time. If the mean level is not zero, then the mean level will raise or lower the line.

Frequency when a rotating object stops

If we have an object rotating around a circle that then stops, in the time after it has stopped, it will have zero frequency. However, if the object is not exactly at 0 degrees on the circle when it stops, its y-axis and x-axis position will not be the same as if it had always had zero frequency. On the Sine wave graph derived from the object’s movement, the y-axis value will remain fixed at the height of the object when it stopped, and on the corresponding Cosine wave graph, the y-axis value will remain fixed at the object’s x-axis position on the circle at the moment that the object stopped.



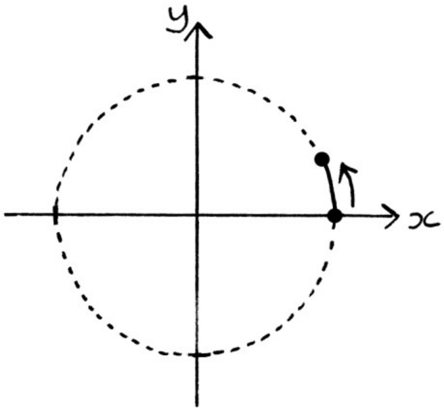
A potential source of confusion

One potential source of confusion with zero frequency is that zero frequency refers to the frequency of an *existing* object that is stationary on the circle. If there is no object there, then there will be nothing that is “rotating” around the circle at 0 cycles per second. Sometimes, people might presume that there is always an object that has a frequency of zero – this is because they see a list of frequencies and think something should be at zero cycles per second. At this point in this explanation, you can tell that that is obviously a mistake, and it might seem odd that anyone would even consider it. However, later on when things are more complicated, someone who has not really thought about zero frequency properly might mistakenly introduce the idea unnecessarily.

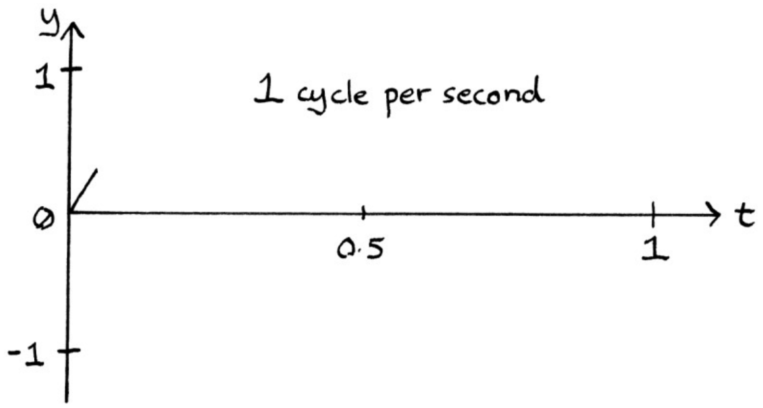
Frequency and partial revolutions

An object rotating around a circle does not need to complete a full rotation to have a frequency, in the same way that a vehicle does not need to complete a full unit of distance to have a speed. If a vehicle travels 1 centimetre at 100 kilometres per hour, its speed is still 100 kilometres per hour. If we see an object rotate only 1 degree at 20 cycles per second, its frequency for that time is still 20 cycles per second.

Strictly speaking, there is no minimum length of time for an object to have a frequency (or a vehicle to have a speed), as long as that time is greater than zero. This means that we could have an object moving around a circle at 10 cycles per second, but it only moves around a tiny part of the circle, and then it stops or disappears. It would still have a frequency of 10 cycles per second for that portion of the circle.



As another example, we could observe only the start of a wave such as this, where the frequency is still 1 cycle per second for the drawn section, even though the wave does not complete one cycle:



In the above graph, it is difficult to tell that the frequency is 1 cycle per second because we cannot see when it would repeat, but its curve is still different from that of a wave of any other frequency.

Conclusion

This chapter contained more complicated aspects of frequency than those in Chapter 6. The concepts would have made Chapter 6 harder to understand, but at this point in the book, they should be more straightforward.

Most of the concepts introduced in this chapter are easier to understand if you think of circles, and if you think of how frequency is analogous to the speed of a vehicle. If you are ever confronted with something to do with frequency that you do not understand, it is always helpful to think of circles and vehicles.

In the rest of this book, unless I specifically say otherwise, all frequencies will be unvarying and positive.

Chapter 12: Recreating the circle

This chapter is designed to emphasise the relationship between a circle and its waves. Some of this chapter will be useful later in this book, and the concepts explained here will give you a better insight into waves.

In Chapter 3, we saw how to recreate the circle that produced a wave by using just one of its two derived waves. We can do this even if we do not know the formula of the wave (presuming the circle is centred on the origin of the axes). Now that we know about the four attributes of a wave, we will look at the idea in more depth. This will involve repeating some of what we learnt in Chapter 3.

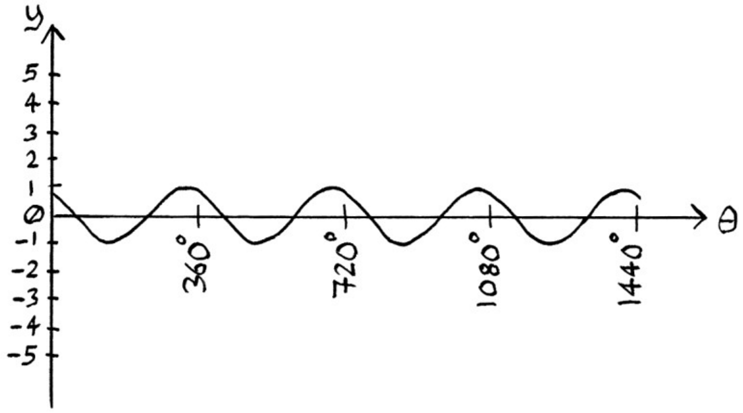
Obviously, to recreate the circle from a wave, we could just look at the wave and figure out the amplitude, frequency, phase, and mean level, and then draw a circle that matched. However, it is better to have a more general method. Not only does a general method sometimes save a lot of time, but it also means that a computer program can use the method. In a way, it is as if we did the process blindfolded and asked someone else to read aloud as little information as possible. If we have a method to recreate the circle, we also have a method of summarising all the attributes of a particular wave and its corresponding twin in a succinct way. This can be useful.

In this chapter, to simplify matters, I will concentrate mainly on waves relating to *angle* rather than *time*. If a wave involves time, then it is slightly harder to read *angles* from it. [However, on such a time-based wave, it is possible to tell what the frequency is by how often it repeats a cycle in a second, and from that, we can work out how many degrees are passed in a second. This enables us to know the relevant angle at any particular time.]

Sine and Cosine

In these explanations, it pays to remember that Sine refers to the y-axis values of a circle's edge, and Cosine refers to the x-axis values of a circle's edge. Sine and Cosine are the same thing but with a difference in phase. The meanings of the formulas "$y = \sin(\theta + 90)$" and "$y = \cos\theta$" are identical. In recreating the circle, it is helpful to treat the Sine and Cosine waves as representing the different axes. We can think of them as the "x-axis wave" and the "y-axis wave". This means that if we have a wave where we do not know if it represents the x-axis values or the y-axis values of a circle, we have to choose one or the other arbitrarily.

Therefore, this wave...



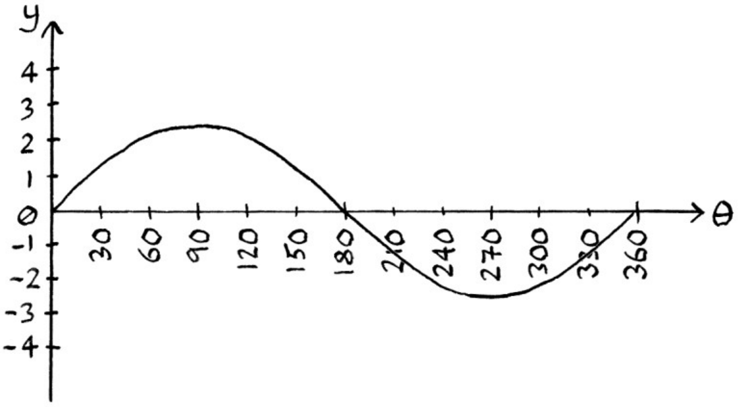
... could be said to be a Sine wave, in which case it represents the y-axis values of the circle, or it could be said to be Cosine wave, in which case it represents the x-axis values of the circle. The same is true for any wave where we do not know its original formula or from which aspect of a circle it came. If we decide that a wave will be treated as a Sine wave, then its twin will be a Cosine wave. If we decide that a wave will be a Cosine wave, then its twin will be a Sine wave.

Amplitude and radius: method 1

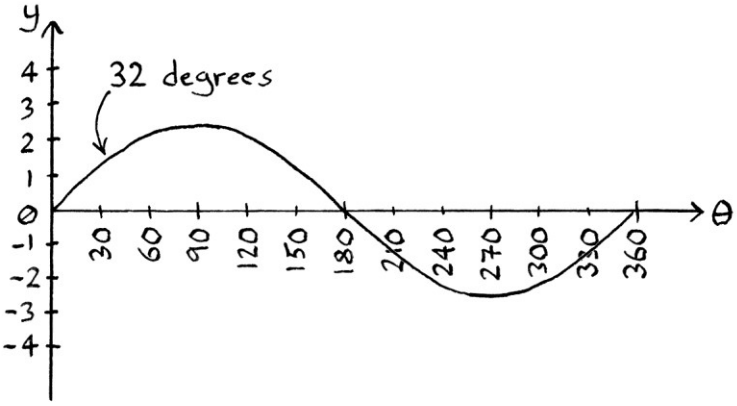
If we are only interested in the amplitude of the wave, and therefore, the radius of the circle, then the following three methods are sufficient to recreate the circle from which a given wave was, or could have been, derived.

If the circle from which a wave is derived is centred on the origin of the axes, the wave has zero phase, and the only attribute of the wave that we are interested in is the amplitude, then we can recreate the circle with just one non-zero point from the graph. We take any non-zero y-axis value from the wave, and plot it on the circle chart at the angle at which it appears in the wave graph.

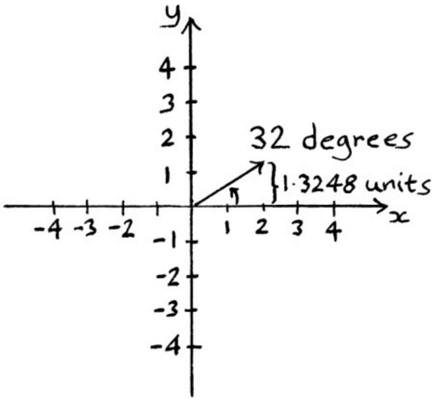
For example, if we have this angle-based Sine wave:



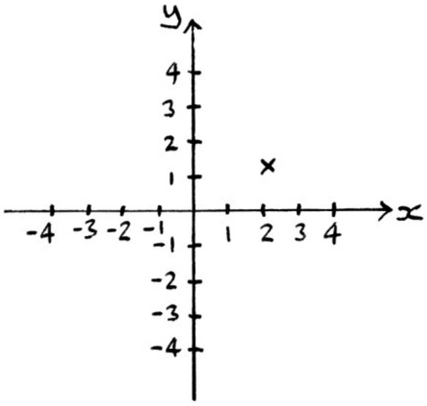
...then we read off any non-zero y-axis value, and the angle at which it appears. In this case, we will choose the y-axis value at the arbitrarily chosen angle of 32 degrees, which is 1.3248 units:

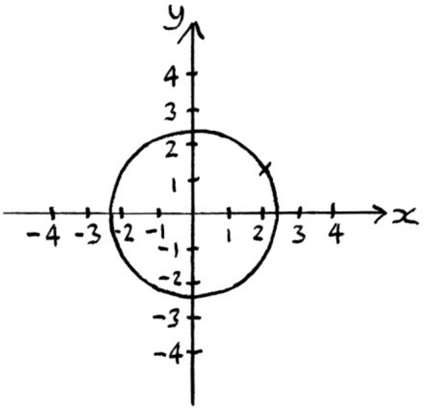


We then draw a line from the origin of our empty circle chart at the angle of 32 degrees until the height of that line is 1.3248 units.



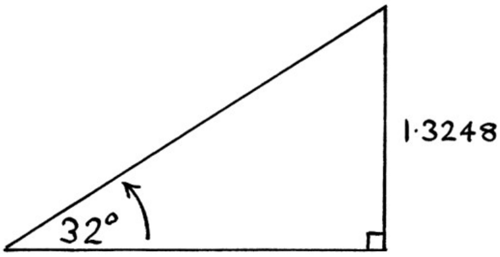
We know that all points on a circle's edge are the same distance from the centre of the circle, which means that we can draw a circle centred at the origin of the axes, with a circumference that goes through that point.





The circle is now the one from which that wave was, or could have been, derived. We can check it is correct by measuring the amplitude of the wave and comparing it to the radius of the circle. They are both 2.5 units, so the circle is correct. In reality, the wave's formula was "y = 2.5 sin θ", which again confirms the method works.

We could also have calculated the radius of the circle by thinking of a right-angled triangle. The angle of interest will be 32 degrees, and the opposite side will be 1.3248 units.



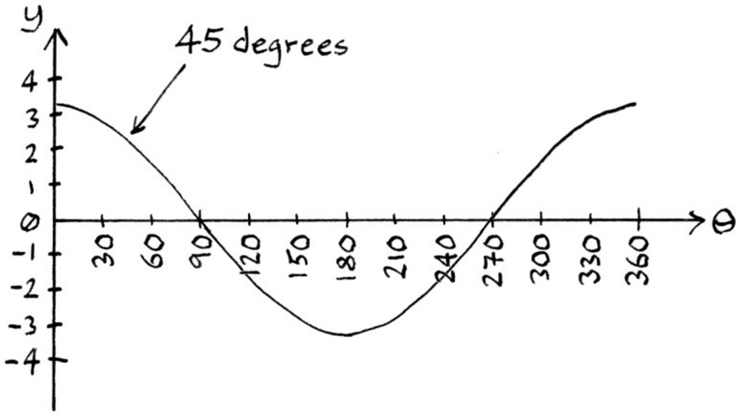
To work out the length of the hypotenuse, we use the formula:
opposite = hypotenuse * sin θ.

We arrange it to be:
hypotenuse = opposite ÷ sin θ
... which means we have:
hypotenuse = 1.3248 ÷ sin 32
hypotenuse = 2.5 units.

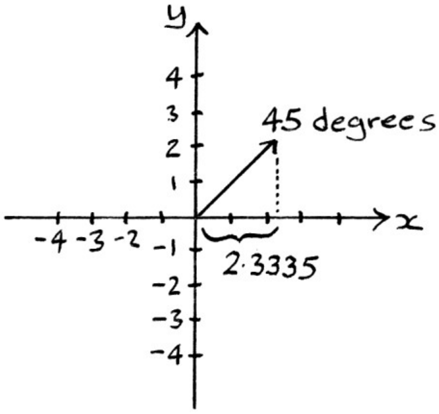
Therefore, the amplitude of the wave and the radius of the circle are 2.5 units, which matches our earlier answer. We can then draw a circle with this radius.

Cosine

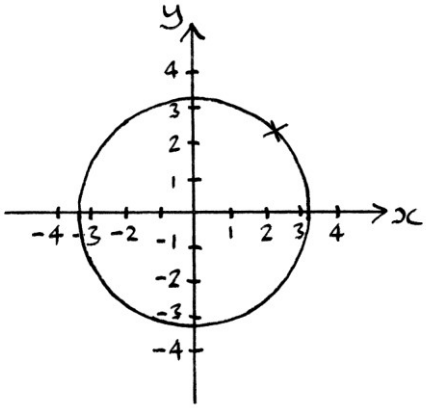
For a Cosine wave, the process is the same, but when drawing the line on the circle chart, we draw it until its *horizontal* distance out is equivalent to the y-axis value from the graph. As an example, for the following wave, we will pick the point at 45 degrees:



We could have picked any non-zero point, so 45 degrees is as good as any other. The y-axis value here is 2.3335. Therefore, we start drawing a line on the future circle chart at 45 degrees, and we stop drawing it when the x-axis value of the end of the line is at 2.3335 units.

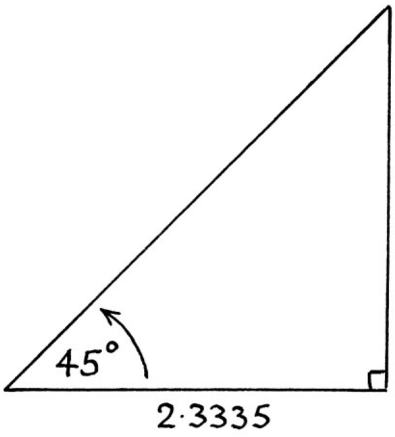


We then draw a circle, centred on the origin of the axes, with a circumference that passes through that point.



We can check the circle. The circle has a radius of 3.3 units, which matches the wave's amplitude of 3.3 units. The actual wave was "y = 3.3 cos θ", so this is all correct.

If we had done this by thinking of right-angled triangles, we would have had an angle of 45 degrees and an *adjacent* side of 2.3335 units.



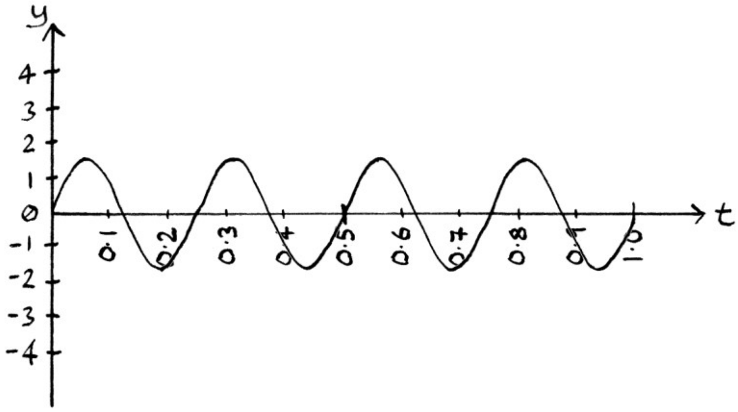
We would have used the formula:
adjacent = hypotenuse * cos θ
... which we would rearrange to be:
hypotenuse = adjacent ÷ cos θ.

This gives us:
hypotenuse = 2.3335 ÷ cos 45
hypotenuse = 3.3 units.

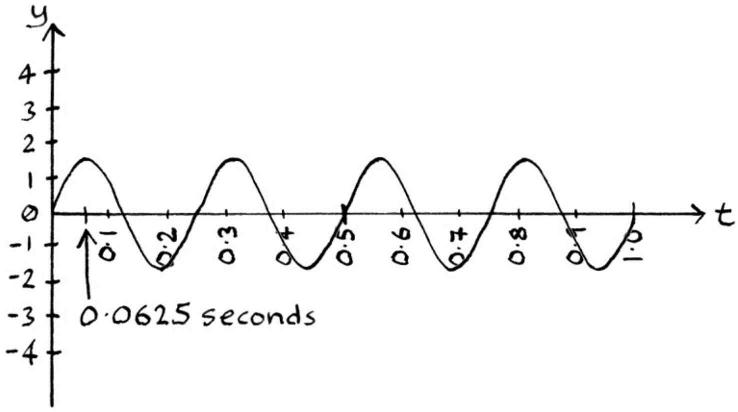
Therefore, the amplitude of the wave and the radius of the circle are 3.3 units, which is the same answer as before. We then draw a circle with this radius.

Time

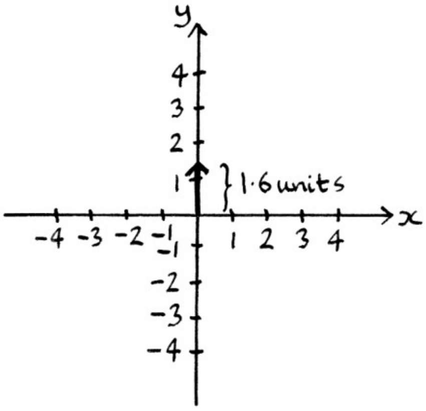
If our wave graph had related to time, then the method works in the same way. For example, we will look at this time-based Sine wave graph:



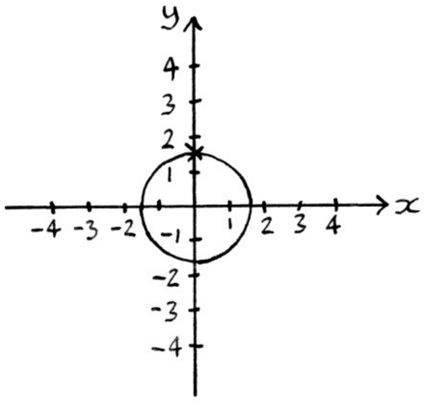
The wave repeats itself 4 times a second. In other words, its frequency is 4 cycles per second. The period is 0.25 seconds. From this, we know that the object rotating around the circle from which this wave is derived completes 360 degrees in 0.25 seconds. If we wanted to pick the point at 90 degrees (quarter of a circle), this would occur when the time is 0.25 ÷ 4 = 0.0625 seconds. [Note that we can pick the point at any angle for which the y-axis value is not zero. In this example, 90 degrees is just an arbitrary choice.] At 0.0625 seconds, the y-axis value is 1.6 units.



Therefore, on our future circle chart, we start to draw a line at 90 degrees and stop when the end has a y-axis value of 1.6.



We then draw a circle that goes through this point.



We can then check the circle to see if it is correct. The radius of this circle is 1.6 units, and the amplitude of the wave is 1.6, so this is correct. The actual formula for the wave was “y = 1.6 sin (360 * 4t)”, which confirms this.

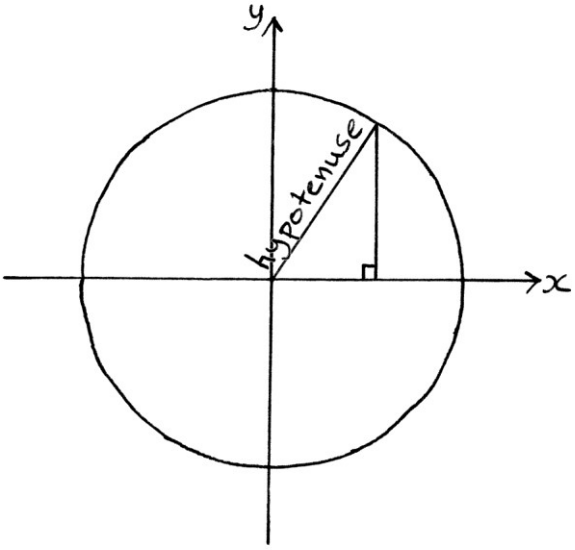
Amplitude and radius: method 2

This method works on *any* wave with no mean level, even if it has a non-zero phase, however, it still only finds the amplitude of the wave, which is enough to draw the basic circle. This method relies on Pythagoras's theorem. It is a simple method, but the reason why Pythagoras's theorem is relevant takes some explanation.

In a right-angled triangle, the square of the hypotenuse is equal to the sum of the square of the opposite side and the square of the adjacent side:
$h^2 = o^2 + a^2$

If we treat the triangle's sides with reference to the triangle's angle, then we can say that the square of the hypotenuse is equal to the sum of the square of the Sine of the angle and the square of the Cosine of the angle. For a right-angled triangle with a *unit long hypotenuse*, the relevant formula is this:
$h^2 = (\sin \theta)^2 + (\cos \theta)^2$

If we draw a right-angled triangle inside a circle, then the hypotenuse of the triangle is the same length as the circle's radius. Strictly speaking, the hypotenuse *is* the radius.



Therefore, we can say that for a triangle in a *unit-radius* circle, the square of the *radius of the circle* is equal to the square of the Sine of the angle added to the square of the Cosine of the angle:
$\text{radius}^2 = (\sin \theta)^2 + (\cos \theta)^2$

Given that the Cosine of a value is equal to the Sine of that value +90 degrees, we can rewrite that formula as:
$\text{radius}^2 = (\sin \theta)^2 + (\sin \theta + 90)^2$

A Sine wave graph is just a drawing of the Sines of a series of consecutive angles. A Sine wave graph is also derived from a particular circle and is a representation of the y-axis values of points at evenly spaced angles from the centre to the circle's edge. Therefore, it does not matter if the list of Sines is portrayed in a wave graph or portrayed within a circle – it is the same information in two different forms. Therefore, the above formula can be used to say that the square of the *radius* of the unit circle is equal to the square of the Sine of a value taken from a wave graph, added to the square of the Sine of a value +90 degrees later from that graph.

This means we can take *any* two values from a unit-amplitude Sine wave graph that are 90 degrees apart, and put them into the formula and we will have the radius of the circle, which is also the amplitude of the wave, which will be 1 unit.

Significantly, this method works for Sine waves with any amplitude and any phase. The idea put in a slightly mathematical form is:
$\text{radius}^2 = (\text{any value from the Sine wave graph})^2 + (\text{a value from 90 degrees later})^2$

A more mathematical formula is:
$\text{radius}^2 = (A \sin (\theta + \phi))^2 + (A \sin (\theta + \phi + 90))^2$
... where:

- "θ" refers to *any* chosen angle.
- "ϕ" is the phase of the wave.

An explanation in words is:

> "If we read any two values from a wave graph that are 90 degrees apart, square each of them, add those squares, then square root the sum, we will end up with the amplitude of the wave, and also the radius of the circle from which that wave is, or could have been, derived (presuming the circle was originally centred on the origin of the axes)."

Cosine

If the wave graph is a Cosine wave, the idea still works, but we take one value and then a second value 90 degrees *earlier*. It might be apparent that this is actually identical to taking one value and then a second value 90 degrees later. There is no difference in the ultimate meaning.

The formula for a unit-amplitude wave would look like this:
$\text{radius}^2 = (\cos \theta - 90)^2 + (\cos \theta)^2$

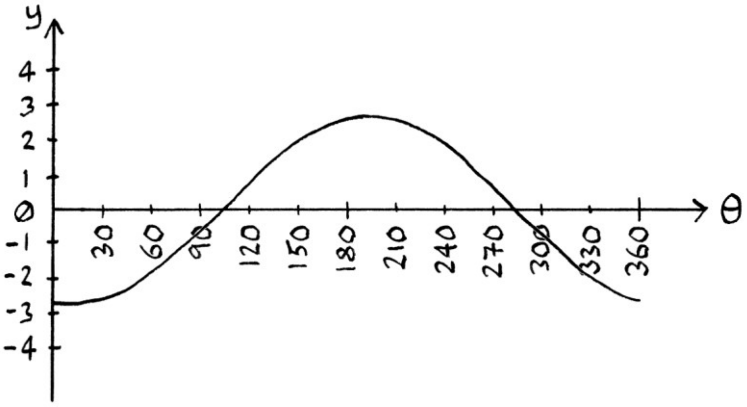
The more complicated formula would look like this:
$\text{radius}^2 = (A \cos (\theta + \phi - 90))^2 + (A \cos (\theta + \phi))^2$

We can see that this still refers to two points on the graph that are 90 degrees apart. It does not matter whether the wave is a Sine wave, a Cosine wave, or any pure wave with a phase – taking any two points that are 90 degrees apart will work in this way to find the radius of the circle from which the wave is, or could have been, derived.

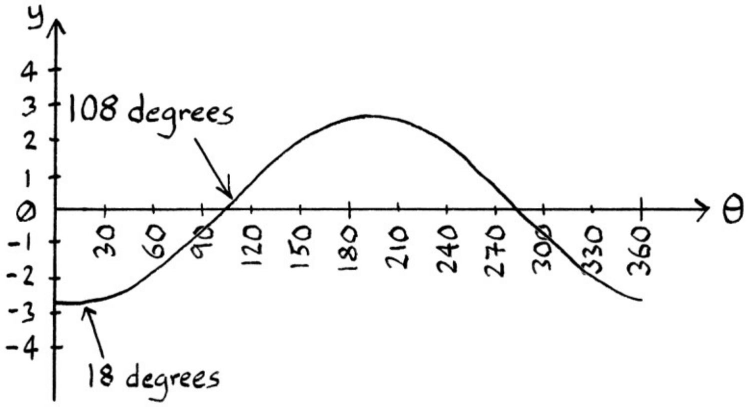
Example

As an example, we will use a Sine wave graph with the formula:
"$y = 2.7 \sin (\theta + 256)$".

We will pretend that we have forgotten its formula. The graph looks like this:

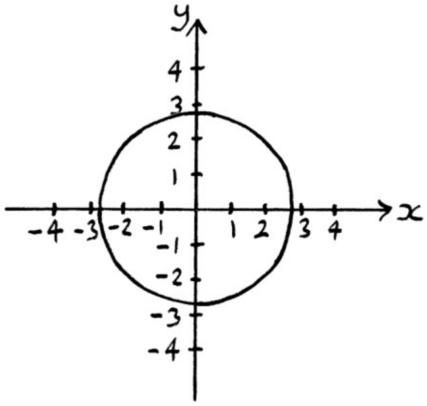


We will pick *any* point on the graph – it does not matter if it is zero or not. We will pick the point at θ = 18 degrees. The y-axis value is −2.6934. We then pick a point that is 90 degrees later, so in other words at θ = 108. The y-axis value is 0.1883.



We put them into the formula:
radius^2 = −2.6934^2 + 0.1883^2
radius^2 = 7.28986045
radius = 2.7000 (to 4 decimal places).

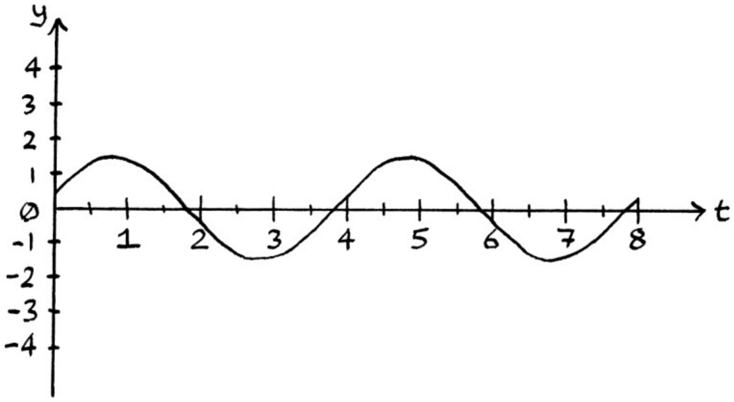
We have calculated that the radius is 2.7 units, which is correct.
We can now draw a circle centred on the origin of the axes with a radius of 2.7 units.



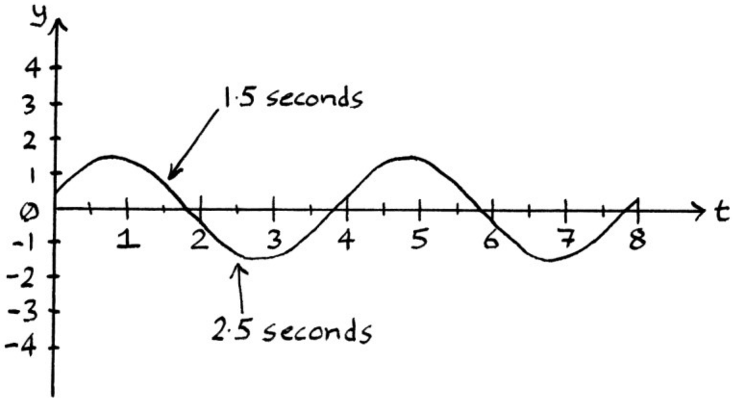
To demonstrate how we can use any two points 90 degrees apart, we will take another two points: 87.2 degrees and 177.2 degrees. The y-axis values are −0.7804 and 2.5848. We put these into the formula:
radius^2 = −0.7804 ^{2} + 2.5848^2
radius = 2.7000 (to 4 decimal places), which is the same result as before.

Time example

We can still use this method if we have a time-based wave. As an example, we will work with the formula “y = 1.5 sin ((360 * 0.25t) + 15)”. The graph looks like this:



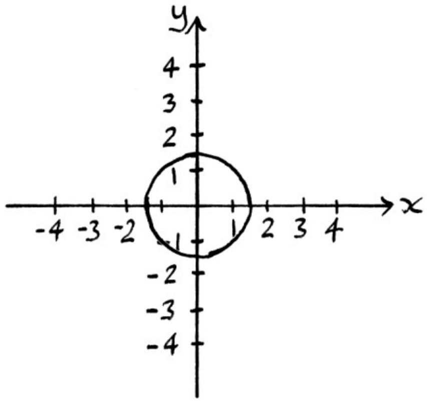
We will pretend we have forgotten the formula. From the graph, we can see that the wave repeats a cycle once every 4 seconds. Therefore, its period is 4 seconds and its frequency is 0.25 cycles per second. From that, we know that an object rotating around the circle would complete 360 degrees in 4 seconds. Therefore, to take any two points that are 90 degrees apart (a quarter of a circle), we will need to take any two points that are a quarter of 4 seconds apart. Therefore, we will read off any two y-axis values that are 1 second apart. We will take a reading at t = 1.5 when “y” is 0.75, and at t = 2.5, when “y” is −1.2990.



Therefore:
radius^2 = 0.75^2 + (−1.2990)^2
radius^2 = 2.2499^2
radius = 1.5000 (to 4 decimal places).

We have calculated that the radius of the circle is 1.5 units and the amplitude of the wave is 1.5 units.

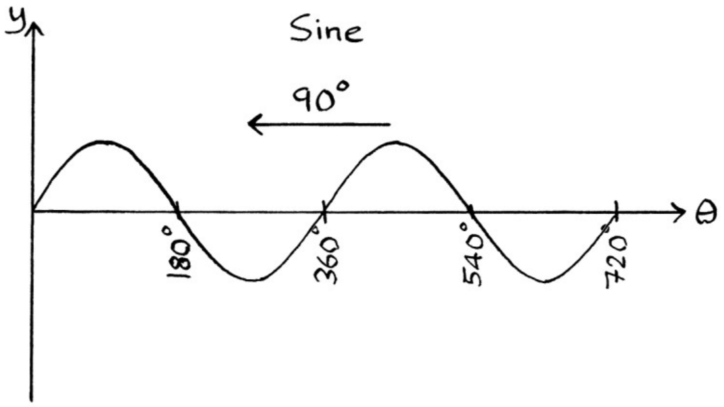
These answers match with the original formula. We can therefore draw a circle with a radius of 1.5 units.

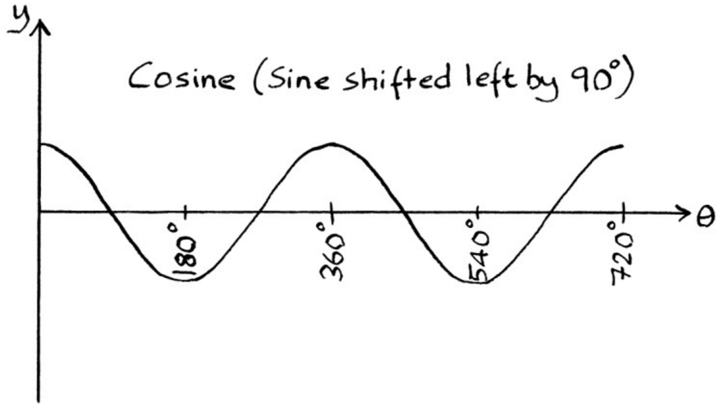


Amplitude and radius: method 3

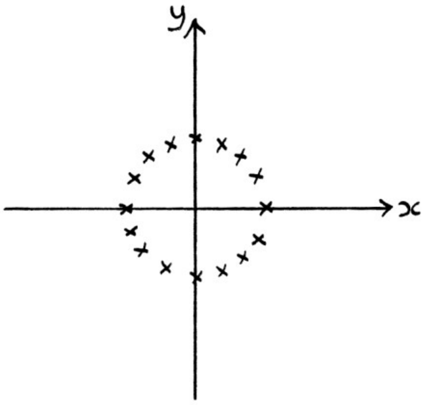
This method is, in essence, identical to method 2, but it is another way of thinking about the underlying concept. As with method 2, the wave can have any amplitude or phase, but it must have a zero mean level. The phase will not be shown in the resulting circle.

If we have a Sine wave, we know that it has the same shape as a Cosine wave that has been slid to the right by 90 degrees. Therefore, we can recreate the Sine wave's corresponding Cosine wave graph, by sliding it 90 degrees to the *left*.

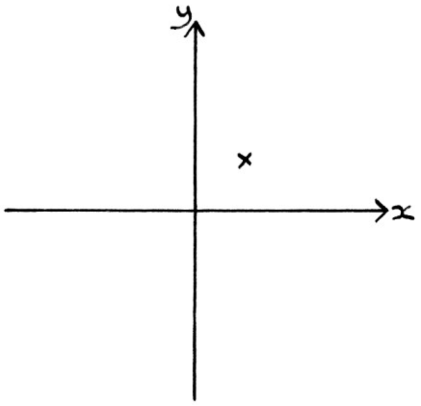


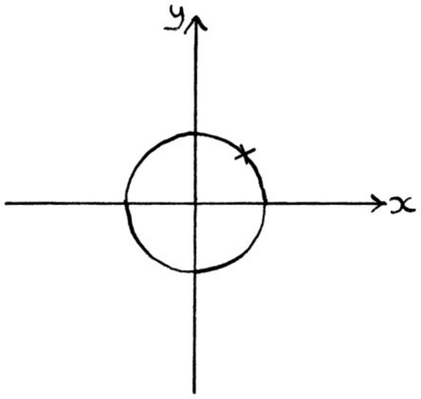


A Sine wave shows the y-axis values of points on the circle's edge at evenly spaced angles from the centre, and a Cosine wave shows the x-axis values of these points. If we have both the Sine wave and the Cosine wave, we can read off corresponding values from each and use them as coordinates to draw points round the circle's edge. The x-axis part of each coordinate comes from the y-axis value of the Cosine wave; the y-axis part of each coordinate comes from the y-axis value of the corresponding angle for the Sine wave.

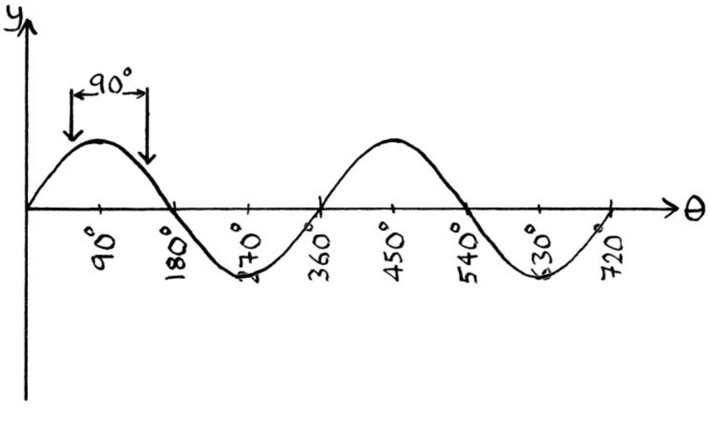


Therefore, if we recreate the Cosine wave for our Sine wave by sliding the Sine wave, we can use corresponding points from both to draw the circle. However, we only actually need one coordinate on the circle's edge as every other point on the circle's edge will be the same distance from the centre.

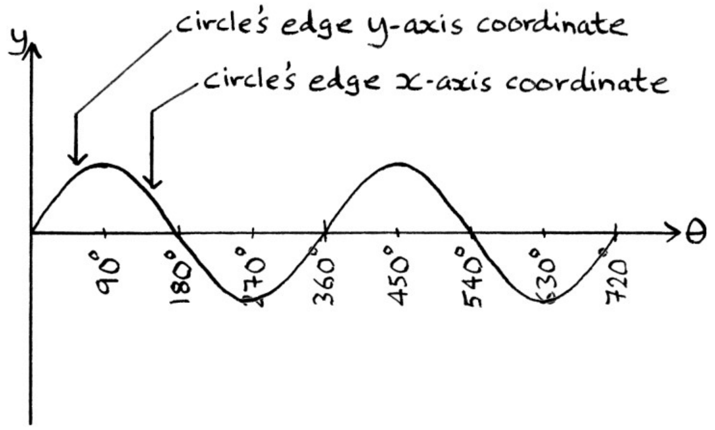




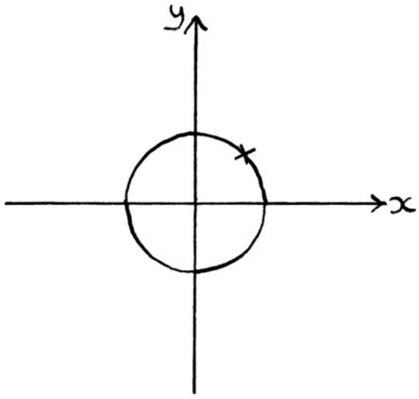
We only need to read one value from the Sine wave and one value from the Cosine wave to have that one coordinate. We created the Cosine wave by sliding the Sine wave 90 degrees to the *left*, which, if you are not confused by phase, you will know is equivalent to adding +90 degrees to the formula for the Sine wave. Therefore, if we only need one value from the Cosine wave, we could just as easily take that equivalent value from the Sine wave, but at a point 90 degrees to the *right* of the Sine wave's coordinate value.



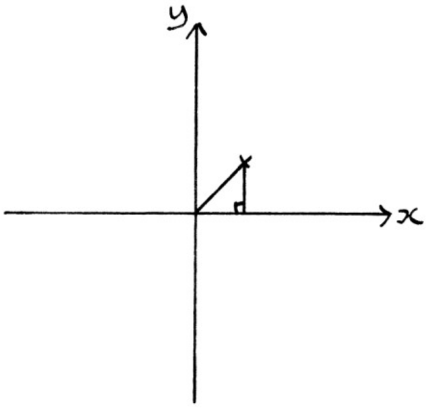
This means that we are taking *any* two points from the Sine wave graph that are 90 degrees apart. This is exactly what we were doing in Method 2. Those two points are treated as coordinates of a point on the circle's edge.


circle's edge y-axis coordinate
circle's edge x-axis coordinate
y
θ
90°
180°
270°
360°
450°
540°
630°
720

We can draw the wave's circle with its centre on the origin, and its circumference passing through that point:



... or, we can treat those coordinates as indicating the end of the hypotenuse of a right-angled triangle:



... in which case, we can use Pythagoras's Theorem to calculate the hypotenuse, which will be equal to the radius and the amplitude:

$\text{hypotenuse}^2 = (\sin \theta)^2 + (\sin \theta + 90)^2$
... or to put it another way:
$\text{radius}^2 = (\sin \theta)^2 + (\sin \theta + 90)^2$

...and this is exactly the same conclusion that we reached in Method 2.

Cosine

If our wave is a Cosine wave, we would take one point and then the point 90 degrees beforehand. This is still the same as taking any two points 90 degrees apart.

Recreating the circle with frequency

The circle on its own cannot indicate frequency, which means it is not possible to use a wave to create a circle that does so. It would be possible to recreate a *helix* from a wave, and that helix would contain a representation of the frequency. However, we will leave this idea for now.

Recreating the circle with phase

In methods 2 and 3 of the section on recreating the circle with amplitude, I showed how taking any two points from a wave that are 90 degrees apart is sufficient to calculate the radius of the circle (and thus the wave's amplitude). It is therefore enough to draw the circle from which that wave is, or could have been, derived (as long as the circle is centred on the origin of the axes). The method works well for calculating the amplitude, but it does not help us calculate the phase of the wave.

The phase in a wave graph is most apparent when the angle on the graph is zero (if the graph is showing angles), or when the time on the graph is zero (if the graph is showing time). We can calculate the phase if the first of the two points we read is the very first point on the graph. For a Sine wave, the second point will be 90 degrees *later*. For a Cosine wave, the second point will be 90 degrees *earlier*.

As before, by using Pythagoras's theorem on the two points, we can calculate the radius of the circle.

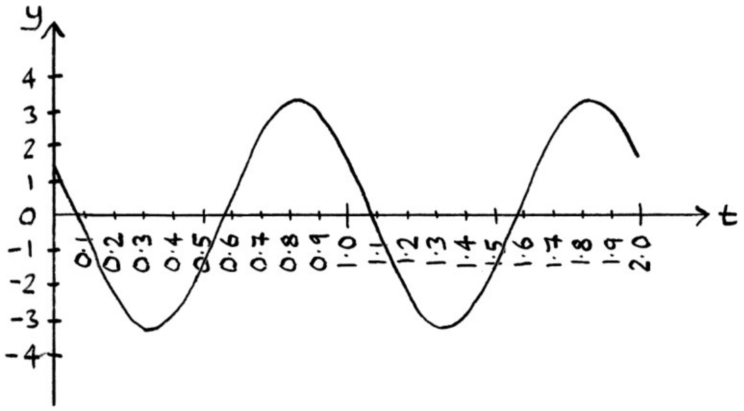
What is different now is that when we treat those two points as coordinates, with the first point being the y-axis coordinate, and the second point being the x-axis coordinate, we will be marking the place on the circle at which the object rotating around it will start. In other words, the two points make up the coordinates of the phase point.

Whether we are dealing with Sine or Cosine waves, the reading from the earlier angle or time will always be the y-axis coordinate of the phase point, and the reading from the later angle or time will always be the x-axis coordinate of the phase point.

Sine wave example

As an example for Sine waves, we will use the graph of the wave:
"y = 3.3 sin (360t + 151)"

We will pretend we have forgotten the actual formula, but we can read the drawings of the graph very accurately. The graph looks like this:

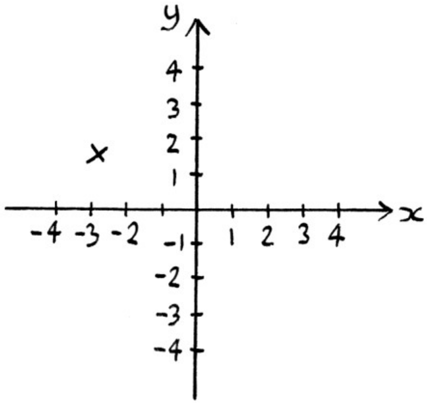


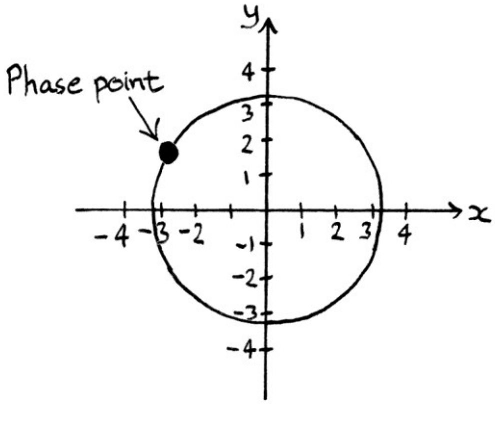
The graph repeats every second, so it has a frequency of 1 cycle per second. This makes it much easier to tell at which angle the object rotating around the circle is at any particular time.

We read off the y-axis value when t = 0. This is 1.5999. Then we read off the y-axis value 90 degrees later. As the frequency is 1 cycle per second, 90 degrees later (or quarter of a cycle or circle later) is equivalent to a quarter of a second later. Therefore, we read the y-axis value when t = 0.25. This is −2.8862.

We will use the reading from the *earlier* time as the y-axis value, and the reading from the *later* time as the x-axis value. Therefore, the reading at t = 0 will be the y-axis coordinate on the circle, and the reading at t = 0.25 will be the x-axis coordinate. These become the coordinates of the phase point of the circle:
(−2.8862, 1.5999).

Either we can plot that point on the circle chart, in which case the point will not only indicate a point on the circumference of the circle, but also the phase:





... or, using Pythagoras's theorem, we can work out that the radius of the circle is 3.3000 units to 4 decimal places. Therefore, the circle has a radius of 3.3 units, and the wave has an amplitude of 3.3 units, so we can draw a circle of 3.3 units. We can then mark the phase on the circle's edge at the point (−2.8862, 1.5999) as before.

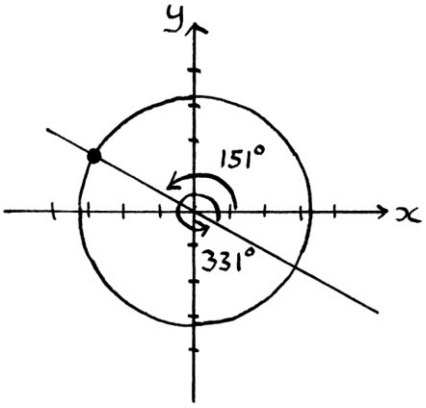
The circle, therefore, shows the amplitude and the phase of the wave.

We can check that the circle is correct. To work out the angle of the phase point, we could just draw the circle and measure the angle, or we could use maths.

If we think of the coordinates as the adjacent and opposite sides of a right-angled triangle, we can use the formula:
“tan θ = opposite ÷ adjacent”
... which we can rephrase as:
“θ = arctan (opposite ÷ adjacent)”.
Therefore:
θ = arctan (1.5999 ÷ −2.8862)
θ = arctan (−0.5543)

A typical calculator will give the result as −29.0008 to 4 decimal places, which we will call −29 degrees. We will convert this to a positive angle: −29 + 360 = 331 degrees.

Remember that the tangent of two different angles in a circle will give the same gradient. The other angle that would have a gradient of −0.5543 is:
331 − 180 = 151 degrees.

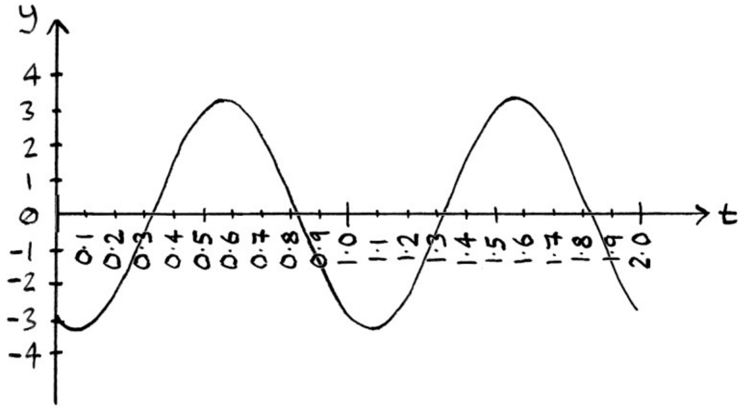


From thinking about the position of the point (−2.8862, 1.5999), we know it is in the top left quarter of the circle. Therefore, the angle we are looking for will be 151 degrees, and the 331 degrees result is irrelevant. Therefore, our point is at 151 degrees, so the phase of the circle is 151 degrees, and the phase of the original wave is 151 degrees. This matches what we know about the formula.

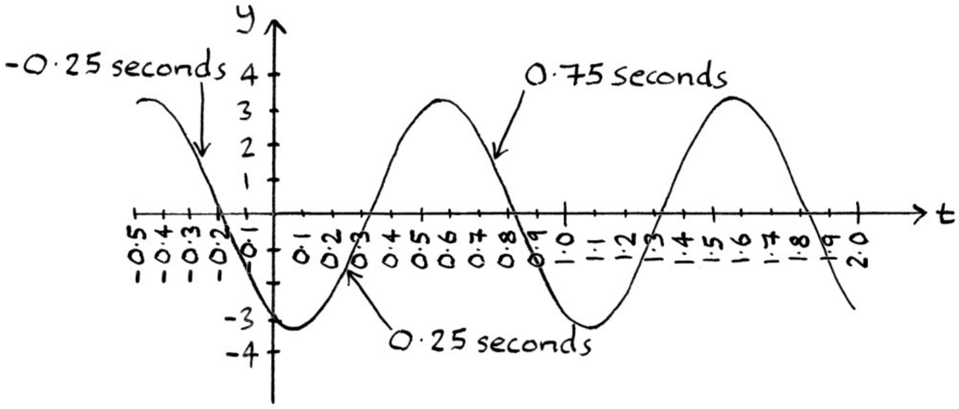
We have created a circle that mentions the phase of the wave using just two points from the Sine wave graph. If we wanted, we could now use the circle to calculate the corresponding Cosine wave for the original wave. Given that we know the Sine wave’s frequency, amplitude and phase, we would not really need to because the Cosine wave would have the same characteristics.

Cosine wave example

For this example, we will use the previous example's corresponding Cosine wave: "y = 3.3 cos (360t + 151)". We will pretend we have forgotten the actual formula. The graph looks like this:



We read off the y-axis value when t = 0. This is −2.8862 to 4 decimal places. Then we read off the y-axis value from 90 degrees *earlier*. As the wave has a frequency of 1 cycle per second, this will be quarter of a second earlier, so we will read the y-axis value at −0.25 seconds. Given that I only drew the positive time values on the graph, either we could extend the graph back in time, or we could take advantage of the wave's repetitiveness and take the value at 0.75 seconds instead. Both values would be the same.



The y-axis value at −0.25 seconds (or 0.75 seconds) is 1.5999 to 4 decimal places.

For the Cosine wave, the y-axis value at 0 seconds will be the x-axis part of the phase point's coordinated on the circle, and the y-axis value at −90 degrees (or −0.25 seconds or +0.75 seconds) will be the y-axis part of the phase point's coordinates on the circle. To express this in another way, the reading from the *earlier* time will be the y-axis value; the reading from the *later* time will be the x-axis value – this will always be the case, whether we are dealing with Sine waves or Cosine waves. The coordinates of the phase point on the circle will be (−2.8862, 1.5999). This is exactly the result we had before in the Sine wave example, and so every other calculation will be identical.

[Supposing we had taken the value at t = 0 and t = 0.25, as we did with the Sine wave, and used the reading from the earlier time as the y-axis value, and the reading from the later time as the x-axis value, we would have ended up with the coordinates (−1.5999, −2.8862). This would have given us the correct amplitude, but the wrong phase. We would have ended up with a phase of 241 degrees, instead of 151 degrees.]

Coordinates to indicate phase and amplitude

Given everything so far, it should be apparent that we can indicate the phase and the amplitude of a circle (and therefore, those of its two derived waves) with just one pair of coordinates on the circle – as long as the circle is centred on the origin of the axes. These coordinates can be thought of as the starting point of an object about to rotate around a circle. They indicate the phase point of the circle.

For a circle with a 3 unit radius and a phase of 47 degrees, it is sufficient just to give the coordinates, (2.04600, 2.1941) – in other words, 3 * cos 47 and 3 * sin 47. That point marks the starting position of an object about to rotate around the circle, and therefore it gives the phase for the Sine and Cosine waves derived from that circle. The point also indicates a place on the circumference of the circle, and therefore indicates the radius of the circle and the amplitude of the Sine and Cosine waves derived from that circle.

The meaning of the coordinates is very easy to understand on the circle, but when thinking of the wave graphs derived from the circle, the meaning is slightly obscured. Therefore, as always, it helps to think about the circles.

Recreating with mean level: method 1

In this section, we look at recreating the mean levels of the circle from reading points off one wave graph.

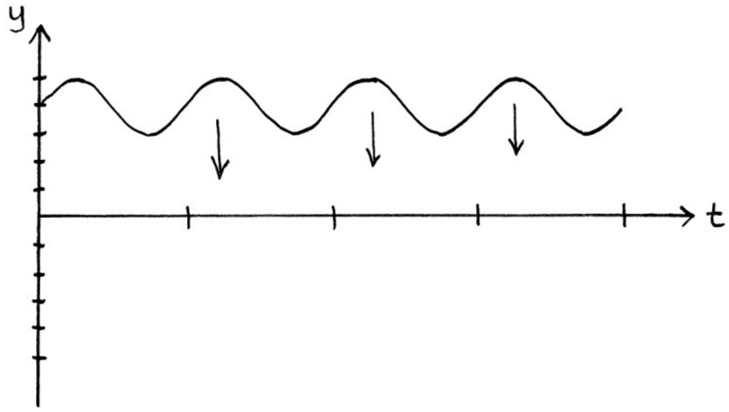
If we only have one wave, we cannot know the mean level of the corresponding wave. Therefore, the position of the recreated circle can only be known for one axis. Ideally, if we do not know the mean level of the corresponding wave, we should treat it as undefined and set it to zero. Otherwise, any result will imply that we did know the mean level of the corresponding wave. Of course, if we have two waves, then we can recreate the circle perfectly.

If a wave has a non-zero mean level, then using the previous methods of calculating the amplitude and phase will not work. As the circle is no longer centred on the origin of the axes, the extra height of the wave distorts the reading of both the amplitude and the phase, so it is no longer accurate. Therefore, other methods need to be used.

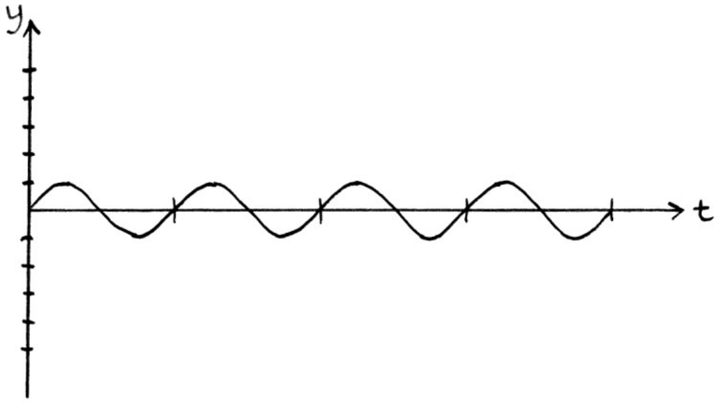
If we are content to have just one mean level on the circle, then there are two methods we can use to recreate the circle.

The most obvious method for a Sine wave is to calculate the mean level of the wave, then shift the wave so it is centred around y = 0. If we use our previous method for working out the amplitude and phase, we can draw the circle as normal, but raise it up the y-axis by the amount of the Sine wave's mean level.

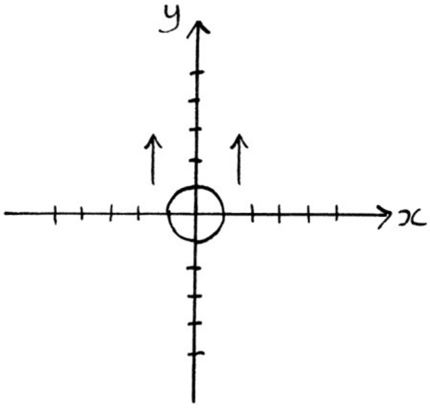
Therefore, we shift the Sine wave up or down to be centred around the x-axis (θ-axis or time axis):



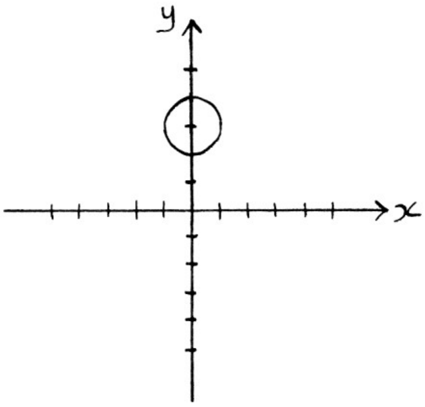
... to end up with this:



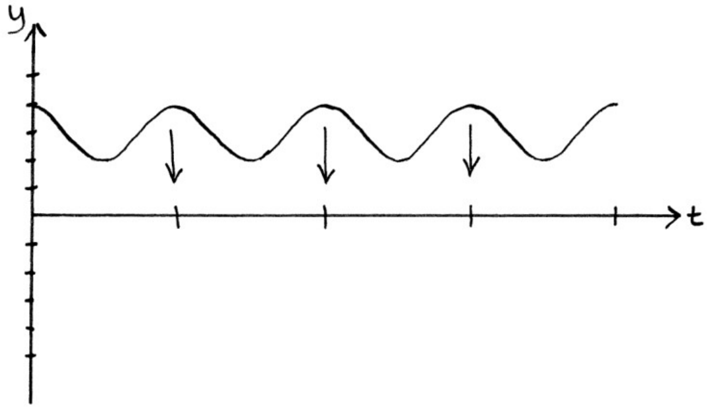
Then we work out what the circle will look like from this centred Sine wave, and then slide the completed circle up or down the y-axis by the amount that we raised or lowered the original Sine wave:



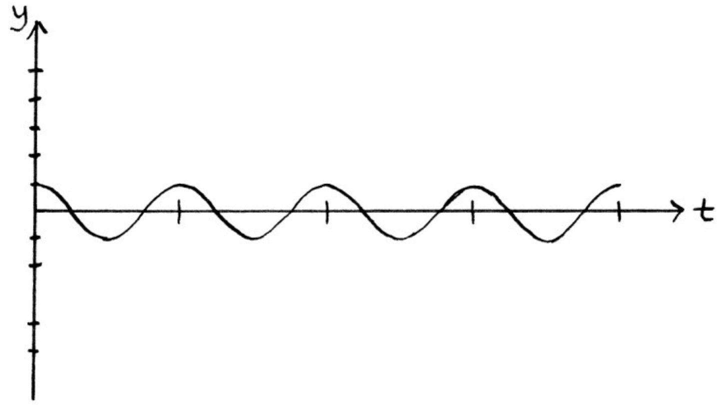
... to become this:



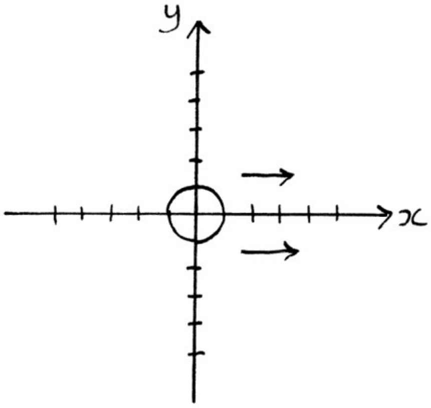
For a Cosine wave, we calculate the mean level, then shift the wave so it is centred around y = 0:



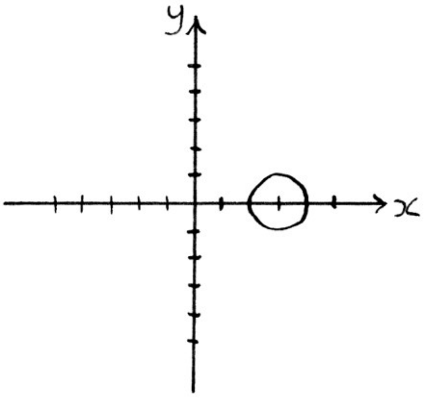
... like this:



We calculate the amplitude and phase, and then draw the circle, which we then shift left or right by the amount we raised or lowered the original Cosine wave:



... to end up with something like this:

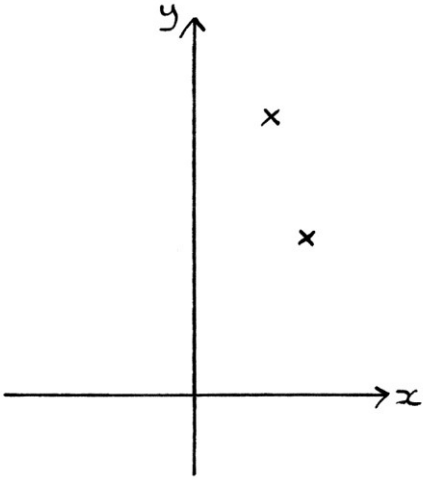


Recreating with mean level: method 2

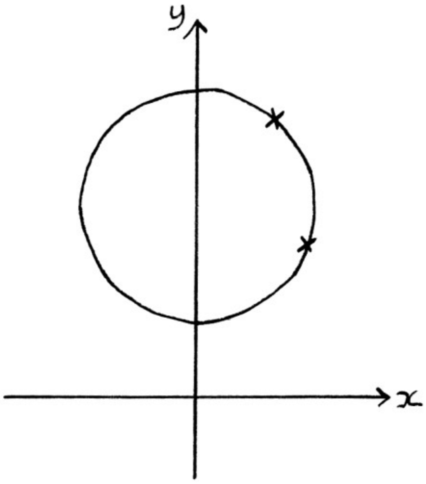
Some basic facts about circles

Ignoring waves for a moment, to describe a circle that is centred somewhere along the y-axis, it is only necessary to know two points on its circumference. The circumference of the circle must pass through both those points, and the centre of the circle is somewhere on the y-axis. Therefore, those two points contain all the information we need for this type of circle.

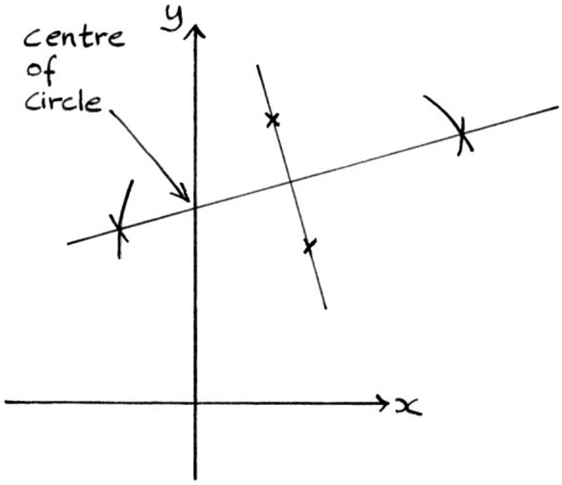
These two points:



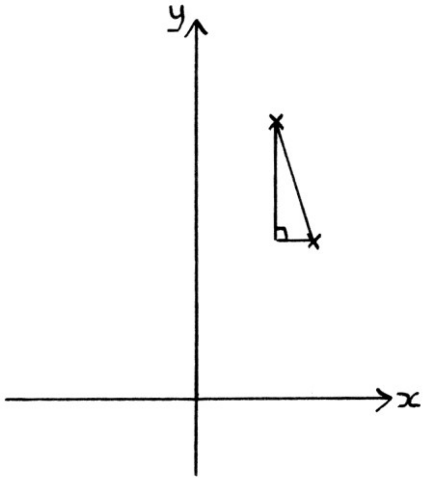
... can only be points on the edge of the following circle (because we know the circle is centred somewhere on the y-axis):



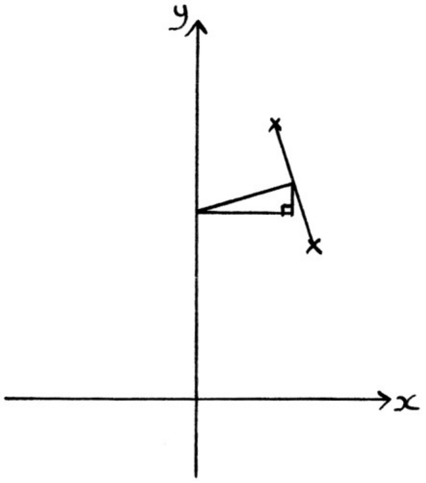
If we have two points and a y-axis drawn on a piece of paper, it is possible to use a compass to find the centre of the circle and the radius. We can use trial and error by expanding and contracting the compass until it is the right size and in the right position. Alternatively, we can draw a straight line through the two points, and then create a second line that is perpendicular to that line and equidistant from the two points. Where that line crosses the y-axis will be the centre of the circle.



We could also use maths to work out the centre of the circle. We do this by considering the two coordinates as the ends of the hypotenuse of a right-angled triangle. We then work out the angle of the hypotenuse using arctan.

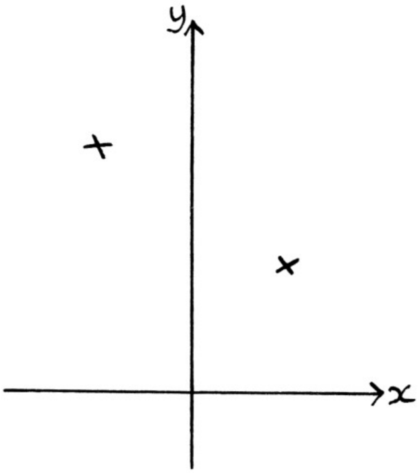


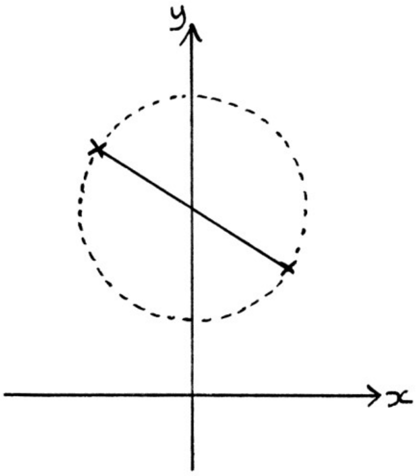
Then we work out the midpoint of the two coordinates – this is just the average of the two "x" coordinates, and the average of the two "y" coordinates. Then, we work out the angle of a line drawn perpendicularly to that midpoint. Then, we think of the midpoint of the two coordinates as the end of the hypotenuse of a right-angled triangle that is at the angle of the perpendicular line. The end of the hypotenuse is a point on the y-axis, which indicates the centre of the circle:

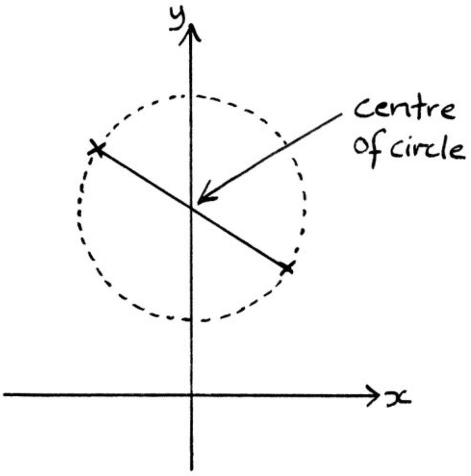


We do not yet know the length of the hypotenuse of this triangle, but we *do* know the adjacent side of this triangle – one end is at x = 0 (because it is on the y-axis), and the other end shares the same y-axis value as the centre of the line drawn between the two original coordinates. Using the angle of the hypotenuse and the length of the adjacent side, we can calculate the opposite side. The length of the opposite side subtracted from the height of the midpoint will be the y-axis value of the centre of the circle. Therefore, we will have calculated the coordinates of the centre of the circle.

We can avoid all of this work if the two original coordinates are from opposite sides of the circle. In that case, we just calculate the midpoint between them, and that is the centre of the circle.



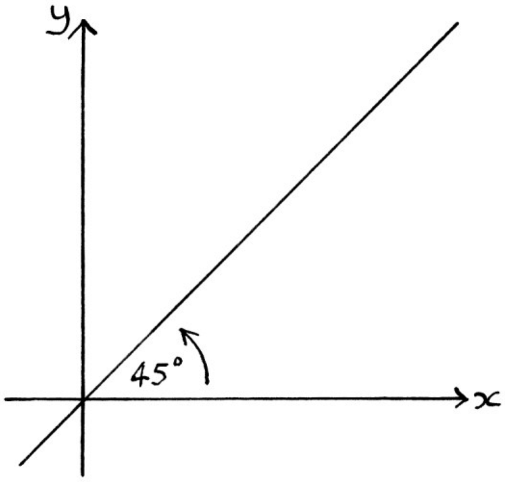




Knowing all of the above will be helpful in recreating a circle from a wave.

Mean level on the Sine wave

When recreating a circle from a wave that has a non-zero mean level, we will encounter one major problem. The original method of recreating a circle with phase creates coordinates by assuming that the Cosine wave is identical to the Sine wave after the Sine wave has been slid to the left by 90 degrees. If the Sine wave has a non-zero mean level, then this will involve us assuming that the Cosine wave also has the same mean level. This is unavoidable. The circle, from which a Sine and Cosine wave with the same mean level are derived, will be centred somewhere on a line where the x-axis is identical to the y-axis. In other words, for both waves to have the same mean level, the circle will be centred somewhere on a 45-degree line. By presuming that the mean level on the missing Cosine wave is the same as that of the Sine wave, we are saying that the circle will be centred somewhere along the 45-degree line.

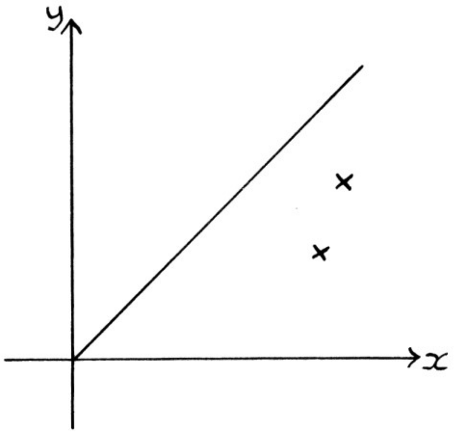


Given that, we can choose to do either of two things here:

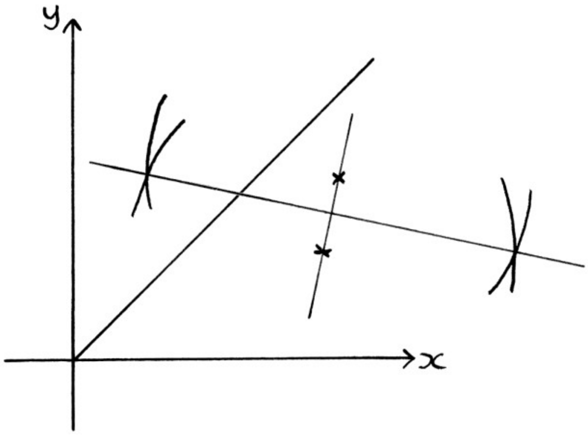
- We can calculate the height of the centre of the circle (which will be on the 45-degree line), and then say that the Cosine wave has no mean level, so we then re-centre the circle over the y-axis at the same height.
- We can calculate the centre of the circle, and just say that the Cosine wave has the same mean level.

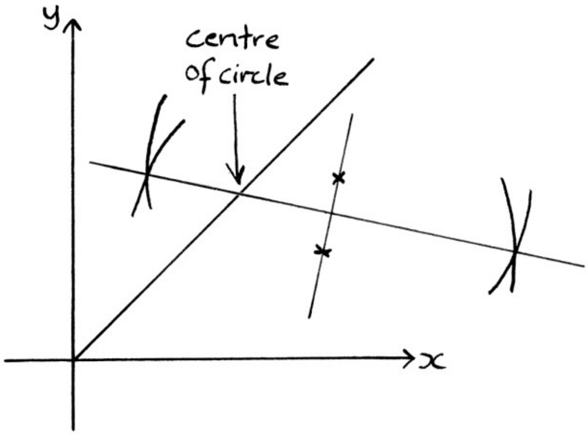
Personally, I think it is best to treat the corresponding Cosine wave as having no mean level. This means we will need to shift the circle so it is centred on the y-axis after our calculations.

In the first part of this section when we were dealing with two pairs of coordinates, the point where the perpendicular line in the calculations crossed the y-axis was the centre of the circle. Now, the point where the perpendicular line crosses the *45-degree line* will be the centre of the circle.

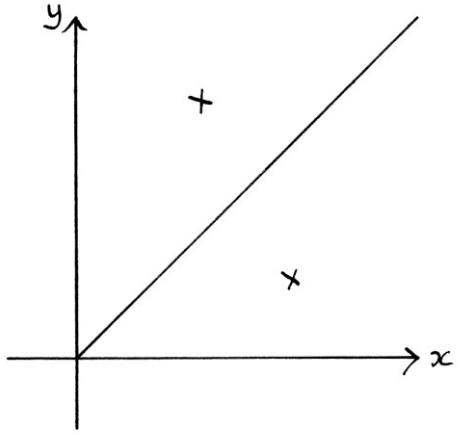


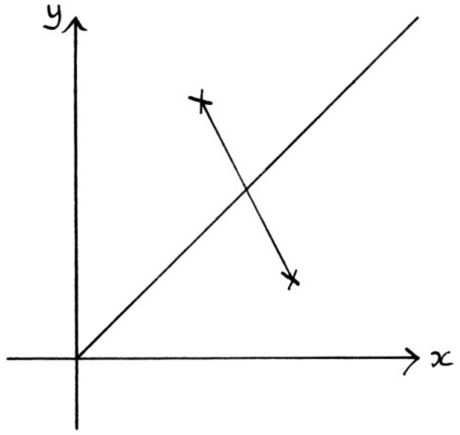
This is still very easy to calculate using a compass and ruler – we just draw a line perpendicular to the middle of a line joining the two coordinates. Where that line meets the 45-degree line is the centre of the circle:

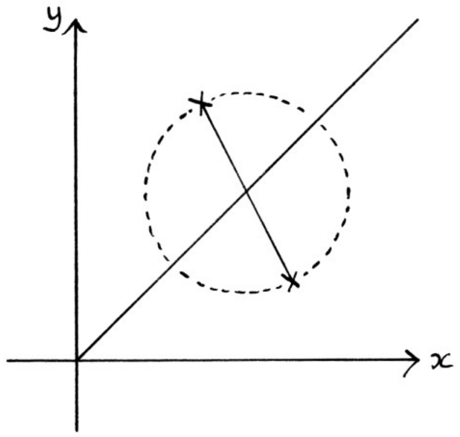




It is slightly more effort to do this using maths. We are really finding the point on the perpendicular line where the x-axis value is the same as the y-axis value. Fortunately, we do not need to worry about this as we can pick two coordinates that are at 180 degrees around the circle. The centre will be on the 45-degree line, but we do not need to do any maths, save for averaging the coordinates to find the centre.







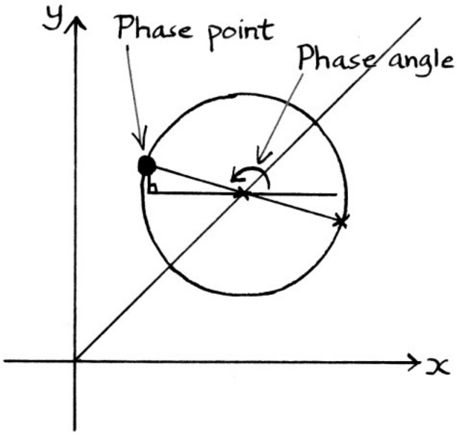
The basic process

For an angle based wave, we take the y-axis value on a Sine wave graph when θ = 0, and then the y-axis value 90 degrees later. For a time-based wave, we take the y-axis value when t = 0, and then the y-axis value 90 degrees later. These make up our first pair of coordinates on the circle. [As always, the reading from the *earlier* angle or time will be the y-axis coordinate, and the reading from the *later* angle or time will be the x-axis coordinate.] Then we take a second pair of values 180 degrees later, in other words, when the angle is 180 degrees and when the angle is 270 degrees. [Again, the reading from the earlier angle or time will be the y-axis coordinate, and the reading from the later angle or time will be the y-axis coordinate]. These will make up the second pair of coordinates. Our two pairs of coordinates are at opposite sides of the circle.

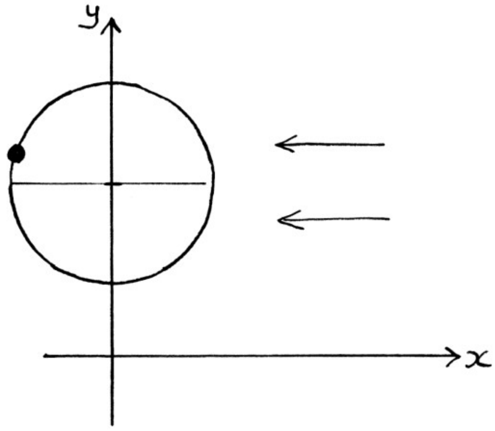
By calculating the midpoint of the two coordinates, we will find the centre of the circle. The circle will be centred on a line at 45 degrees from the origin of the x and y-axes. The height of the centre of the circle is equal to the mean level of the graph.

As we know the centre of the circle and two points on its edge, we can draw the circle that represents the wave. The distance between either of the points and the centre is the radius (obviously) and therefore the amplitude of the wave. The coordinates of the first point mark the phase point.

We can calculate the angle of the phase point with respect to the centre of the circle, by imagining a right-angled triangle with the centre of the circle and the phase point at either ends of its hypotenuse.



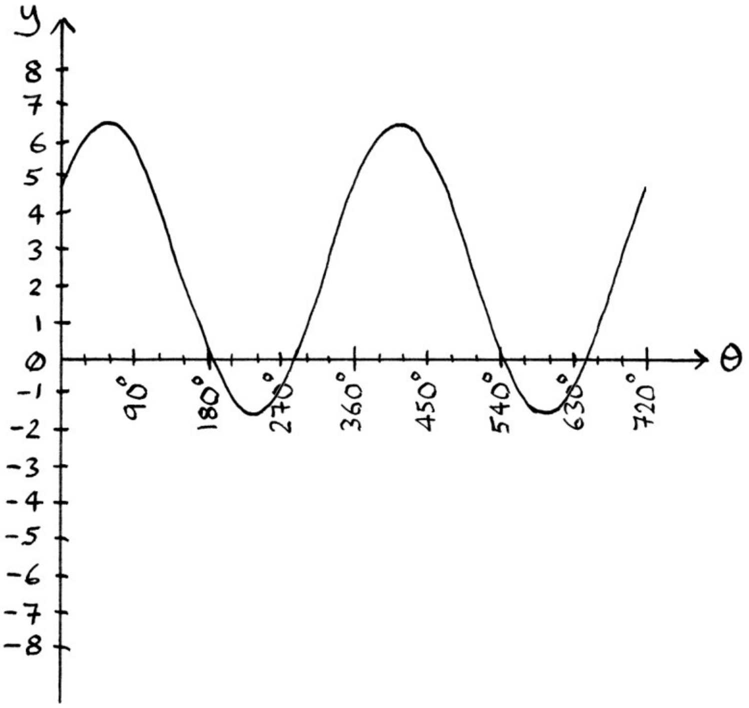
To make the circle accurately represent the mean level for our Sine wave, and therefore have zero mean level for the Cosine wave, we need to shift it horizontally, so its centre is over the y-axis. We shift the phase point by the same amount:



Sine wave example

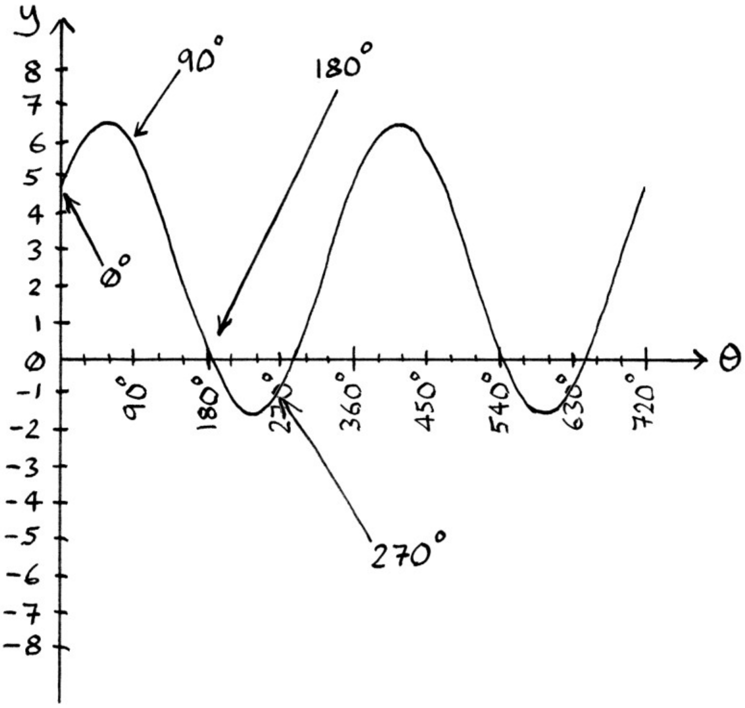
In this example, we will use the graph for the angle-based wave:
"y = 2.5 + 4 sin (θ + 34)".

We will pretend that we have forgotten the formula, but that we can read the values directly off the graph. The graph looks like this:

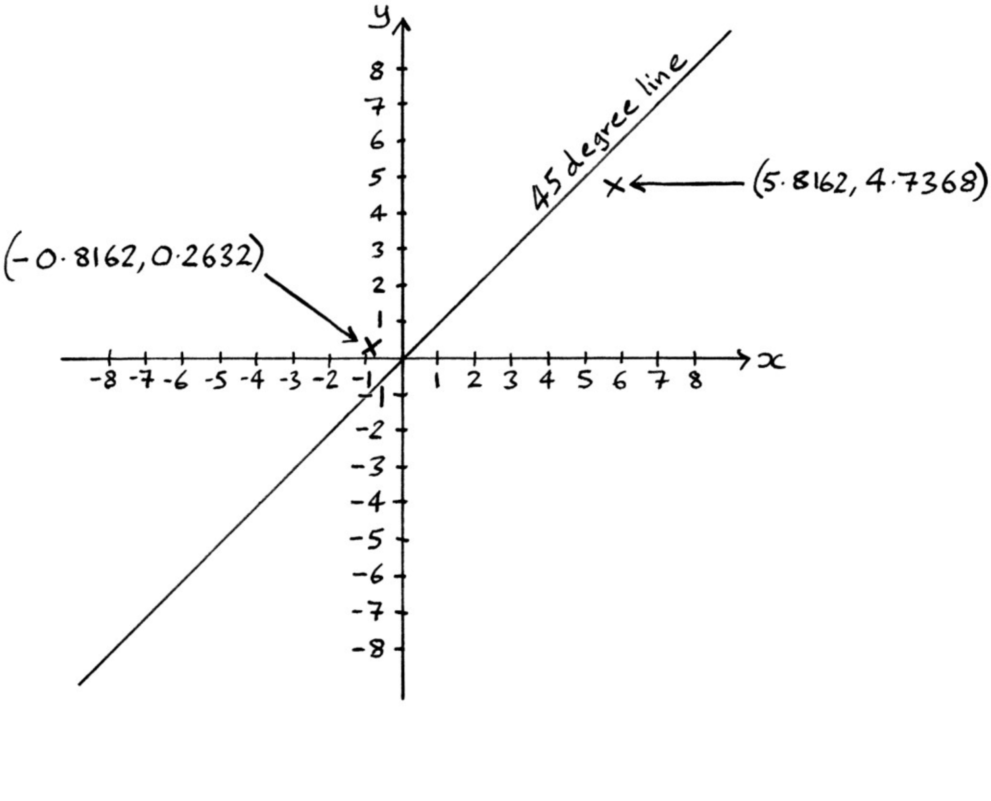


We read the y-axis points at θ = 0 and θ = 90. These are 4.7368 and 5.8162.

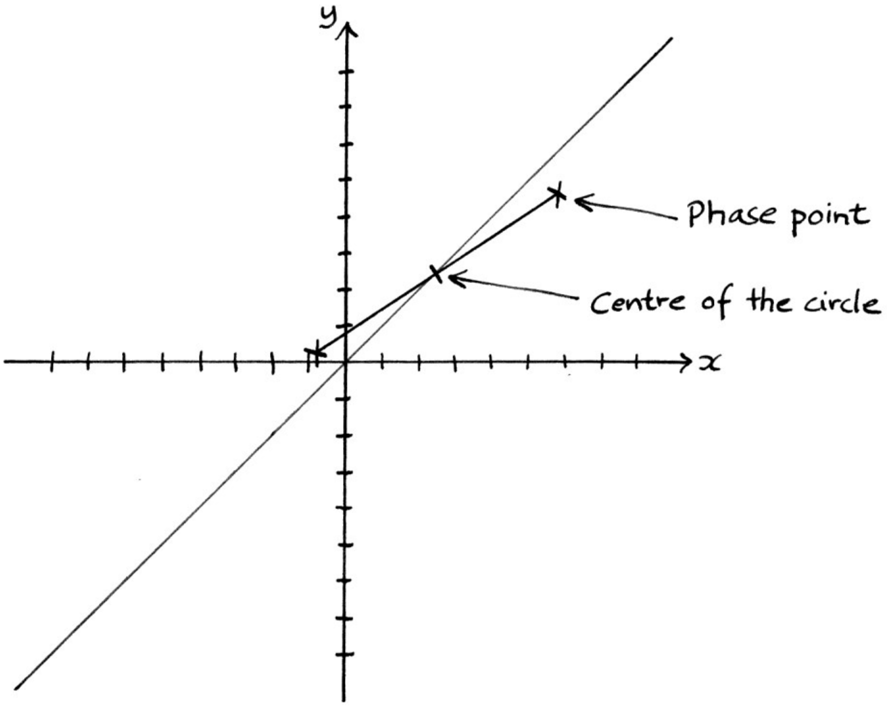
We then read off the points at θ = 180 and θ = 270. These are 0.2632 and −0.8162.

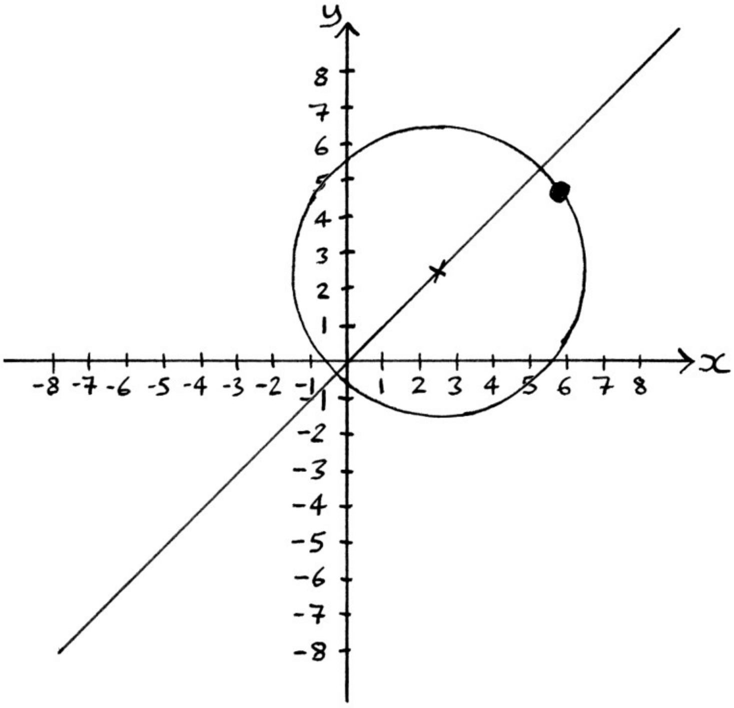


Therefore, the pairs of coordinates of our two points are (5.8162, 4.7368) and (−0.8162, 0.2632). We can plot these on our future circle chart:



The centre of the circle is the midpoint of these two pairs of coordinates, which will also be a point on the 45-degree line. The height of the midpoint is the mean level of the graph. The first pair of coordinates indicates the phase. We, therefore, have a circle that indicates the amplitude, phase and mean level of the graph:





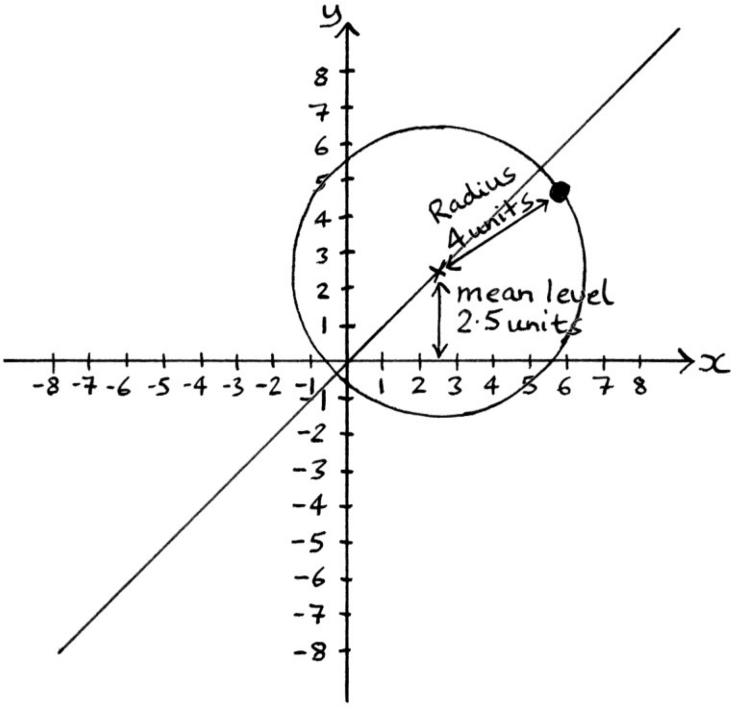
We can check that the circle is correct using maths. The centre of the circle is the midpoint between the two coordinates, which can be calculated as the average of each corresponding value: the x-axis value is 0.5 * (5.8162 + −0.8162) = 2.5; the y-axis value is 0.5 * (4.7368 + 0.2632) = 2.5. Therefore, the midpoint, which is the centre of the circle, is at (2.5, 2.5). The mean level of the wave graph is therefore also 2.5 units. Remember that the centre is on the 45-degree line.

We can calculate the circle's radius by finding the distance from the midpoint to either of the two original coordinates. We will calculate the distance from (5.8162, 4.7368) to (2.5, 2.5). We can use Pythagoras's theorem on the difference in the x-axis and y-axis values:

$\sqrt{(5.8162 - 2.5)^2 + (4.7368 - 2.5)^2}$

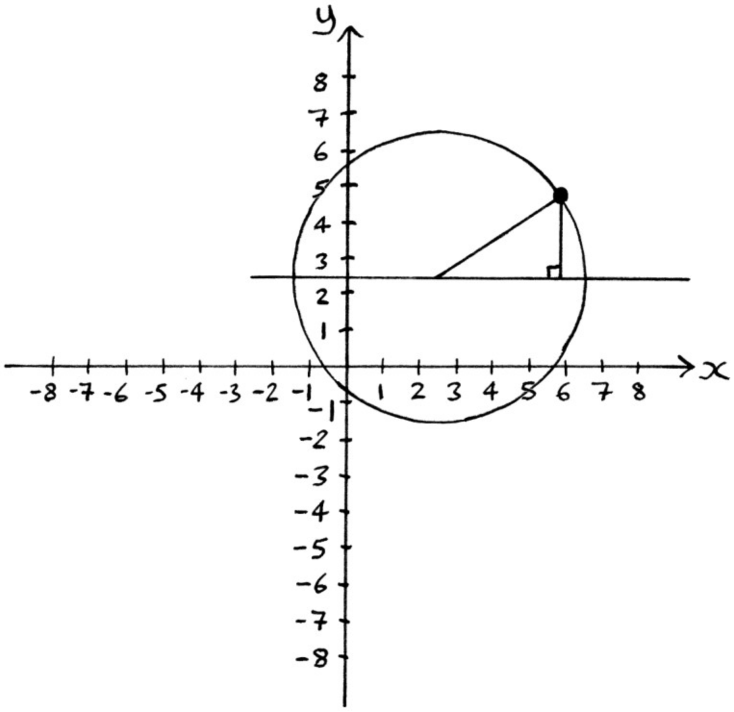
$= \sqrt{3.3162^2 + 2.2368^2}$

= 4.0001 (to 4 decimal places), which given the accuracy we have been using, we will round to 4 units. This is also the amplitude of the wave.



To find the angle of the phase point, we need to calculate it from the centre of the circle (not from the centre of the axes). To do this, we think of a right-angled triangle with the phase point and the circle's centre at either ends of its hypotenuse.

The triangle is essentially the same as we used to obtain the radius.



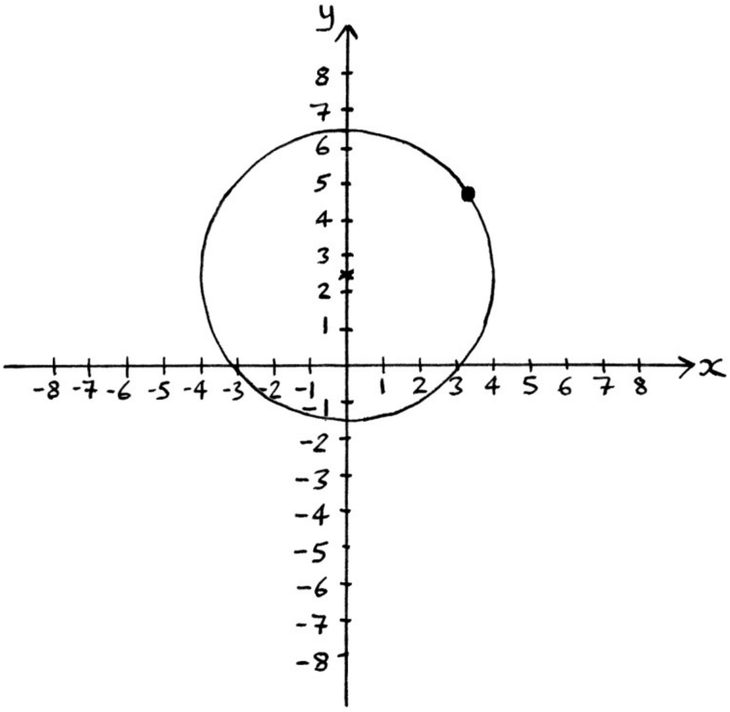
The opposite side will be 5.8162 – 2.5 = 3.3162 units. The adjacent side will be 4.7368 – 2.5 = 2.2368 units. We already know that the hypotenuse is 4 units long as it is the same length as the radius of the circle.

We can calculate the angle using:
"tan θ = opposite ÷ adjacent"
... which we will rephrase to:
"θ = arctan (opposite ÷ adjacent)"

This gives us:
"θ = arctan (2.2368 ÷ 3.3162) = 33.9999 degrees (to 4 decimal places)
... which we will say is 34 degrees. [Of the two possible results, this is the one we want – the other one would be at 34 + 180 = 214 degrees, which would be in the wrong quadrant of the circle].

An easy mistake to make here is to divide the adjacent side of the triangle by the opposite side, instead of dividing the opposite side by the adjacent side. Another possible mistake is to confuse the quadrants of the *whole chart* with the quadrants of the *circle we are dealing with*. [If the circle is centred on the origin of the axes, then they will be the same thing, but if the circle has a mean level, then they will not be the same. We always want to pay attention to the quadrants of the *circle* we are dealing with.]

We need to shift our circle to be centred on the y-axis, so it only has a mean level for the Sine wave. [Technically, we do not *have* to, but we cannot know what the corresponding Cosine wave's mean level is, so it is better to say that it has a zero mean level, instead of assigning it anything else].

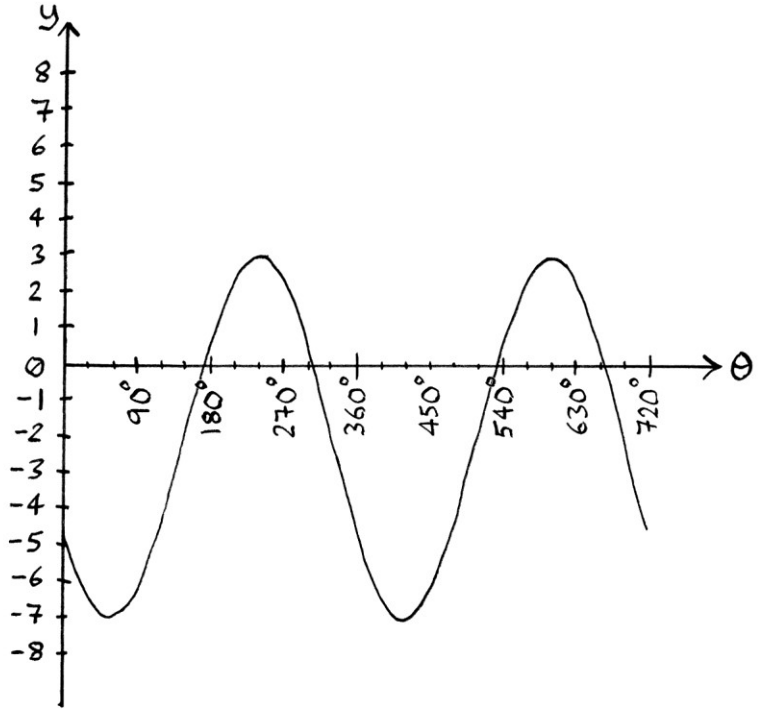


We have shown that the circle has a radius of 4 units, a height of 2.5 units, and a phase of 34 degrees. This means that the Sine wave has an amplitude of 4 units, a mean level of 2.5 units, and a phase of 34 degrees. This matches the formula that we started with: "y = 2.5 + 4 sin (θ + 34)". The corresponding Cosine wave would have the same characteristics, but we would be unable to know its mean level. We could say that its formula is "$y = h_c + 4 \cos(\theta + 34)$", where we do not know the value of "h_c".

Cosine wave example

To recreate the circle for a Cosine wave using this method, we read off the points at θ = −90 and θ = 0 (which we could also phrase as θ = 270 and θ = 0), and then the points at θ = 90 and θ = 180. The second set of points are at 90 and 180, and not at 180 and 270 (as they would be with a Sine wave), because we want them to be 180 degrees later than the first two points.

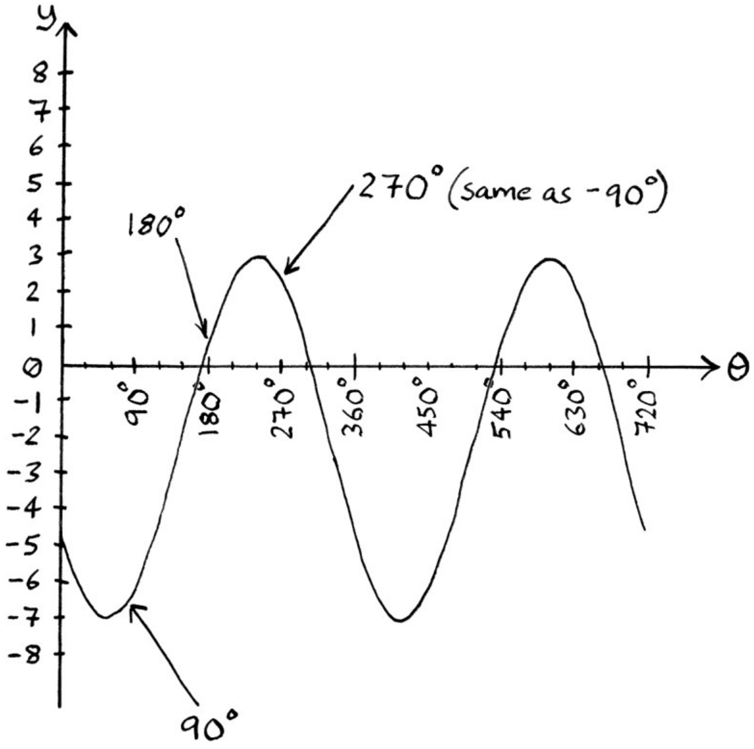
As an example, we will look at the Cosine wave graph for "y = −2 + 5 cos (θ + 120)". We will pretend we have forgotten the formula, and that we can read the drawing with great accuracy. The graph looks like this:



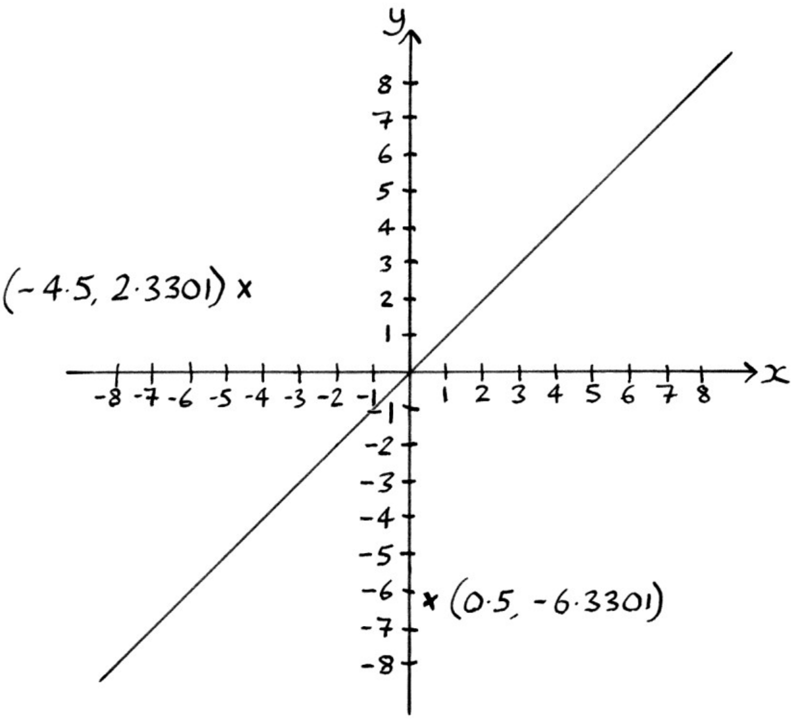
We then want to read off the points at θ = −90 and θ = 0, but to save having to draw the negative part of the wave, we will read from +270 (which will have the same y-axis value as that at θ = −90) and θ = 0.

The first reading (at −90 degrees or 270 degrees) will be the y-axis part of the first pair of coordinates and the second reading (at 0 degrees) will be the x-axis coordinate. We end up with (−4.5, 2.3301).

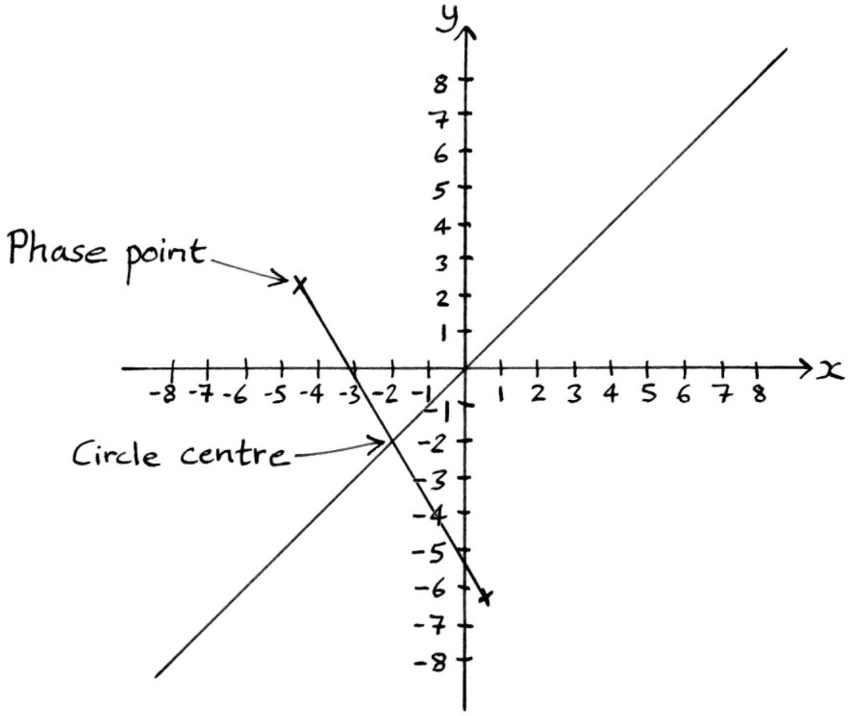
We then read the vales at θ = 90 and θ = 180, and these make up the "y" and "x" parts of our second pair of coordinates, which become (0.5, −6.3301).

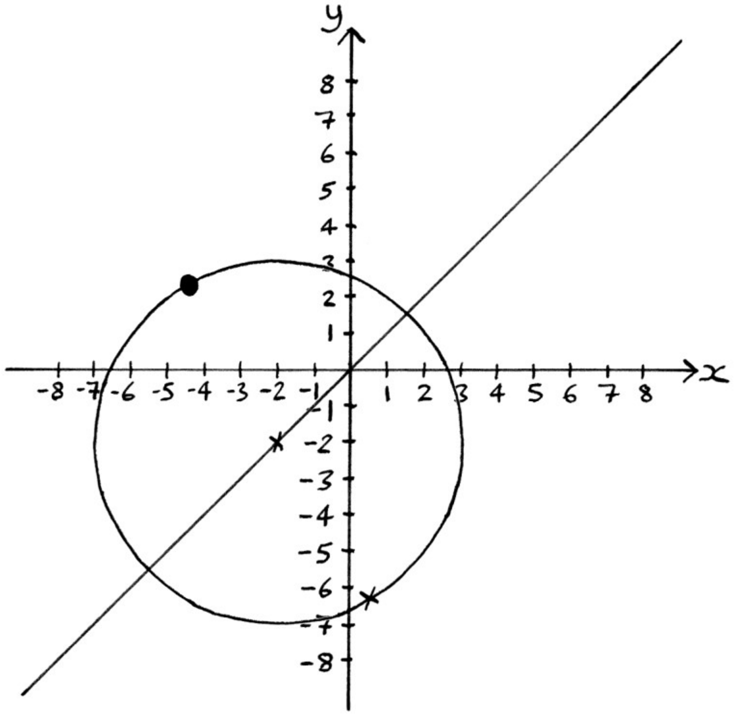


We then plot those points on our future circle chart:



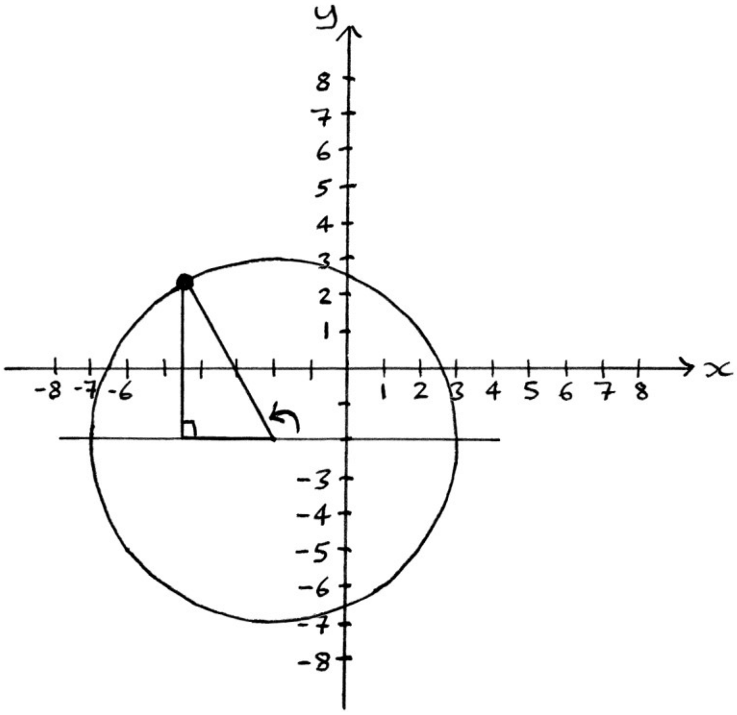
The two pairs of coordinates, which are (−4.5, 2.3301) and (0.5, −6.3301), indicate points at either side of a circle centred on the 45-degree line. The first pair of coordinates indicates the phase point of the circle. The centre of the circle will be in between the two points. The x-axis coordinate will be (−4.5 + 0.5) ÷ 2 = −2. The y-axis coordinate will be (2.3301 + −6.3301) ÷ 2 = −2. Therefore, the centre is at (−2, −2). This is on the 45-degree line, which confirms that it is correct.



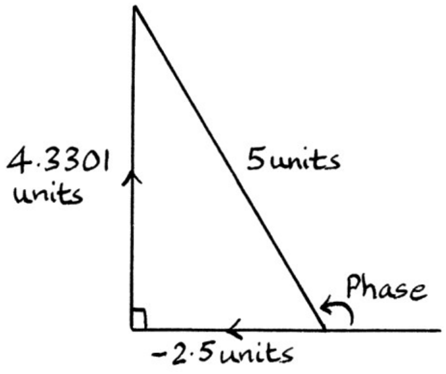


The mean level of the wave will be the x-axis value of the centre of the circle, so is −2 units.

The radius of the circle, and therefore the amplitude of the wave, will be the distance from the midpoint to either of the pairs of coordinates. The distance from (−4.5, 2.3301) to (−2, −2) can be calculated by treating the distance as the hypotenuse of a right-angled triangle, and using Pythagoras's theorem:

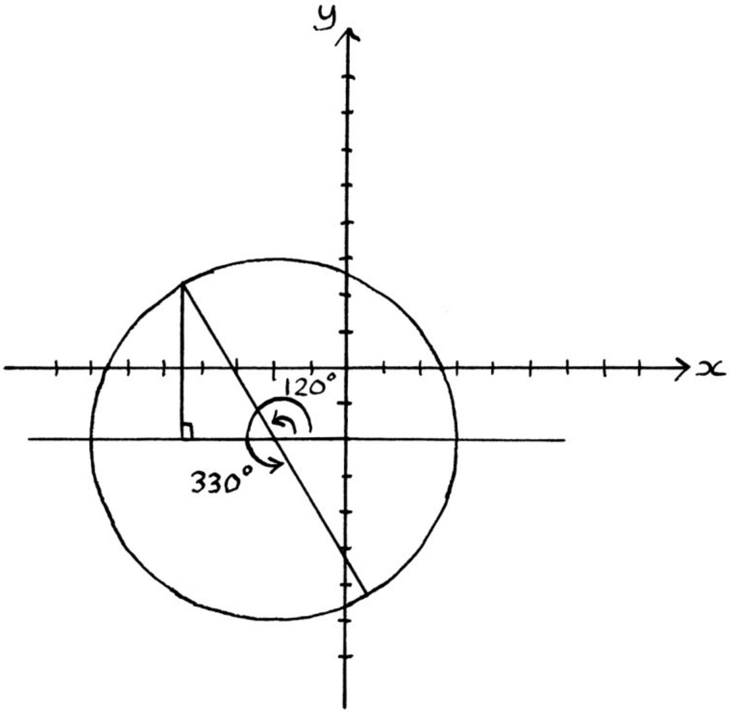


The adjacent side is −4.5 − −2 = −2.5 units long; the opposite side is 2.3301 − −2 = 4.3301 units long. Pythagoras's theorem gives us the hypotenuse as 5.0000 units (to 4 decimal places), which we will call 5 units. This is the radius of the circle and the amplitude of the wave.



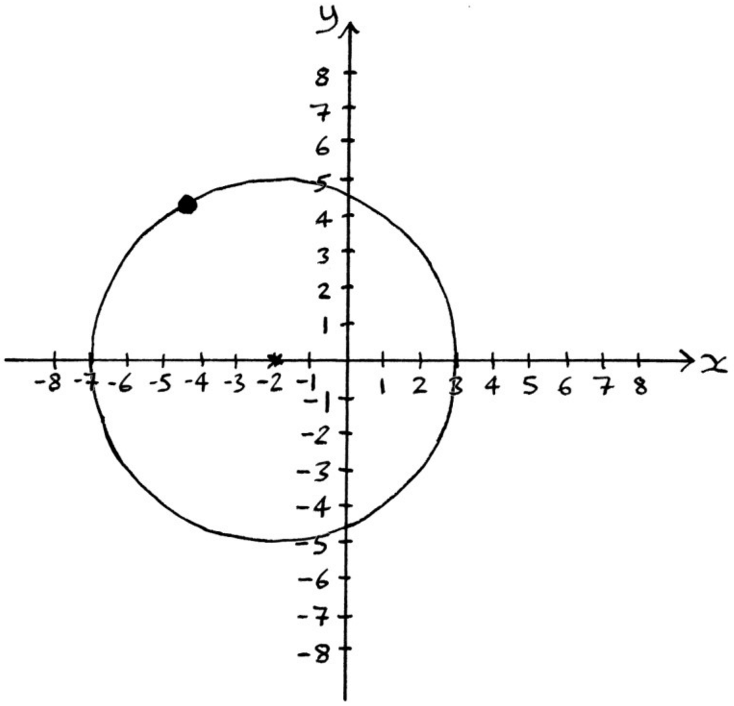
The phase of the circle and the wave will be the angle of the phase point from the centre of the circle, which is also the angle of the hypotenuse of that triangle starting from the outside, because the triangle is the wrong way around.

The gradient is 4.3301 ÷ −2.5 = −1.73204. The arctan of this is −59.9998 degrees, which we will round to −60 degrees. Converting this to a positive angle, this is the same as +300 degrees. There are two possible angles that will produce a gradient of −1.73204, and these are +300 degrees and 300 − 180 = +120 degrees. As our phase point is in the top left hand quarter of the circle, we know that we want the 120-degree answer, so our phase is 120 degrees. [Again, we want the correct quarter or quadrant of the *circle*, and not necessarily the correct quarter or quadrant of the whole chart].



We have now worked out the amplitude, phase and mean level of the wave.

As with Sine, the circle will be centred on a 45-degree line from the origin of the x and y-axes. If we want the circle to indicate that its derived Sine wave has a zero mean level, and thus avoid implying that we know what its value is, we have to shift the circle up or down so that it is centred on the x-axis. In this way, the circle shows the mean level of the Cosine wave, and treats the mean level of the corresponding Sine wave as zero. As explained before, we cannot deduce the mean level of a corresponding wave, so setting the Sine wave's mean level to zero is probably the best thing to do.



The angle of the phase point from the centre of the circle is 120 degrees. The radius of the circle is 5 units. The x-axis mean level is −2 units. These all match our original formula, which was "y = −2 + 5 cos (θ + 120)".

Summary

If a Sine wave is derived from a circle centred on the origin of the axes (in other words it and its corresponding Cosine wave have no mean level), then it is possible to recreate the circle from which it is derived, and have that circle represent the amplitude and phase, by reading the y-axis values from just two points from its graph. These will be the y-axis value at 0 degrees on the θ-axis (which is t = 0 on a time based wave), and the y-axis value 90 degrees later. These two values are used as a pair of coordinates, with the reading from the earlier angle or time being the y-axis coordinate, and the reading from the later angle or time being the x-axis coordinate. The coordinates indicate the phase point of the circle, which is enough information to know the radius of the circle, and the amplitude and phase of the wave.

If a Cosine wave is derived from a circle centred on the origin of the axes, then it is possible to recreate its circle by reading just two points from its graph. These points are the y-axis value at −90 degrees and the y-axis value at 0 degrees (which is at t = 0 on a time-based wave). These two values are used as one pair of coordinates, with the reading from the earlier time or angle being the y-axis coordinate, and the reading from the later time or angle being the x-axis coordinate. The coordinates allow us to know the radius of the circle and the amplitude and phase of the wave. [If we only have the positive part of the graph, then instead of reading the y-axis value at −90 degrees, we can read the y-axis value at +270 degrees because it will be the same. This will still be the future coordinate's y-axis value.]

Both of these ways of recreating a circle also mean that it is possible to describe the amplitude and phase of a pair of waves with just one pair of coordinates.

If a Sine wave has a non-zero mean level, it is possible to recreate the circle from which it is derived, and have the circle represent the amplitude, phase and mean level, by reading the y-axis values from four points from its graph. These are the y-axis values at 0 degrees and 90 degrees, and the y-axis values at 180 degrees and 270 degrees. These become two pairs of coordinates that indicate two points on either side of the circle, from which can be deduced the centre and the radius, and therefore the amplitude. The first pair of coordinates indicates the phase point. After the calculations, it is best to shift the circle so its centre is over the y-axis, or else it will suggest that the derived Cosine wave has a known mean level too.

If a Cosine wave has a non-zero mean level, it is also possible to recreate the circle from which it is derived, and have the circle represent the amplitude, phase and mean level. The method is the same as for a Sine wave, except the y-axis values needed to make up the two pairs of coordinates are those at −90 degrees (or 270 degrees) and 0 degrees, and at 90 degrees and 180 degrees. Again, it is best to shift the circle so its centre is over the x-axis or else it will suggest that the derived Sine wave has a known mean level.

Both these methods mean that it is possible to represent the amplitude, phase and mean level of a pair of waves on a circle with just two pairs of coordinates on the circle chart – one pair indicates the phase point, and the other pair indicates the point on the opposite side of the circle (at an angle of 180 degrees to the phase point).

Chapter 13: Addition with waves

In the next five chapters, we will look at basic maths involving Sine and Cosine waves. In this chapter and Chapter 14, I will give a basic explanation of addition. In Chapter 15, I will look at how addition can be represented in the frequency domain, and in Chapters 16 and 17, I will give a basic explanation of multiplication. In these chapters, I will not give an exhaustive explanation, but enough to have a reasonable grasp of the subject. There will be a lot that I will leave out.

It is good to know how to do some basic maths on waves, as the results have characteristics that will be useful later on.

For these chapters, it can be useful to have a basic graphing calculator app, so that you can visualise the waves more easily. You do not have to have one, but if you do, it will allow you to experiment with waves to become used to how everything works. Note that many graphing calculator apps work in radians and not degrees. [I will explain radians in Chapter 22]. If this is the case for your graphing calculator, you will need to convert all angles into radians to see the correct curves. For time-based waves, you only need to convert the phase. To convert an angle in degrees into radians, you divide the angle by 360 to see the portion of a circle that that angle represents, and then multiply the result by 2π to convert that portion into radians. The number 360 within the Sined part of the formulas in these chapters will need to be replaced by 2π. The basic idea is that if we have a wave such as this:
“y = 2 + 3 sin ((360 * 4t) + 56)”
... then it will need to be rephrased to be this:
“y = 2 + 3 sin ((2π * 4t) + (2π * (56 / 360)))”

Addition

This chapter on addition is really an overview of adding waves together. There exist rules and reasonably complicated formulas that can help in adding waves, but I am not including them here, as I think it is better to have a more intuitive understanding of addition. Knowing that the result of an addition can be rephrased in a more complicated form does not really help someone understand the basics, and such formulas separate the learner from what is really happening.

If maths on waves seems complicated to start with, just remember that we are really performing maths on all the individual points of a wave. For example, if we add 10 to a wave, we are really adding 10 to every individual y-axis value of the curve along the time axis. If we multiply two waves, we are really multiplying every y-axis value from one wave with the y-axis value for the same time from the second wave. It also helps to remember that the waves are portraying aspects of a circle, and that that circle can also be thought of as countless right-angled triangles. Circles and triangles are much easier to visualise than waves.

When we perform maths on two waves, the basic idea is to do the maths on every corresponding (by time) y-axis value for an infinite amount of time. If we try to do this in practice, we will discover that this is impossible for two reasons – first, it would be impossible even to write down *every* y-axis value for just one microsecond, as there is no limit to how far one can zoom into a wave to see more y-axis values. There are really an infinite number of y-axis values in any moment of time. Second, we cannot perform maths on an infinite time's worth of values because it would take forever. The solution to the first problem is to take evenly spaced y-axis values from each wave – for example a hundred y-axis values per second. This will make the task possible, but at the expense of reducing the accuracy of any results. The solution to the second problem is to perform the maths on only a small section of time. Ideally, this amount of time should include an integer number of cycles. In this way, the results of any calculations will be identical to the cycles afterwards. This will also cause any calculations of mean levels to be accurate.

[If we were using analogue electronics to perform the maths, then it would be possible to perform maths on every y-axis value.]

In these chapters, I will often pretend that we *can* do maths on every y-axis value and for an infinite amount of time.

In these explanations, if I give an example with a Sine wave, the process will be identical for a Cosine wave. Although these examples involve time, the methods are the same for waves that just deal with angles.

Superposition

Addition is a very important part of the study of waves, mostly due to what is called “superposition”. For radio waves or sound waves in the real world, if they exist at exactly the same time and at exactly the same place, they become added together from the point of view of an observer at that time and place.

As a very simplified example, if one radio transmitter is transmitting a wave with the formula:
“y = 2 sin (360 * 252,000t)”
... and a second transmitter is transmitting a wave with the formula:
“y = 3 sin (360 * 252,000t)”
... then a receiver that can pick up both signals equally well, with no loss of power and both with the same perceived phase, will receive a signal with the formula:
“y = 5 sin (360 * 252,000t)”.
[In practice, the amplitude of the wave would have diminished by the time it reached the receiver, so would be less than this total.]

The waves do not have to be of the same frequency – any detectable radio waves, no matter where they come from, become added together from the point of view of a receiver. The same is true of sound waves.

When a radio receiver is switched on, it is not just receiving one frequency of radio wave – it is really receiving *every* frequency of radio wave that is being transmitted within range at the same time, all added together into one gigantic signal. A Long Wave radio receiver, for example, is detecting radio waves from Long Wave transmitters, from Medium Wave transmitters, from FM transmitters, from aeroplanes, from taxis, from household electric wiring, from nearby motor engines, from chemical reactions, from the sun, just to name a few sources, all added together. Fortunately, many of these waves are too faint to be significant – their amplitude is very small. It is the radio’s job to filter out the other unwanted frequencies, and then decode the frequencies that it needs. The first step in this filtering is the antenna – the radio’s antenna will be of a design that ideally reduces frequencies that the radio does not need, and passes on frequencies the radio does need. The next steps in filtering out the unwanted frequencies will be in the electronics of the radio.

When it comes to sound, at any moment, your ears are receiving sounds from air movements, echoes, birds, fridges, voices, music, fans, mice, bats, insects, clocks, engines, volcanoes, earthquakes and so on, all at the same time as one combined signal. All the waves are added together. Again, many of these sounds will be too weak to be noticed. The mechanism of your ears filters out a great deal of the other sounds in this combined signal based on frequency – your ears can only process sounds within a particular frequency range. Your brain then processes the received frequencies to distinguish between useful sounds and unwanted sounds.

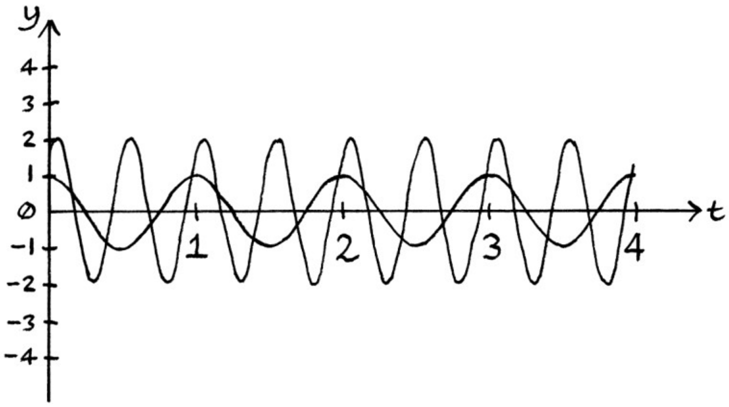
The addition of waves in this way is called "superposition". In a general sense, the word "superposition" means the placing of one thing on top of another. The Latin word "super" means "above" or "over", and the English word "position", in this sense, means "placing" or "arranging". In the academic field of waves and signals, "superposition" can be thought of as placing one wave on top of another so that they are added together.

Not all real-world waves exhibit the behaviour of superposition, but the two most commonly analysed waves, electromagnetic waves (radio waves, light waves and so on) and sound waves do.

Superposition is most obvious when there are waves of the same frequency occurring at the same time. If the phases match, the waves will be added together to produce a louder sound or a stronger signal. If the phases are 180 degrees apart, the waves will cancel each other out. You can witness superposition behaving in this way if you are ever near a sports stadium. The sound of the cheering of the spectators is many, many times louder than the sound that any one individual could achieve on their own. This is because the spectators are all cheering within a certain frequency range, so there are more occasions when there are matching frequencies and phases coming from their voices, and this means the sound waves become added together to produce a louder sound. [Theoretically, it would be possible to arrange the spectators in a stadium so that their cheering cancelled out to nothing from the point of view of someone listening in a particular place.]

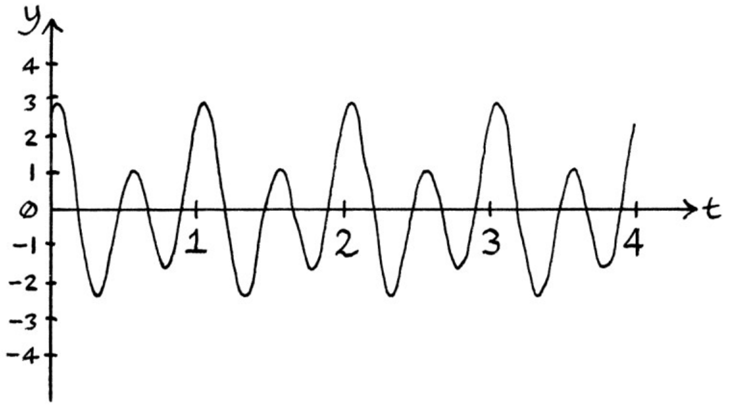
It is important to note that not all waves are electromagnetic waves or sound waves, and not all types of waves add together when existing at the same time and place. An obvious example would be waves used to describe the motion of a pendulum. You can put as many pendulums as you want next to each other, but they will not affect each other in any way.

If we had two waves of a type that was not affected by superposition, they would be independent of each other. An example of such waves drawn on a graph is as so:



[The graph is showing the detection of the waves at a particular place over time.]

However, if the above waves were of a type that was affected by superposition, then they could not exist in this way. Instead, they would become added together as in the following graph (from the point of view of a receiver at a particular place):



In explanations of waves, it is common to see two or more separate waves drawn on the same graph, but this is usually done just to illustrate a difference in their characteristics – it is easier to see differences between waves if they are drawn over each other, than if they are drawn on two separate graphs. However, it is important to remember that if they were radio or sound waves, they could not appear like that in reality – instead they would become added together.

The concept of real-world waves being added together, that is to say, superposition, is one of the most important factors in all signal processing. A huge amount of signal processing involves trying to "un-sum" summed waves by identifying and isolating the individual component waves.

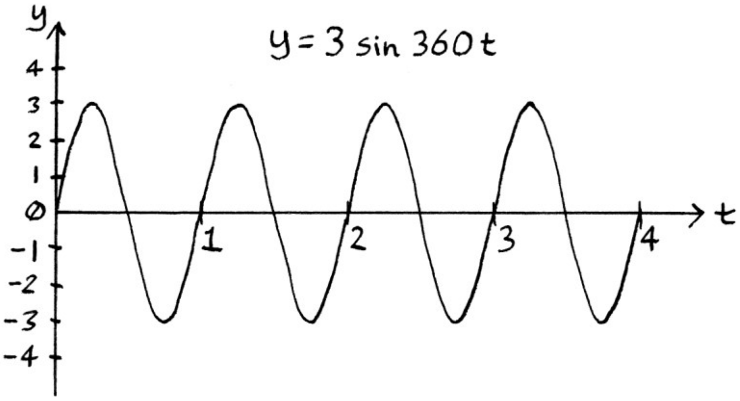
Addition of waves and numbers

Addition

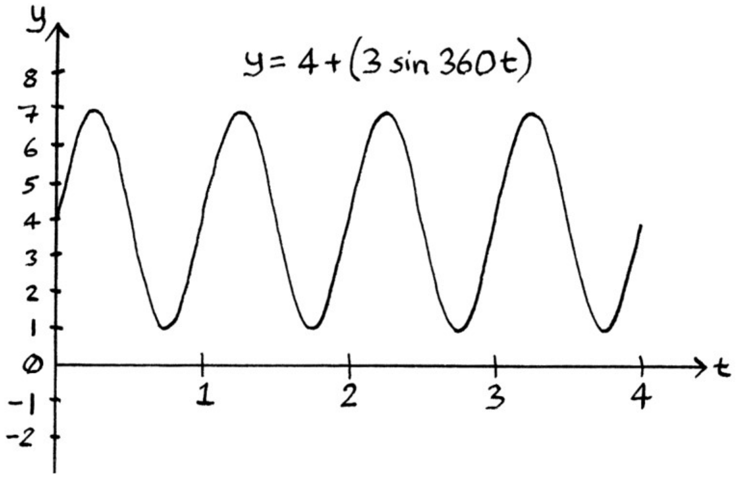
If we add a fixed number to a wave, we are really adding that number to every possible y-axis value. The wave will be shifted up the y-axis by that amount. We are actually changing the mean level of the wave.

As an example, 4 added to “y = 3 sin 360t” results in: “y = 4 + (3 sin 360t)”. Every value on the wave is raised up the y-axis by 4 units:

“y = 3 sin 360t”:



“y = 4 + (3 sin 360t)”:

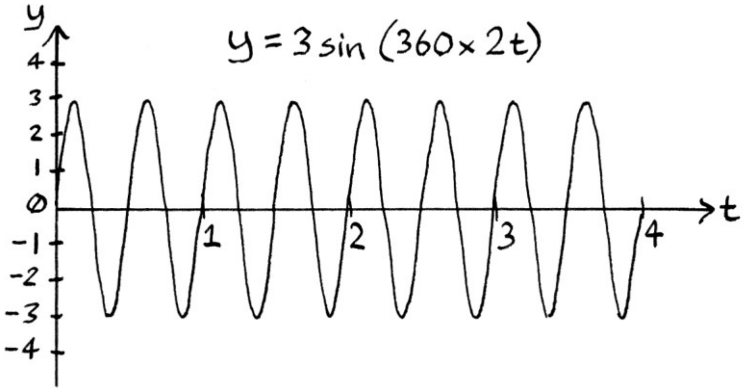


Subtraction

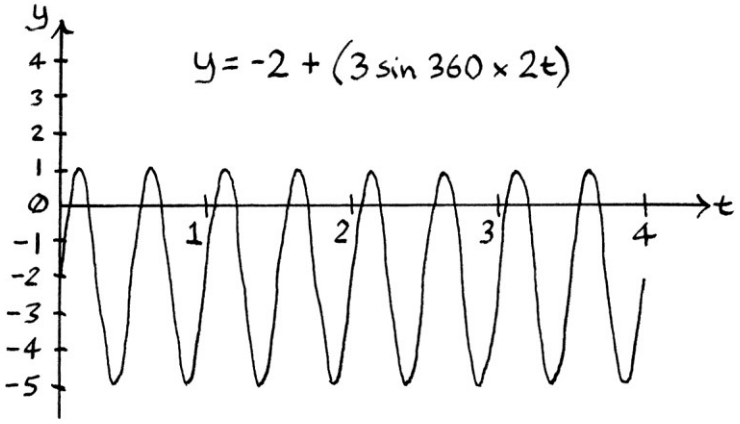
Subtraction of a fixed number is similar – the whole curve becomes lower on the y-axis.

For example, “y = 3 sin (360 * 2t)” minus 2 results in “y = −2 + (3 sin (360 * 2t)”. Every point on the graph becomes 2 units lower.

“y = 3 sin (360 * 2t)”:

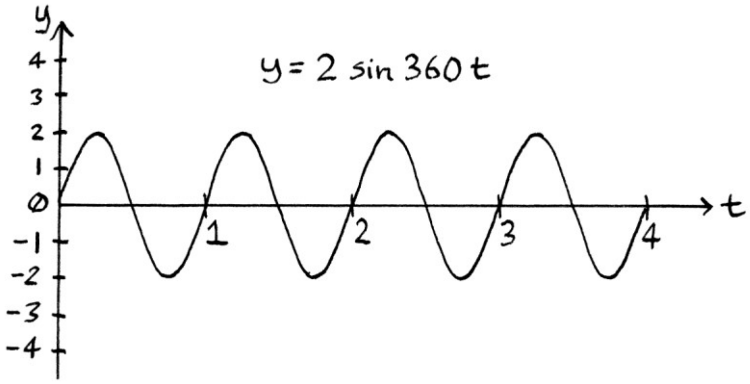


“y = −2 + (3 sin (360 * 2t)”:

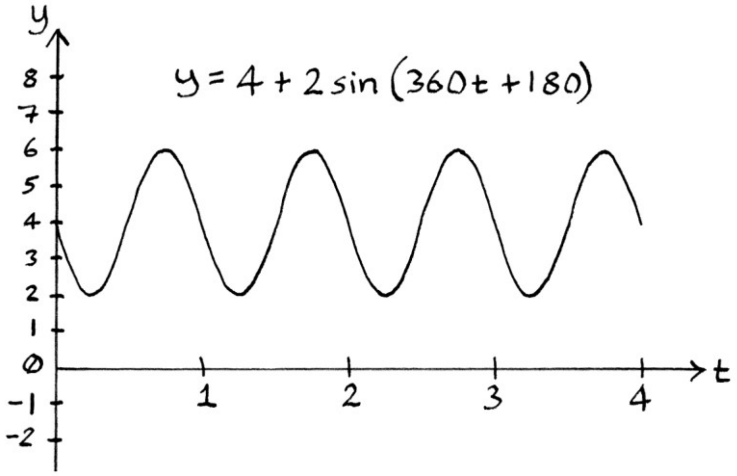


Subtraction of a wave *from* a fixed number results in the curve becoming shifted up or down the y-axis by the amount of that number, at the same time as being inverted. For example, 4 – (2 sin 360t) results in 4 + 2 sin (360t + 180):

“y = 2 sin 360t”:



“y = 4 + 2 sin (360t + 180)”:



The reason for this is clearer if we think about how:
“4 – (2 sin 360t)”
... is the same as:
“4 + (−2 sin 360t)”.

In other words, we can think of the subtraction as really being an addition of a negative-amplitude Sine wave (which is an inverted Sine wave). While learning about waves, it can be good to convert negative amplitudes to positive amplitudes to reinforce how the amplitude of a wave is the same as the radius of the circle it is, or could have been, derived from. To do this, we add 180 degrees to the phase.

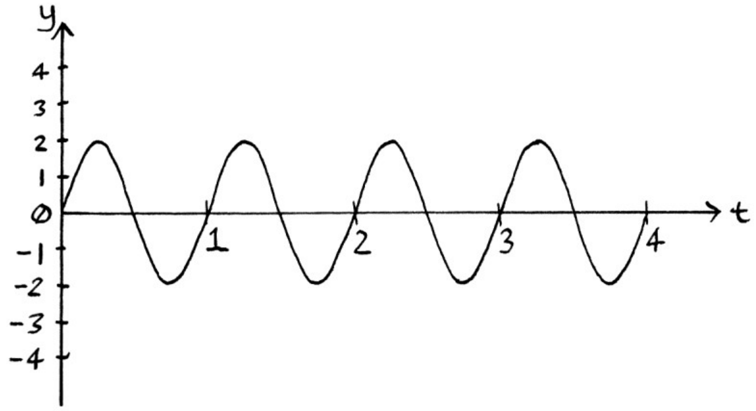
Therefore:
“4 + (−2 sin 360t)”
... can be rephrased to be:
“4 + (2 sin (360t + 180)”.

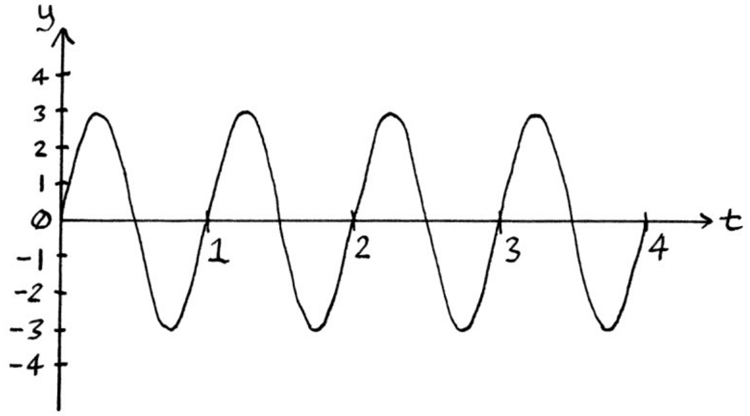
From this, we can see that the shifted (or inverted) Sine wave is raised up the y-axis by 4 units.

As a rule, any time we are *subtracting* a wave from anything, we can rephrase the calculation to be an *addition* of that wave with a phase of +180 degrees. [Or an addition of that wave with a phase of −180 degrees as that is the same thing.] Often, doing this will make the calculation easier to do and easier to understand, but there will be occasions when it is unnecessary.

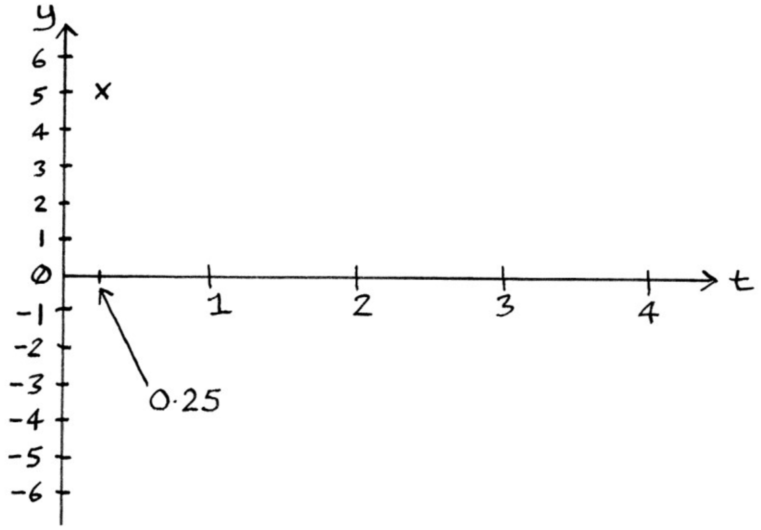
Addition of two waves

If we add two waves together, we are really adding every corresponding (by time) y-axis value of the waves together. As an example, we will add these two waves:





On the first wave, the y-axis value at 0.25 seconds is 2 units. That is added to the y-axis value at 0.25 seconds from the second wave, which is 3 units. The result of the sum (5 units) becomes the y-axis value of the resulting wave at 0.25 seconds:



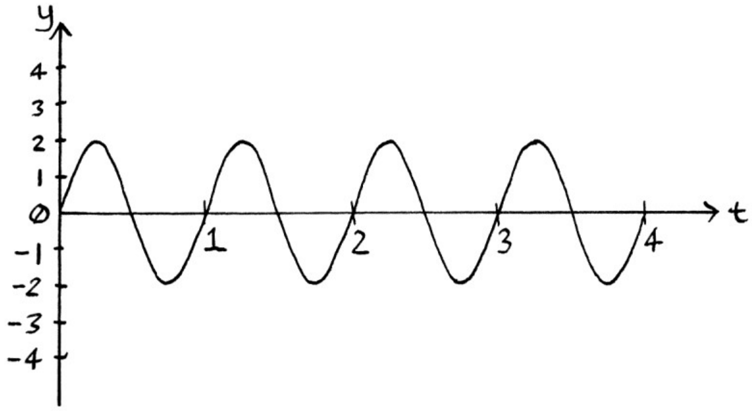
If we wanted an approximate resulting wave, we could read and calculate the result at evenly spaced points along the time axis. The more points we measure and add, the more accurate the resulting graph. In the above example, we would only need to add the points from one cycle of either wave, as the waves repeat at the same rate. All the later cycles of the result would be the same.

When adding two waves, the resulting wave will have characteristics dependent on those of the two added waves, and how the cycles of the two waves line up.

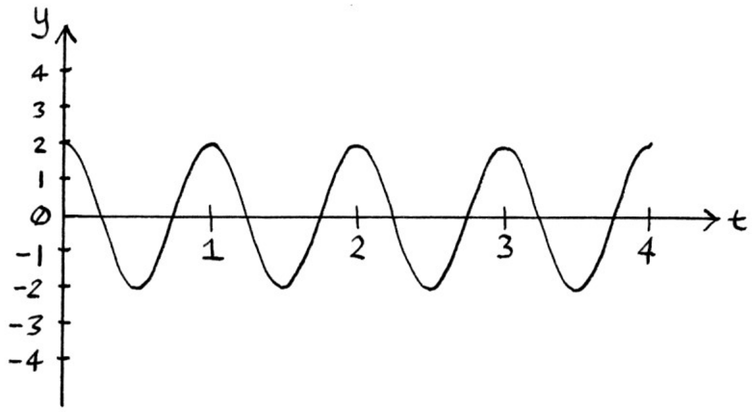
If we are using formulas instead of reading y-axis values from graphs, then adding two waves together is reasonably straightforward if those two waves have an identical frequency, as we will see shortly. If the frequencies are different, then much more thought is needed.

Two observations on addition

A Sine wave with zero phase and zero mean level will always be zero when the time is zero. Therefore, no matter how many Sine waves *with zero phase and zero mean level* are added together, the result will always be zero at t = 0, because 0 + 0 = 0.



A Cosine wave with zero phase and zero mean level will always be non-zero at t = 0. Therefore, no matter how many Cosine waves *with zero phase and zero mean level* are added together, the result will never be zero at that point.



These facts might never be any help whatsoever with addition, but they can sometimes be useful to tell if you have made a mistake.

Addition of amplitudes and mean levels

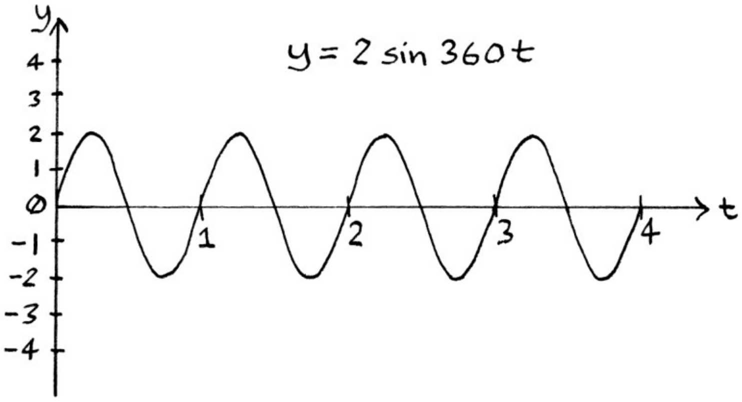
Same frequency and phase; *zero* mean level

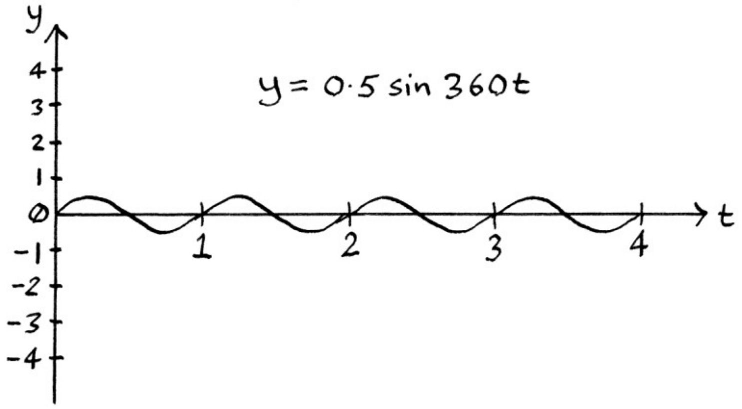
If we are given the formulas for two waves to add together that have the same frequency, the same phase and zero mean level, we can work out the resulting formula by just adding the amplitudes together. The resulting wave will have the same frequency, phase, and mean level, but the sum of the amplitudes.

As an example, we will add:
"y = 2 sin 360t"
... and:
"y = 0.5 sin 360t".

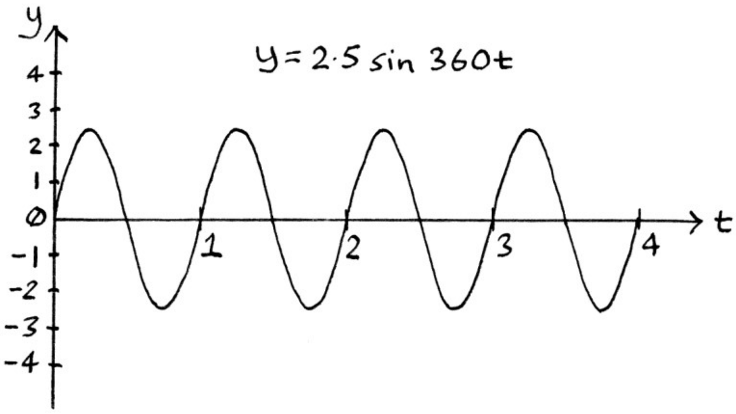
The waves are identical save for the amplitudes, which are 2 and 0.5. We add them together to produce 2.5, and our resulting wave is:
"y = 2.5 sin 360t".

The frequency and phase are unchanged. The graphs for the waves we are adding are as so:





The result ("y = 2.5 sin 360t") looks like this:



If we added:
"y = 7 cos ((360 * 3.54t) + 15)"
... to:
"y = 3 cos ((360 * 3.54t) + 15)"
... the result would be:
"y = 10 cos ((360 * 3.54t) + 15)".

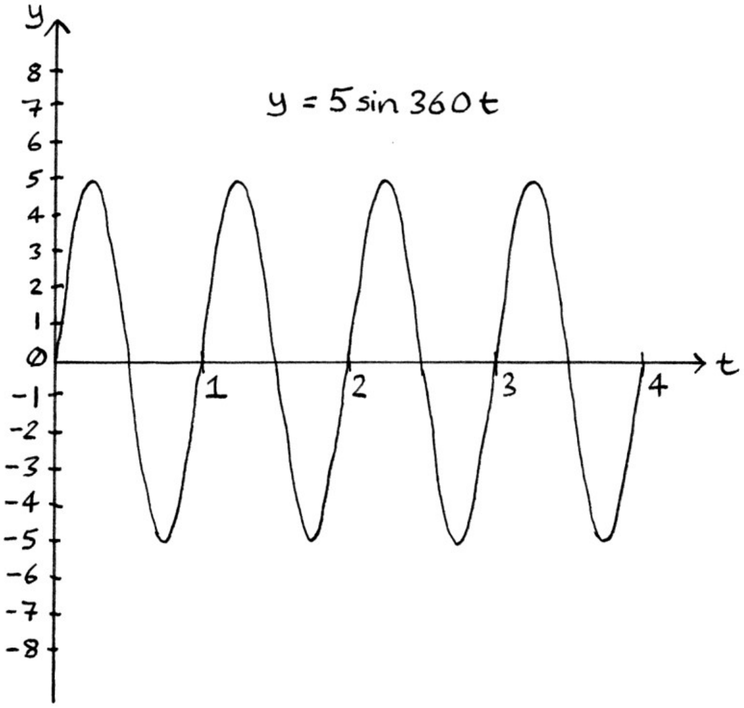
Subtraction

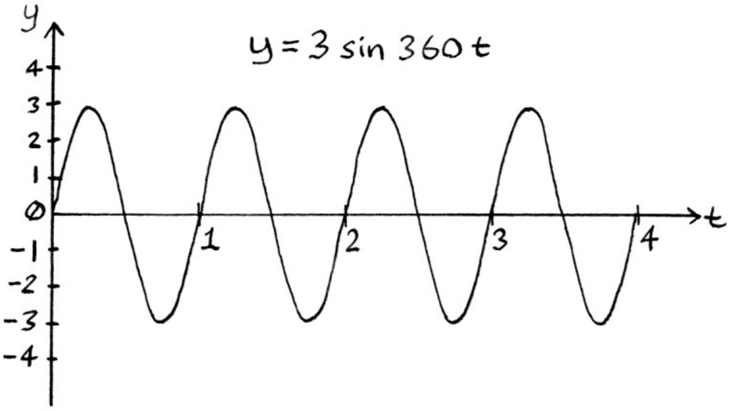
Subtraction works in the same way.

If we subtract:
"y = 3 sin 360t"
... from:
"y = 5 sin 360t"
... we end up with:
"y = 2 sin 360t".

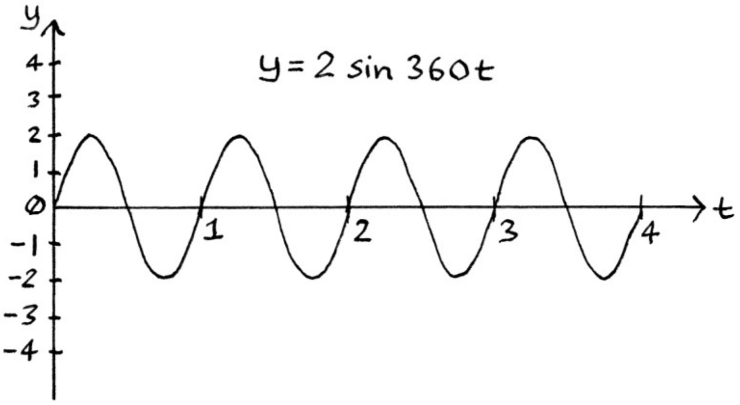
In this particular example, there is no advantage in turning the subtraction of a wave into an addition of the same wave with a +180 degree phase.

The waves that we are dealing with look like this:



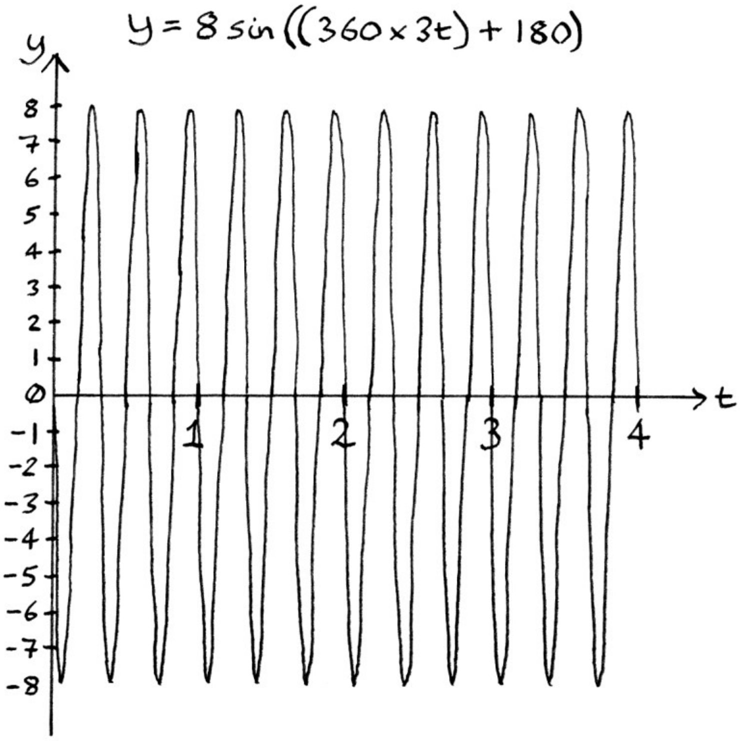


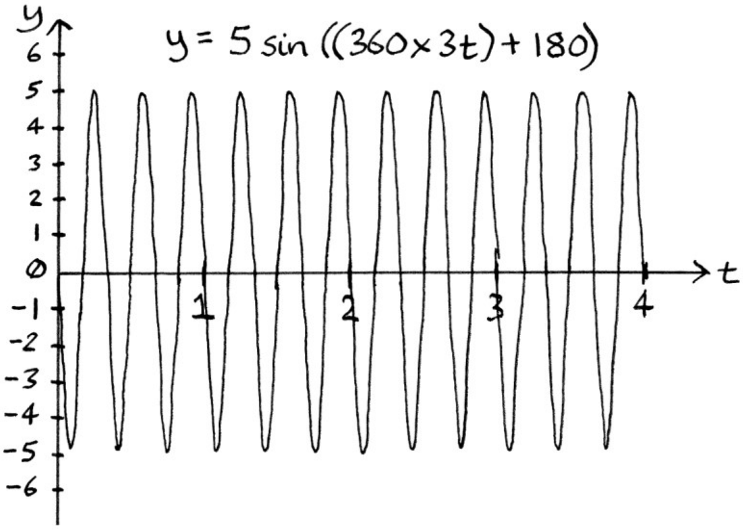
The result ("y = 2 sin 360t") looks like this:



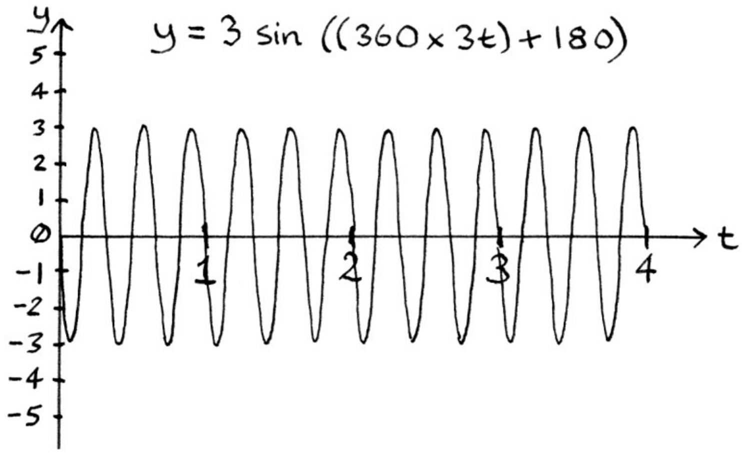
If we subtracted:
“y = 5 sin ((360 * 3t) + 180)”
... from:
“y = 8 sin ((360 * 3t) + 180)”
... we would end up with:
“y = 3 sin ((360 * 3t) + 180)”

The graphs for these waves are as follows:





The result: “y = 3 sin ((360 * 3t) + 180)”:



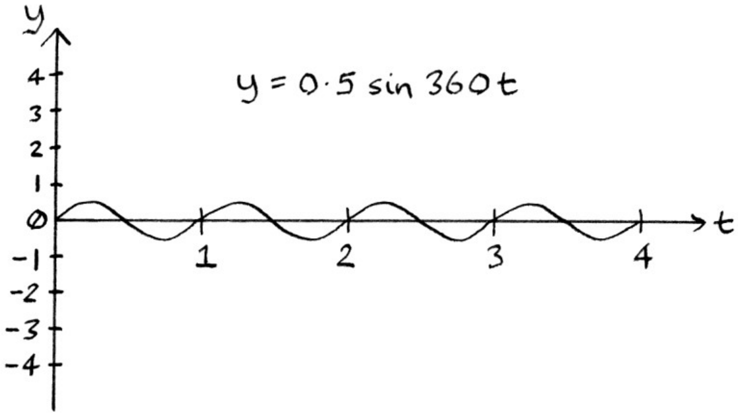
If we subtract a larger amplitude from a smaller amplitude, the resulting wave will be upside down (which is the same as being shifted either left or right by 180 degrees).

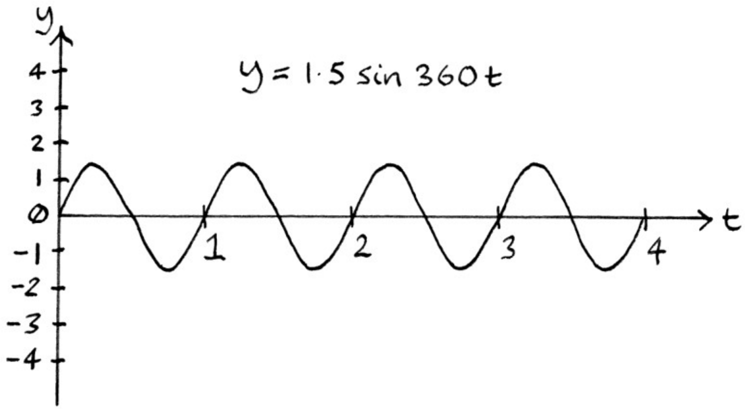
For example, if we subtract:
“y = 1.5 sin 360t”
... from:
“y = 0.5 sin 360t”
... we end up with:
“y = −1 sin 360t”.

To be consistent with how we are treating waves as derived from circles, we will turn this negative amplitude into a positive amplitude with a 180-degree phase. Therefore, this result can be rephrased to be:
“y = 1 sin (360t + 180)”
... or, more succinctly:
“y = sin (360t + 180)”.

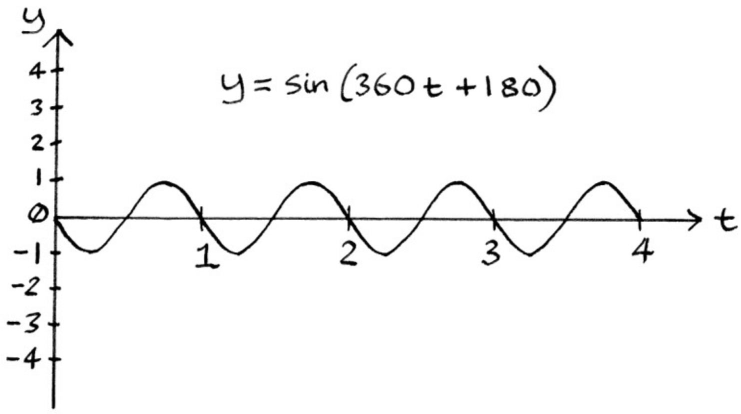
[In this particular example, if we had turned the subtraction into an addition of a wave with a +180 degree phase, it would have made everything more complicated. Turning subtractions into additions is not always helpful.]

The graphs for these waves are as follows:





The result: "y = sin (360t + 180)":



If we subtract one wave from an identical wave, the result will just be a horizontal line at y = 0. This might be obvious if we think about how it will have zero amplitude, so every point will be zero.

For example:
“y = 3.5 sin 360t”
... minus:
“y = 3.5 sin 360t”
... is:
“y = 0 sin 360t”.

This means that for all values of “t”, the result will be zero. The formula of the result could also be written as:
“y = 0”
... or, if we want to emphasise that the graph still relates to time, as:
“y = 0t”.

Same frequency and phase; non-zero mean level

If the waves have the same frequency and phase, and one or both has a *non-zero* mean level, then addition is still straightforward. The result will be a wave with the same frequency and phase, but with a changed amplitude, and the wave will be shifted up or down the y-axis. To calculate the resulting wave, we take the two original waves, add the amplitudes together, and then add the mean levels together.

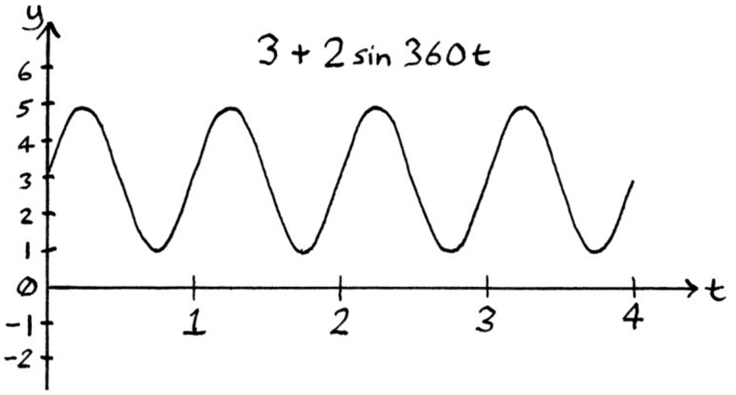
As an example, we will add:
“y = 3 + 2 sin 360t”
... to:
“y = 1 + 1.5 sin 360t”.

We add the amplitudes:
2 + 1.5 = 2.5
... and then we add the mean levels:
3 + 1 = 4.

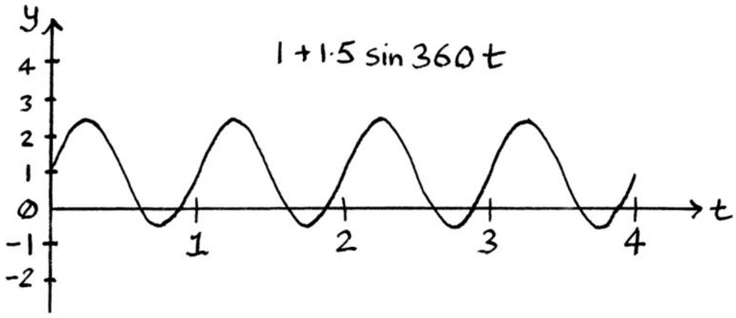
Therefore, our resulting wave is:
“y = 4 + 2.5 sin 360t”.

The graphs for these waves look like this:

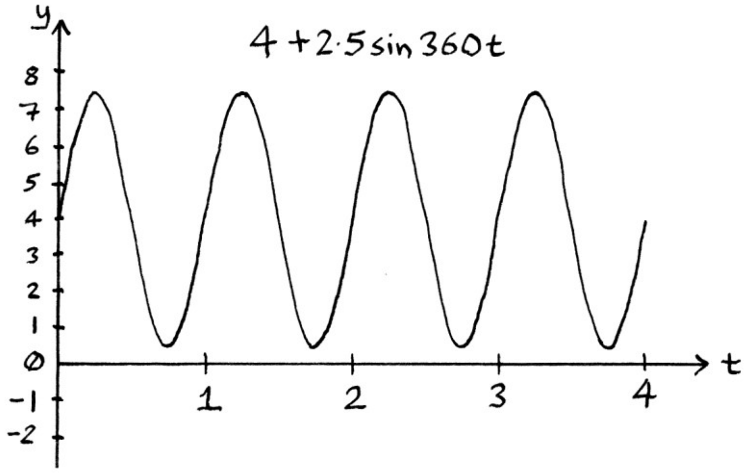
“y = 3 + 2 sin 360t”:



“y = 1 + 1.5 sin 360t”:



The result: “y = 4 + 2.5 sin 360t”:



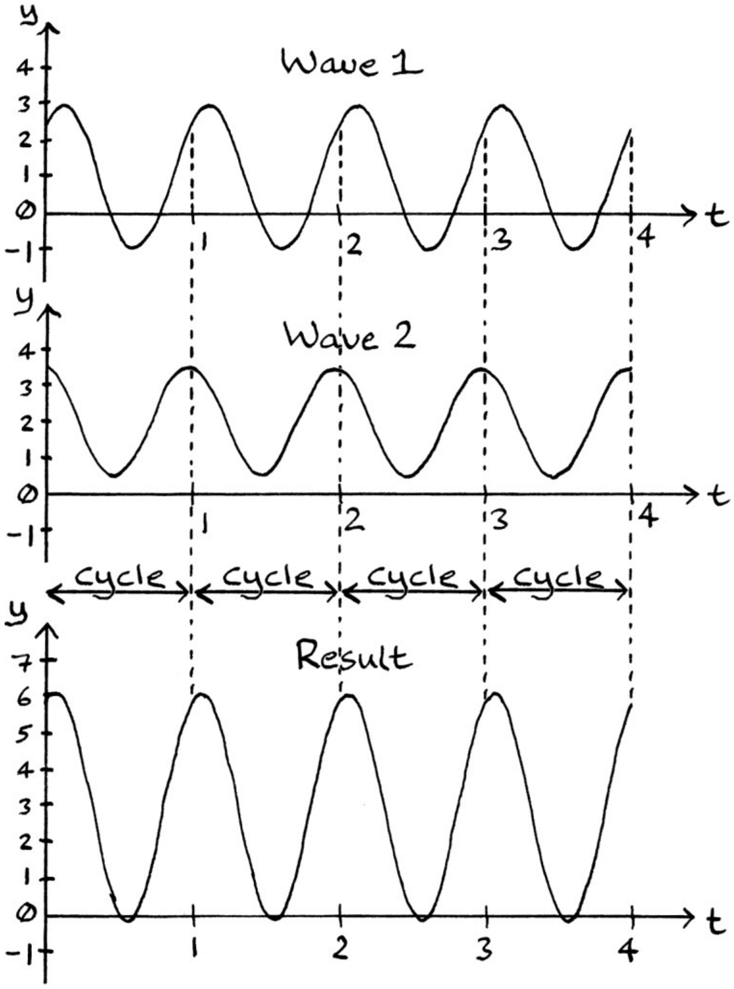
Supposing we added:
“y = −3 + 12 sin ((360 * 5t) + 271))”
... to:
“y = 6 + 24 sin ((360 * 5t) + 271))”
... we would end up with:
“y = 3 + 36 sin ((360 * 5t) + 271))”.

Subtraction works in the same way. If we do the calculation:
“y = 3 + 25 sin 360t”
... minus:
“y = 1.2 + 17 sin 360t”
... we end up with:
“y = 1.8 + 8 sin 360t”.

Addition of different phases

Addition becomes slightly more complicated when we add waves with different phases. In this section, we will only look at waves with non-zero phases that have an identical frequency.

If the two waves have the same frequency, then addition will produce a wave with that frequency too. In the first cycle of the first wave, the y-axis values will be added to the y-axis values of the first cycle of the second wave. As the frequencies are the same, the cycles are the same length. Therefore, the second cycle of the first wave will be added to the second cycle of the second wave, and the result will be identical to the first cycle. Therefore, the result will repeat with the same frequency as that of the two original waves. The phase, amplitude and mean level might be different, but the frequency will be the same. Knowing all this does not particularly help in calculating the sum of two waves with a different phase, but it gives a clue as to how the result will look.



Significantly, the result will still be a pure wave. In other words, it will still be possible to describe it using the formulas:
"$y = h_s + A \sin (360ft + \phi)$"
... or:
"$y = h_c + A \cos (360ft + \phi)$".

This has been true for all the examples so far, but it is not true if the frequencies are different. An important rule to know is that the addition of two or more waves with the same frequency will *always* result in a pure wave that has that frequency.

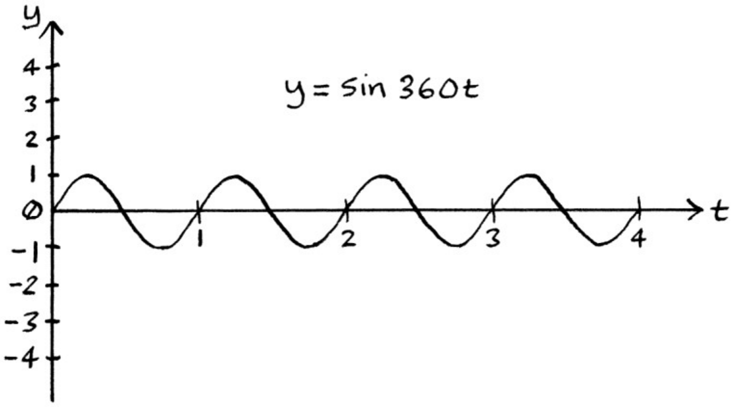
As an example, we will add:
"$y = \sin 360t$"
... and:
"$y = \sin (360t + 45)$".

The resulting wave still has a frequency of 1. Although it would be hard to guess using the information I have given so far, the resulting wave turns out to be:
"$y = 1.8478 \sin (360t + 22.5)$".

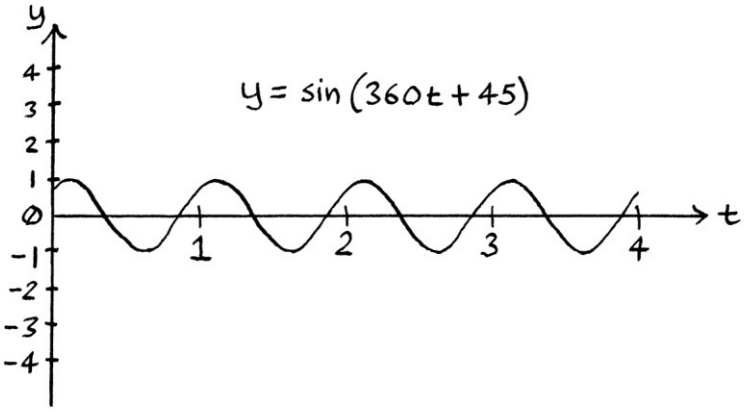
This is still a Sine wave. I will explain how to calculate such a result in a way that is easy to understand in the next chapter on adding circles.

The graphs for these waves are as follows:

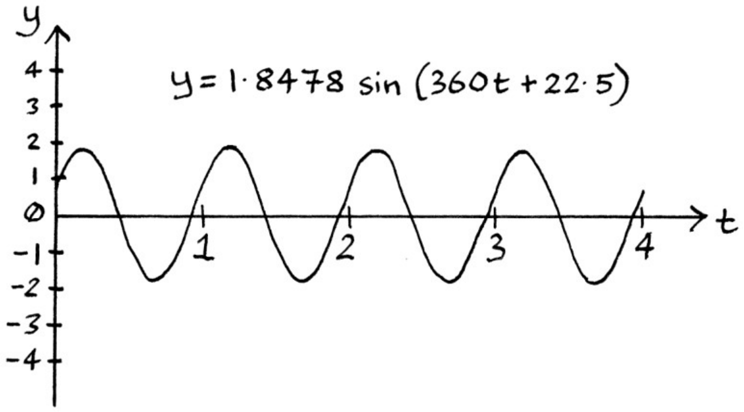
“y = sin 360t”:



“y = sin (360t + 45)”:



Result: “y = 1.8478 sin (360t + 22.5)”



180-degree phases

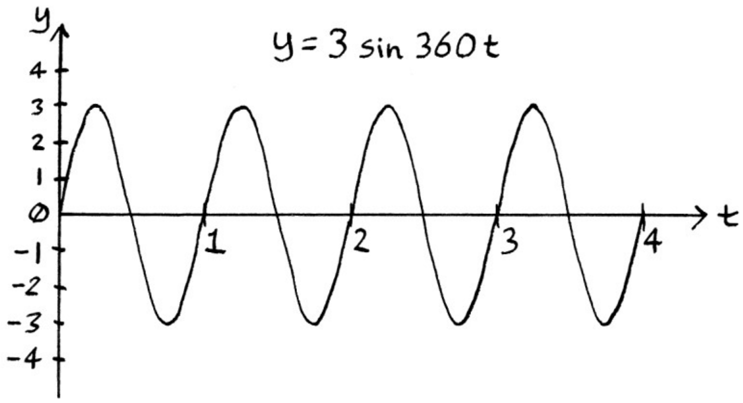
If we add two waves with phases that are 180 degrees apart, the process is identical to subtracting one from the other as if the phases were the same. This is because one wave will be an inverted version of the other (but perhaps with a different amplitude and mean level).

For example, if we add:
“y = 3 sin 360t”
... and:
“y = 2 sin (360t + 180)”
... we end up with:
“y = sin 360t”.

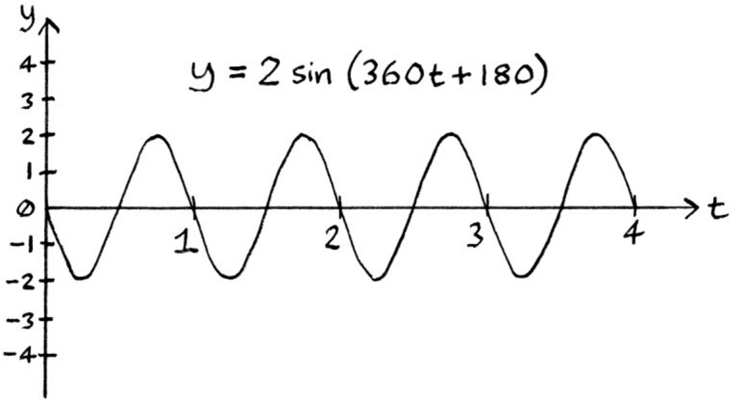
The calculation is identical to subtracting:
“y = 2 sin 360t”
... from:
“y = 3 sin 360t”.

The graphs for these waves are as follows:

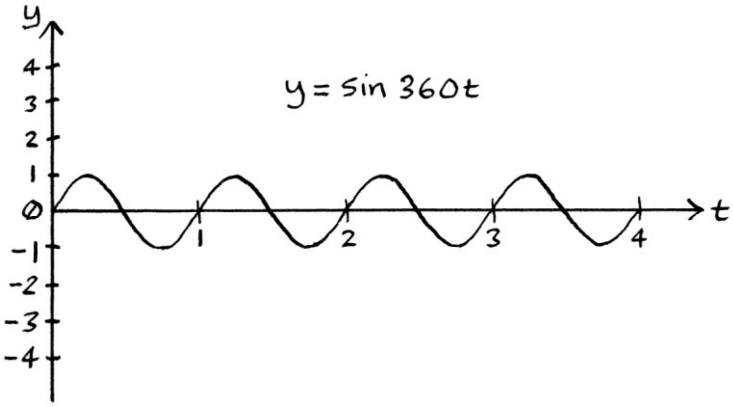
“y = 3 sin 360t”:



“y = 2 sin (360t + 180)”:



Result: “y = sin 360t”:

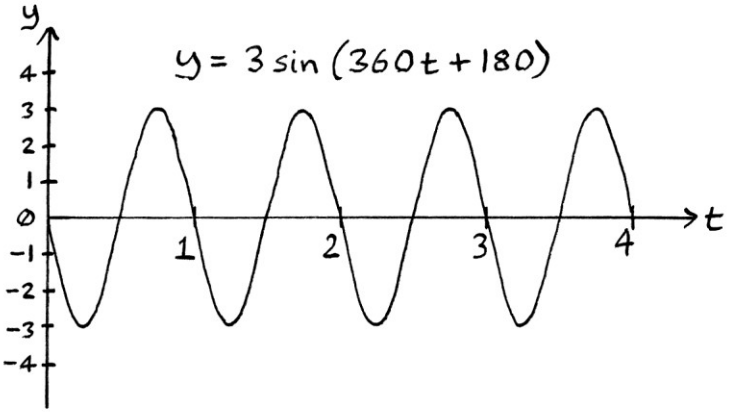


As a sort of “trick” example, if we add:
“y = 3 sin (360t + 180)”
... and:
“y = 1 sin (360t – 180)”
... we end up with:
“y = 4 sin (360t + 180)”.

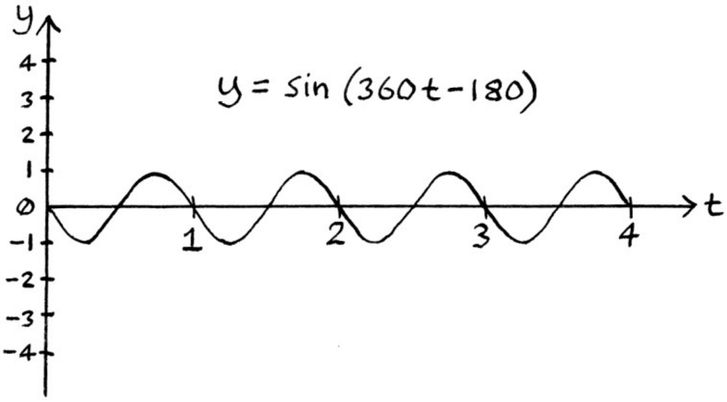
The reason for this result is that the phase of +180 degrees is the same as the phase of −180 degrees. The waves actually have the same phase, so the calculation is simpler than it first looks.

The waves are as follows:

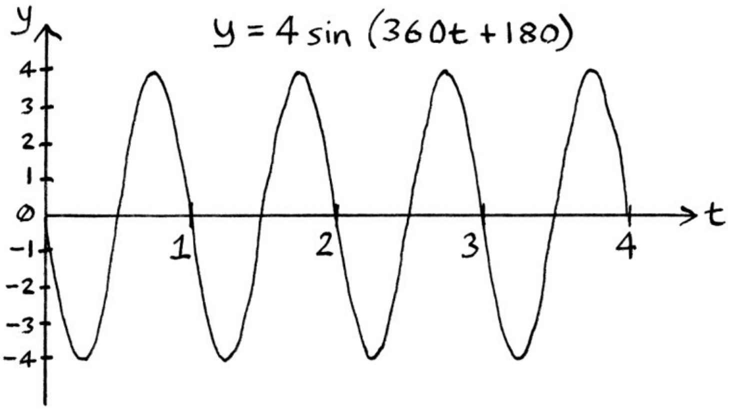
“y = 3 sin (360t + 180)”:



“y = 1 sin (360t – 180)”:



The result: “y = 4 sin (360t + 180)”:



If we add waves with phases that are 180 degrees apart, and that have the same amplitude, the two waves will cancel each out, and we will end up with y = 0 for all time. The resulting graph will be a straight line at y = 0.

For example:
“y = 2 sin 360t”
... added to:
“y = 2 sin (360t − 180)”
... produces:
“y = 0t”.

In the resulting graph, whatever the value of “t”, “y” will always be zero. The result could also be given as:
“y = 0”
... or:
“y = 0 sin 360t”.

The reason for the result is made clearer by thinking about how a phase of −180 degrees in a wave formula can also be thought of as the same wave with zero phase and a negative amplitude. Therefore, the sum in the above example could be thought of as:
“y = (2 sin 360t) + (−2 sin 360t)”
... or:
“y = (2 sin 360t) − (2 sin 360t)”
... which results in:
“y = 0 sin 360t”
... which is equivalent to:
“y = 0t”

Note how the graph will not be blank. There will still be a line on the graph to indicate that “y” is zero at all times.

[It does not particularly matter whether the result is given as “y = 0” or “y = 0t”, although leaving the “t” in the formula reinforces how the formula represents something happening over a length of time.]

Earlier, I said that adding two waves of the same frequency results in a wave of that same frequency. This section suggests that that is not true if the waves cancel each other out – the frequency would be zero. However, we could just as easily give the resulting wave formula in a way that kept the original frequency. In the above example of:

“y = 2 sin 360t” added to “y = 2 sin (360t − 180)”
... which amounts to:
“y = 2 sin 360t” added to “y = −2 sin 360t”
... we could give the result as:
“y = 0 sin 360t”.

This still has a frequency of 1 cycle per second, but the amplitude is zero, so anything produced by “sin 360t” is multiplied by zero to produce nothing but a flat line at y = 0. The rule is still true, but the formula might be written as “y = 0t”, which hides what the zeroed frequency is.

Sine and Cosine

One particular addition of which to be aware is that of a Sine wave and a Cosine wave with the same frequency. This is actually adding two waves with the same frequency but a different phase – in other words, it is the same as adding a Sine wave to another Sine wave with a phase +90 degrees higher in its formula. Therefore, the result will be a wave with the same frequency.

As an example, we will add:
“y = sin 360t”
... and:
“y = cos 360t”.

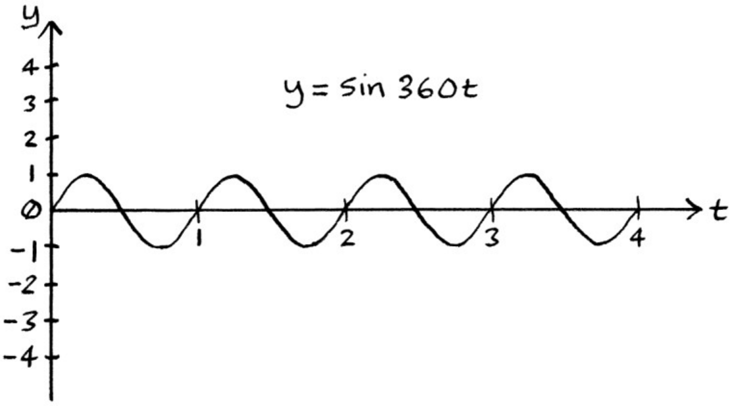
This is the same as adding:
“y = sin 360t”
... and:
“y = sin (360t + 90)”.

The result turns out to be:
“y = 1.4142 sin 360t + 45”.

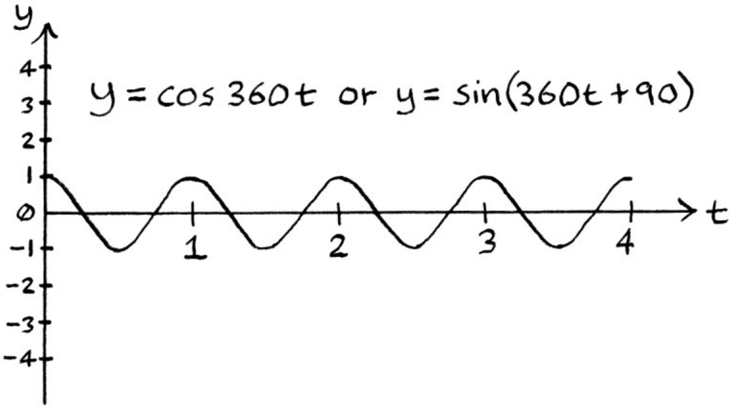
The reason why the amplitude and phase end up like this will be explained in the next chapter.

The graphs for these waves are as follows:

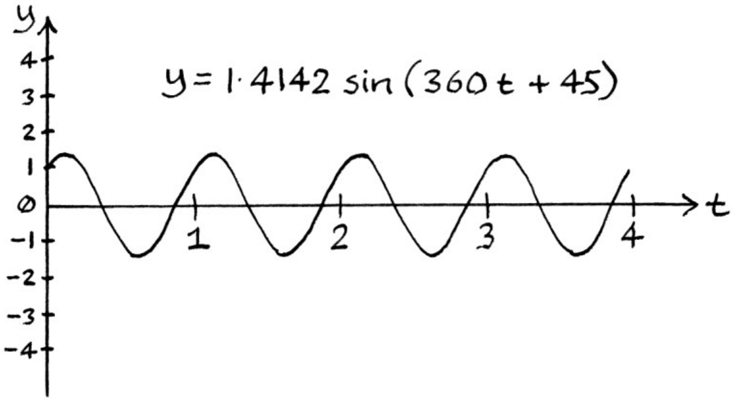
“y = sin 360t”:



“y = cos 360t”, which is the same as “y = sin (360t + 90)”:



The result: “y = 1.4142 sin 360t + 45”:



Phases spaced evenly over one cycle

As we have seen, if we add two waves with the same amplitude and frequency, but with phases that are 180 degrees apart, the process is identical to subtracting one from the other as if the phases were the same. The result is "y = 0t" for all time.

If we add *three* waves with the same amplitude, and with phases of 0, 120 and 360 degrees, then the result will also be "y = 0t" for all time.

If we add *five* waves with the same amplitude, with phases of 0, 72, 144, 216 and 288 degrees, the result will, again, be "y = 0t" for all time. The angles 0, 74, 144, 216 and 288 are evenly spaced angles from 0 to 360 degrees.

In fact, if we add *any* number of waves that have the same amplitude and frequency, but phases spaced evenly from 0 to 360 degrees, they will cancel each other out, and we will end up with "y = 0t". The reason for this will make more sense when we look at circles in the next chapter.

The rule for same frequency, different phase

Calculating the resulting wave from adding two waves with different phases is not difficult at all. However, it is easier to visualise how and why it works when looking at circles. Therefore, we will examine the method in the next chapter.

Addition of different frequencies

The most complicated form of addition is when the two waves have a different frequency. The result will have a more varied shape, and most importantly, it will *not* be a pure wave. In other words, it will not be possible to describe the result using the formulas:
"$y = h_s + A \sin (360ft + \phi)$"
... or:
"$y = h_c + A \cos (360ft + \phi)$".

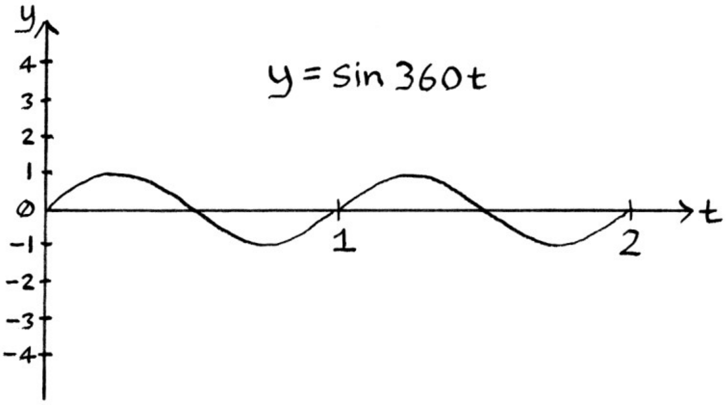
Although calculating the actual result is harder, *visualising* what is happening is not particularly difficult, especially when we consider the circles, which we will do in the next chapter.

It is still helpful to think of the addition of waves of different frequencies as being the y-axis points from one wave added to the corresponding (by time) y-axis points from the second wave.

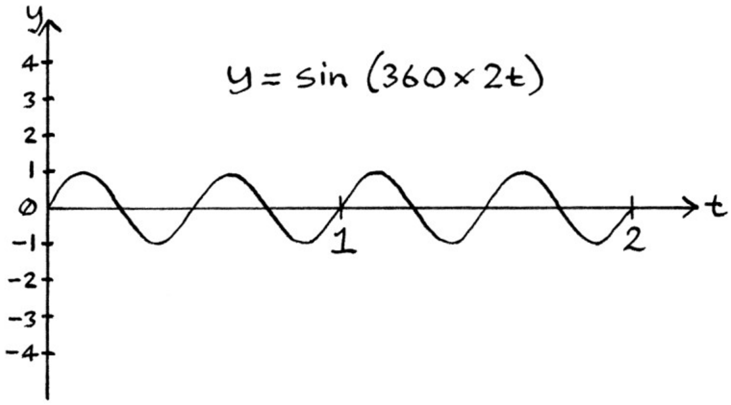
As an example of adding waves with different frequencies, we will add:
"y = sin 360t"
... and:
"y = sin (360 * 2t)"

The first wave has a frequency of 1 cycle per second, and the second wave has a frequency of 2 cycles per second.

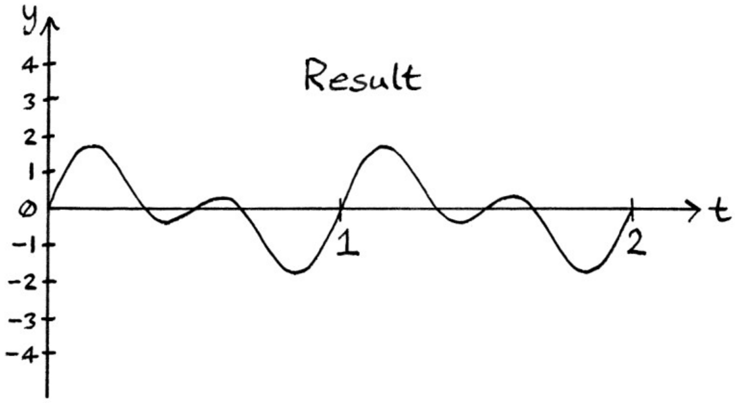
"y = sin 360t":



"y = sin (360 * 2t)":



The result looks like this:



Things to notice about the result are:

- Its shape is different from either of the waves that were added together.
- The result is not a pure wave. It would be impossible to describe this using either of the formulas "y = h_s + A sin (360ft + ϕ)" or "y = h_c + A cos (360ft + ϕ)". The most succinct way of describing the result is by saying it is:
  "y = sin (360t) + sin (360 * 2t)".
  In other words, the best way to describe it is by saying which waves were added to create it. Personally, I think it is better to call the result a "signal" rather than a "wave" to emphasise the fact it is not a pure wave.
- Given all that, the resulting signal does not represent the y-axis values of a *circle*. Instead, it represents the y-axis values of a "shape".
- Its mean level is still zero. The average height of all the y-axis values for 1 repetition of the signal is zero.
- The resulting signal repeats its pattern every 1 second. This means that its frequency is 1 cycle per second, and its period is 1 second. Note that because the resulting signal is not a pure wave, this frequency does not apply to the resulting *formula*. As I said before, the formula for the resulting signal is "y = sin (360t) + sin (360 * 2t)". The signals created from adding waves of different frequencies will repeat their cycles at a frequency that cannot necessarily be deduced from the formula, and probably will not even be mentioned in the formula. It is common to refer to the frequency of an impure signal as the "fundamental frequency". In this sense, the term "fundamental frequency" can be thought of as meaning "the overall frequency". The term "fundamental frequency" can be useful to avoid

confusing the overall frequency of a signal with the frequencies of the waves that were added (or could have been added) to create it.

- Because the resulting signal is not a pure wave, the use of the term "amplitude" when discussing it needs more thought. For this particular signal, the y-axis values reach up to +1.7602 and down to −1.7602. If we were looking at a *pure* wave, we would say that the amplitude was 1.7602 units, by which we would mean that the "*overall* amplitude" was 1.7602 units. This would also mean that the circle from which such a pure wave was, or could have been, derived would have a radius of 1.7602 units. However, the resulting signal here is *not* a pure wave, and it does not and cannot be derived from a circle. Therefore, the term amplitude to mean "overall amplitude" makes no sense here. The signal cannot be said to have an overall amplitude.

  We could still say that the signal has a "peak amplitude" in that its maximum y-axis value is 1.7602 units, but in my view it is better not to use the term "peak amplitude" as it still hints that the signal shares properties of a pure wave. [Other people might disagree]. I prefer the terms "maximum value" and "minimum value" instead of "peak amplitude". In this particular example, the maximum and minimum values are the negative of each other (+1.7602 and −1.7602), but in many cases when two waves of different frequencies (and zero mean level) are added together, the maximum and minimum values will be different. The reason for this becomes clearer when we look at circles in the next chapter. The *average* y-axis value will still be zero if the original waves had zero mean levels, but the maximum and minimum values might not be the negative of each other.

  When looking at this signal, people might still use the term "amplitude" to mean "instantaneous amplitude" as in the y-axis value at a chosen moment in time. Generally, I think it is better to be careful using the term "amplitude" with impure waves to avoid confusion about what they represent. However, the term "instantaneous amplitude" can be a useful way to mean "the y-axis value at a particular time", especially when a signal is not being portrayed on a graph with a y-axis labelled as just a y-axis.

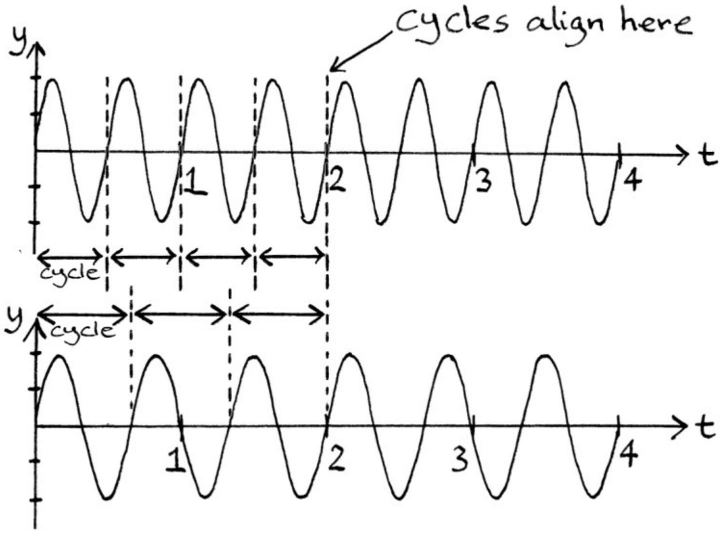
- As with amplitude, I would say that the result of adding two waves of different frequencies cannot be said to have a phase. If we were to imagine an object moving around the shape from which such a signal is, or could have been, derived, the object would start at 0 degrees on what we were calling the circle chart, so one might say that the resulting signal had a phase of 0 degrees. However, when we look at adding circles in the next chapter, we will see that such an object might pass over the same angle more than once on its journey. Therefore, the idea of phase ceases to be a useful indicator of exactly where an object starts. The object still has a starting position – it starts at a particular place. However, that place needs to be specified with more than just its angle. When thinking of the resulting shape, we could give the coordinates of that point, but doing that is not really apt when thinking about wave graphs. All of this will be clearer in the next chapter.

- Not only can we not use the term “phase” when describing the signal, but we should also be very wary of using the term “phase shift” when referring to such a signal if it is shifted left or right along the time axis. The idea of shifting the signal left or right by a certain *angle* is much more complicated or even meaningless. A shifted version of the signal is best only described as having been shifted by a number of seconds, and even then, one has to have a reference to say from where it was shifted. This will all make more sense in the next chapter too.

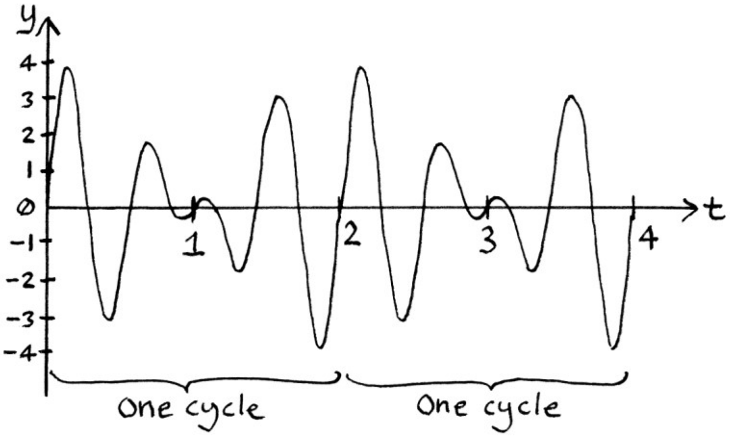
The effect on cycles

If two waves being added have different frequencies, the first cycle of the faster wave will be added to just part of the first cycle of the slower wave. The second cycle of the faster wave will be added to the next part of the slower wave’s first cycle, or to the rest of that cycle, or to the rest of that cycle and a bit of the next cycle. This all depends on the frequencies of the two added waves. Usually, there will be a point where the cycles of the waves coincide again, and that will indicate the frequency of the resulting signal.

In the following picture, the first wave (2 cycles per second) has a faster frequency than the second wave (1.5 cycles per second. The cycles of the waves do not match, but the cycles still coincide an integer number of times at 2 seconds. At 2 seconds, the first wave has completed 4 cycles; the second wave has completed 3 cycles.

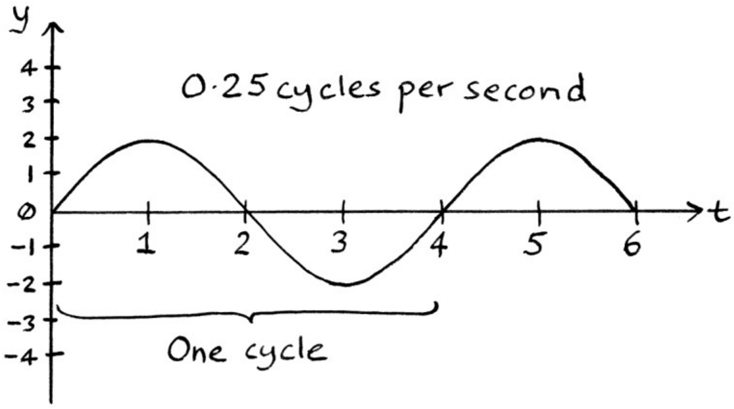


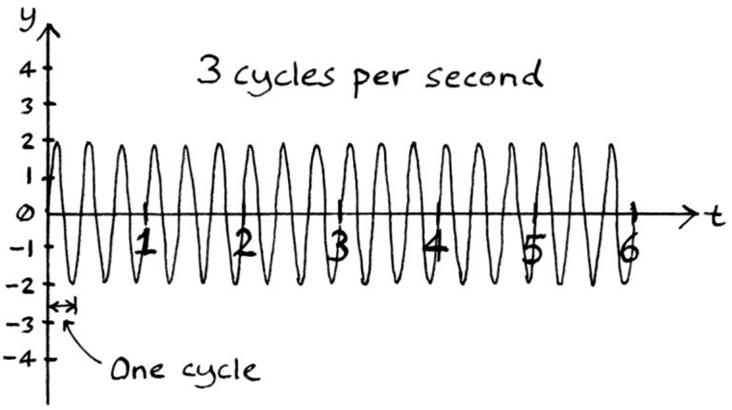
What this means for addition is that the cycles of the first wave are not added to the same point of a cycle of the second wave until 2 seconds have passed. At 2 seconds, the two waves will have the same relationship to each other as they did at 0 seconds. This means that the resulting summed signal will only repeat its shape once every two seconds. The resulting signal only repeats when the cycles of both of the added waves align with each other again. The resulting signal in this case looks like this:

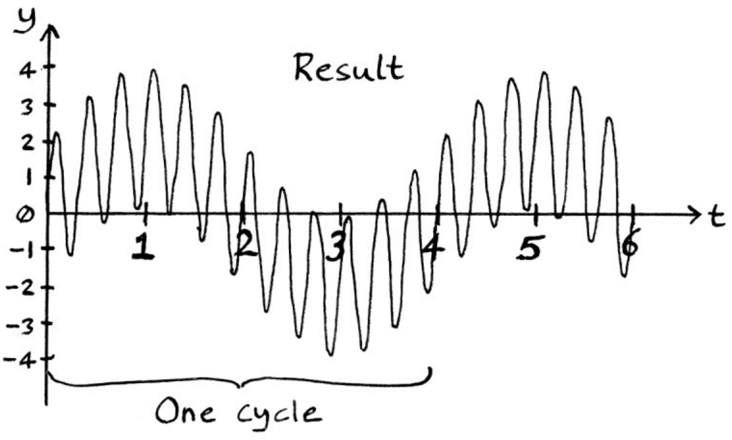


The signal has a frequency of 0.5 cycles per second. Its period is 2 seconds.

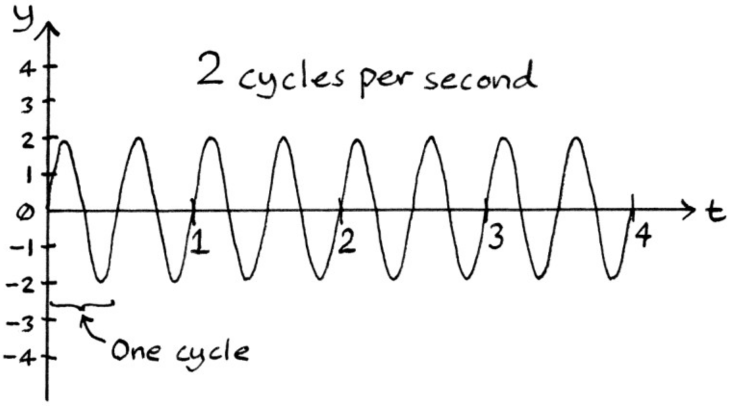
If, say, the first wave has a frequency of 0.25 cycles per second, and the second wave has a frequency of 3 cycles per second, their cycles will not coincide with each other until after 4 seconds. After 4 seconds, the first wave will have completed 1 full cycle; the second wave will have completed 12 full cycles. Therefore, the resulting signal will not repeat its shape until after 4 seconds – it will have a frequency of 1 ÷ 4 = 0.25 seconds.

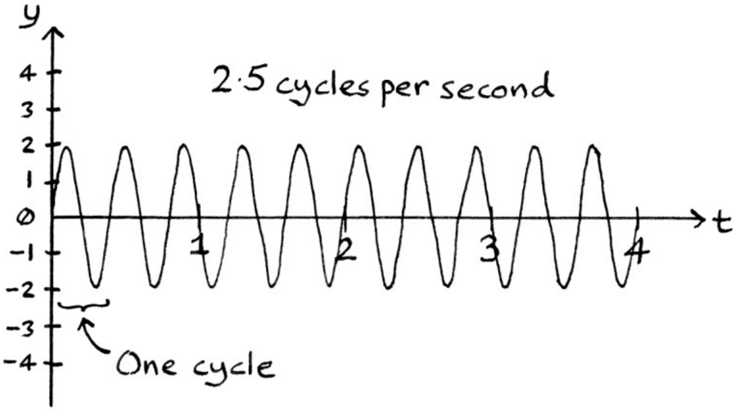




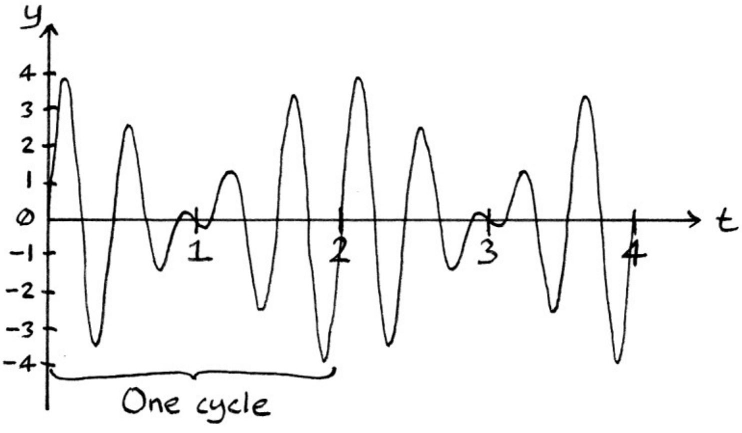


If, say, the first wave has a frequency of 2 cycles per second, and the second wave has a frequency of 2.5 cycles per second, their cycles will not coincide with each other until after 2 seconds. After 2 seconds, the first wave will have completed 4 full cycles, and the second wave will have completed 5 full cycles. This means that the shape of the resulting signal will repeat itself every 2 seconds – it will have a frequency of 1 ÷ 2 = 0.5 seconds.





The result:



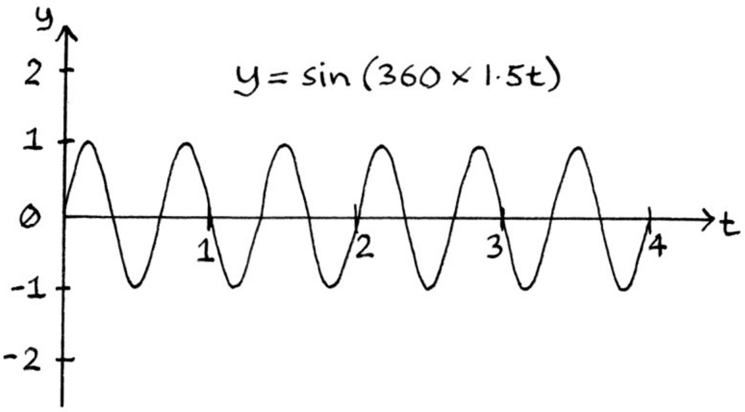
Thinking about the cycles coinciding is helpful in understanding why the sums of two waves with different frequencies look the way they do.

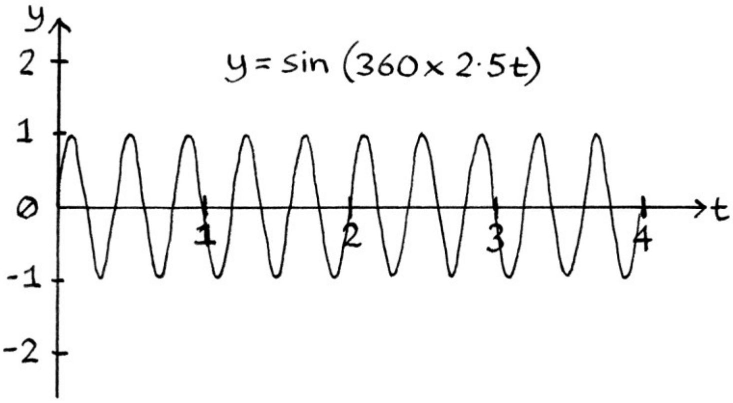
Calculating the frequency of the resulting signal

Calculating the frequency of the result of a sum of two waves with different frequencies is easy to understand, and *reasonably* easy to do. The important thing to remember is that the result of the addition is *not* a pure wave, and therefore, the frequency of the resulting signal cannot, and will not, appear in a single formula for the resulting signal. The formula for the resulting signal will be two pure waves added together – it cannot be a single pure wave.

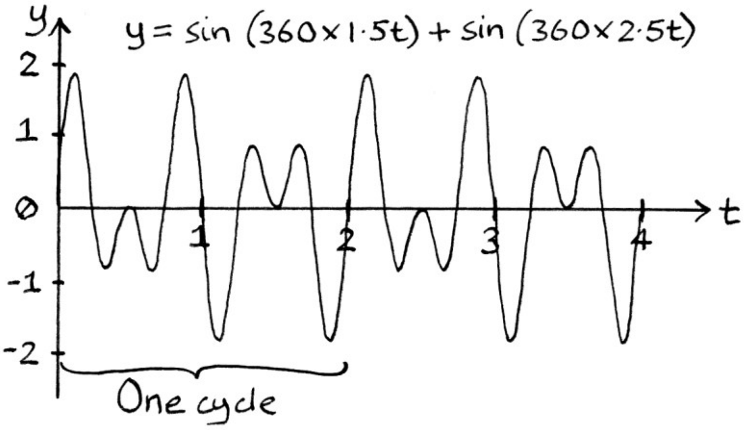
As an example:
“y = sin (360 * 1.5t)”
... added to:
“y = sin (360 * 2.5t)”
... results in the signal:
“y = sin (360 * 1.5t) + sin (360 * 2.5t)”.

The overall frequency of the resulting signal is 0.5 cycles per second (in the sense that its shape repeats once every 2 seconds), but that frequency does not appear in the resulting formula. The two added waves look like this:





The sum of the two waves looks like this:



The main rule for adding frequencies

If we add two waves with different frequencies, the frequencies of each of the added waves will always be integer multiples of the resulting frequency. [They might not be *integers*, but they will be *integer multiples*.] The resulting frequency will be the highest number for which this is true.

To put this another way, the frequency of the resulting signal will be the highest number that divides into each of the added frequencies a whole number of times. It will be the highest common divisor. When either of the frequencies of the two added waves is divided by the resulting frequency, the result will be an integer.

As an example, if we add waves with frequencies of 6 and 9 cycles per second, the resulting signal will have a frequency of 3 cycles per second. The numbers 6 and 9 are both integer multiples of 3, and 3 is the highest number for which 6 and 9 are both integer multiples.

If we add waves with frequencies of 3 and 9 cycles per second, the resulting signal will also have a frequency of 3 cycles per second. The numbers 3 and 9 are both integer multiples of 3, and 3 is the highest number for which 3 and 9 are both integer multiples.

If we add waves with frequencies of 44 and 55 cycles per second, the resulting signal will have a frequency of 11 cycles per second. The numbers 44 and 55 are both integer multiples of 11, and 11 is the highest common divisor of 44 and 55.

If we add waves with frequencies of 4.4 and 5.5 cycles per second, the resulting signal will have a frequency of 1.1 cycles per second. The numbers 4.4, 5.5 and 1.1 are not integers, but the numbers 4.4 and 5.5 are integer multiples of 1.1. The number 1.1 is the highest number for which 4.4 and 5.5 are both integer multiples.

If we add waves with frequencies of 10 and 20 cycles per second, the resulting signal will have a frequency of 10 cycles per second. The numbers 10 and 20 are both integer multiples of 10. They are also integer multiples of 5, but 5 is not the highest number for which 10 and 20 are both integer multiples.

The rule works not just for adding two waves, but also for adding any number of waves. Therefore, if we add waves with frequencies of 1, 2, 3, and 4 cycles per second, the resulting signal will have a frequency of 1 cycle per second. The values 1, 2, 3 and 4 are all integer multiples of 1. The number 0.5 would also be a number for which 1, 2, 3 and 4 are integer multiples, but it is not the *highest* number. The resulting signal will repeat at 1 cycle per second.

If we add the frequencies 4, 6, 8, and 10 cycles per second, the resulting signal will have a frequency of 2 cycles per second. The numbers 4, 6, 8 and 10 are all integer multiples of 2, and 2 is the highest number for which this is the case.

If we add the frequencies 0.5, 1.5, 2.5 and 5.5 cycles per second, then the resulting frequency will be 0.5. The numbers 0.5, 1.5, 2.5 and 5.5 are all integer multiples of 0.5. The number 0.5 is the highest number for which this is the case.

If we add the frequencies 1.5, 2.5 and 5.5 cycles per second, then the resulting frequency will again be 0.5. The numbers 1.5, 2.5 and 5.5 are all integer multiples of 0.5. The number 0.5 is the highest number for which all the added frequencies are integer multiples.

Helpful shortcuts

There are several useful observations that help in calculating the overall frequency of the signal resulting from adding two waves of different frequencies. [If we want to add more than two waves, we can add two waves, and then add the result to the next wave, then the result of that to the next wave and so on].

Observation 1 – integer multiples

If the faster frequency of the two waves is an integer multiple of the slower frequency, then the resulting signal will have the frequency of the slower frequency. This is because the faster wave will repeat an integer number of times in the time that the slower wave repeats once.

Another way of thinking about this is that if the ratio of the faster frequency to the slower frequency can be scaled to be an integer to 1 (for example, something such as 4 : 1), the frequency of the resulting signal will be equal to the slower frequency. [By "scaled to an integer to 1", I mean that if we start with the ratio 6 : 2, then that ratio is the same as the ratio 3 : 1. To end up with 3 : 1, we divide both sides of the ratio by a number that will result in the second value being 1. We have scaled the ratio so that its right hand side value is 1, while keeping the proportions the same]. There is not much point in using a ratio in this way for this observation as it is obvious if one value is an integer multiple of another value. There is no need to use a ratio to confirm it. However thinking about the ratios will help later on, and it is good to introduce the idea now.

Examples are:

For a wave of 3 cycles per second added to a wave of 1 cycle per second, the number 3 is obviously an integer multiple of 1, so the result will be the slower of the two frequencies, which is 1 cycle per second. The first wave will repeat itself 3 times in the time the second repeats itself once (which it does in one second). Therefore, the resulting signal will repeat itself every second. Using ratios here does not really help, but the ratio is 3 : 1, which shows that the first frequency is an integer multiple of the other frequency, which was obvious.

If we add a wave of 4 cycles per second to a wave of 2 cycles per second, we can see that 4 is an integer multiple of 2. Therefore the slower frequency (2 cycles per second) will be the frequency of the sum. The first wave will repeat itself 4 times in the time it takes the second wave to repeat itself twice, or to put it another way, the first wave will repeat itself twice in the time it takes the second wave to repeat once. It takes 0.5 seconds for the waves to coincide again, so the resulting signal repeats itself twice a second. Again, ratios do not particularly help here, but the ratio is 4 : 2, which scales to 2 : 1, which tells us that 4 is an integer multiple of 2, and therefore the slower frequency (2 cycles per second) will be the frequency of the resulting sum.

If we add a wave of 9 cycles per second to a wave of 3 cycles per second, we know that 9 is an integer multiple of 3, so the slower frequency (3 cycles per second) will be the frequency of the resulting signal. [Ratios do not really help here either, but the ratio would be 9 : 3, which scales to 3 : 1, showing that 9 is an integer multiple of 3.]

If we add a wave with a frequency of 34,626 cycles per second to a wave with a frequency of 2 cycles per second, we can see that 34,626 is an integer multiple of 2, so the resulting frequency will be the slower of the two frequencies, which is 2 cycles per second.

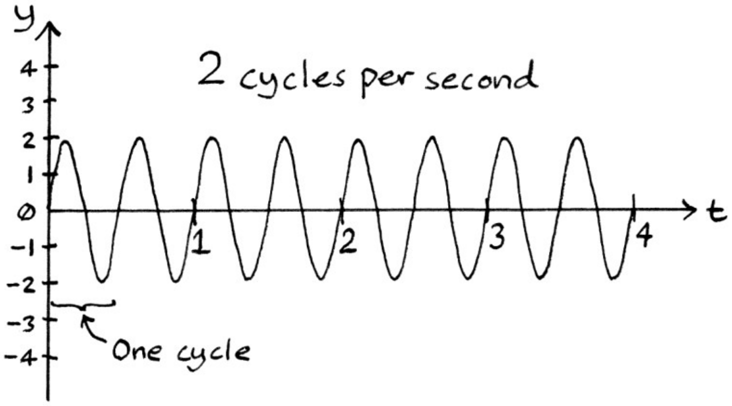
Adding a wave of 11.7 cycles per second and one of 2.34 cycles per second results in a signal with a frequency of 2.34 cycles per second. The number 11.7 is an integer multiple of 2.34. The ratio of 11.7 : 2.34 scales down to become 5 : 1. The first wave repeats itself 5 times for every time the second wave repeats itself once.

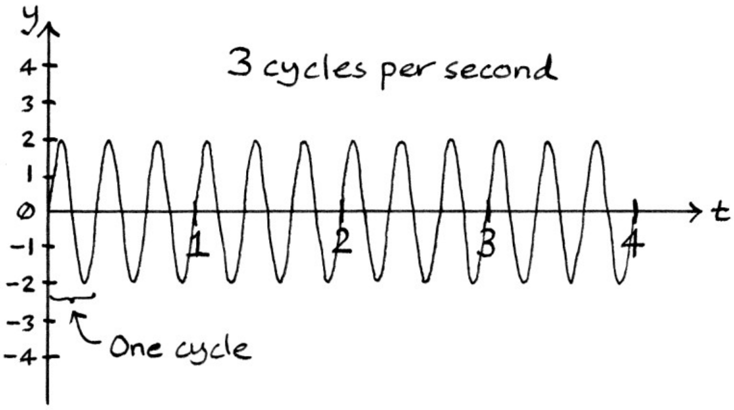
Observation 2 – integers per second

If the frequencies are both integers per second, then the cycles will definitely align every second, *but maybe also sooner*. We can tell that they will align every second because both waves will have completed a whole number of cycles at one second. To know if they align sooner takes more thought, though.

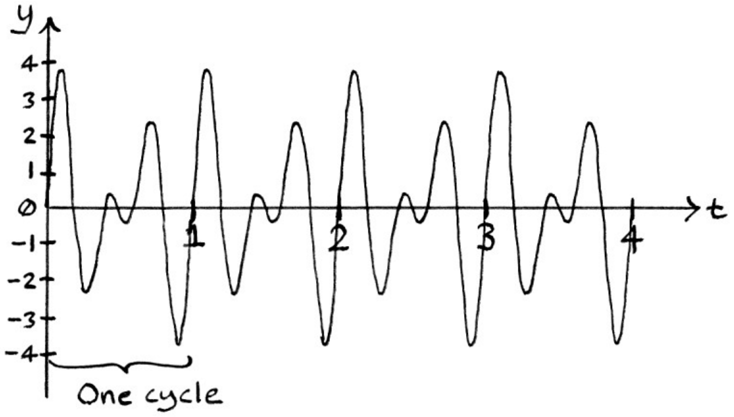
For example, adding waves of 2 cycles per second and 3 cycles per second will result in a signal with a frequency of 1 cycle per second. The first time that the waves align is at 1 second.

Two such waves look like this:

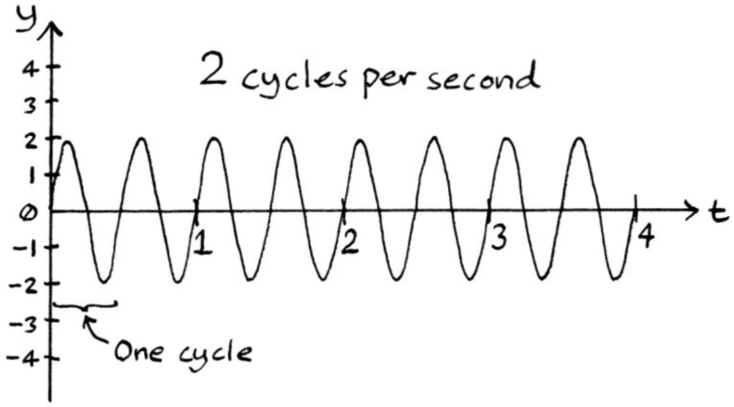


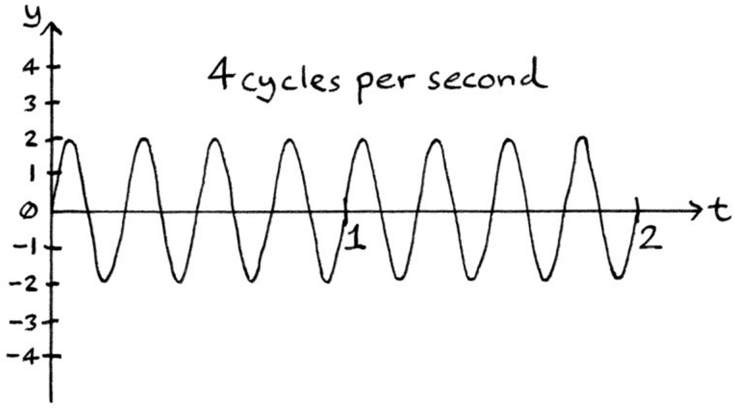


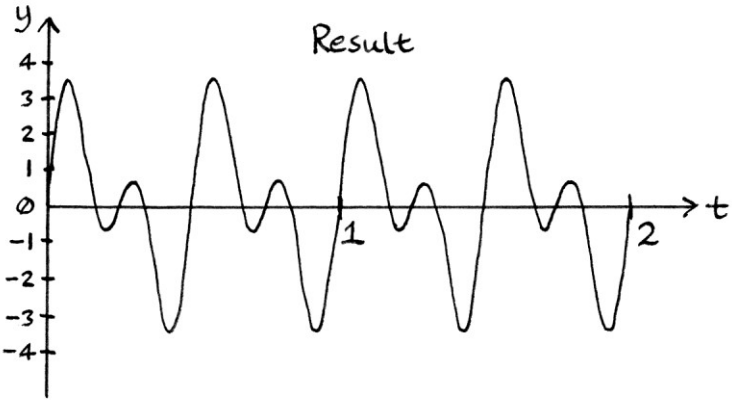
The result of adding them looks like this:



As another example, adding waves with frequencies of 2 cycles per second and 4 cycles per second will result in a signal with a frequency of 2 cycles per second. The cycles do align at 1 second, but they also align sooner – at 0.5 seconds. In this case, the first observation is useful to remember. The number 4 is an integer multiple of 2, so the frequency of the sum will be 2 cycles per second.







Observation 3 – A method for calculating added frequencies

The following is a simple method for calculating the resulting frequency from adding two waves. [It is probably not the best method, but it is simple and helps in understanding the underlying principles.] It uses the ratio of one frequency to the other. For example, if we are adding waves with the frequencies of 7 and 5 cycles per second, then the ratio will be 7 : 5. We can scale down ratios by dividing both sides by the same value to end up with a ratio containing 1, such as 1.4 : 1. The method works for any two frequencies, but it is only worth using if the ratio of one frequency to the other cannot be scaled down to be an *integer* to one. If the ratio *can* be scaled to be an integer to one, then using Observation 1 will be quicker.

If the ratio of frequencies cannot be scaled to be an integer to 1, then we must find the lowest integer multiple of the ratio where both values *are* integers. To do this, we first calculate the ratio to 1. Then we multiply both sides of the ratio by ever-increasing integers until we find a resulting ratio that is an integer to an integer. Then, to calculate the resulting frequency, we divide either of the original frequencies by their corresponding ratio value.

This is much easier to understand with examples.

We will add waves with frequencies of 3.192 cycles per second and 2.28 cycles per second. The ratio for the frequencies is 3.192 : 2.28, which scales down to 1.4 : 1. This ratio is not an integer to 1. Therefore, we multiply both sides by ever-increasing integers, starting at 2, until we stumble across the ratio that has integers on both sides.

In other words:
1.4 : 1 multiplied by 2 is 2.8 : 2. These are not both integers, so we keep looking.
1.4 : 1 multiplied by 3 is 4.2 : 3. These are not both integers.
1.4 : 1 multiplied by 4 is 5.6 : 4. These are not both integers.
1.4 : 1 multiplied by 5 is 7 : 5. These are both integers, so we can stop looking. Our final ratio is 7 : 5.

We then divide either original frequency by the corresponding value in the new ratio:
3.192 ÷ 7 = 0.456
... or:
2.28 ÷ 5 = 0.456

The frequency of the signal created by adding the two waves will be 0.456 cycles per second. This is slower than the two added frequencies, which should be expected.

As another example, we will add the frequencies: 5.25 and 2.1 cycles per second. The ratio is 5.25 : 2.1. The ratio scales down to 2.5 : 1. These are not both integers, so we scale the ratio up until we find a pair of integers. It turns out that 2.5 : 1 multiplied by 2 works. It is 5 : 2. We can find the resulting frequency by calculating:
5.25 ÷ 5 = 1.05
... or:
2.1 ÷ 2 = 1.05

Our resulting frequency will be 1.05 cycles per second. The period of the resulting signal will be 1 ÷ 1.05 = 0.9524 seconds.

As another example, we will add the frequencies: 1.6 and 1.5 cycles per second. The ratio is 1.6 : 1.5. This scales down to 1.06667 : 1, which is a not a ratio with two integers. Therefore, we start the multiplying process:

Multiplying by 2, we have 2.1333 : 2
Multiplying by 3, we have 3.2 : 3
4 gives us 4.2667 : 4
5 gives us 5.3333 : 5
6 gives us 6.4 : 6
7 gives us 7.4667 : 7
8 gives us 8.5333 : 8
9 gives us 9.6 : 9
10 gives us 10.6667 : 10
11 gives us 11.7333 : 11
12 gives us 12.8 : 12
13 gives us 13.8667 : 13
14 gives us 14.9333 : 14
15 gives us 16 : 15. These are both integers so we have found our final ratio.

We find the resulting frequency by calculating:
1.6 ÷ 16 = 0.1
... or:
1.5 ÷ 15 = 0.1

Therefore, the resulting frequency is 0.1 cycles per second. The resulting period is 1 ÷ 0.1 = 10 seconds.

Now, we will add waves with frequencies of 3 and 2 cycles per second. We know that the cycles coincide at 1 second by thinking of the second observation, but we will use this method anyway. The ratio, 3 : 2 scales down to 1.5 : 1. Multiplying this by an integer to obtain an integer ratio we end up with what we started with, 3 : 2. Therefore, we divide 3 by 3 to produce 1, and we divide 2 by 2 to produce 1. The frequency of the resulting signal is 1 cycle per second. Its period is 1 second.

As another example, we will add waves with frequencies of 211 and 7. The first thing to notice about these two frequencies is that they are both integers per second. Therefore, they will definitely align after 1 second, but they might align sooner.

These have the ratio 211 : 7, which scales down to 30.1429 : 1. We will multiply this by a series of increasing integers until we end up with an integer ratio.

Multiplying by 2 produces 60.2857 : 2. These are not integers so we continue.
Multiplying by 3 produces 90.4286 : 3
Multiplying by 4 gives 120.5714 : 4
5 gives 150.7143 : 5
6 gives 180.8571 : 6
7 gives 211 : 7. These are both integers so we can stop. These values are also the ratio of the original frequencies, which means that the two cycles do not align earlier than 1 second. Therefore, they must align at 1 second, and therefore, the resulting frequency will be 1. We can confirm this by the way that 211 ÷ 211 = 1, and 7 ÷ 7 = 1.

To show the method working on an obvious combination, we will look at the addition of the frequencies 4 and 2. The first thing to notice is that these are both integers per second, so we know they will align at 1 second or sooner. The second thing to notice is that one of the frequencies is an integer multiple of the second, so the resulting frequency will be the frequency of the slower frequency (in other words, 2). Ignoring that, we will use the ratio method. The ratio is 4 : 2, which scales down to 2 : 1. These are both integers already. Therefore, we do not need to multiply the ratio by integers. Dividing each frequency by the corresponding value in the ratio, we have 4 ÷ 2 = 2, and 2 ÷ 1 = 2. Therefore, the resulting frequency will be 2 cycles per second (which we knew already).

There might be better ways to calculate the resulting frequency, but doing it with ratios emphasises how we are looking for where the cycles of the two waves coincide.

Larger fractional ratios

In the last section, we calculated the resulting frequency, or where the cycles of the two waves align, by scaling up a ratio until both sides were integers. An important idea to notice is that adding two frequencies where the scaled-up ratio consists of a single digit to a single digit (e.g. 3 : 2) results in a faster frequency, or shorter period, than if the scaled up ratio consists of more than one digit to more than one digit (e.g. 31 : 23). The integer ratio gives a clue to the resulting frequency and period. The more digits in the final scaled integer ratio, the slower the resulting frequency, and the longer the period. This is because more digits in the ratio mean that the cycles of the two original waves will take longer to align.

For example:
Adding the frequencies 7 and 3.5 gives us the ratio 7 : 3.5, which scales down to 2 : 1. The resulting frequency is 3.5 cycles per second (7 ÷ 2 = 3.5 and 3.5 ÷ 1 = 3.5), and the resulting period is 0.2857 seconds per cycle.

Adding the frequencies 7 and 4 gives the ratio 7 : 4 which scales down to 1.75 : 1. Multiplying this by consecutive integers to find an integer ratio takes us back to 7 : 4. Therefore, the resulting frequency is 1 cycle per second (7 ÷ 7 = 1 and 4 ÷ 4 = 1). The resulting period is 1 second.

Adding the frequencies 7 and 4.1 gives the scaled down ratio of 1.7073 : 1. The lowest integer multiple of this is 70 : 41. Therefore, the resulting frequency is 0.1 cycles per second (7 ÷ 70 = 0.1, and 4.1 ÷ 41 = 0.1). The resulting period is 10 seconds.

Adding the frequencies 7 and 4.31 gives the scaled down ratio 1.6241 : 1. The lowest integer multiple of this is 700 : 431. Therefore, the resulting frequency is 0.01 cycles per second (7 ÷ 700 = 0.01, and 4.31 ÷ 431 = 0.01). The resulting period is 100 seconds.

Things become more interesting when it is impossible to find a final ratio where both values are integers. This happens when the first scaled down ratio is an irrational number to 1.

An irrational number is one that cannot be portrayed using an integer ratio. (The word “ratio” appears in the word “irrational”, so we can think of “irrational” as being “un – ratio – able”). To put this another way, an irrational number is one that cannot be portrayed as a fraction with an integer numerator and an integer denominator. To explain what this means, 0.5 is a *rational* number – it can be expressed as a fraction with integers: $\frac{1}{2}$. However, the square root of 0.5 is an *irrational* number – it cannot be expressed as a fraction consisting of integers. The number $\sqrt{0.5}$ is 0.707106781186... . We could create an approximate fraction, for example, $\frac{7{,}071}{10{,}000}$, but this does not represent the square root of 0.5 particularly accurately. If the fraction were $\frac{70{,}710{,}678}{100{,}000{,}000}$, it would be closer, but still not equal. In reality, there is no fraction consisting of integers that is equal to $\sqrt{0.5}$.

One characteristic of irrational numbers is that the digits after the decimal point go on forever. However, this is not exclusive to irrational numbers – for example, $\frac{1}{3}$ has digits that go on forever (0.33333333...), but the difference is that 0.33333333... can be expressed as an integer divided by another integer, while an irrational number cannot.

Irrational numbers are often seen in maths when dealing with circles. For example:

- $\sqrt{0.5}$ (0.7071...), which is Sine 45 or Cosine 45.
- π (3.1415...), which is the length of the circumference of a circle divided by its diameter.
- e (2.7182...), which if raised to an Imaginary power, indicates the position of a point on a unit radius circle's circumference.
- $\sqrt{2}$ (1.4142...), which is the radius of a circle for which the x-axis or y-axis value of a point at 45 degrees on the circumference is 1. In other words, $\sqrt{2}$ * sin 45 = 1, and $\sqrt{2}$ * cos 45 = 1.

When adding frequencies together, if the ratio between the frequencies scales down to be an irrational number to 1, then there will be no possible integer that can be multiplied against the ratio to make both sides equal to an integer. If there were an integer that could be multiplied by the ratio to make an integer ratio, then the ratio would not be an irrational number to 1. In trying to find an integer to multiply against the ratio to make it into an integer ratio, we would be trying higher and higher numbers forever without a result. A hypothetical fraction that represented such a ratio would require an infinite number of digits in the numerator and denominator. This means that the period of the resulting signal would be infinitely long. The signal would never repeat, and its frequency would be zero.

[Note that the concept of zero frequency for a summed *signal* refers to how it never repeats. The concept of zero frequency for a *pure wave* refers to how the object rotating around the circle from which the wave is, or could have been derived, is not moving, and so the "wave" is a straight line.]

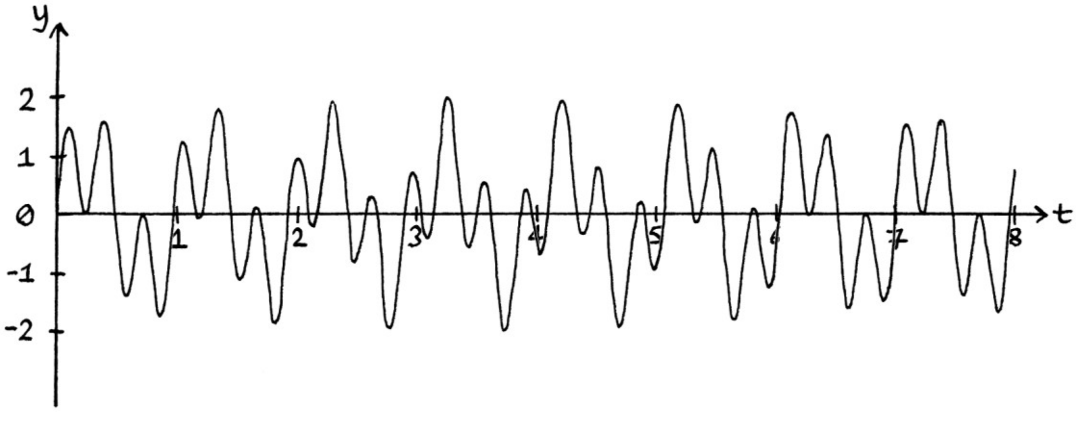
We will say that we want to add the frequencies π cycles per second (or in other words, 3.1415926535897932384626...) and 1 cycle per second. The ratio is π : 1. If we were to calculate the combined frequency in the normal way, we would start by multiplying the ratio by ever-increasing integers:

If we multiply the ratio by 2, we will have 6.2831 : 2
If we multiply by 3, we will have 9.4248 : 3
If we multiply by 4, we will have 12.5664 : 4
5 gives us 15.7080 : 5
6 gives us 18.8496 : 6
7 gives us 21.9911 : 7

... and if we skip a bit...
112 gives us 351.8584 : 112
113 gives us 354.99997 : 113

... and if we skip a bit more...
10,000,000 gives us 31,415,926.5359 : 10,000,000

We will never end up with an integer ratio. This means that the period of the signal resulting from adding the frequencies π and 1 will be infinitely long. The signal will never repeat. The frequency of the signal will be zero. The resulting graph will resemble the sum of two waves that has been skewed differently along its length:



The resulting signal has zero frequency because it never repeats, but because it is not a pure wave, its graph is not a horizontal line (which pure waves with zero frequency are).

It is important to note that it is the *ratio* between the frequencies that is important and not the frequencies themselves. If we have two waves with frequencies of 2π and π cycles per second, the ratio is 2 : 1. This ratio, obviously, *can* be expressed as a fraction with integers. In this case, the faster wave is an integer multiple of the slower wave, so the resulting signal will have a frequency equal to the slower wave. Therefore, the resulting signal will have a frequency of π cycles per second.

Frequency and amplitude

The complexity of maximums and minimums

Calculating the maximum and minimum y-axis values of a signal that is the sum of two waves of different frequencies is not particularly straightforward. Depending on the way two waves engage with each other, the resulting maximum and minimum might be equal to the sum of the amplitudes of the two waves, or they might be some value lower. [They will never be more, which might be obvious if you think about how the y-axis values of each graph will never be more than those wave's amplitudes, and therefore if they are added together, they can never be more than the sum of those two waves' amplitudes]. The maximum and minimum y-axis values of a signal created from adding waves of different frequencies might be the negative of each other, or they might not be.

The resulting maximum and minimum are dependent on:

- The amplitude of each wave being added.
- The ratio between the two frequencies.
- The ratio between the portion of total amplitude assigned to each frequency.
- The phase of the two waves (although we will ignore this for now).

For *any* pair of waves that have the same ratio between the frequencies, and the same portion of the total amplitude split between each frequency (and zero phase), the resulting maximum of the sum will be the same fraction multiplied by the sum of the two amplitudes (as will be the resulting minimum, but the fraction might not be the negative of the one for the maximum). The complicated part in calculating the resulting maximum and minimum for the sum of two waves is calculating the fractions that are relevant for a particular combination of frequency and amplitude ratios.

Examples

The idea is best explained with examples. If we have a frequency ratio of 2 : 1, and the amplitudes are the same for each wave (and thus have a ratio 0.5 : 0.5 in the proportion assigned to each frequency), the maximum y-axis value of the resulting signal will *always* be 0.8801 multiplied by the sum of the two original amplitudes.

For example, the two waves “y = 2 sin (360 * 2t)” and “y = 2 sin (360 * t)” have a frequency ratio of 2 : 1. The total amplitude (2 + 2) is split evenly between the two waves, so we can say the amplitudes have a ratio of 0.5 : 0.5. The sum of the two waves will result in a signal that has a maximum y-axis value of 0.8801 * (2 + 2) = 3.52.

The waves “y = 7 sin (360 * 4t)” and “y = 7 sin (360 * 2t)” also have the frequency ratio of 2 : 1, and the total amplitude is split evenly between those frequencies, so we can say the amplitudes have a ratio of 0.5 : 0.5. The maximum y-axis value of the signal created by adding them together will therefore be 0.8801 * (7 + 7) = 12.32.

However, if we have the waves “y = 3 sin (360 * 2t)” and “y = 1 sin (360 * t)”, then the frequency ratio is still 2 : 1, but the amplitude ratio is 0.75 : 0.25. Therefore, (without me showing how I know this), the fraction for these two waves will not be 0.8801, but instead 0.9317. This means the resulting maximum y-axis value will be 0.9317 * (3 + 1) = 3.7268. It is the case that *any* two waves that have the frequency ratio 2 : 1, and the amplitude ratio 0.75 : 0.25, will also produce a wave with a maximum y-axis value equal to 0.9317 multiplied by the sum of the original amplitudes.

If we have the waves “y = 2 sin (360 * 5t)” and “y = 8 sin (360 * 1t)”, the frequency ratio is 5 : 1, and the amplitude ratio is 0.2 : 0.8. The fraction for these ratios is 1. In other words, the resulting signal will have a maximum y-axis value equal to
1 * (8 + 2) = 10. Any two waves with the same ratios will have the same fraction of 1. In fact, for this particular frequency ratio, *any* amplitude ratio will also have the same fraction of 1.

Calculating the fraction is the difficult part. The reason why there is a consistent fraction for the same frequency and amplitude ratios becomes apparent when we look at addition with circles in the next chapter. The reason why the resulting maximum is not always the sum of the two original amplitudes also becomes obvious. [Two waves of different amplitudes and different frequencies do not create a circle, but instead a shape. That shape might not be as high as it is wide, and therefore the maximum y-axis value from that shape will not necessarily be the sum of the two original amplitudes. There is a consistent fraction for the same frequency and amplitude ratios because the same ratios create the same shape, although not necessarily of the same size.]

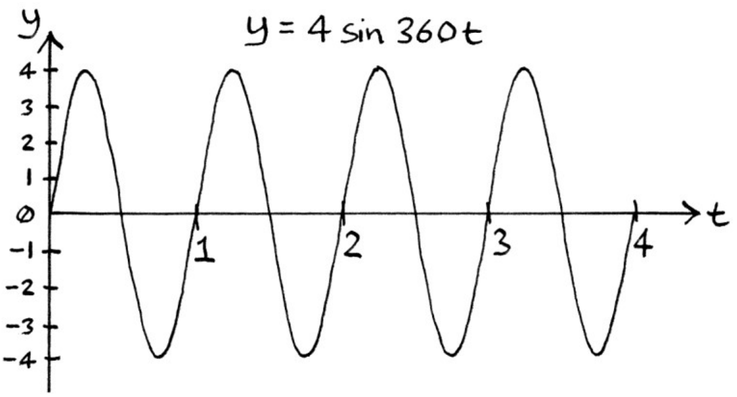
Significantly different amplitudes

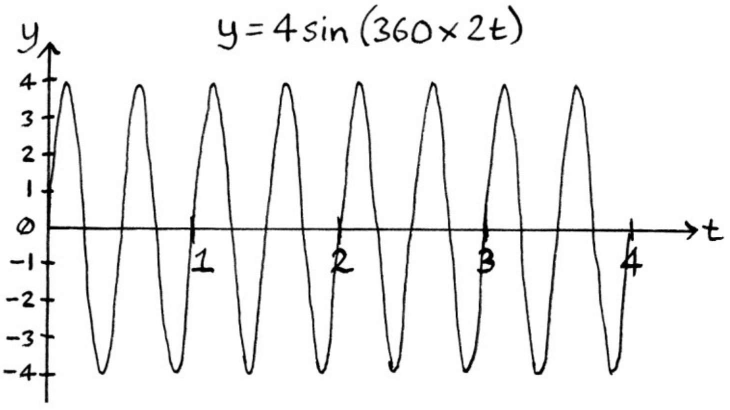
If we add waves with different frequencies, but where one wave’s amplitude is significantly larger than the other’s, the smaller amplitude signal will have less of an influence on the appearance of the sum. The higher the ratio between the amplitudes, the more the resulting signal will resemble the larger amplitude wave.

To illustrate this, we will start by adding:
“y = 4 sin 360t”
... and:
“y = 4 sin (360 * 2t)”

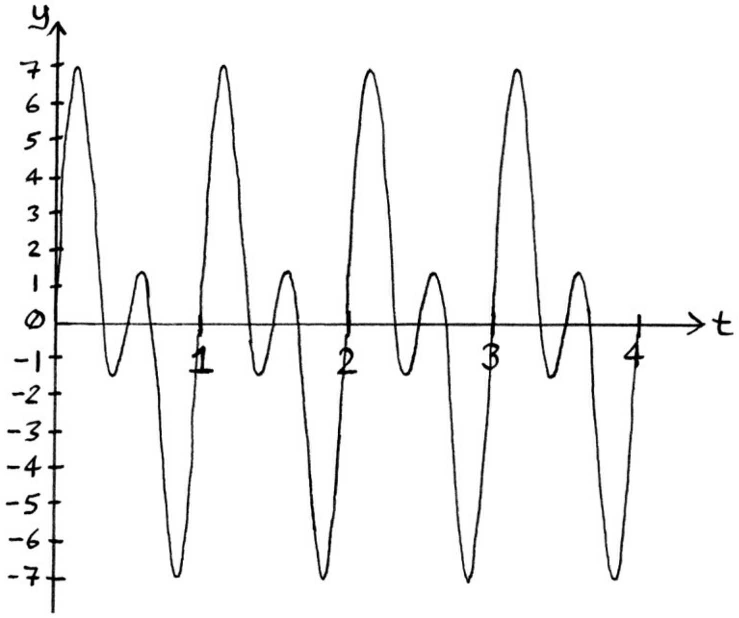
These have the same amplitude.

The two added waves look like this:



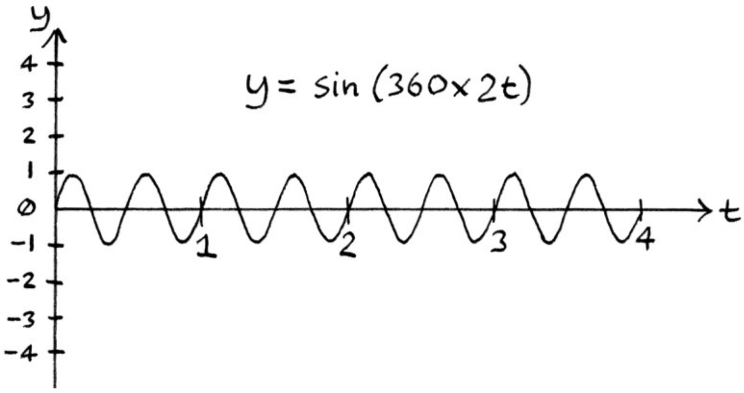


The resulting signal looks like this:

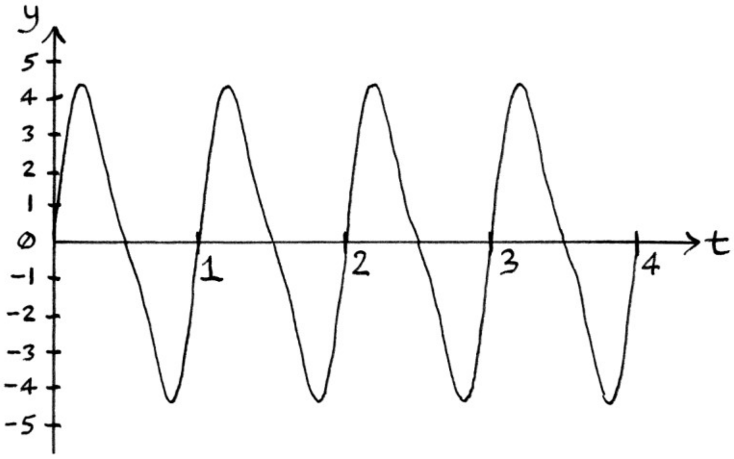


Now, we will add:
“y = 1 sin (360 * 2t)”
... to:
“y = 4 sin 360t”

The wave “y = 1 sin (360 * 2t)” looks like this:

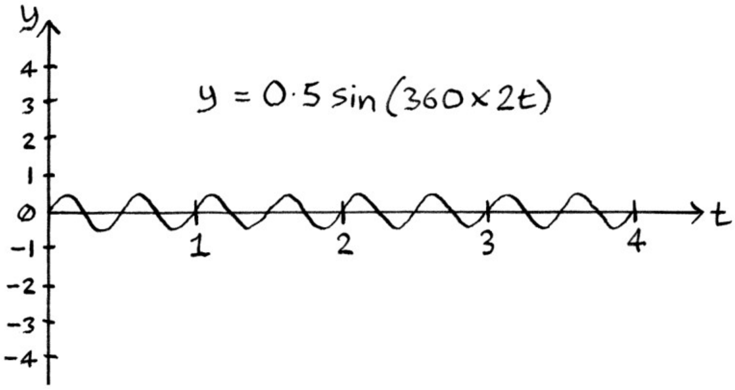


The result of the sum looks like this:

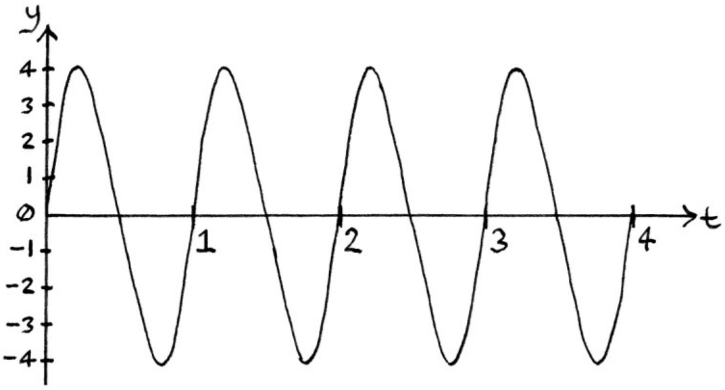


Now, we will add:
“y = 0.5 sin (360 * 2t)”
... to:
“y = 4 sin 360t”

The wave “y = 0.5 sin (360 * 2t)” looks like this:

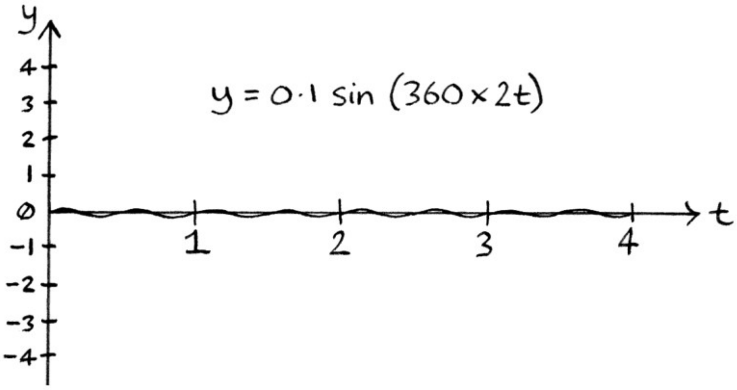


The result of the sum looks like this:

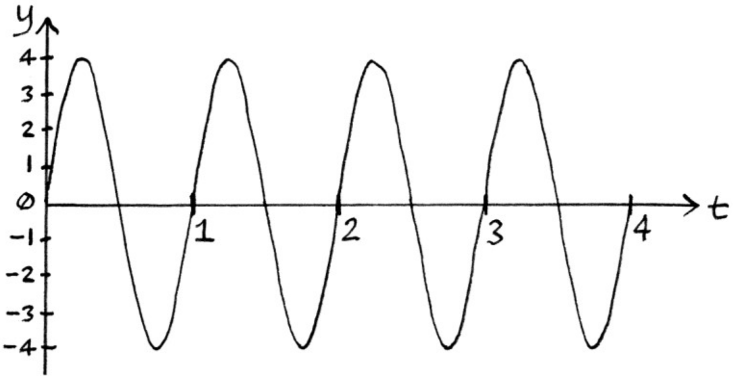


Next, we will add:
“y = 0.1 sin (360 * 2t)”
... to:
“y = 4 sin 360t”

The graph of “y = 0.1 sin (360 * 2t)” looks like this:



The result of the sum looks like this:



As the amplitude of one wave decreases, the resulting signal looks more and more like the other wave. This might be obvious because the y-axis values from the second wave become smaller and smaller until they cease to be significant in the sum.

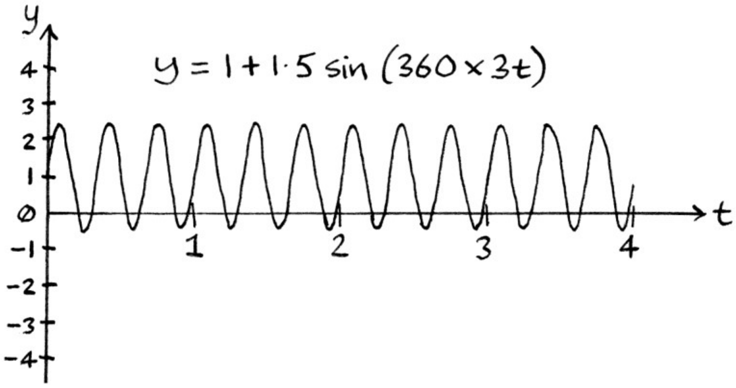
Frequency and mean levels

The mean level aspect of adding two waves with different frequencies is easy to understand. The resulting signal will have a mean level equal to the sum of the mean levels of the two original waves, in the same way as if we were adding two same-frequency waves. If there are non-zero mean levels, it is helpful to calculate the resulting mean level, and then deal with the complicated differences in frequency, amplitude and phase afterwards.

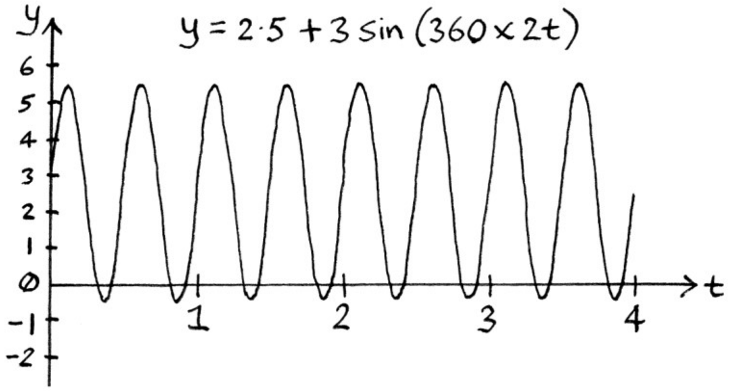
If we add the waves:
“y = 1 + 1.5 sin (360 * 3t)”
... and:
“y = 2.5 + 3 sin (360 * 2t)”
... then the resulting signal will have a mean level of: 1 + 2.5 = 3.5 units.

This can be confirmed by the formula for the resulting signal, which is:
“y = 1 + 1.5 sin (360 * 3t) + 2.5 + 3 sin (360 * 2t)”
... which can be rephrased as:
“y = 3.5 + 1.5 sin (360 * 3t) + 3 sin (360 * 2t)”

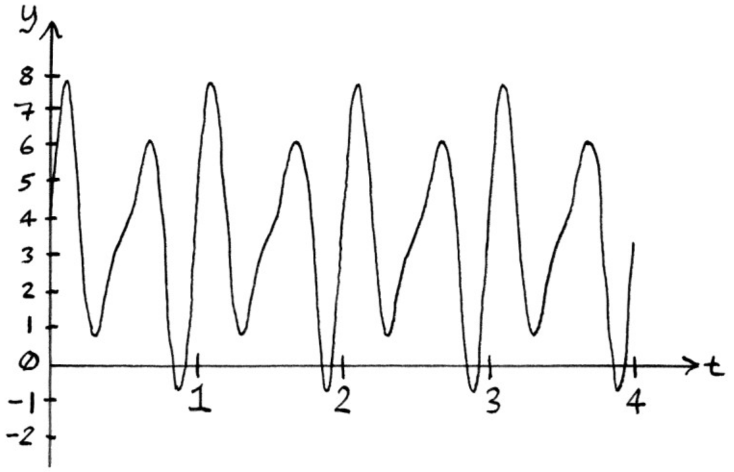
The wave “y = 1 + 1.5 sin (360 * 3t)” looks like this:



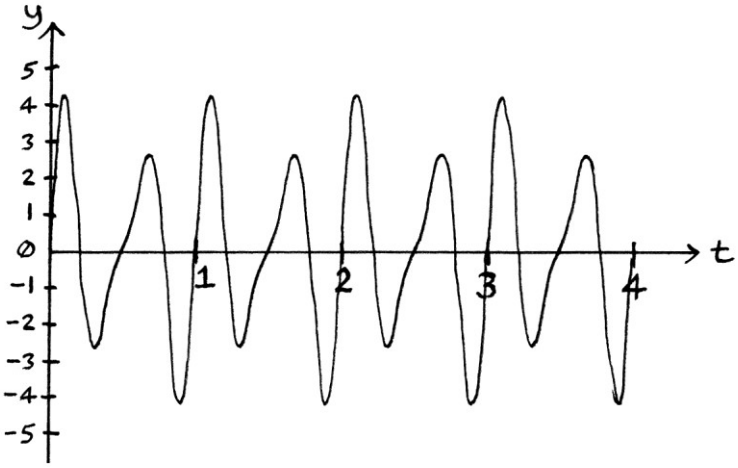
The wave “y = 2.5 + 3 sin (360 * 2t)” looks like this:



The result, “y = 3.5 + 1.5 sin (360 * 3t) + 3 sin (360 * 2t)”, looks like this:



The centre of this signal is at y = 3.5 units. [If we average all of the y-axis values for one cycle (or every cycle), we will end up with 3.5 units]. Supposing the mean level in both added waves were zero, then the resulting signal would be:
“y = 1.5 sin (360 * 3t) + 3 sin (360 * 2t)”
... which looks like this:

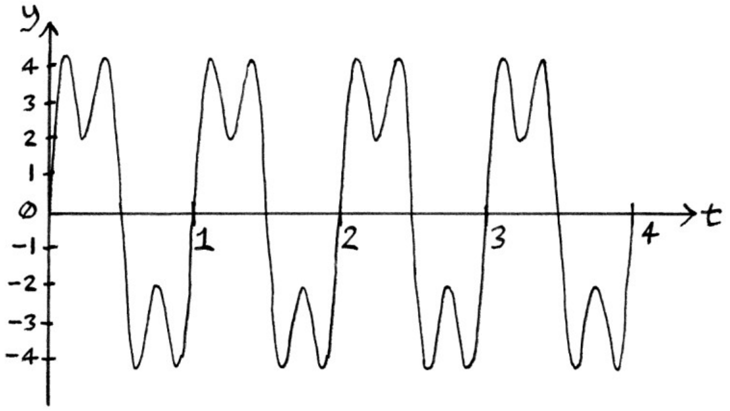


The graph is identical, except it is centred on y = 0, instead of on y = 3.5. Mean levels do not affect the *shape* of the resulting signal – they only affect the y-axis position of its centre.

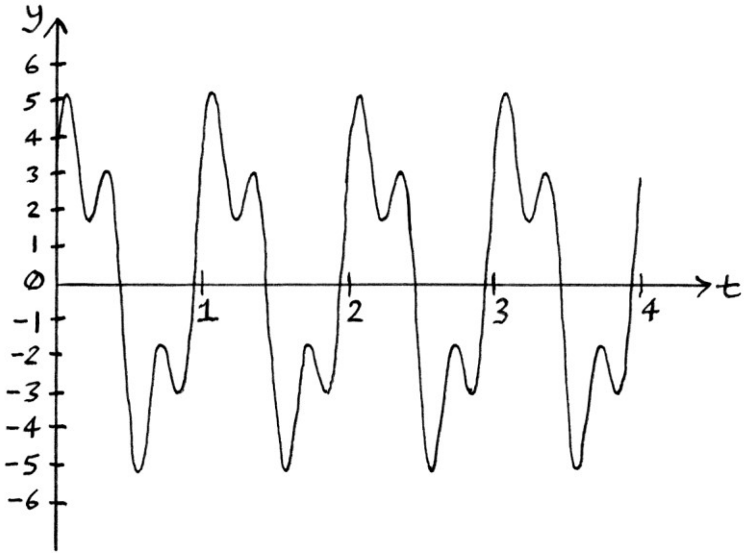
Frequency and phase

Dealing with different frequencies and different phases is less intuitive than dealing with mean levels, but easier to understand than dealing with amplitude. It is much easier to understand frequency and phase when looking at circles, which we do in the next chapter. We will look at some examples.

If we add the two waves:
“y = 4 sin 360t”
... and:
“y = 2 sin (360 * 3t)”
... we end up with this signal:

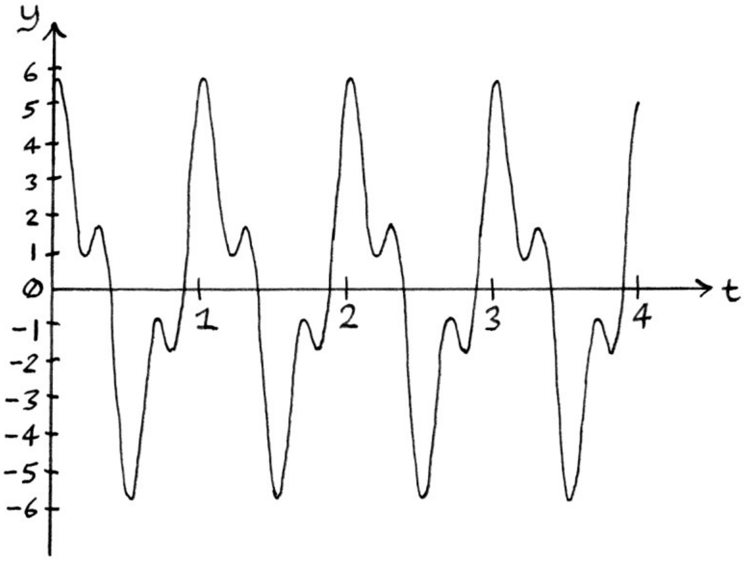


However, if we add the same phase to each of the added waves, the result is not a shifted version of the resulting signal. For example, the result of:
“y = 4 sin (360t + 30)”
... added to:
“y = 2 sin ((360 * 3t) + 30)”
... has a different shape altogether:

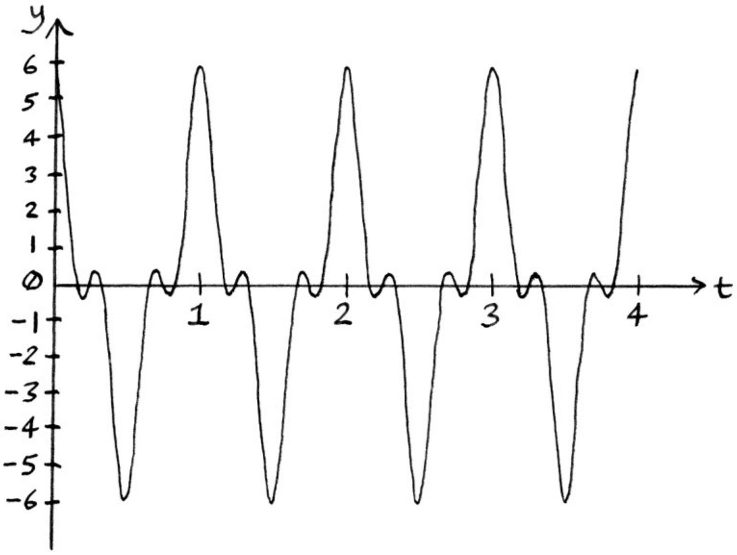


This is because the waves do not repeat at the same rate. A full set of degrees on one wave is completed in a different amount of time to a full set of degrees in the other wave. Therefore, adding the same phase value to each wave's formula causes each added wave to be shifted to the left by a different amount. This changes how the cycles align with each other, and therefore, which y-axis values are added to which y-axis values. Changing the phase of both waves skews the resulting signal.

We will give each wave a phase of 60 degrees. We will be adding:
"y = 4 sin (360t + 60)" and "y = 2 sin ((360 * 3t) + 60)"
... to produce this signal:



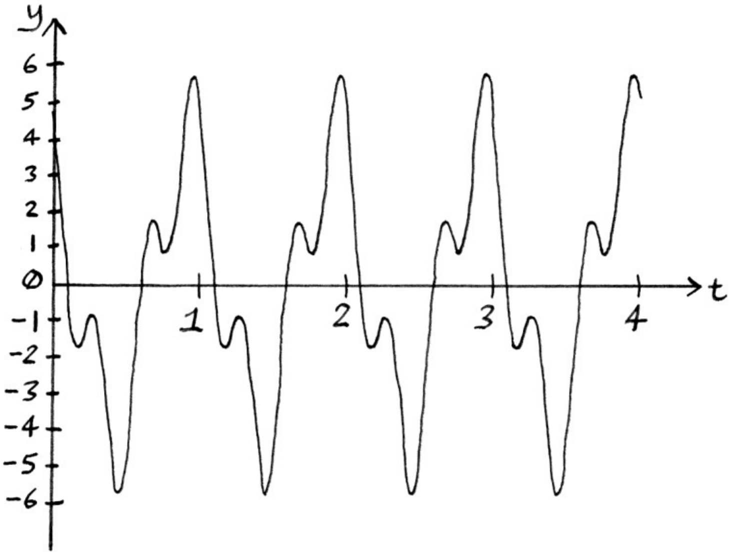
If we give each wave a phase of 90 degrees, we will be adding:
"y = 4 sin (360t + 90)" and "y = 2 sin ((360 * 3t) + 90)"
... to produce this signal:



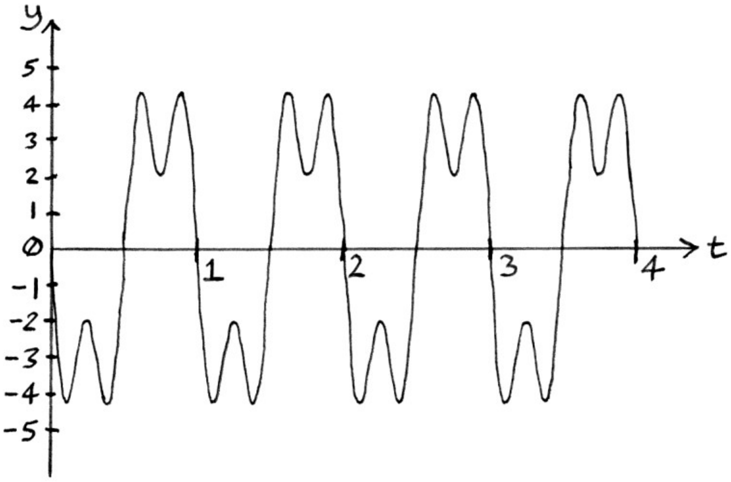
With a phase of 90 degrees, the skew of the resulting signal is the maximum it can be. Note that the sum is the same as “y = 4 *cos* 360t” added to “y = 2 *cos* (360 * 3t)”.

With a phase of 120 degrees, the waves are:
“y = 4 sin (360t + 120)” added to “y = 2 sin ((360 * 3t) + 120)”.

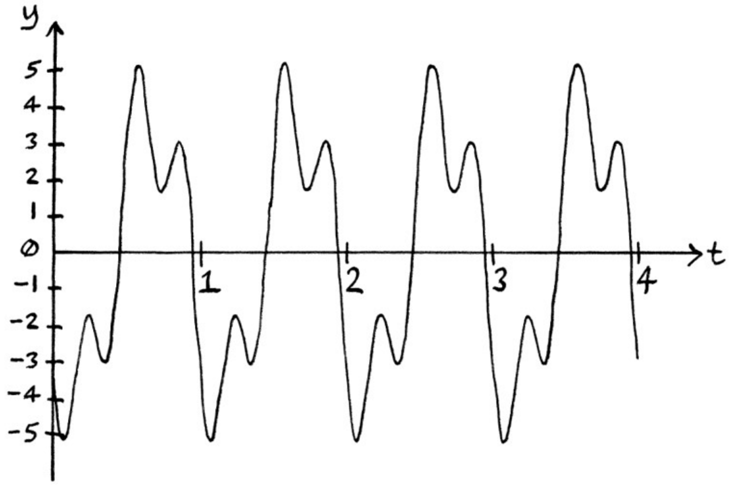
The resulting signal is this:



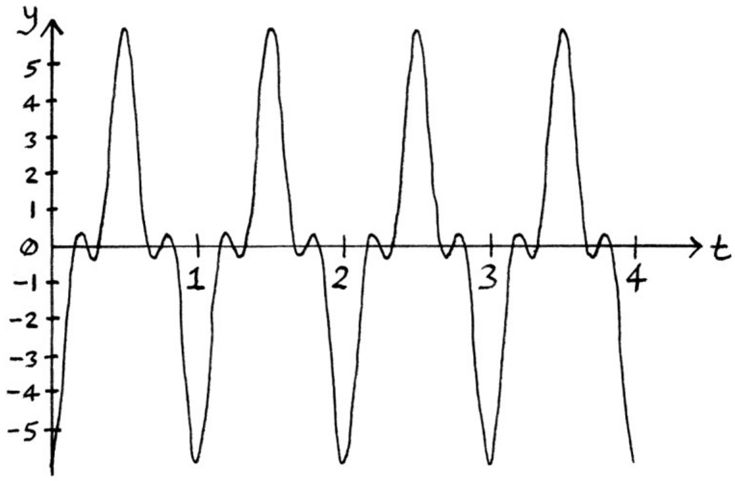
With a phase of 180 degrees, we will have the following signal. Note how this is an upside-down version of the original zero-phase result:



With a 210 degree phase:



With a 270 degree phase: (note how this is an upside down version of the 90 degree phase):



From all of this, it should be clear that if we wanted to shift our resulting signal along the time axis by a certain amount, it would not work if we shifted each of the two waves that were summed to make that signal by the same angle. If we did that, we would just end up with a skewed version of the signal.

The amount of the skew is related to the ratio between the two phases, and how that relates to the ratio of the frequencies. For example, for our two waves:
“y = 4 sin 360t”
... and:
“y = 2 sin (360 * 3t)”
... the ratio between the frequencies is 3 : 1.

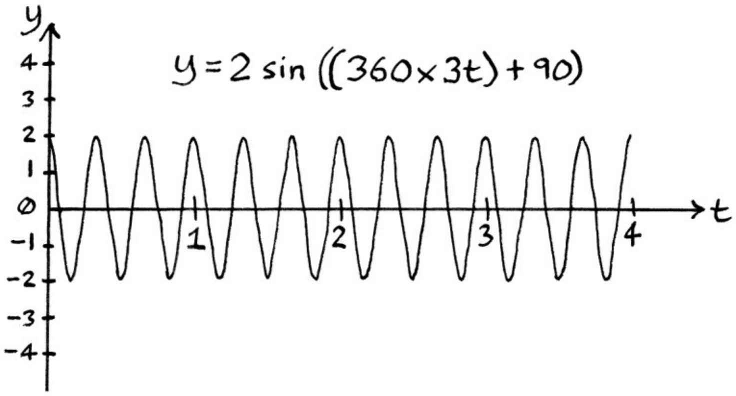
If we want the resulting signal to be slid along the time axis and maintain its shape, the phases added to each wave must also have a ratio of 3 : 1. In other words, the wave with the frequency of 3 needs to be shifted 3 times further than the wave with the frequency of 1. The reason for this might be apparent if you remember that on each wave, a cycle of 360 degrees is completed in a different time. If we want the cycles to coincide together, we must shift the two waves a proportional amount so that the cycles still coincide. The wave that completes 360 degrees 3 times in one second needs to be shifted 3 times as much as the wave that completes 360 degrees once per second.

We will say that we want the resulting signal from the above example [“y = 4 sin 360t” and “y = 2 sin (360 * 3t)”] to be slid to the left without losing its shape, and that we want to add 30 degrees to the phase of the one-cycle-per-second wave. The waves we would need to add are:
“y = 2 sin ((360 * 3t) + 90)”
... added to:
“y = 4 sin (360t + 30)”

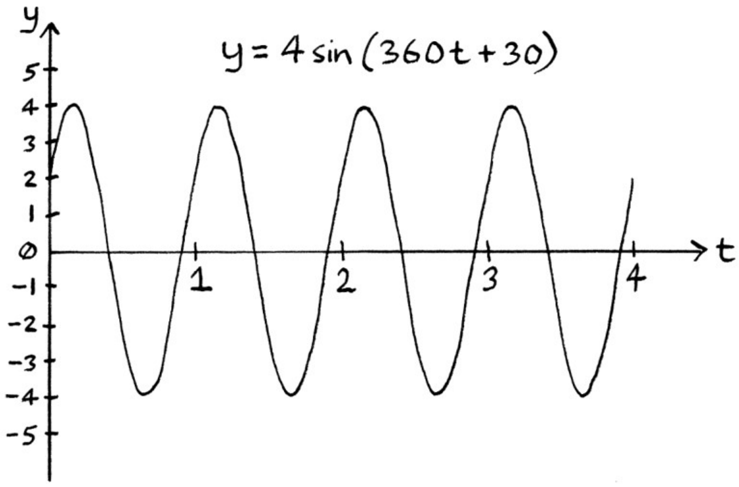
The phases are 90 degrees and 30 degrees, which have a ratio of 3 : 1 to match the frequency ratio of 3 : 1. The wave with the 3-cycle-per-second frequency is shifted 3 times as much as the wave with the 1-cycle-per-second frequency.

The graphs are as follows:

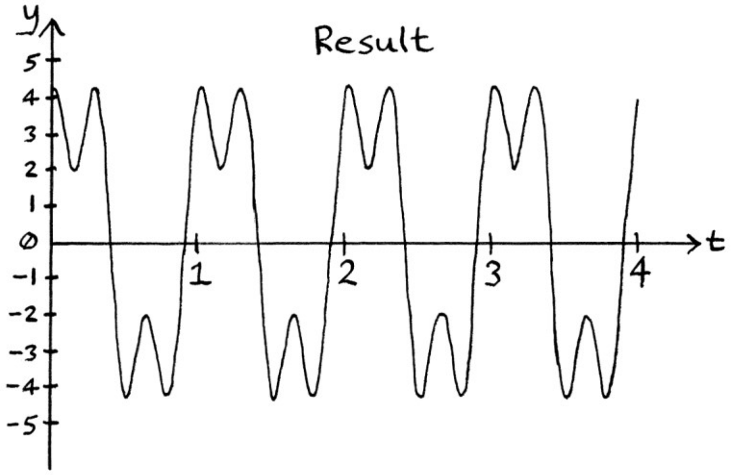
“y = 2 sin ((360 * 3t) + 90)”:



"y = 4 sin (360t + 30)":



The result:

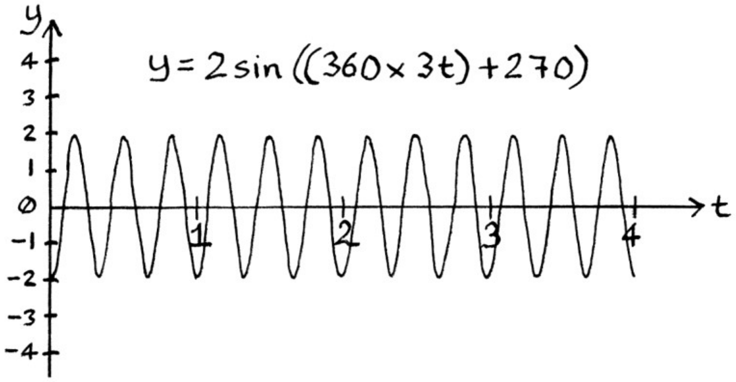


If we want to slide the resulting signal further to the left by adding 90 degrees to the phase of the one cycle-per-second wave, then the formulas we need are:
"y = 2 sin ((360 * 3t) + 270)"
... added to:
"y = 4 sin (360t + 90)"

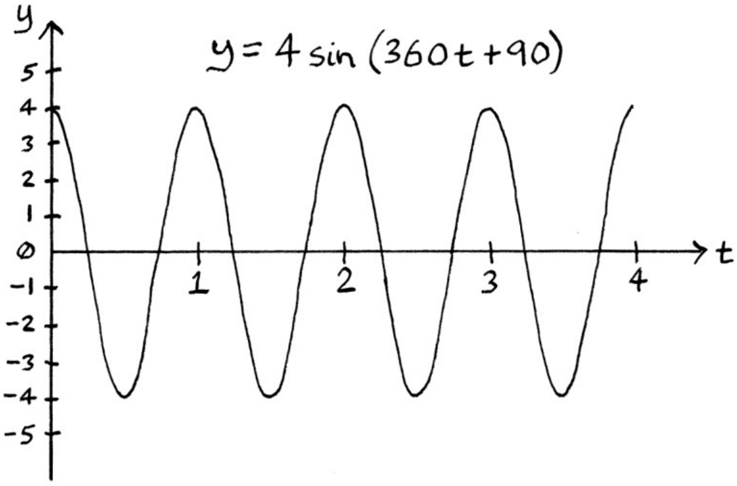
The phases still have a ratio of 3 : 1.

The graphs are as follows:

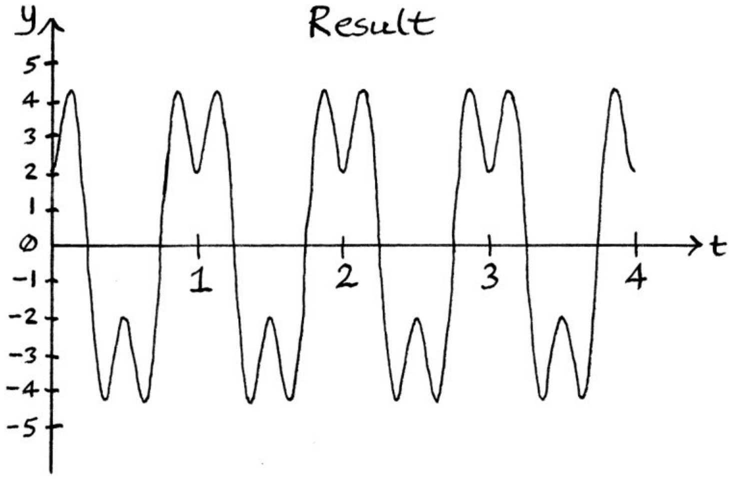
“y = 2 sin ((360 * 3t) + 270)”:



“y = 4 sin (360t + 90)”:



The resulting signal:



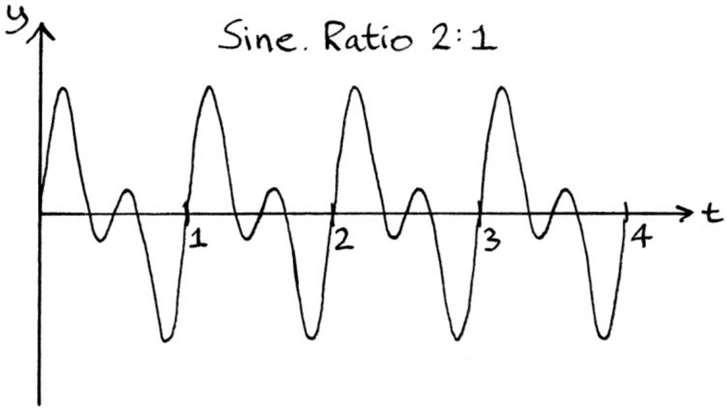
Phase and circles

The idea of phase becomes much more intuitive when looking at circles, which we will do in the next chapter.

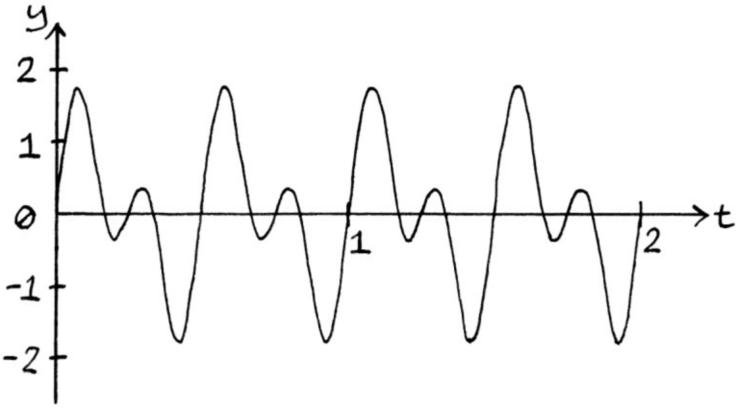
Addition of frequencies: patterns for waves

When adding waves with zero phase, similar amplitudes, but different frequencies, it is easy to see certain patterns depending on the ratio between the frequencies. There are countless patterns to recognise, which are independent of the actual frequencies.

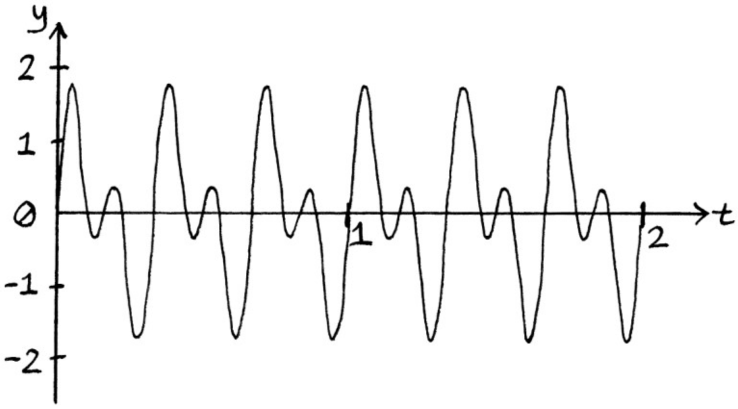
If the frequency ratio is 2 : 1, and the amplitudes are identical, then the resulting signal will look like this for added Sine waves:



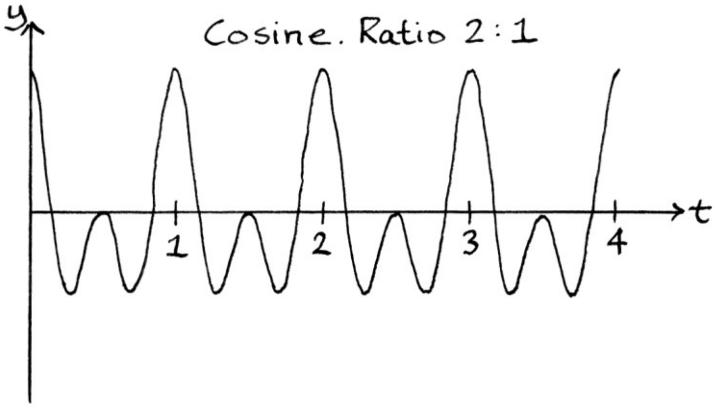
If the amplitudes are slightly different, then the basic shape will be slightly different, but still close enough to be recognisable. The shape will be seen, no matter what the frequencies. For example:
“y = sin (360 * 2t)” added to “y = sin (360 * 4t)” looks like this:



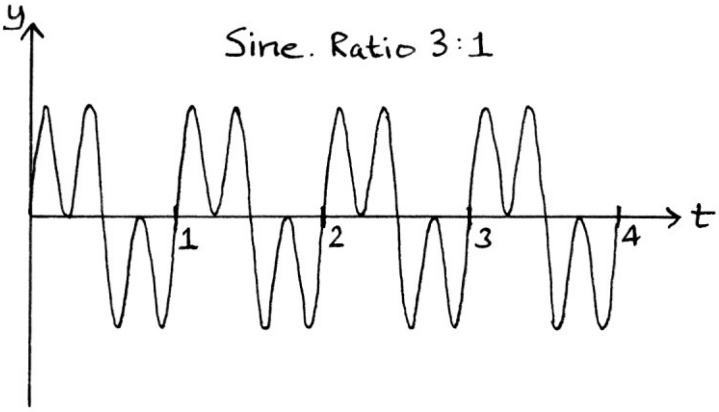
If we add:
“y = sin (360t * 3t)”
... to:
“y = sin (360 * 6t)”
... the result looks like this:



For added Cosine waves with a frequency ratio of 2 : 1, the signal will have a shape similar to this:

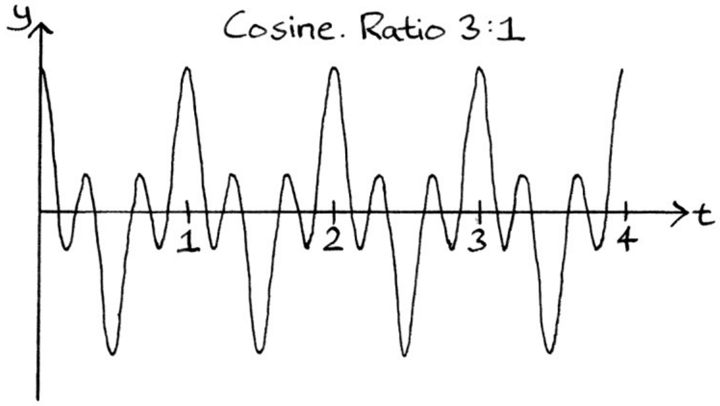


If the ratio is 3 : 1, the resulting signal will look like this for added Sine waves:

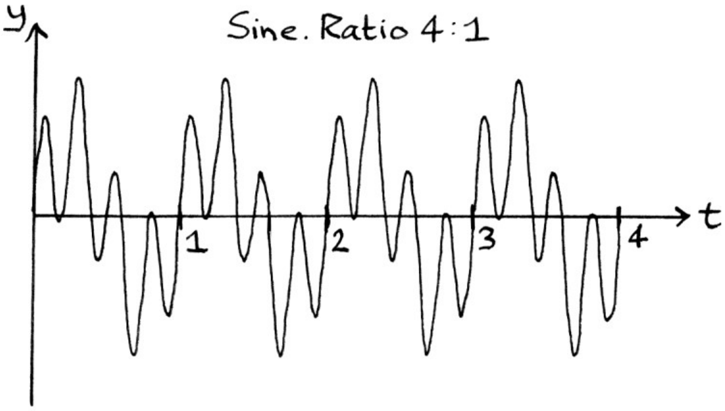


The above ratio is the easiest ratio to recognise. It can be helpful to remember its shape in case you see it produced by any other calculations. The way to remember it is by noticing how it looks like the letters “M” and “W” repeated over and over again.

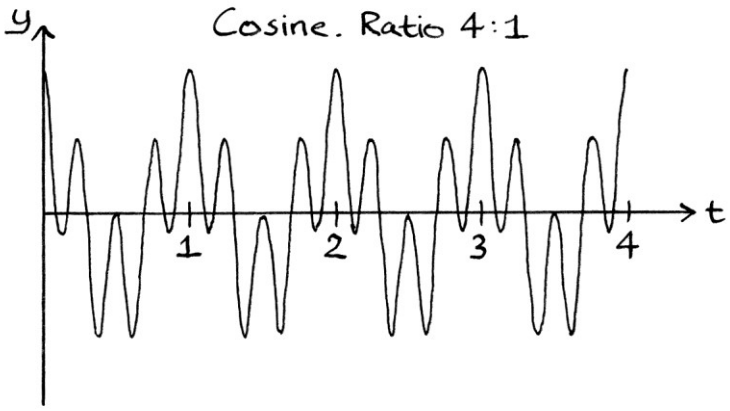
For added Cosine waves, a ratio of 3 : 1 produces a signal that looks like this:



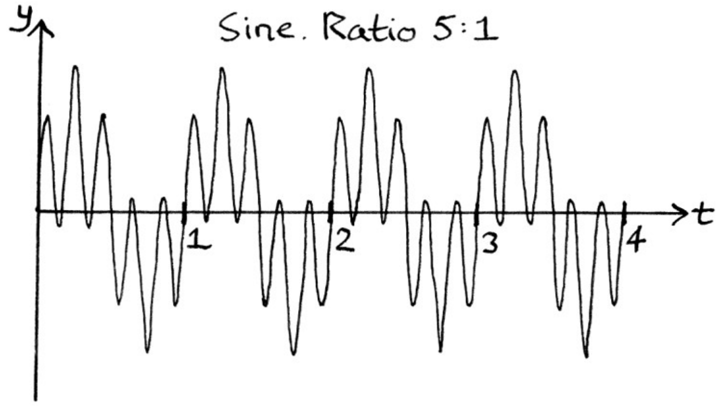
If the ratio is 4 : 1, the resulting signal will look like this for added Sine waves:



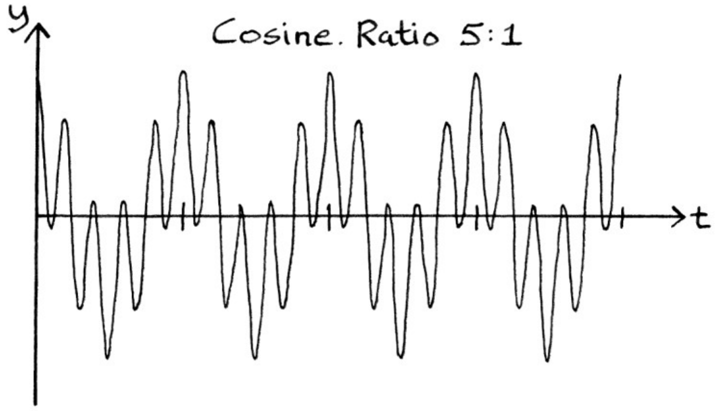
For added Cosine waves, the signal looks like this:



If the ratio is 5 : 1, the signal will look like this for added Sine waves:



For added Cosine waves, the signal looks like this:



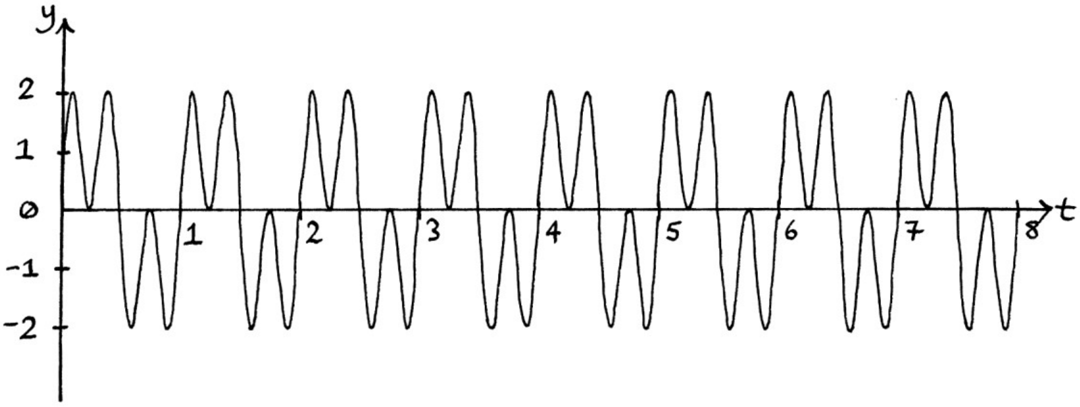
Irrational ratio patterns

As discussed earlier, a signal that is the sum of two waves that have a frequency ratio that is an irrational number to 1 will never repeat. However, it is still possible to see patterns. The patterns will resemble the patterns in the sum of two waves that have a close, but rational ratio. Instead of the cycles repeating in a signal created from an irrational frequency ratio, each cycle will be a skewed version of the one before, as if the phase of each summed wave were shifting for each cycle.

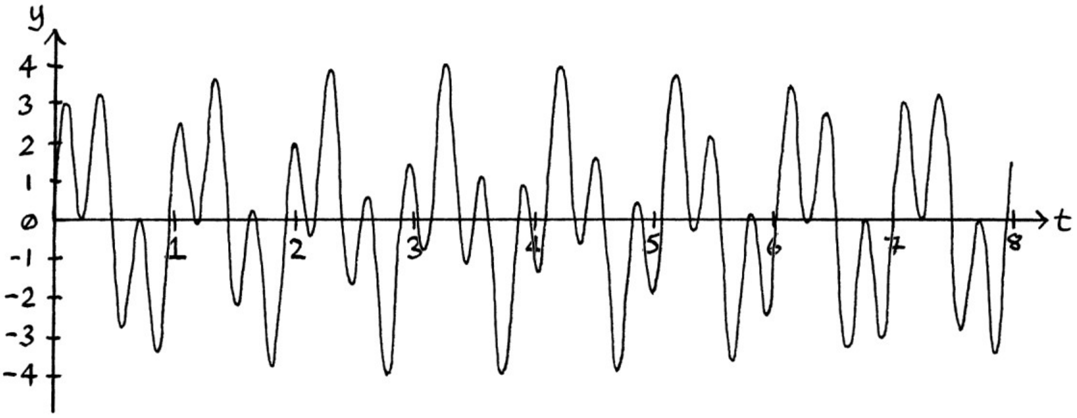
For example, the signal created from:
“y = 2 sin (360 * 1t)”
... added to:
“y = 2 sin (360 * πt)”
... has a ratio of π :1, which is 3.1415926535... : 1.

The resulting signal bears similarities to the sum of:
“y = 2 sin 360t”
... added to:
“y = 2 sin (360 * 3t)”
... [the frequencies of which have a ratio of 3 : 1], but as if the two waves were constantly shifting in phase and never settling down.

This is “y = 2 sin 360t” added to “y = 2 sin (360 * 3t)”, showing the distinctive 3 : 1 “MW” frequency ratio shape:



This is “y = 2 sin 360t” added to “y = 2 sin (360 * πt)”, which resembles a 3 : 1 ratio with zero phases at the very start, but then resembles 3 : 1 ratios with various other phases as it progresses:



It is often possible to recognise an irrational ratio by looking at the graph because of this fact. Note that a graphing calculator or computer program will need to have a good level of accuracy in its programming not to let rounding errors make an irrational-ratio signal appear to repeat.

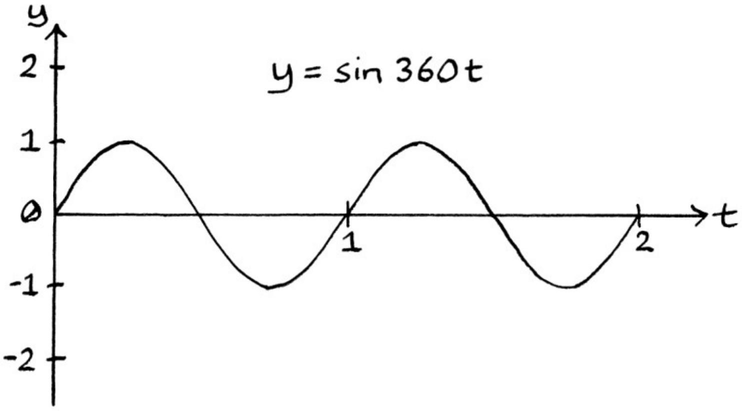
Oddities of adding frequencies

Large frequency ratios

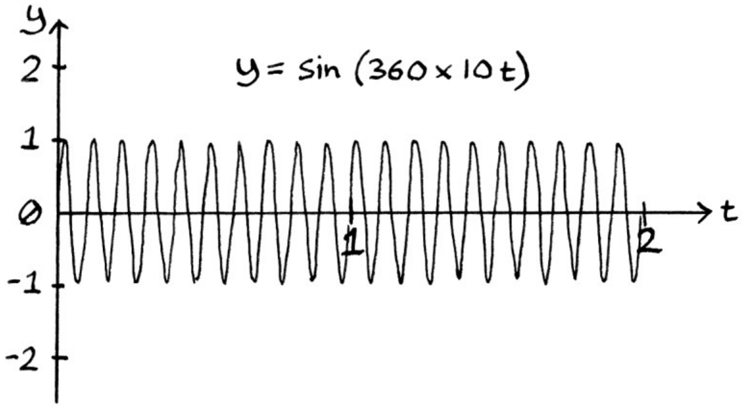
As the ratio of frequencies in two added waves becomes higher, the longer it takes for the cycles to align. However, an interesting consequence is that the shape of the resulting signal will have a pattern that resembles the slower wave with the faster wave incorporated into its curve. As an example, we will add:
“y = sin 360t”
... and:
“y = sin (360 * 10t)”.

The graphs look like this:

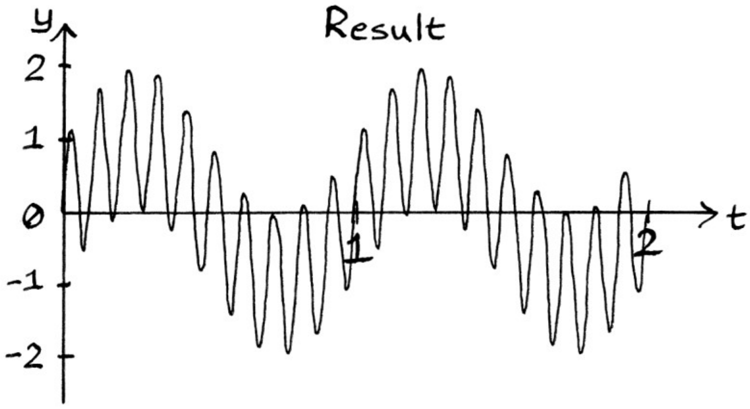
“y = sin 360t”:



“y = sin (360 * 10t)”:



The result of the sum is as so:



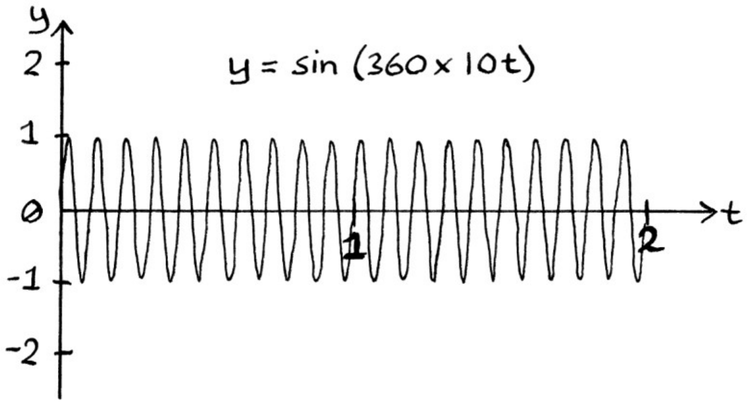
The resulting signal has the overall appearance of the slower frequency, but with ripples that are a result of the higher frequency. The faster frequency wave completes many cycles in the time it takes the slower frequency wave to complete just one. The effect becomes even more obvious (but harder to draw) as the frequency difference increases.

Very similar frequencies

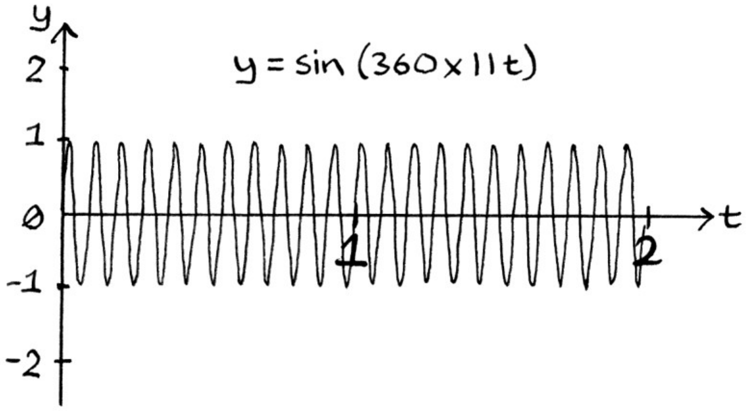
If two waves have very close frequencies, the resulting signal becomes "pulse-like". For example, we will add:
"y = sin (360 * 10t)"
... and:
"y = sin (360 * 11t)".

The graphs are as follows:

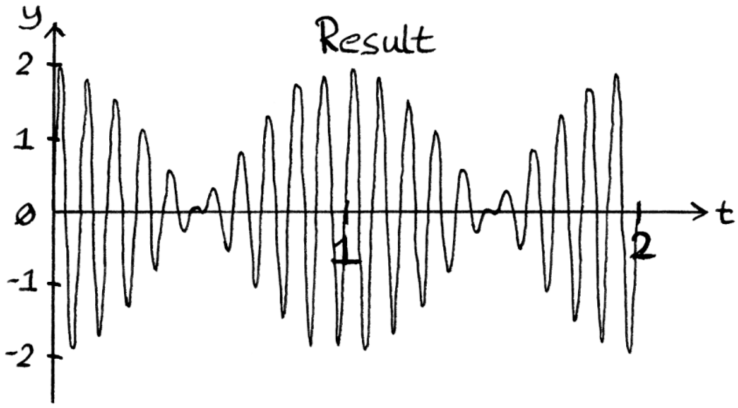
"y = sin (360 * 10t)" :



"y = sin (360 * 11t)":



The result:



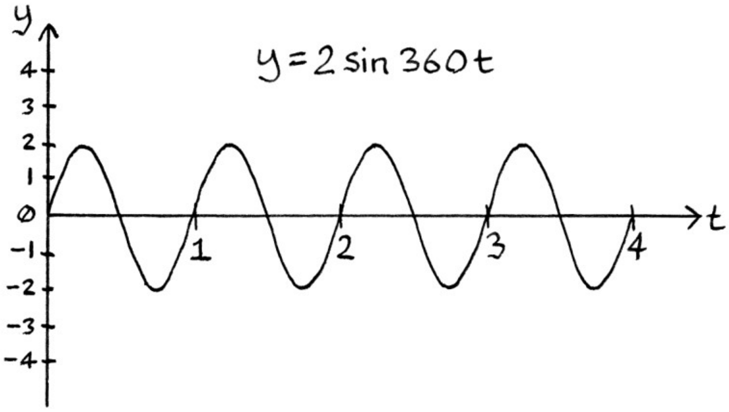
Again, the effect becomes more obvious (and harder to draw), as the two frequencies become closer together.

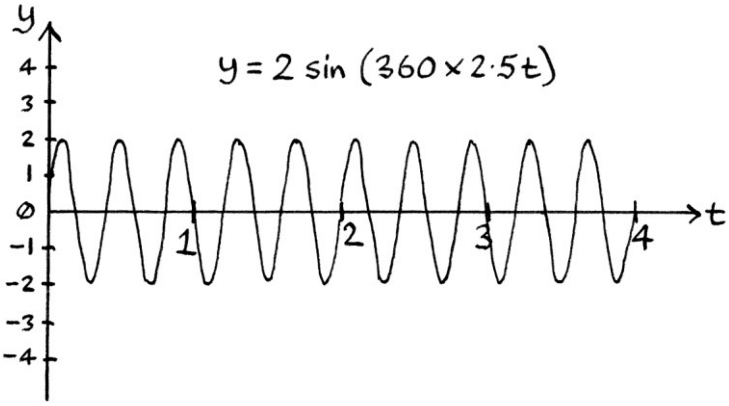
Adding more than two waves

If we are adding three waves together, we first add two of the waves together, and the result will be based around the amplitude, frequency, phase and mean level of those two waves. Then we add the third wave to that result, and we will end up with a signal that is based on the amplitude, frequency, phase and mean level of all three waves.

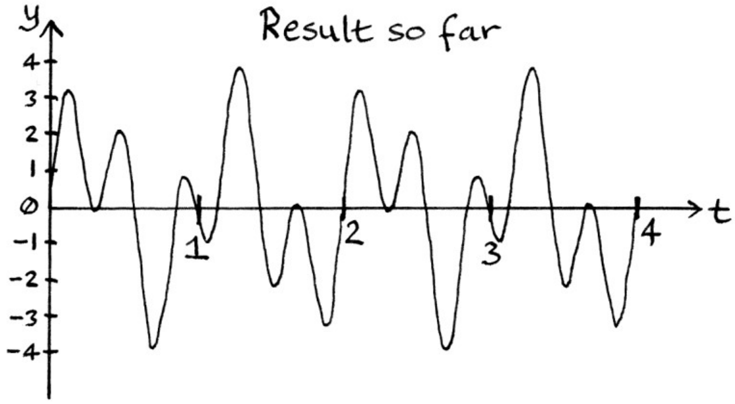
We can add any number of waves like this. For example, we will add:
"y = 2 sin 360t"
"y = 2 sin (360 * 2.5t)"
"y = 2 sin (360 * 3t)"

The first two waves look like this:

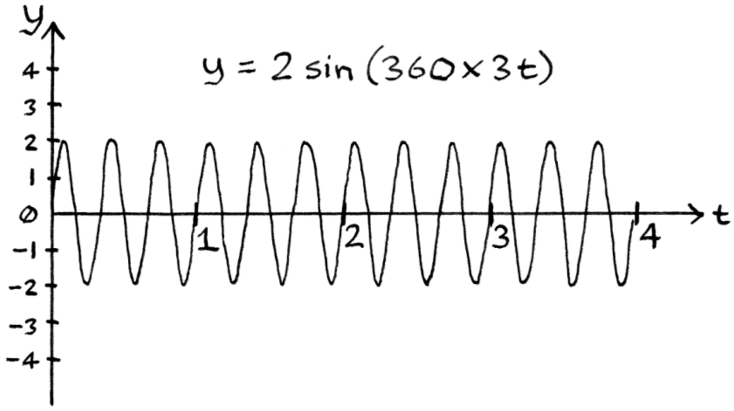




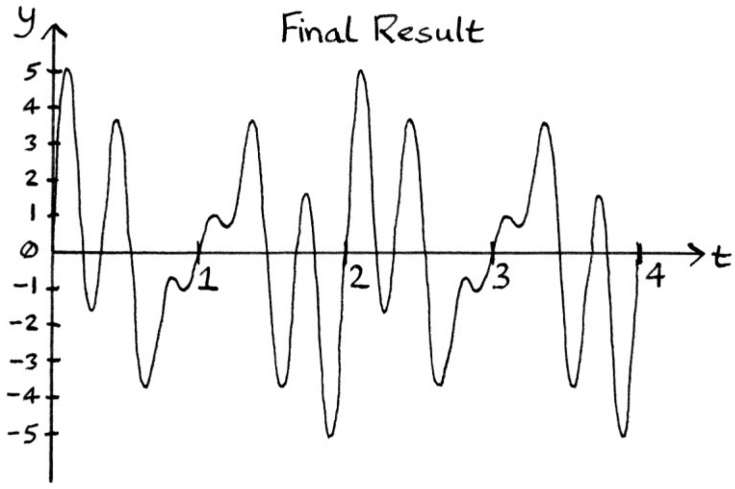
Adding these two waves together, we produce this:



The third wave is this:



... and adding it to the result so far, we end up with this:



The frequencies of the three individual waves are integer multiples of the frequency of the resulting signal, as explained earlier. The frequencies of 1 cycle per second, 2.5 cycles per second, and 3 cycles per second are all integer multiples of 0.5 cycles per second, so the final summed signal has a frequency of 0.5 cycles per second. Its period is 2 seconds.

Harmonic frequencies

The term “harmonic”, in the context of waves, refers to a series of waves that have frequencies that are all integer multiples of one particular “main” or “fundamental” frequency.

As an example, these frequencies are a harmonic series: 1, 2, 3, 4, 5, 6, and 7 cycles per second. They are all integer multiples of 1. The frequency of 1 cycle per second is the main frequency or “fundamental” frequency. If these waves were added together, the resulting signal would have a frequency of 1 cycle per second and a period of 1 second. If we added more waves that had frequencies that were also integer multiples of 1, then the resulting signal would still have the same frequency and period.

This series is also a harmonic series: 3, 6, 9, 12, 15, 18, 21, 24 cycles per second. The frequencies are all integer multiples of 3. The frequency 3 cycles per second is the fundamental frequency. If we added these waves together, the resulting signal would have a frequency of 3 cycles per second, and a period of 0.3333 seconds. If we added more waves that had frequencies that were also integer multiples of 3, the resulting signal would still have a frequency of 3 cycles per second and a period of 0.3333 seconds.

This series of numbers 6, 9, 18, 27, 33 and 660 cycles per second are a subset of the harmonic series 3, 6, 9, 12, 15..., but it is slightly harder to tell. If waves with the frequencies 6, 9, 18, 27, 33 and 660 were added together, the resulting signal would have a frequency of 3 cycles per second. Although 3 does not appear in the list, the numbers are all part of 3’s harmonic series. The numbers 6, 9, 18, 27, 33 and 660 are all integer multiples of 3.

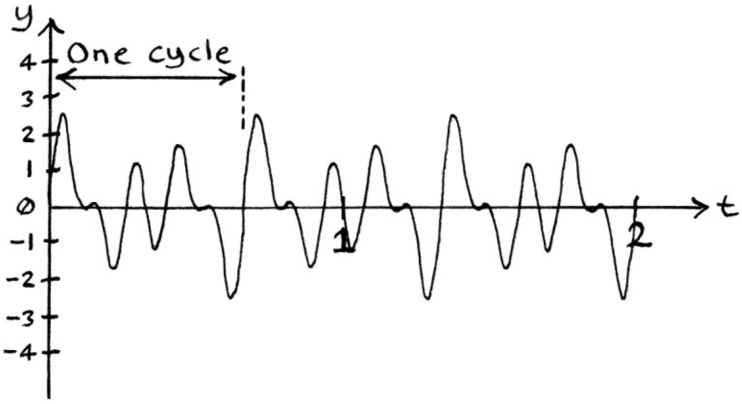
We already knew how to calculate the resulting frequency of a signal made from adding two waves if one of them had a frequency that was an integer multiple of the other. In the last section, I showed how to calculate the resulting frequency of the sum of three or more waves, in which the same rule applies. Thinking of frequencies as being a harmonic series is another way of understanding the rules for adding frequencies.

This next series is also part of a harmonic series: 14, 28, 56, 63, and 70 cycles per second. These are all integer multiples of 7 cycles per second, but the frequency of 7 cycles per second is not in the list. If we added waves with these frequencies together, the resulting signal would have a frequency of 7 cycles per second, and a period of 1 ÷ 7 = 0.1429 seconds. We could add any other waves with frequencies that were an integer multiple of 7, or even 7 itself, and the resulting signal's frequency would remain the same. The full harmonic series from which 14, 28, 56, 63 and 70 come is 7, 14, 21, 28, 35, 42, 49, 56, 63, 70... and so on.

Although in a previous section I gave a method of calculating the frequency and period of a signal created from adding waves of different frequencies by thinking of ratios, the same thing can be achieved (less easily) by finding the frequency for which all the added frequencies are integer multiples. As an example, if we had waves with frequencies of 1.26, 1.89, 2.1 and 2.31 cycles per second, with some thought, we could work out that they were all integer multiples of 0.21 cycles per second. Therefore, the signal resulting from adding them together would have a frequency of 0.21 cycles per second and a period of 1 ÷ 0.21 = 4.7619 seconds. Finding the frequency for which every wave in the list is a multiple is really another way of finding the ratio for a pair of waves.

Knowing all of the above means that if we are presented with a repeating signal that is the sum of two or more waves, we can be sure that all the individual added waves had frequencies that were integer multiples of the frequency of the signal.

As an example, the following signal has a frequency of 1.5 cycles per second:



... and therefore, we know that the waves that were added together to make it would all have had frequencies that were integer multiples of 1.5 cycles per second. The frequencies of the added waves, therefore, could only have come from the following set: 1.5, 3, 4.5, 6, 7.5, 9, 10.5, 12, 13.5, 15, 16.5 and so on.

Using the knowledge we have so far, we cannot tell exactly which frequencies from that set would have been added together to make that signal, but we know that there could not have been any frequencies from outside of that set. If there had been any frequencies from outside of that set, then the resulting signal would not have had a frequency of 1.5 cycles per second. This idea is very important in the analysis of periodic signals.

Observations about different frequencies

Pure waves

The first observation we should make from this chapter is that adding two or more waves of the *same* frequency, no matter what the amplitude, phase or mean level, will result in a pure wave.

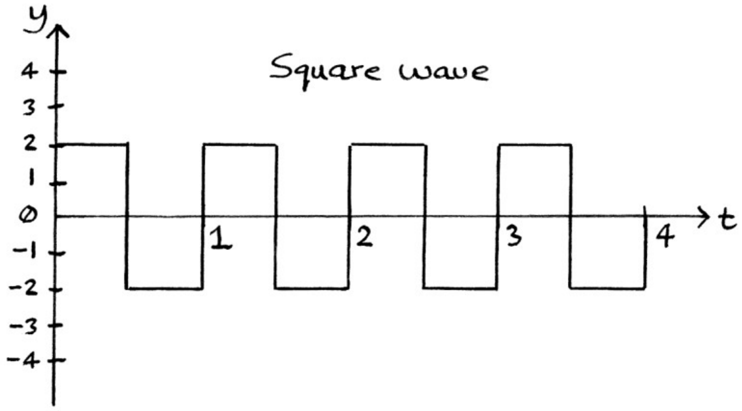
The second observation we should make is that the addition of two or more waves with *different*, non-zero, frequencies results in a signal that is not a pure wave. [If one of the frequencies is zero, the result will be a pure wave, and the zero-frequency wave will act as a mean level.]

A related rule to the second observation that is, in some ways, the reverse of that is:

> "Any periodic signal that is not a pure wave is really the sum of two or more pure waves of different frequencies, and various phases, amplitudes and mean levels".

This rule is not a logical deduction from the second observation – in fact, this rule is not even always true. However, it is *true enough* for it to be useful when dealing with waves. The rule is sufficiently valid to the extent that it is sensible to consider any periodic signal that is not a pure wave as being the sum of two or more pure waves of different frequencies added together.

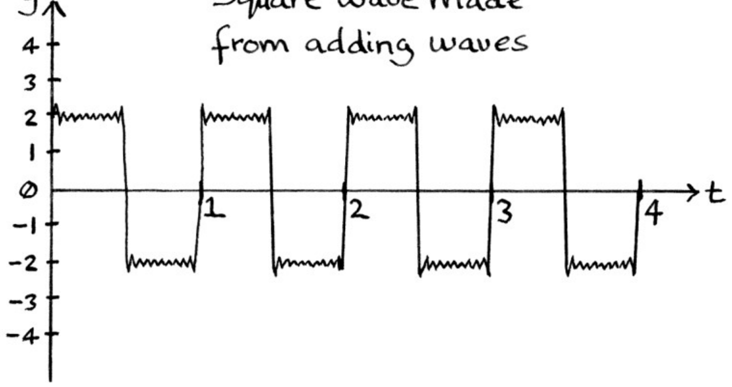
There are exceptions to the rule – the most common being periodic signals that jump instantly upwards or downwards – in other words signals with vertical lines in them such as "square waves":



[Square waves are often used to indicate the switching of an entity between two states – for example, in digital electronics circuits. If I were being pedantic, I would call a square wave, a square *signal*.]

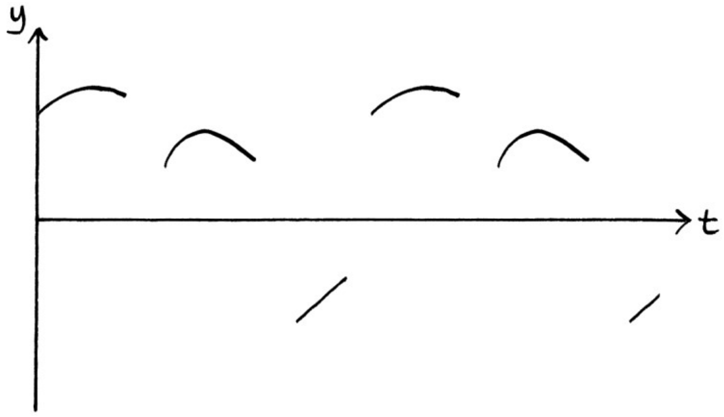
For square waves, it is possible to say that such signals are *approximately* the sum of two or more pure waves added together. We can recreate approximations of these signals by adding pure waves of different frequencies together. For much of dealing with radio or sound waves, being able to approximate such waves is sufficient.

For signals with vertical jumps, adding waves can bring us close to the original wave, but no matter how many waves we add, the result will always be slightly inaccurate. An example of a square wave recreated by adding pure waves is as follows:

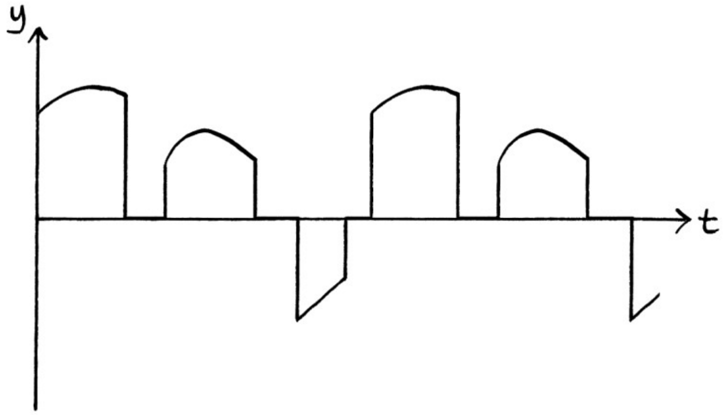


The above square wave is close to being a proper square wave, but it is not exact. It is impossible to make an accurate square wave by adding pure waves, but it is possible to obtain a close approximation.

Some signals that have apparent gaps in their curves are actually signals that jump instantly upwards or downwards, but to y = 0. For the duration of the gaps, the curve is really at y = 0. For example, this signal:



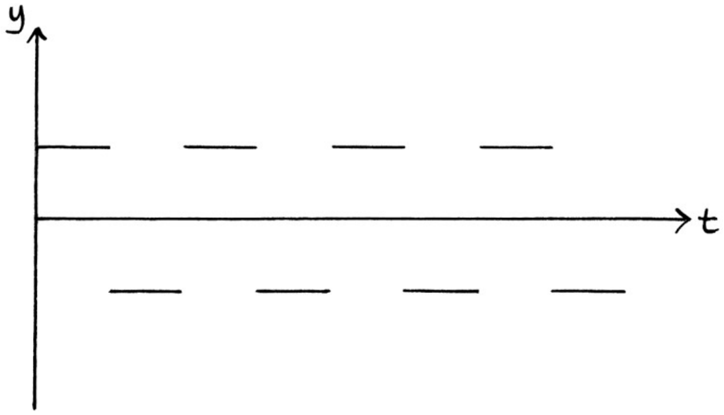
... can also be drawn as in the following picture to emphasise how it jumps instantly up and down to and from y = 0:



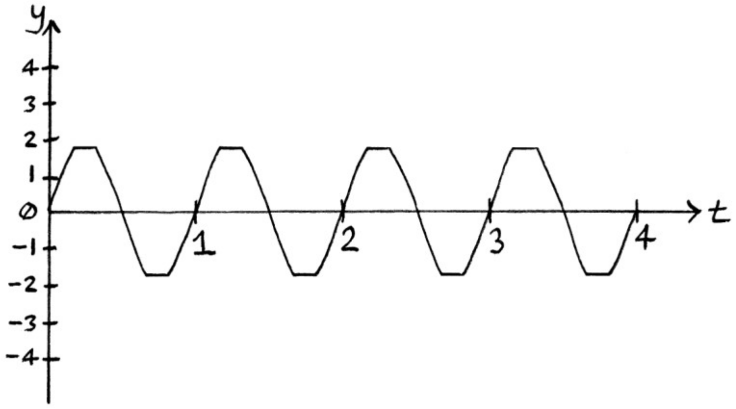
It would not be possible to make an accurate portrayal of the above signal by adding waves together, but it would be possible to create a good approximation.

[Note that there are other signals that have gaps where the y-axis values are *undefined* in that gap, as opposed to being zero. We will ignore such signals for now to keep things simpler].

A square wave could also be drawn without its vertical lines to emphasise its similarities to a signal with gaps in it:



There is an important implication of the rule about any periodic signal that is not a pure wave being equivalent to the sum of two or more pure waves of different frequencies. If we alter a pure wave in any way that stops it being a pure wave, it will instantly become the sum of two or more pure waves of differing frequencies. This idea is very important in signal processing. An example of this idea is that if we take a Sine wave and limit the maximum levels so that the peaks are squashed, the resulting signal will no longer be a pure wave, and therefore, it must instead be the sum of two or more pure waves with different frequencies.



Given all of the above, we can make a more accurate version of the rule:

> "Periodic signals that are not pure waves are the sum, or approximately the sum, of two or more pure waves of different frequencies, and various phases, amplitudes and mean levels."

An even better rule might be:

> "*Treat* all signals that are not pure waves as being the sum of two or more pure waves of different frequencies, and various phases, amplitudes and mean levels."

Frequencies do not mix

The fact that adding pure waves of different frequencies produces a result that is not a pure wave means that in any summed result, it is possible to tell that there were different frequencies in the constituent waves. Conversely, we cannot tell from a summed wave created from waves of the *same* frequency whether it is actually the sum of several waves or just a single wave.

Adding waves with the *same* frequency (and any amplitude, phase and mean level) is analogous to pouring different containers of water into a bucket – we end up with a bucket of water, and looking at the bucket of water will not give us any clues as to how many different containers were used or how big each of those containers were. Adding waves of *different* frequencies is analogous to pouring containers of different unmixable liquids into a bucket. We still might not be able to tell how many containers were used, but we can tell that there are different liquids and the amount of the different liquids in the bucket.

This idea of frequencies still being distinguishable after being added is an important one to remember. It is the reason humans and animals can distinguish between different frequencies of light and sound, and the idea is a huge part of analysing waves and signals.

Computer programming

If you know how to do basic computer programming with arrays, then it is not particularly difficult to write a program to perform addition (or other maths) with waves.

A wave can be put into an array by storing its y-axis values at evenly spaced moments in time. The whole array can be thought of as holding so many seconds of "wave". If you decide to store a wave by noting its y-axis values at every 0.1 seconds, then you would only need 10 values to store one second's worth of a wave. If a wave had a frequency of 1 cycle per second, then one second's worth would be enough to hold one complete cycle.

The smaller the time interval between the y-axis values you store in an array, the more accurate the result, but at the expense of needing a larger array. The more cycles you can store, the easier it will be to add waves of different frequencies. For a basic program, having a thousand y-axis entries for each second, and having an array that can hold 8 seconds of "wave" will allow you to achieve a great deal. This means that there would be 1000 * 8 = 8,000 array entries for one wave. One thing to consider is that the higher the frequencies you want to deal with, the more y-axis entries you will need per second – if there are too few, then the arrays will not be able to represent the waves correctly. [I will explain more about this in Chapter 42].

To fill an array with a wave in C, we could use code similar to this:

```
for(x = 0; x < (samplerate * 8); x++)
{
   mywavearray[x] = meanlevel + (amplitude *
         sin ((2*pi*frequency * (x / samplerate)) + phaseinradians))
}
```

... where:

- "x" is used to count through all the array entries.
- "samplerate" is the number of y-axis values from the wave that we want to store per second. To keep things straightforward, this is best being considerably higher than the frequency of the wave that we will be using. Faster frequencies require more y-axis values per second than slower frequencies. As an arbitrary choice, we can set this to 1000 values per second, while making sure that we never try to store a wave with a frequency faster than 50 cycles per second. [We could set it lower, but it makes everything easier if it is much higher than we need.] If the value is

too high, the program will run slowly; if the value is too low, we will not have enough values to portray the wave correctly. This variable must be the same for all the waves and signals in our program, or else the corresponding (by time) y-axis values will not line up correctly when we perform addition or multiplication.

- "8" is there because we want to store 8 seconds' worth of wave.
- "mywavearray" is the name of the array into which we are storing the wave.
- "meanlevel" is the mean level of the wave.
- "amplitude" is the amplitude of the wave.
- "sin" is the Sine command in C, although depending on the type of numbers you are using, you might need "sinl" instead. The Sine part is on a new line for the sole reason that it makes it easier for us to read it. C treats the whole command in the same way, whether it is on one line or not.
- "pi" is a variable that contains the number π. C requires you to create this variable yourself, so you will need to declare it earlier with something such as:
  ```
  double pi = 3.14159265358979323846 26;
  ```
- "2 * pi" is used because the C programming language works with radians and not degrees, so instead of using formulas such as "y = 2 sin (360 * 4t)", we need to use formulas such as "y = 2 sin (2π * 4t)".
- "frequency" is the frequency in cycles per second.
- The multiplication by "x / samplerate" is there because we want to calculate y-axis values spaced at so many values per second. If it were not for the division by "samplerate", "x" would be having the same effect as the time in seconds. In that case, we would be storing only one y-axis value from the wave every second, which would be far too little. The division by "samplerate" means that we are storing the y-axis values for consecutive fractions of seconds.
- "phaseinradians" is the phase of the wave, but not in degrees but in radians. It is in radians because the C programming language works with radians. To convert an angle from degrees into radians, we divide it by 360 to find out the portion of a circle that the degrees represent, and then multiply that by 2π.

Once we have filled an array with a wave, we can perform basic maths with waves. For example, to perform the addition of two waves, we just add each entry of the array for one wave to the corresponding entry of the array for the other wave.

Technically, when storing a wave as separate y-axis values, we are using what are called "discrete waves". I will explain more about these from Chapter 39 onwards.

If you are reasonably new to programming, the trickiest part of writing a program will be displaying the waves held in the arrays as pictures. Depending on what it is you want to achieve, you can still program with waves without needing pictures, but it makes things clearer if you can see what you are doing. Displaying the arrays as pictures is reasonably easy in programming languages such as Python. If you know how to work with files and binary, one way to make basic pictures in C is to create black and white bitmapped TIFF files, the specification of which can be found on the internet. When creating pictures, the y-axis values from the wave arrays will need to be plotted as points on the picture. If you have a high enough number of points per second (in other words, a high enough sample rate), you will not need to join the points up.

Writing a program to add waves is a very good way to develop your programming skills, and it is also a good way to learn more about waves.

Spreadsheet programs

If you do not know how to program or you do not know how to program arrays, you can perform maths with waves by using a spreadsheet program such as Microsoft Excel or one of the free alternatives. As a basic example for Microsoft Excel:

- First, make sure that the cells you are going to use are set to be numbers with at least 4 decimal points. The quickest way to do this is to right click on the top left corner of the spreadsheet border between "1" and "A", and select "Format Cells...". Then under "Category" choose "Number" and set the decimal places to 4. Then Press "OK". This will turn all the cells in the spreadsheet into numbers.

- Starting in cell "F3", make a column consisting of the numbers:
  0.00
  0.01
  0.02
  0.03
  ... and so on until:
  0.97
  0.98
  0.99
  1.00

These numbers will represent the time over one second, split into intervals of 0.01 seconds. A quick way to enter these values is to type in 0.00 and 0.01, select those cells, hold the mouse cursor over the bottom right hand corner of the selected cells so that it turns into a thinner, black cross, and then drag that cross directly downwards. Excel will fill in the rest of the sequence for as long as you drag the mouse downwards.

- In the next column, in the cell "G3" (to the right of the entry for 0.00), type:

  ```
  =SIN(2 * PI() * F3)
  ```

  ... *and press enter*. The meanings of the parts of this entry are as follows:

  The equals sign tells Excel to treat this as a calculation.

  `SIN(...)` means the Sine of everything within the brackets.

  `PI()` is Excel's way of giving the number π.

  `2 * PI()` is there because Excel works with radians and not degrees, so our formulas have 2π in them and not 360.

  `F3` (the cell containing the entry for 0.00) is acting as the time at t = 0.

- Single left click on cell "G3", and copy it to the clipboard. (It should start to have a moving dotted border). Then select the cells in that column from cell "G4" down to cell "G103", and paste into those cells. Excel will fill the cells with the equivalent formula for each row. After doing that, those cells will contain the y-axis values for the Sine wave with the formula "y = sin 2πt", which are the same as those for the degrees-based Sine wave with the formula "y = sin 360t".

To draw a Sine wave graph, select the cells from "G3" to "G103", click on the "Insert" menu at the top of the screen, click on the "Line Chart" button, and then click on the simplest of the "Line Chart" choices. This will draw a graph for you. [Excel's graphs can often look a bit clumsy, but they are still good enough for the basic viewing of waves.] You can then move the graph to where you want it to be on the spreadsheet.

Note that your version of Microsoft Excel might have slightly different menus and options, but the general idea will be the same. Other spreadsheet programs tend to work in a similar way to Microsoft Excel.

As a more advanced example, supposing we wanted the wave for:
“y = 2 + 3 sin ((360 * 4t) + 30)”
... we would first convert the formula to be in radians as so:
“y = 2 + 3 sin ((2π * 4t) + (30 * 2π / 360))”

Our first y-axis cell (“G3”) would look like this:

```
=2 + (3 * SIN((2*PI() * 4 * F3) + (30 * 2*PI() / 360)))
```

After pressing enter, we select the cell, and copy and paste it into the cells below, from cell “G4” to cell “G103”. We can then view a graph of it as before.

We could have one wave in the column starting in “G3” and one wave in the column starting in “H3”. We could then add the two waves by having the following in cell “I3”:

```
=G3 + H3
```

This tells Excel to add the contents of cell “G3” and cell “H3”. After pressing enter, we would select the cell, then copy and paste it into the cells below, from cell “I4” to cell “I103”. We could then draw a graph of the resulting signal.

If we wanted to multiply the two waves, we would start with “I3” containing:

```
=G3 * H3
```

... and continue as normal.

You can learn a lot about waves by using Excel, and you can also learn a lot about Excel by using it to create waves. In this Excel explanation, a cell in a column in Excel is essentially the same as an entry in an array in computer programming.

Conclusion

Understanding the general ideas behind adding waves is very important. Depending on what you want to achieve, you do not necessarily need to remember particular formulas, but having a basic knowledge of addition will be very useful. In this chapter, I have not explained everything about the addition of waves. There are several formulas for addition that are commonly used in school maths books, but I have intentionally left them out because they do not particularly help in understanding waves.

Chapter 14: Addition with circles

When performing mathematical operations on waves, thinking of the circles from which the waves are derived makes the process much more intuitive.

A static picture of a circle contains the information for an angled-based Sine wave and its corresponding angle-based Cosine wave. The path taken by an object rotating around a circle contains the information for a time-based Sine wave and its corresponding time-based Cosine wave. The path taken by an object rotating around a circle will, itself, be a circle.

In this chapter, and in this book, I will call the path taken by a rotating object over time “a circle” so as to have a succinct way of referring to it. Such an idea could also be called “the object’s circle path over time”, a “circle path” or a “time circle”. Using the single word “circle” makes explanations much easier to read. Any possible confusion with a “circle” meaning a static shape will be avoided by the context. By using the term “circle” in this way, we can say such things as, “A circle contains the information for both a time-based Sine wave and a time-based Cosine wave.”

In this chapter, we will look at “adding” circles to each other, in the sense that we will look at “adding” the circular paths of objects rotating around static circles. This idea needs explaining. In the sense that I am using it here, “adding” two circles involves adding the corresponding (by time) x-axis coordinates of the objects rotating around each circle to each other, and adding the corresponding (by time) y-axis coordinates of the objects rotating around each circle to each other. The result will be a new shape, which may or may not also be a circle.

There is another way of thinking about the process. As a circle can be thought of as a series of coordinates taken from a time-based Cosine wave and a time-based Sine wave, when we “add” two circles, we are really adding the Cosine wave for one circle to the Cosine wave for the other circle, and adding the Sine wave for one circle to the Sine wave for the other circle. We are then using the resulting signals as coordinates for the resulting shape. By adding the underlying waves, we are, in essence, adding the circles.

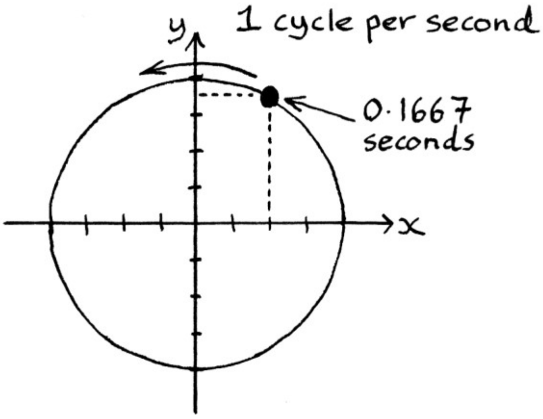
To summarise these two ideas: we can achieve the result of adding two circles either by adding coordinates or by adding the pairs of derived waves and treating the resulting signals as coordinates.

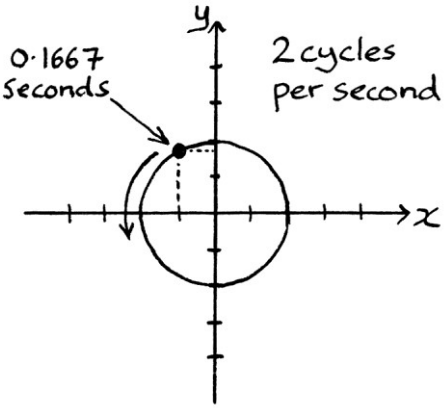
There is another way to add circles that makes certain aspects of the process easier to visualise. In a drawing, we arrange the two circles together, with the outer circle centred on the object rotating around the inner circle. We then imagine an object moving around the outer circle, which, in turn moves around the inner circle. The outer object would draw out a shape that is the sum of the two circles. The position of the outer object at any one time will be the coordinates taken from the sums of the original circles' derived waves.

All of this will become clearer as this chapter progresses.

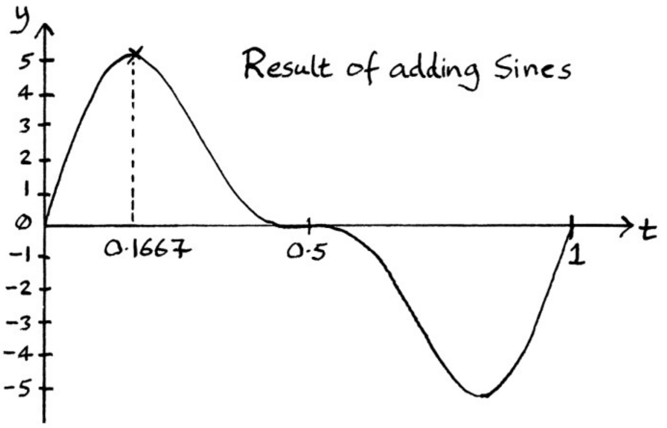
Addition

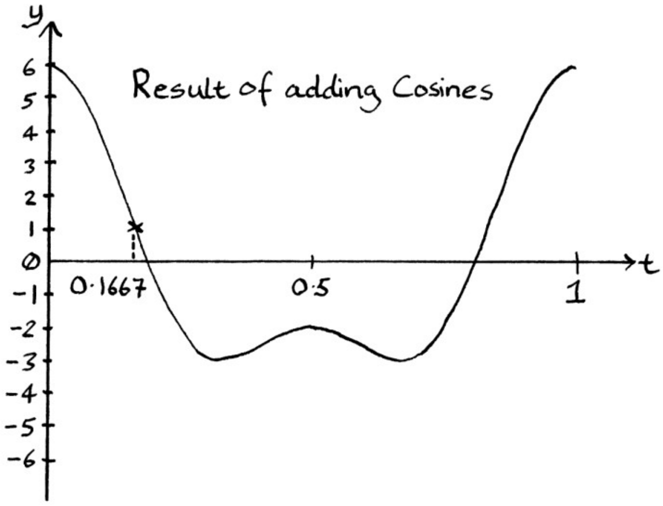
When we "add" two circles, we are really adding together the y-axis coordinates of both objects rotating around each circle at every moment in *time*, as well as adding together the x-axis coordinates of both objects rotating around each circle at every moment in *time*.



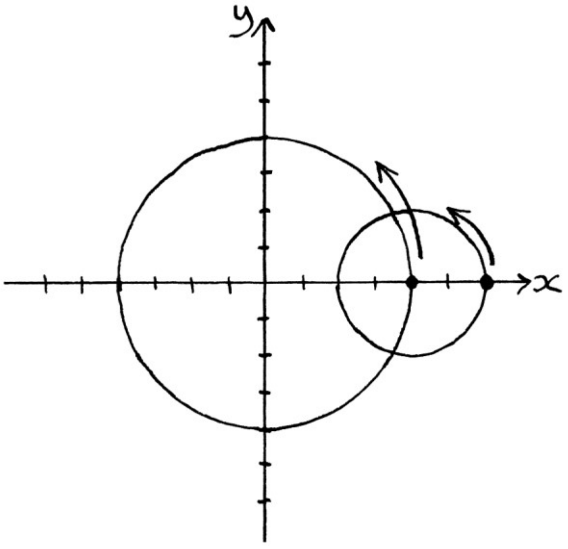


If we graph the sum of the y-axis coordinates over time, we will have a graph that shows the sum of the two Sine wave graphs derived from the original circles. If we graph the sum of the x-axis coordinates over time, we will have a graph that shows the sum of the two Cosine wave graphs derived from the original circles.

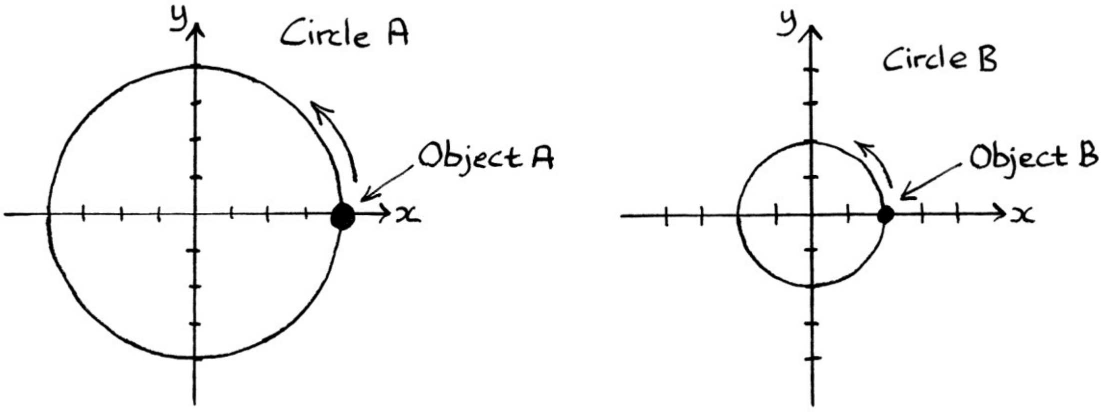




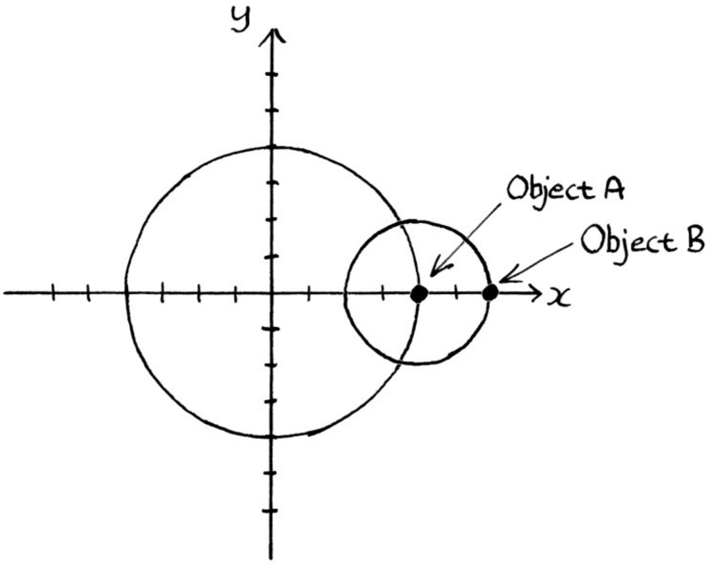
The concept of adding the points from two circles is made much simpler when we combine the circles on the circle chart, so that the centre of one circle is fixed to the object going around the other circle. As the outer object rotates around its circle, its circle also rotates around the inner circle.



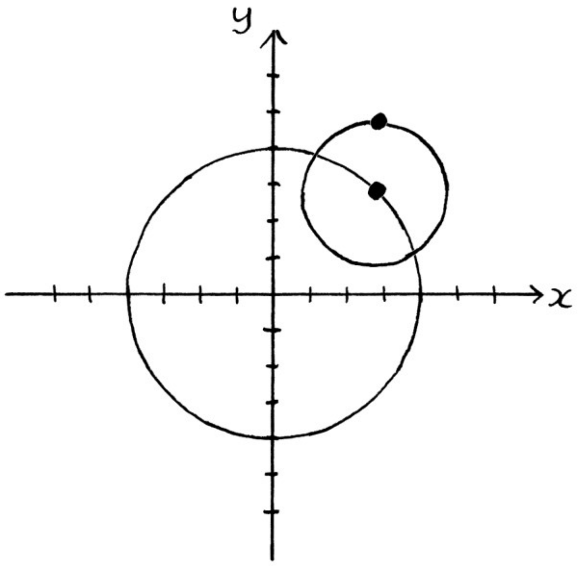
The y-axis and x-axis points of the outer object indicate the sum of the corresponding circles for that moment in time. This is easiest to visualise when we add two circles with significantly different radiuses. We will call the larger circle, Circle A, and the smaller circle, Circle B. We have one object rotating around the larger circle, which we will call Object A, and one object rotating around the smaller circle, which we will call Object B.

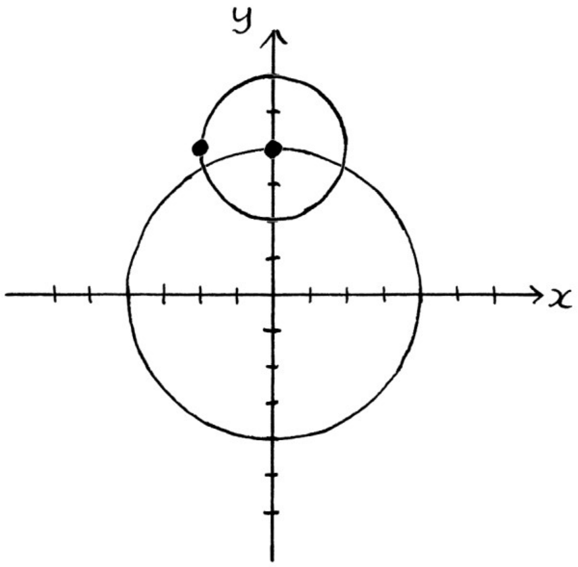


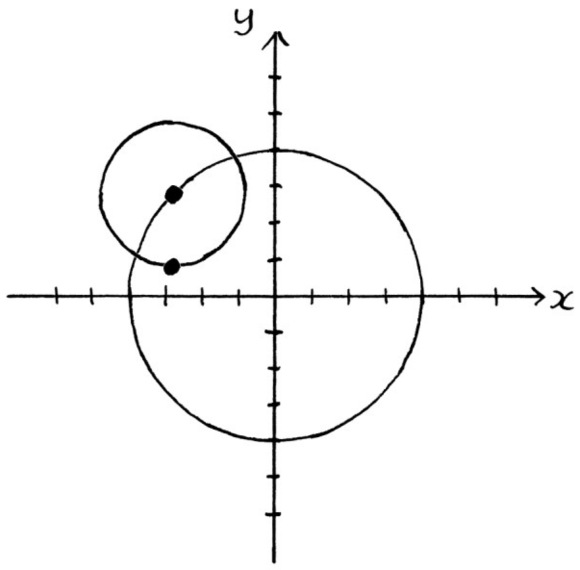
To see how the circles add together, we first centre Circle A on the origin of the axes. We then centre Circle B on *Object A*.

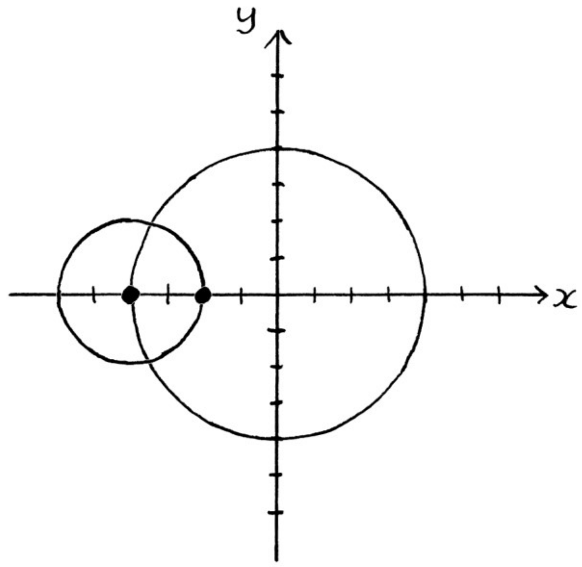


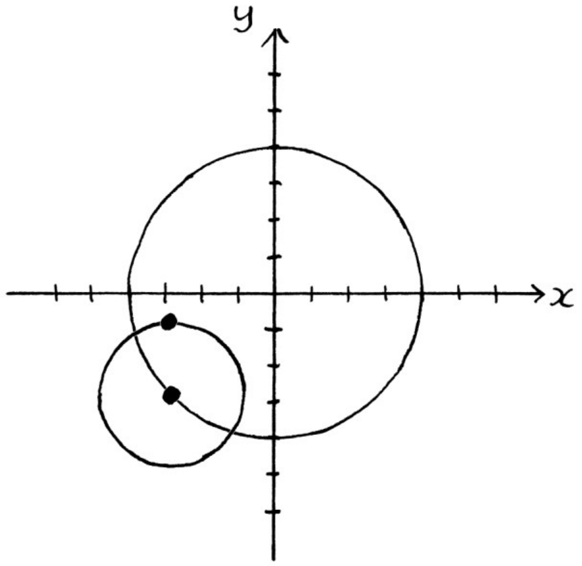
As Object A rotates around its circle, Object B rotates around *its* circle, which, because that circle is fixed on Object A, rotates around Object A's circle.

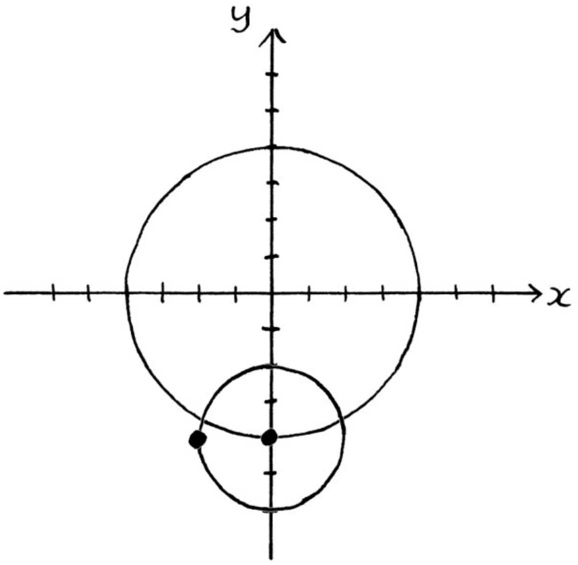


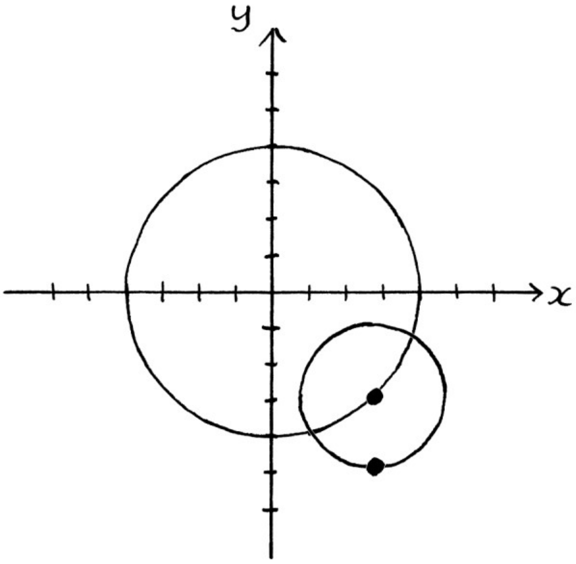


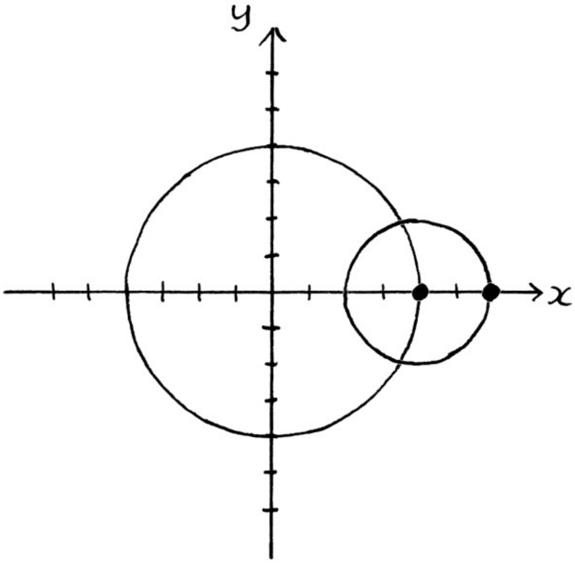



y
x


y
x

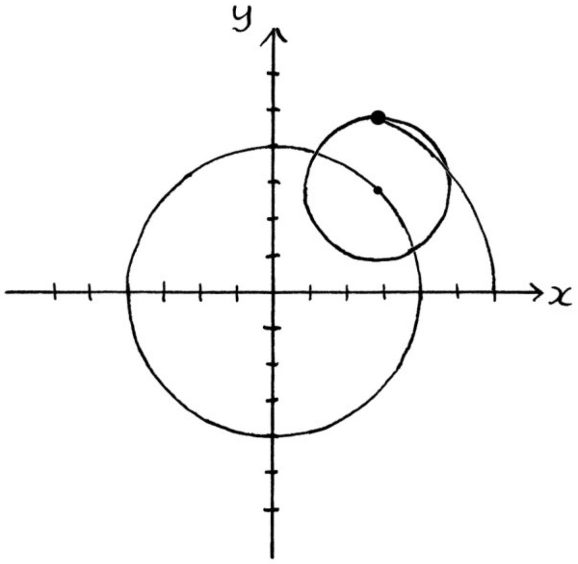

y
x

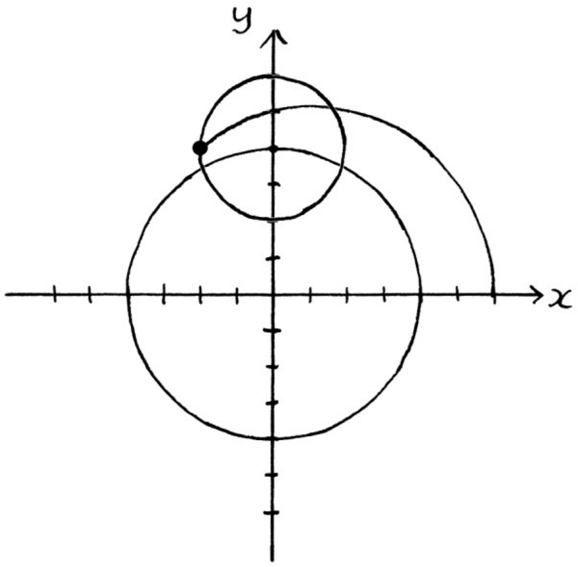


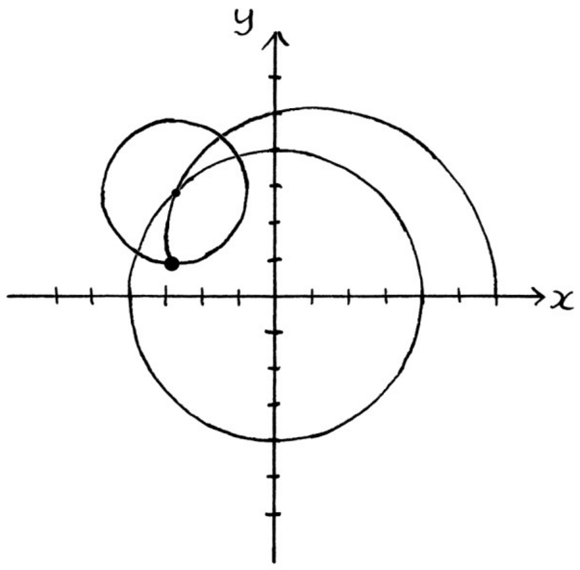


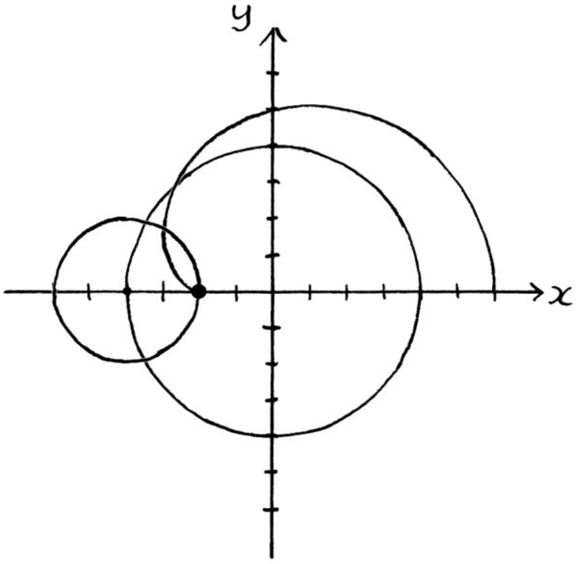
The movement of Object A around its own circle draws out a circle (obviously), and the movement of Object B around *its own* circle also draws out a circle (obviously again). However, the movement of Object B around *the centre of Object A's circle* may or may not draw out a circle. This all depends on the underlying frequencies of the two circles.

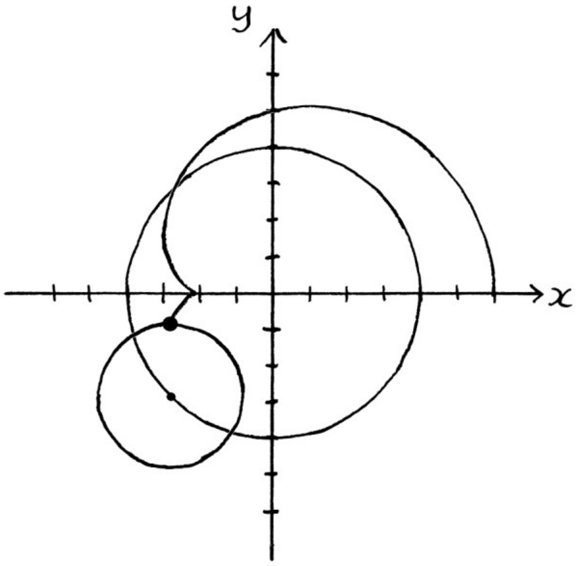
If we let Object B mark the paper as it rotates around its circle, and as that circle rotates around Circle A, we will see this:

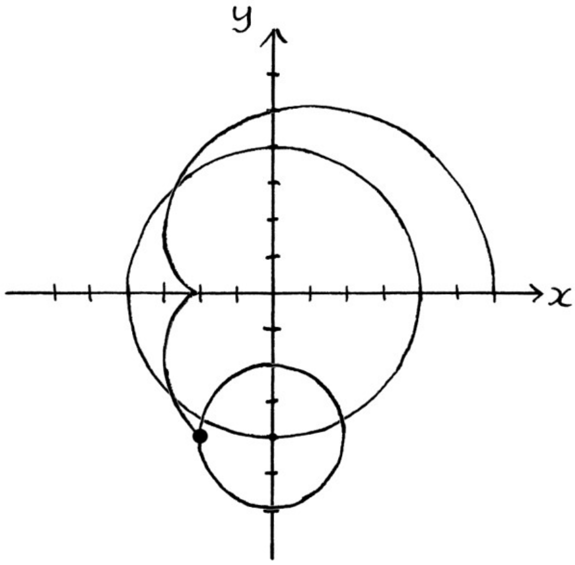


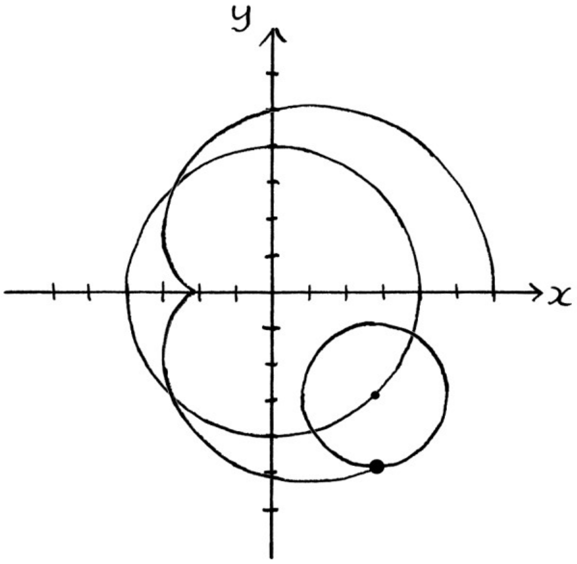


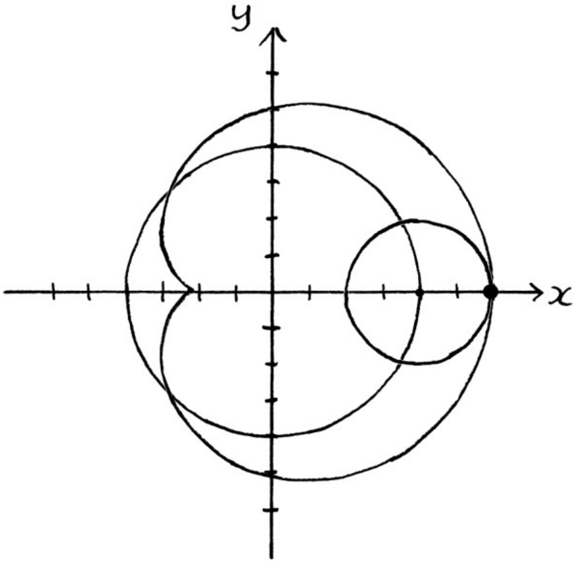



y
x

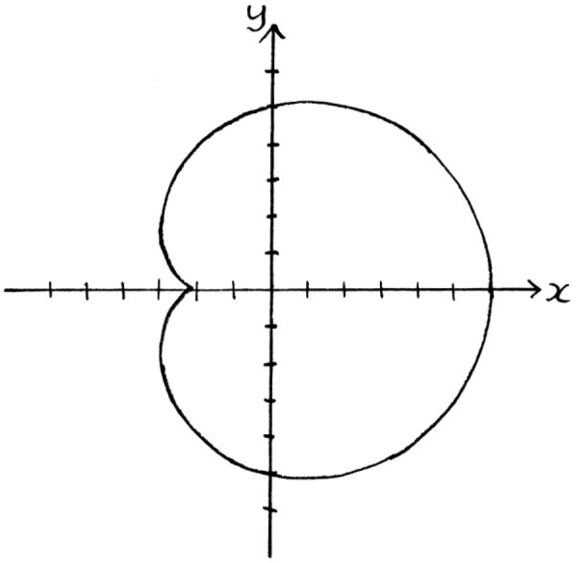

y
x


y
x

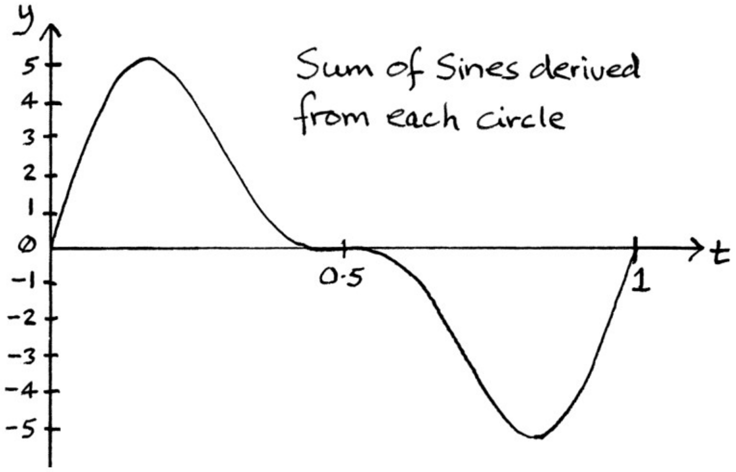

y
x


y
x

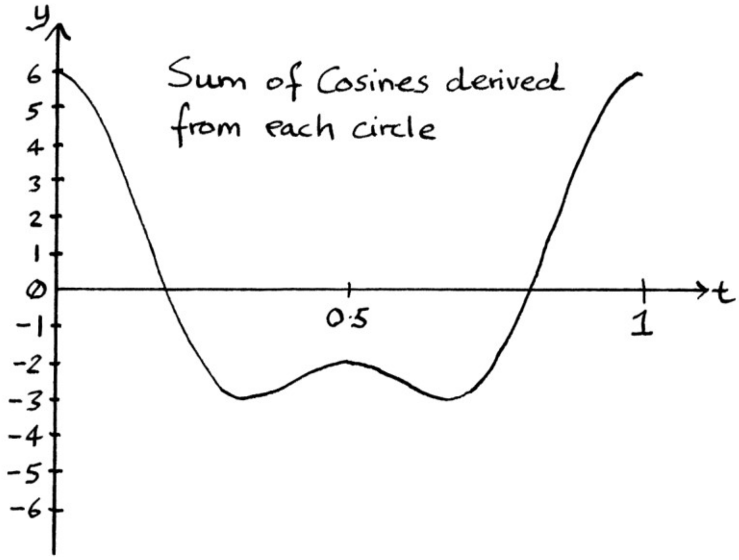
In this particular example, Object B's movement around the origin does not draw out a circle, but instead, what can, at best, be called a "shape".



The y-axis values of Object B, as it moves around the origin of the axes, as plotted over time, will produce a signal equal to the sum of the Sine waves derived from each individual circle:



The x-axis values will produce a signal equal to the sum of the Cosine waves derived from each individual circle:



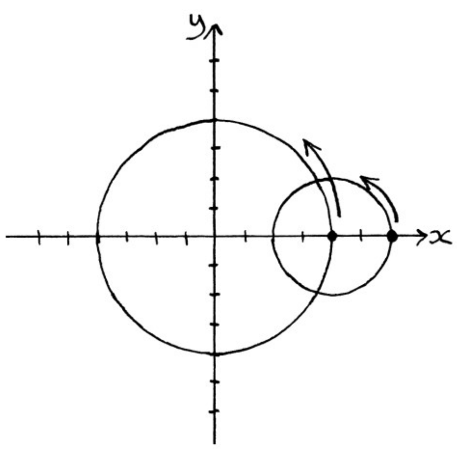
We can create a graph of the y-axis values of Object B over time either by observing its movement around the circle chart or by adding the Sine waves derived from the original circles. We can create a graph of the x-axis values of Object B over time either by observing its movement around the circle chart or by adding the Cosine waves derived from the original circles.

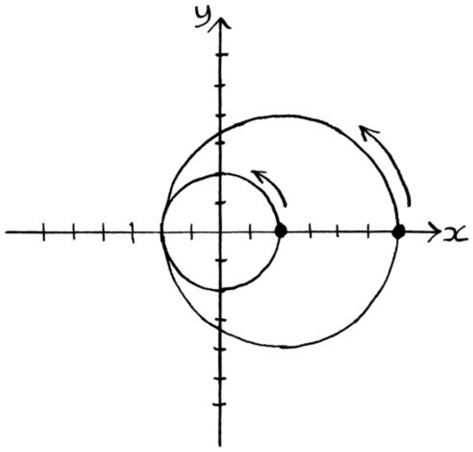
It is important to notice that when adding the circles, we read the y-axis and x-axis values from Object B's movement around the axes for particular *times* and not for particular *angles*. Object B's angle from the origin might be completely independent of the time that it is at that angle. This is always true if the circles have different frequencies, and this idea will become clearer as this chapter progresses.

Another thing to note is that if we are given a completed resulting "shape", it will not be possible to know the times when the object was at particular y-axis and x-axis values on that shape, and therefore, it will not be possible to construct the resulting signals. It is necessary to observe the shape being created as it happens, or to have more information than appears in just a picture of the finished shape. A static drawing of a shape does not indicate any timing.

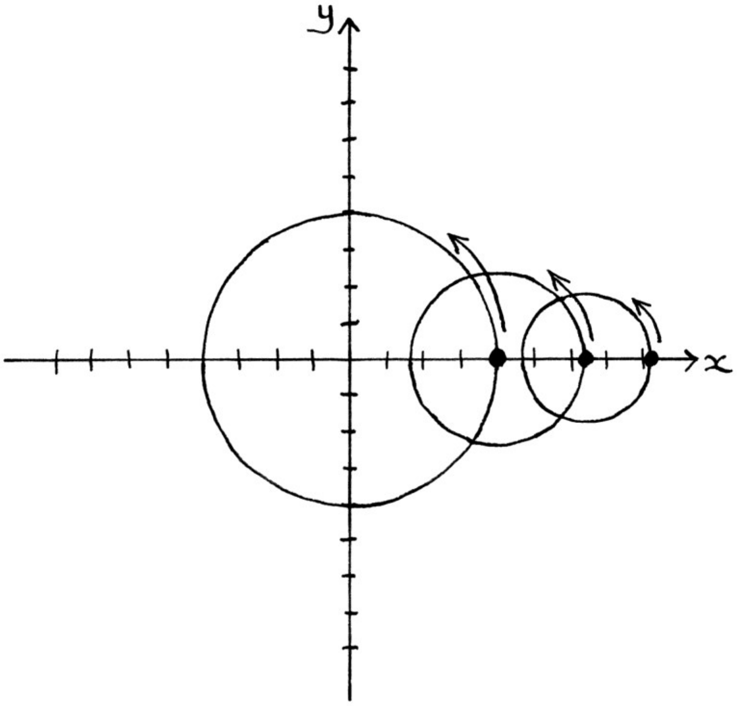
If we are given the two signals resulting from adding the original derived Sine waves and Cosine waves, we can use corresponding (by time) y-axis values from each as coordinates to reconstruct the resulting shape, in the same way that we would use corresponding y-axis values from a Sine wave and a Cosine wave to reconstruct the circle from which they were derived. The x-axis coordinates for the shape come from the sum of Cosine waves and the y-axis coordinates come from the sum of Sine waves.

The two circles can be placed either way around. We can have Circle B rotate around Circle A (in which case, we look at Object B's coordinates), or we can have Circle A rotate around Circle B (in which case, we look at Object A's coordinates). The results will be identical. The only difference between how the circles are arranged is that it is easier to see what is happening if the smaller circle rotates around the larger circle.





We can also have any number of circles rotating around each other, and in any order.



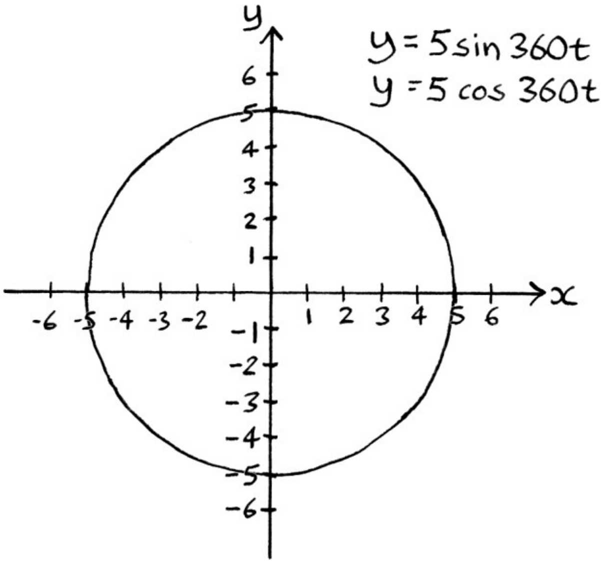
When adding circles in this way, we only pay attention to the movement of the outermost object.

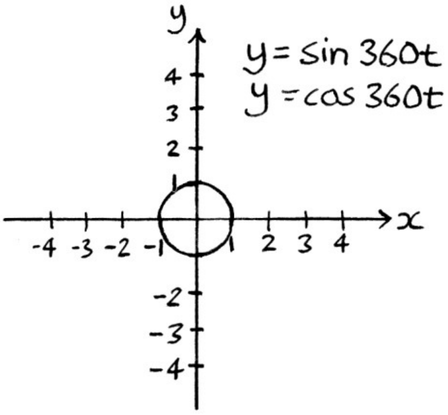
In the same way that we can calculate the Sine and Cosine of an angle by drawing and accurately measuring a circle by hand, we can also calculate the sum of two or more Sine waves, as well as the sum of their corresponding Cosine waves, by drawing by hand and accurately measuring two or more circles rotating around each other. The method involves drawing the circles at every moment in time as they move around each other, or more realistically, at *evenly spaced* moments in time. Doing this is time consuming because we might need to draw a large number of outer circles, depending on the accuracy we want, but the task is not particularly difficult.

Amplitude example

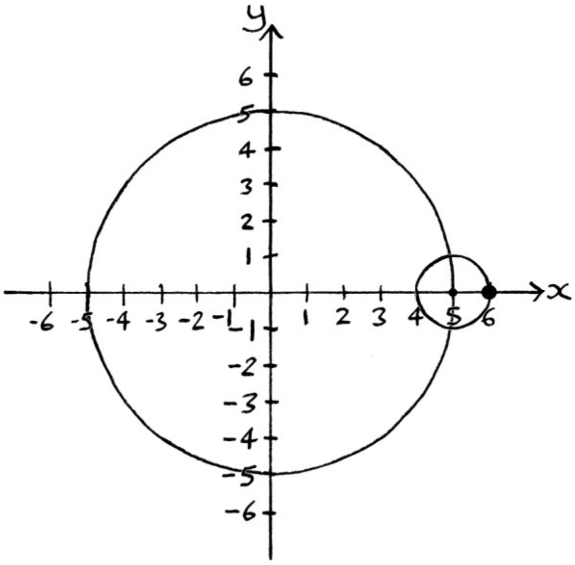
To demonstrate how the process of adding circles works when adding different amplitudes, we will add two circles that represent pairs of waves with the same frequency, phase and mean level, but with different amplitudes. The larger circle will represent the pair of waves:
“y = 5 sin 360t” and “y = 5 cos 360t”
... and the smaller circle will represent the pair of waves:
“y = sin 360t” and “y = cos 360t”.

The circles for these waves look like this:



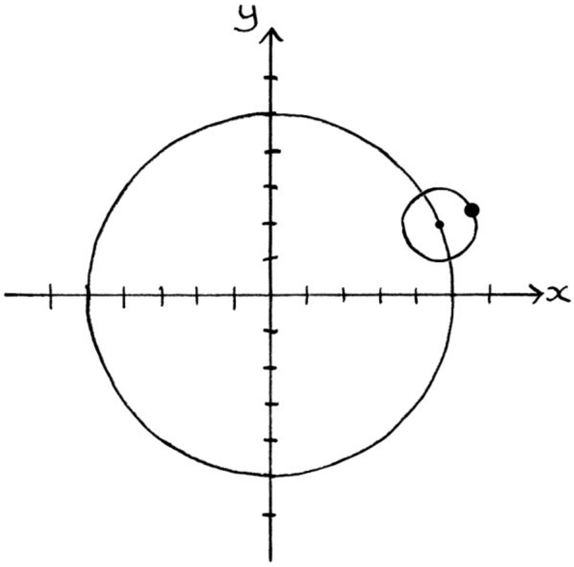


We centre the larger circle on the origin of the axes; we centre the smaller circle on the object rotating around the larger circle. At t = 0, the circles look like this:



As the object rotating around the larger circle rotates, the smaller circle moves with it, and the object rotating around the smaller circle rotates too.

At 0.0625 seconds, the object rotating around the larger circle will be at 22.5 degrees in relation to the origin of the axes. This means that the centre of the smaller circle will be at 22.5 degrees in relation to the origin of the axes, because it is centred on the object rotating around the larger circle. The object rotating around the smaller circle will be at 22.5 degrees in relation to the centre of the smaller circle. In this example, it just so happens that this is also 22.5 degrees in relation to the origin of the axes because the circles have the same frequency. [In the following pictures, I have not drawn numbers on the axes so that the pictures are clearer.]



At this moment in time, the y-axis value of the centre of the smaller circle (which is also the height of the object rotating around the larger circle) is 1.9134 units, which is:
“5 * sin (360 * 0.0625)” or “5 * sin 22.5”.

The vertical distance of the object rotating around the smaller circle *to the centre of the smaller circle* is 0.3827 units, which is:
“sin (360 * 0.0625)” or “sin 22.5”.

The vertical distance of the object rotating around the smaller circle *to the origin of the axes* (i.e. its y-axis value) is 2.2961, which is:
“5 * sin (360 * 0.0625) + sin (360 * 0.0625) = 6 sin (360 * 0.0625)”.

That y-axis value is equal to the result of adding the points from the Sine waves derived from the two circles at that particular moment in time.

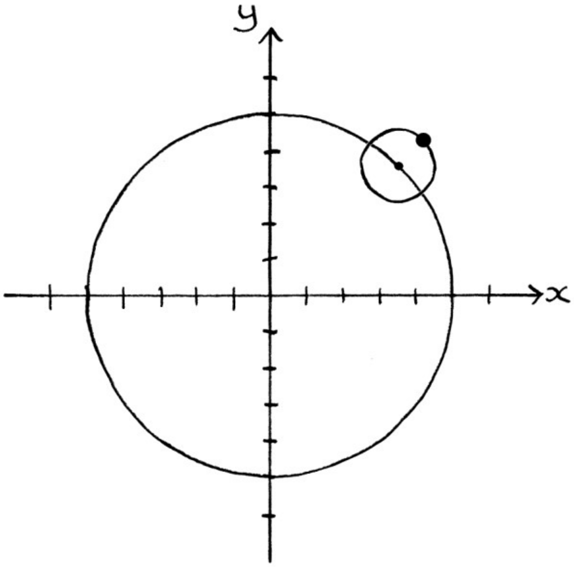
At this moment in time, the x-axis value of the centre of the smaller circle (which is also the x-axis value of the object rotating around the larger circle) is 4.6194 units, which is:
“5 * cos (360 * 0.0625)” or “5 * cos 22.5”.

The horizontal distance of the object rotating around the smaller circle *to the centre of the smaller circle* is 0.9239 units, which is “cos (360 * 0.0625)” or “cos 22.5”.

The horizontal distance of the object rotating around the smaller circle *to the origin of the axes* (i.e. its x-axis value) is 5.5433, which is:
5 * cos (360 * 0.0625) + cos (360 * 0.0625) = 6 cos (360 * 0.0625).

That x-axis value is equal to the result of adding the points from the Cosine waves derived from the two circles at that particular moment in time.

At 0.125 seconds, the centre of the smaller circle will be at 45 degrees in relation to the origin of the axes. The object rotating around the smaller circle will be at 45 degrees in relation to the centre of the smaller circle.



At this moment in time, the y-axis value of the centre of the smaller circle (which is also the height of the object rotating around the larger circle) is 3.5355 units, which is:
“5 * sin (360 * 0.125)” or “5 * sin 45”.

The vertical distance of the object rotating around the smaller circle to the centre of the smaller circle is 0.7071 units, which is:
“sin (360 * 0.125)” or “sin 45”.

The total y-axis value of the object moving around the smaller circle is 4.2426 units. This is also:
5 sin (360 * 0.125t) + sin (360 * 0.125) = 6 sin (360 * 0.125) = 6 sin 45.

In other words, the y-axis value of the outer object on the chart is equal to the sum of the formulas “y = 5 sin 360t” and “y = sin 360t” for that moment in time.

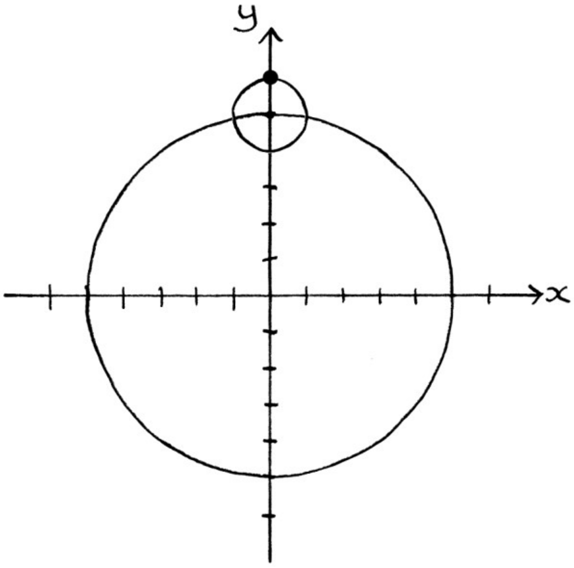
At that same moment in time, the x-axis value of the object on the larger circle is 3.5355 units, which is:
“5 cos (360 * 0.125)” or “5 cos 45”.

The horizontal distance from the object on the smaller circle to the centre of the smaller circle is 0.7071 units, which is:
“cos (360 * 0.125)” or “cos 45”.

The x-axis value of the object on the outer circle is 4.2426, which is:
5 cos (360 * 0.125) + cos (360 * 0.125) = 6 cos (360 * 0.125) = 6 cos 45.

In other words, the x-axis value of the outer object on the chart is equal to the sum of the formulas “y = 5 cos 360t” and “y = cos 360t” for that moment in time.

When the time is 0.25 seconds, the two circles will look like this:



The object on the larger circle will be at 90 degrees, and have a y-axis value of:
“5 sin (360 * 0.25)” = 5 units.

The object on the smaller circle will be at 90 degrees in relation to the centre of the smaller circle, and will have a vertical distance from the centre of the smaller circle of:
“sin (360 * 0.25)” = 1 unit.

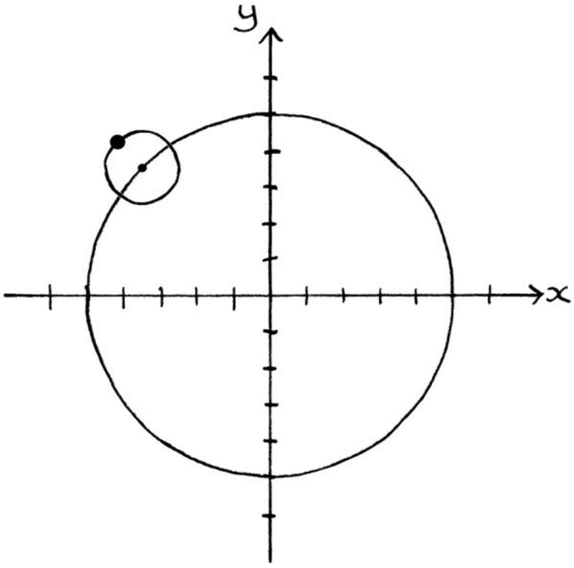
The y-axis value of the object on the smaller circle will be:
5 sin (360 * 0.25) + sin (360 * 0.25) = 6 sin (360 * 0.25) = 6 units.

The object on the larger circle will have an x-axis value of:
"5 cos (360 * 0.25)" = 0 units.

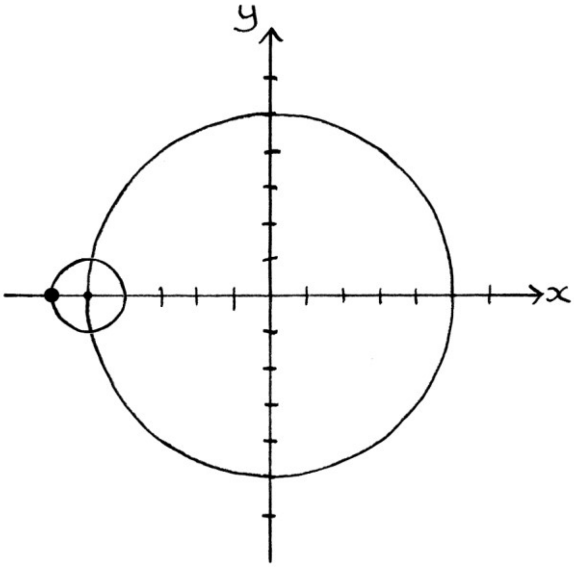
The object on the smaller circle will have a horizontal distance from the centre of the smaller circle of:
"cos (360 * 0.25)" = 0 units.

The x-axis value of the object on the smaller circle will be:
5 cos (360 * 0.25) + cos (360 * 0.25) = 0 units.

When the time is 0.375 seconds, the two circles will look like this:



When the time is 0.5 seconds, the two circles will look like this:



The y-axis value of the object rotating around the larger circle will be:
"5 sin (360 * 0.5)" = 0 units.

The vertical distance from the object rotating around the smaller circle to the centre of the smaller circle will be:
"sin (360 * 0.5)" = 0 units.

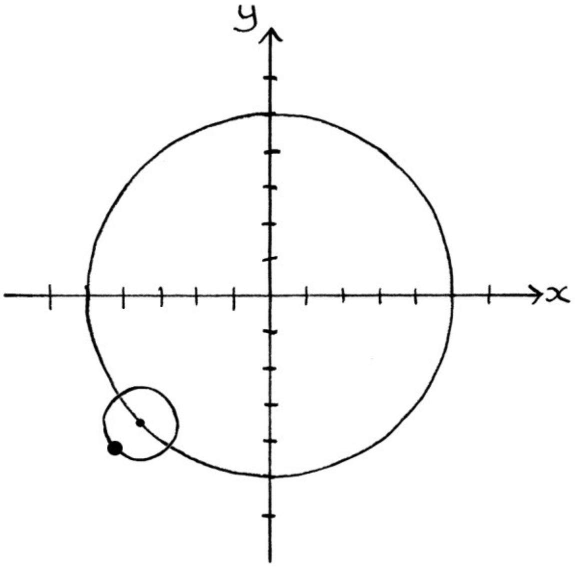
Therefore, the y-axis value of the object rotating around the smaller circle will be:
5 * sin (360 * 0.5) + sin (360 * 0.5) = 6 * sin (360 * 0.5) = 0 units

The x-axis value of the object rotating around the larger circle will be:
"5 cos (360 * 0.5)" = −5 units.

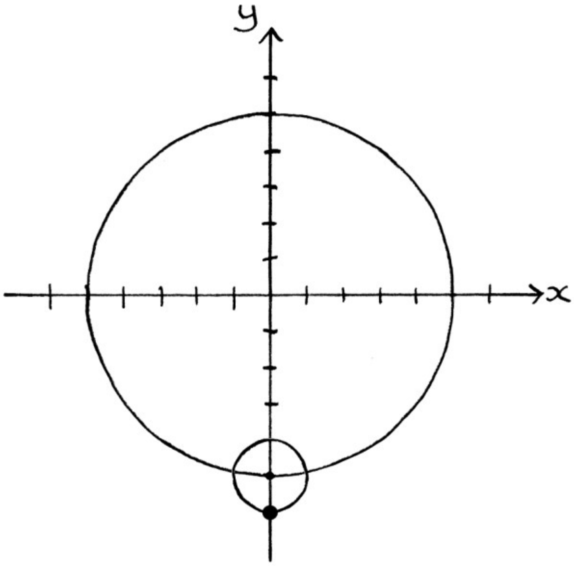
The horizontal distance from the object rotating around the smaller circle to the centre of the smaller circle will be:
"cos (360 * 0.5)" = −1 unit.

Therefore, the x-axis value of the object rotating around the smaller circle will be:
5 * cos (360 * 0.5) + cos (360 * 0.5) = 6 * cos (360 * 0.5), which is −6 units.

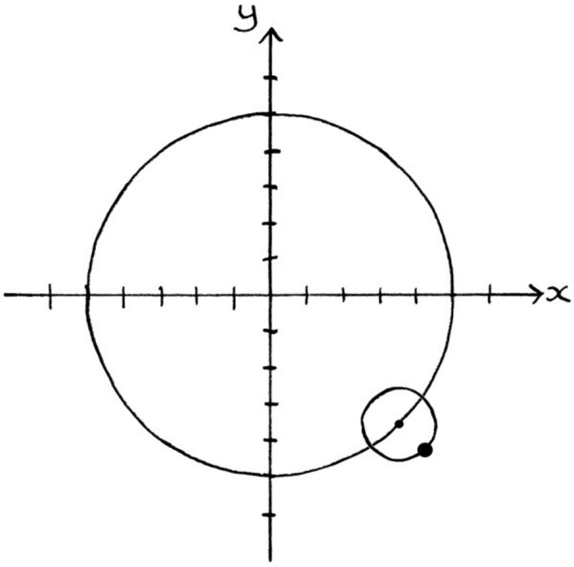
At 0.625 seconds:



At 0.75 seconds:

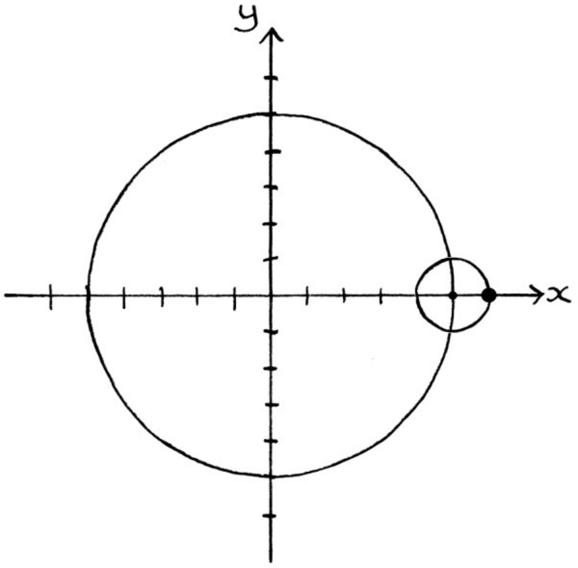


At 0.875 seconds:

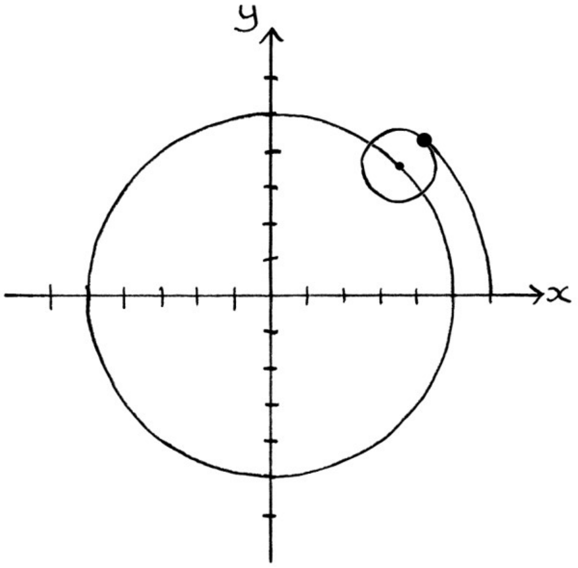


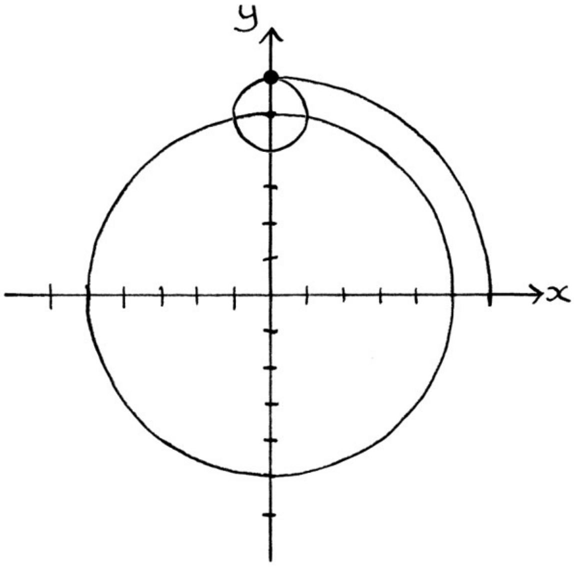
When t = 1 second, the y-axis value of the outer object will be:
“5 sin (360 * 1) + sin (360 * 1)” = 0.

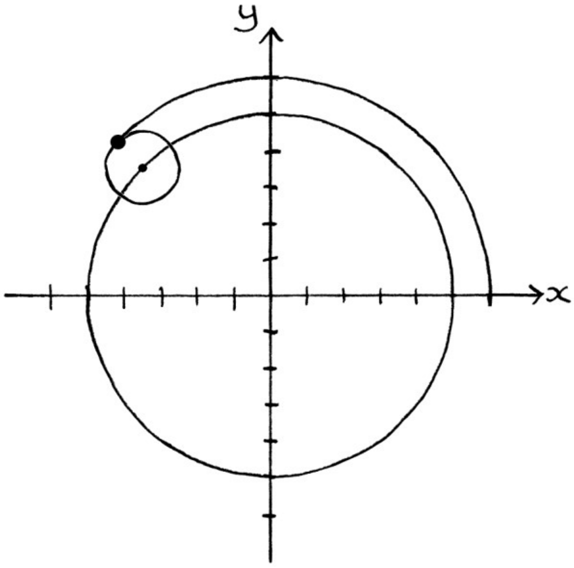
The x-axis value at this time will be:
“5 cos (360 * 1) + cos (360 * 1)” = 6.

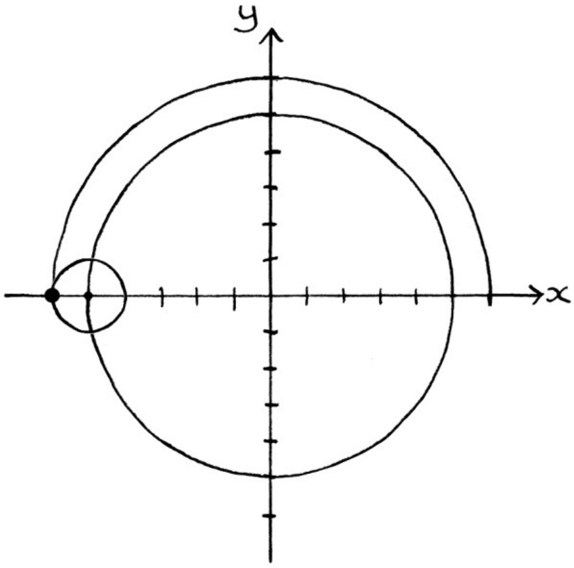


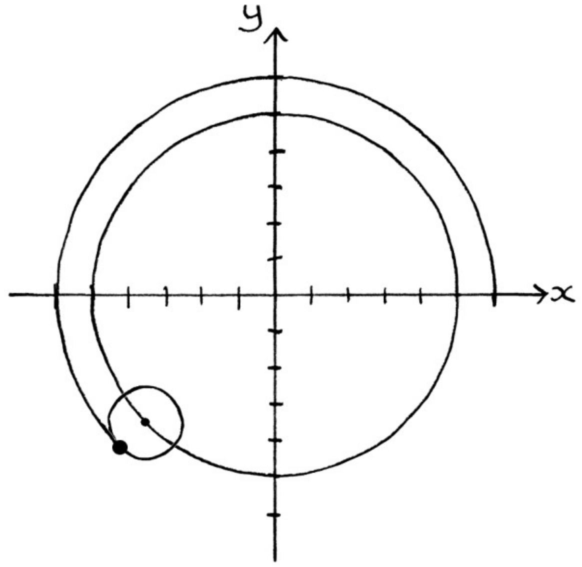
If we followed the movement of the outer object over time, it would draw out a shape:

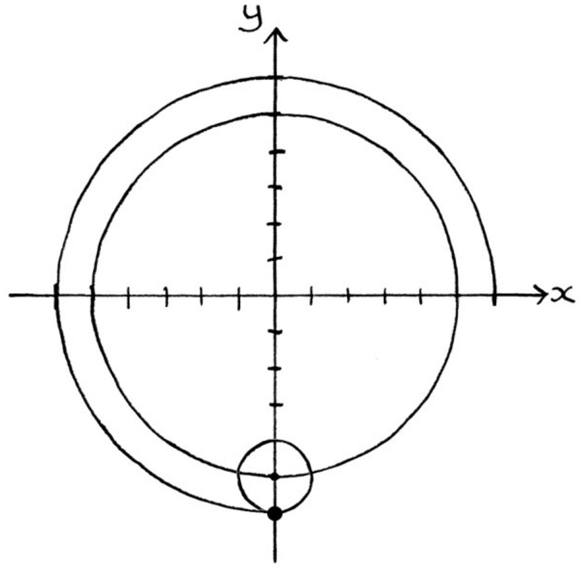


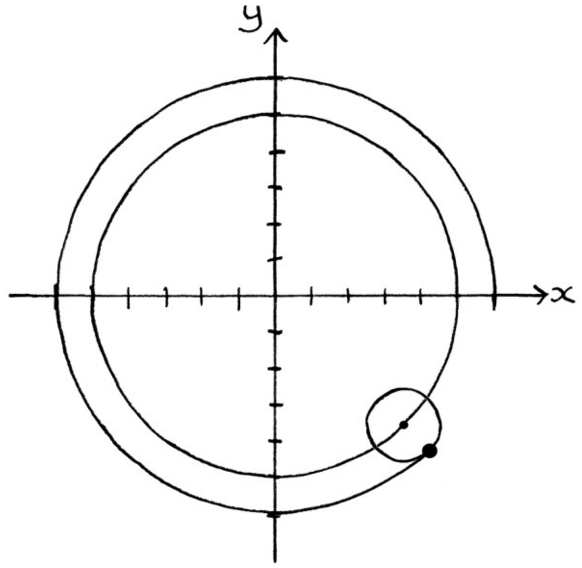

y
x

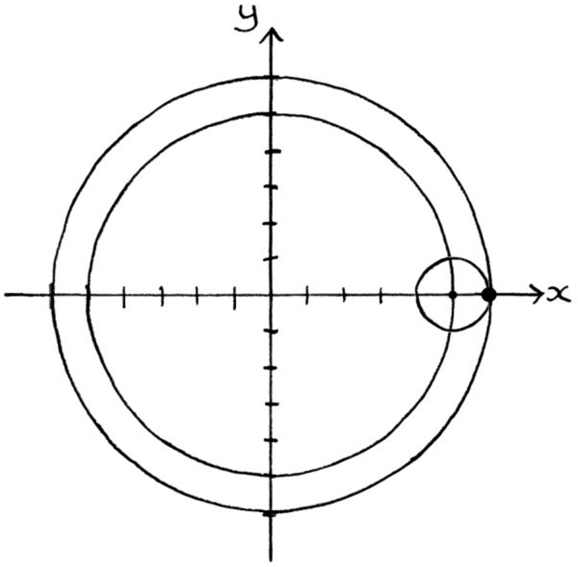

y
x


y
x

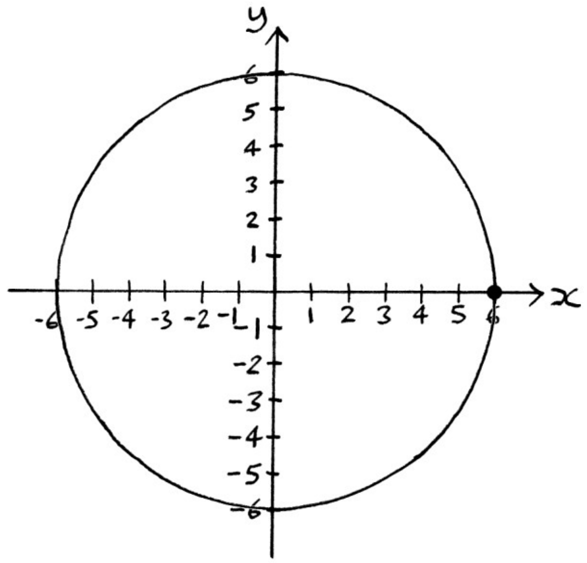

y
x


y
x


y
x



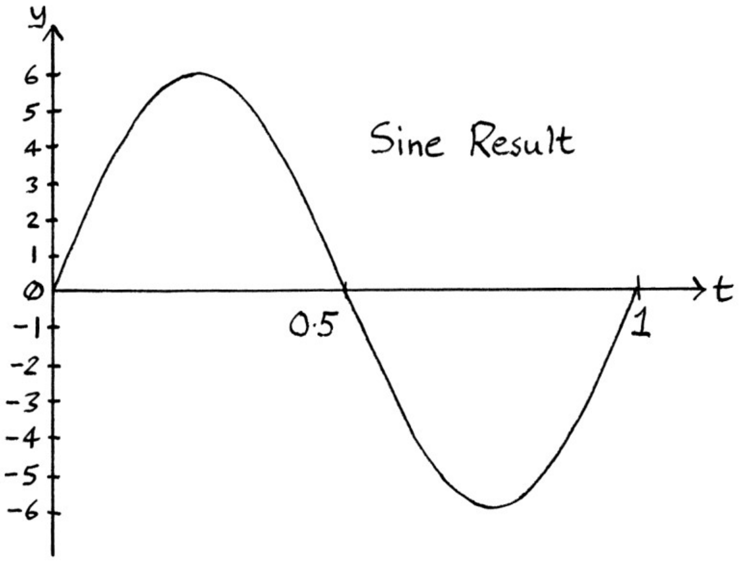
The resulting shape looks like this:



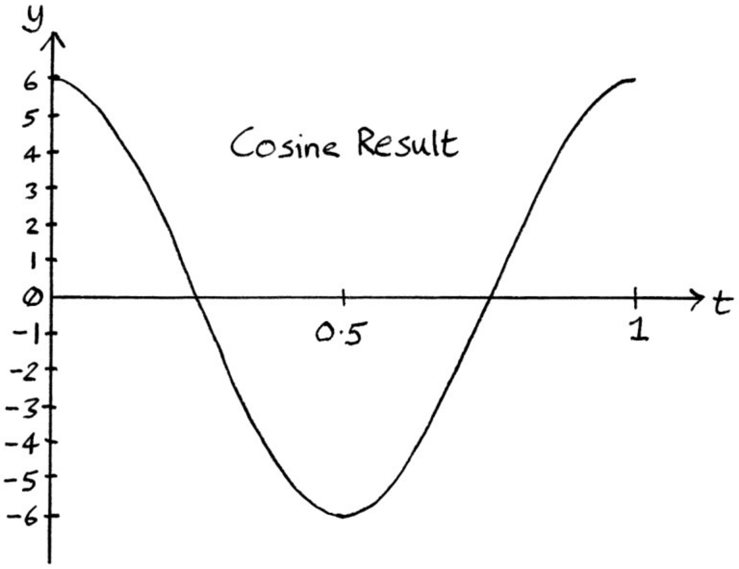
We have ended up with a circle. This circle represents the result of adding the two original Sine waves, and also the result of adding the corresponding Cosine waves. The circle has a radius of 6 units, a phase of zero degrees, and is centred at (0, 0). We know that its phase is 0 degrees by how when the outer object started rotating at t = 0, it was at 0 degrees in relation to the centre of the inner circle.

Because the two objects are rotating around their circles at the same rate, they are always at the same angle from their own circles' centres. Therefore, the outer object draws out a perfect circle. The waves derived from a circle are pure waves. This is a good visual way to see that adding a pure wave to a pure wave of the same frequency will result in a pure wave.

If we plotted the y-axis positions of the outer object against time as it moved around this resulting circle, we would end up with a graph that showed the result of adding "y = 5 sin 360t" to "y = sin 360t", so in other words, "y = 6 sin 360t".



If we were to plot the x-axis values of the outer object on the chart against time as it moved around the outer circle, we would end up with a graph that showed the result of adding "y = 5 cos 360t" and "y = cos 360t", so in other words, "y = 6 cos 360t":



If we took every point from the graphs we just drew, with the points from the Cosine wave sum graph as x-axis coordinates, and the corresponding points from the Sine wave sum graph as y-axis coordinates, and plotted those points, we would end up drawing the resulting circle again.

By making one circle rotate around the other, we have found a more intuitive way to calculate the sums of waves. The circles can be drawn on paper and accurately measured to find results, but for most purposes, the idea is useful as a way of visualising the process of adding waves. The method works for adding waves that have any of the four attributes (amplitude, frequency, phase and mean level), and it can also work for any number of waves being added together. The order of the circles is irrelevant – any order will achieve the same result, however it is much clearer when the smaller circles move around the larger circles.

As I said before, if we are given the resulting shape (which was a circle in this case), and no other information, we will not be able to derive the signal that was made from adding Sine waves or the signal that was made from adding Cosine waves. We would need to know the time that the outer object was at a particular place on its journey to know where on the graph time lines that the y-axis and x-axis positions should be marked. This is the same as before we were doing addition, but the idea becomes more obvious when the resulting shape is not a circle, as happens when adding different frequencies.

Converting a wave into a circle

The process of adding circles together produces a result that contains the sum of the derived waves together. Adding circles automatically adds the derived waves together. Sometimes, addition with circles is simpler or easier to visualise than addition with waves. Therefore, if we want to add two waves, it can sometimes be helpful to convert them into circles first, and then add the two circles. To do this, we pick the circles from which the waves were, or could have been, derived.

It is best to think of there being only one type of circle that contains both types of wave – a Sine wave and a Cosine wave. As we know, a circle implies a Sine wave and a Cosine wave with the same amplitude, frequency and phase. Therefore, if we are given a Sine wave or a Cosine wave, we automatically know its Cosine or Sine twin, and we also know the circle that would have created it (presuming there is no mean level).

If we are given the wave, “y = 2 sin 360t”, then the circle from which it was derived (or the circle from which it can be said to be derived) must have a radius of 2 units and the phase point will be at 0 degrees. Therefore, we will let that circle represent the Sine wave in additions. Adding two circles from which two Sine waves were derived will also result in the corresponding Cosine waves being added too. [As we know, the circle for “y = 2 sin 360t” is also the circle for “y = 2 cos 360t”.] To have a result consistent with the type of the original wave (a Sine wave in this case), the resulting shape would need to have its *y-axis* values read over time to produce the required resulting signal.

If we are given the wave, “y = 2 cos 360t”, then the circle from which it was derived will also have a radius of 2 units and the phase point will be at 0 degrees. That circle will represent the Cosine wave in additions. Adding the two circles from which two Cosine waves were derived will also result in the corresponding Sine waves being added too. To have a result consistent with the original wave (a Cosine wave in this case), the resulting shape would need to have its *x-axis* values read over time to produce the resulting signal.

From these examples, we can see that we use the same circle for Sine or Cosine. The result is read from the appropriate axis. For Sine (which is the y-axis wave), we read the y-axis position of the outer object on the resulting shape over time; for Cosine (which is the x-axis wave), we read the x-axis position of the outer object on the resulting shape over time.

If we are given the wave, “y = 4.6 sin ((360 * 7t) + 34)”, then the circle from which it is derived will have a radius of 4.6 units and a phase point at 34 degrees. That circle will represent the Sine wave and its Cosine twin in additions. When it comes to reading the result of adding that circle to another circle, we read the y-axis values.

If we are given the wave, “y = 4.6 cos ((360 * 7t) + 34)”, then the circle will also have a radius of 4.6 units and a phase point at 34 degrees. That circle will represent the Cosine wave in additions. When it comes to reading the result of adding that circle to another circle, we read the x-axis values.

If we are given a Sine wave with a non-zero mean level, then the circle will be raised up or down the y-axis on the circle chart accordingly. We cannot deduce the x-axis mean level for the circle, so it is best to presume it is zero. If we are given a Cosine wave with a non-zero mean level, then the circle will be shifted left or right on the circle chart accordingly. Again, we cannot deduce the y-axis mean level for this circle, so it is best to presume that it is zero.

The only tricky situation is if we want to add a Sine wave and a Cosine wave. In that case, we must convert them both into the same form. The Cosine wave must become a Sine wave with 90 degrees added on to its phase, or the Sine wave must become a Cosine wave with 90 degrees subtracted from its phase. Then, the relevant circles can be chosen. The result will be based on the positions of the outer object over time as read on the *relevant axis.* If we made both into Sine waves, we would read the y-axis positions of the object over time; if we made both into Cosine waves, we would read the x-axis positions of the object over time.

For example, if we want to add:
“y = 2 sin (360 * 5t)”
... and:
“y = 4 cos (360 * 7t)”
... then we can choose the two circles that we will add as the ones from which these waves are derived:
“y = 2 sin (360 * 5t)” and “y = 2 cos (360 * 5t)”
... and:
“y = 4 sin ((360 * 7t) + 90)” and “y = 4 cos ((360 * 7t) + 90)”
... in which case, the result would be read off the y-axis positions of the outer object over time.

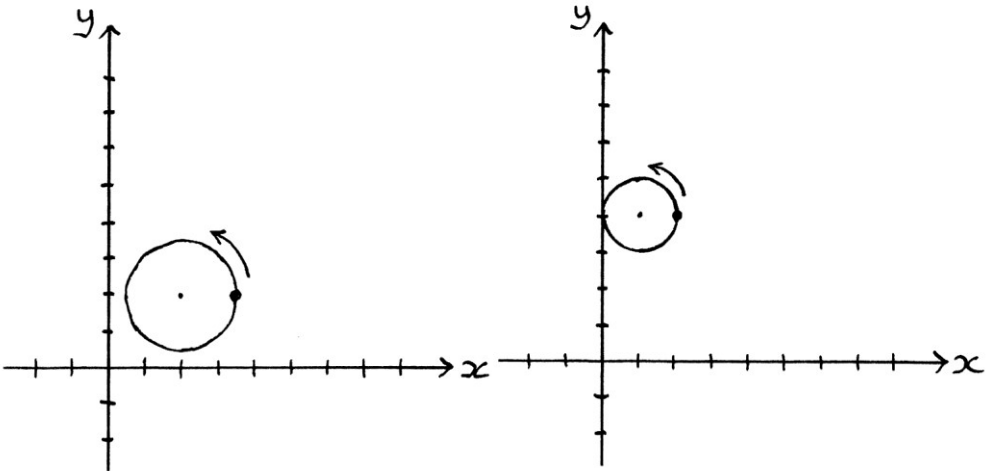
Alternatively, we could choose the two circles that we will add as the ones from which these waves are derived:
“y = 2 sin ((360 * 5t) − 90)” and “y = 2 cos ((360 * 5t) − 90)”
... and:
“y = 4 sin (360 * 7t)” and “y = 4 cos (360 * 7t)”
... in which case, the result would be read off the x-axis positions of the outer object over time.

Generally, in this chapter, I will show the addition of circles and not the addition of waves made into circles.

Mean levels

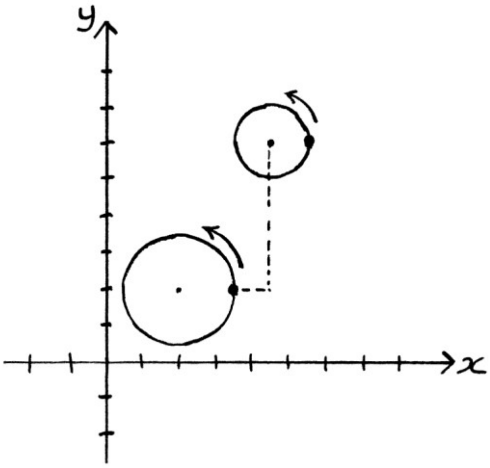
Calculating the sum of two circles with the same frequency, any amplitude, any phase, and non-zero mean levels is straightforward.

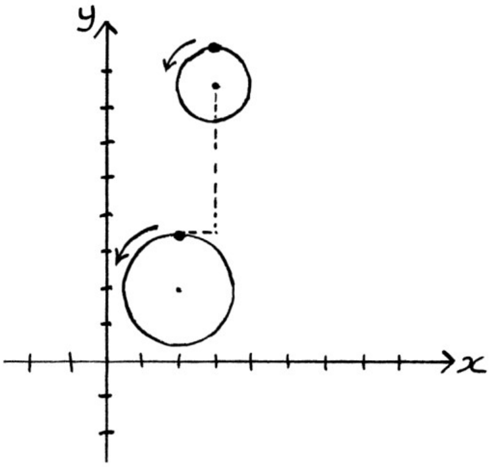
If we have two circles with mean levels, such as these:

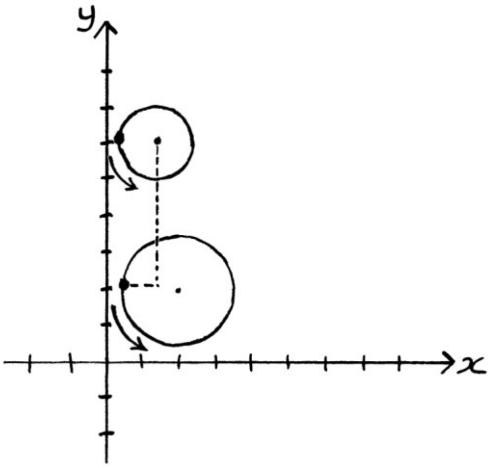


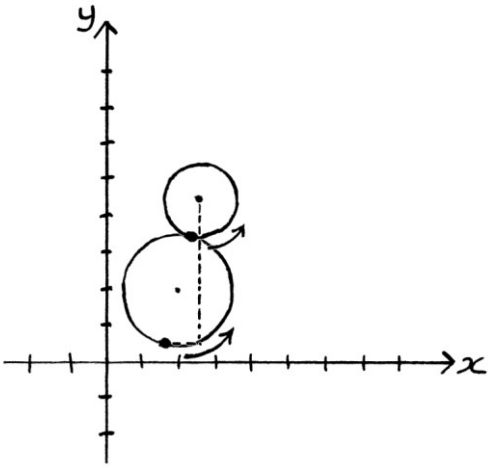
... then there are two ways to deal with them:

- We can keep each circle’s mean levels with the relevant circle. If we are adding two circles and both have mean levels, then the first circle will be moved away from the origin of the axes, and the second circle will be moved away from the object rotating around the first circle.

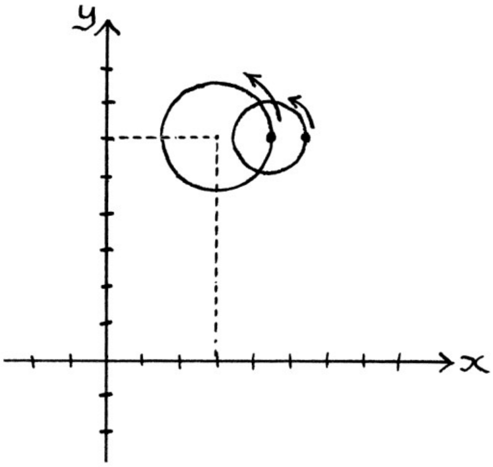


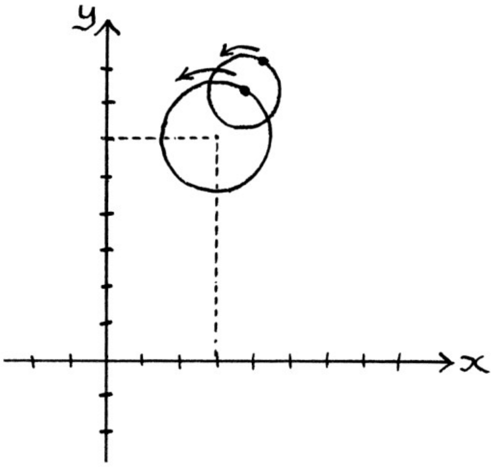

y
x


y
x


y
x

- We can group the mean levels together and apply them to the first circle. If we are adding two circles with mean levels, the mean levels from the second circle are added to the first, and the first circle is moved away from the origin of the axes to that extent. The second circle is centred directly over the object moving around the first circle. The outer object rotates around the outer circle, which in turn rotates around the inner circle.



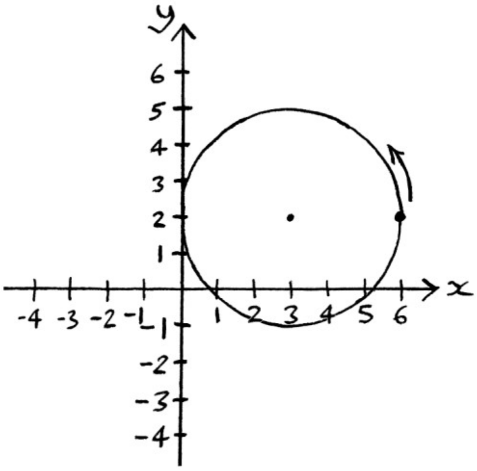


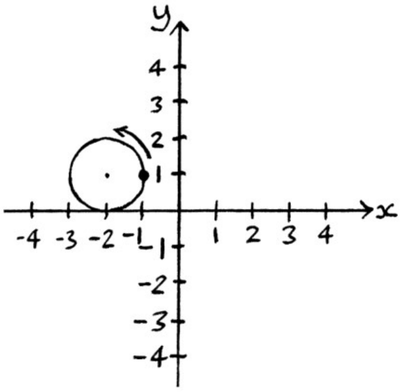
Both ways have identical results if we follow the path of the outer object. The first way looks less tidy as there are gaps between the circles. The second way is tidier, but we cannot see which circle has which mean level. It does not particularly matter which method we choose, but the second method is clearer.

Example

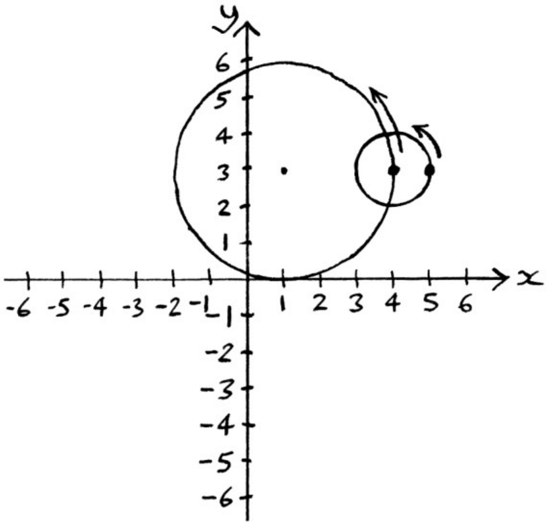
We will look at the sum of the two circles that represent the wave pairs:
“y = 2 + 3 sin (360 * 0.5t)” and “y = 3 + 3 cos (360 * 0.5t)”
“y = 1 + sin (360 * 0.5t)” and “y = −2 + cos (360 * 0.5t)”

The individual circles are as so:

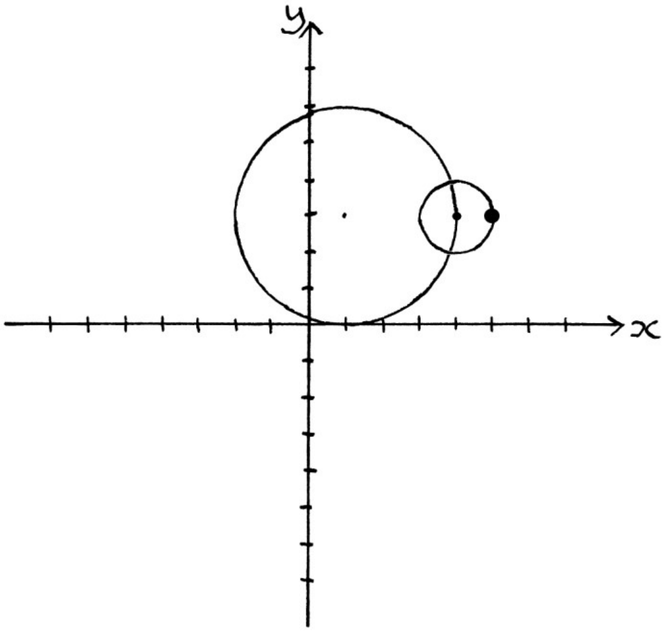




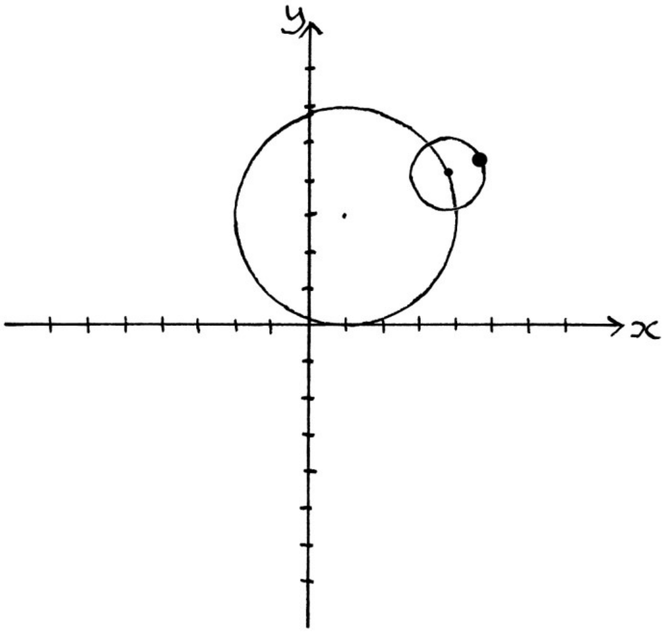
We will combine the mean levels and centre the larger circle at that point. Therefore, the larger circle is centred at (1, 3). The smaller circle is centred on the object rotating around the larger circle.



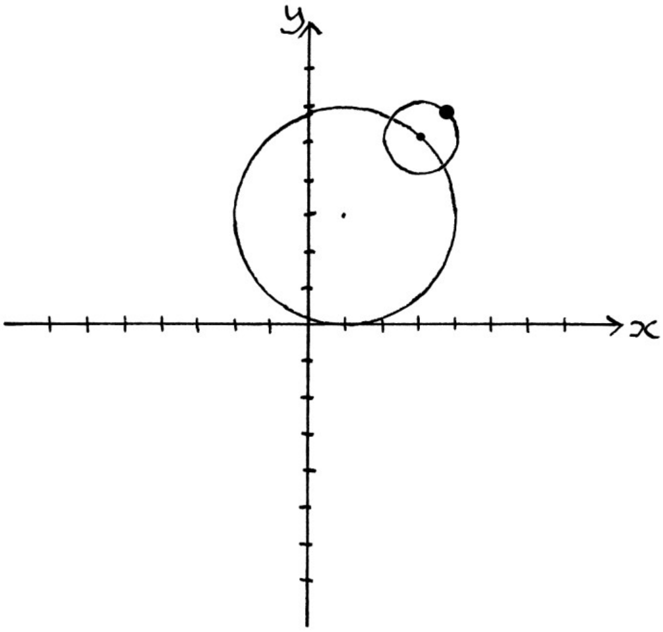
At t = 0, our combined circles look like this:



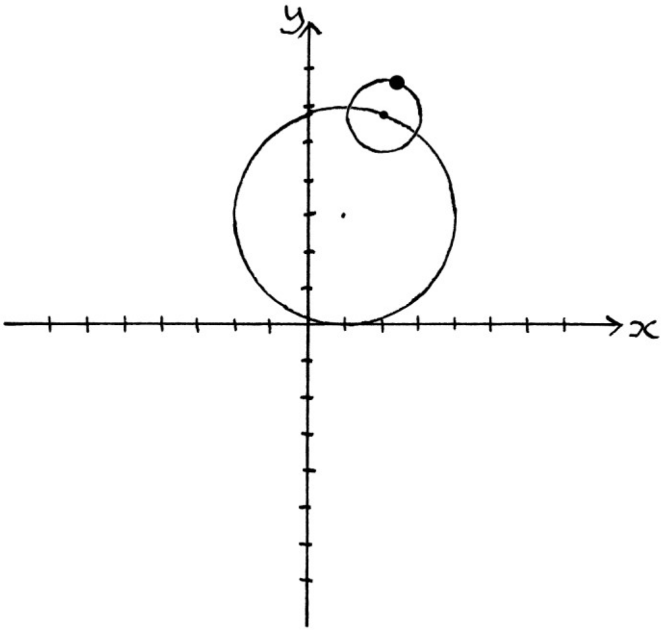
At 0.125 seconds, the circles are like this:



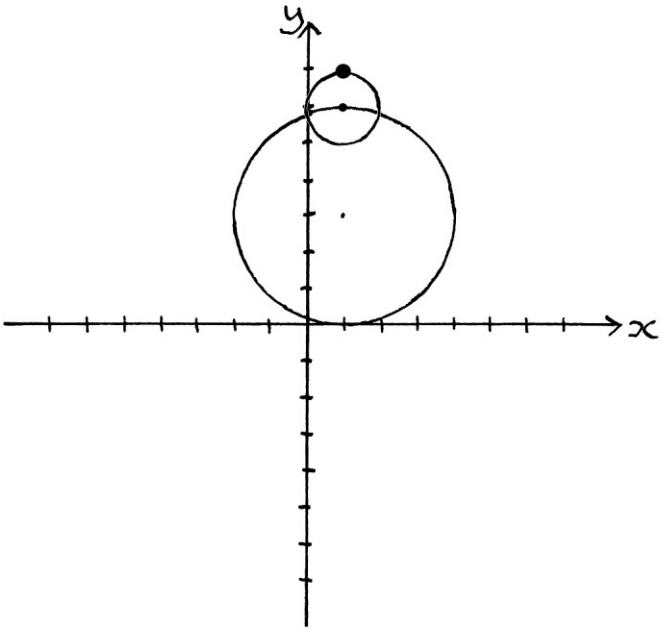
At 0.250 seconds, like this:



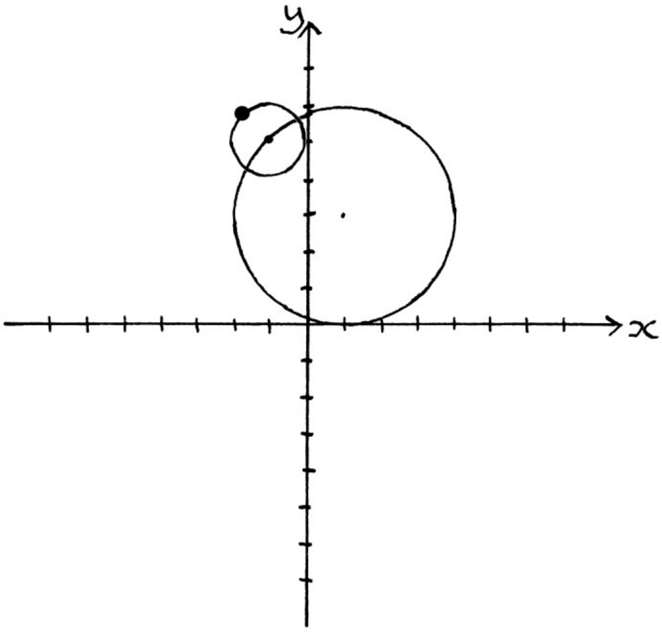
At 0.375 seconds, like this:



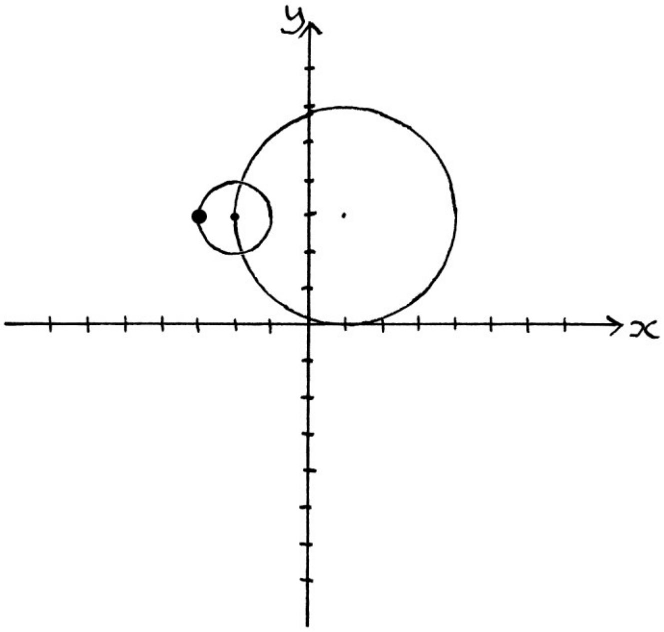
At 0.5 seconds, like this:



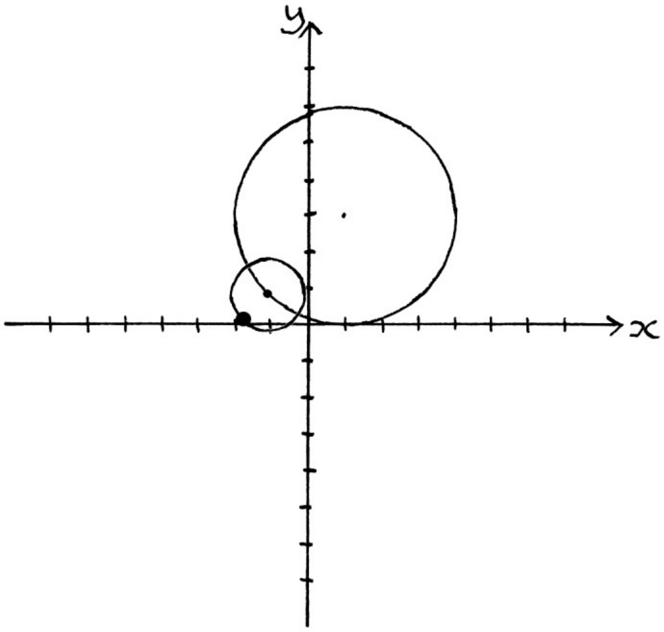
At 0.750 seconds like this:



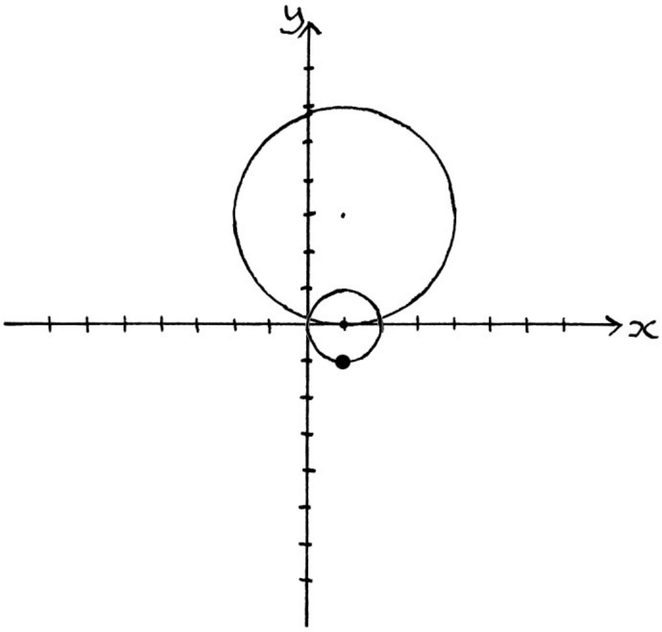
At 1 second like this:



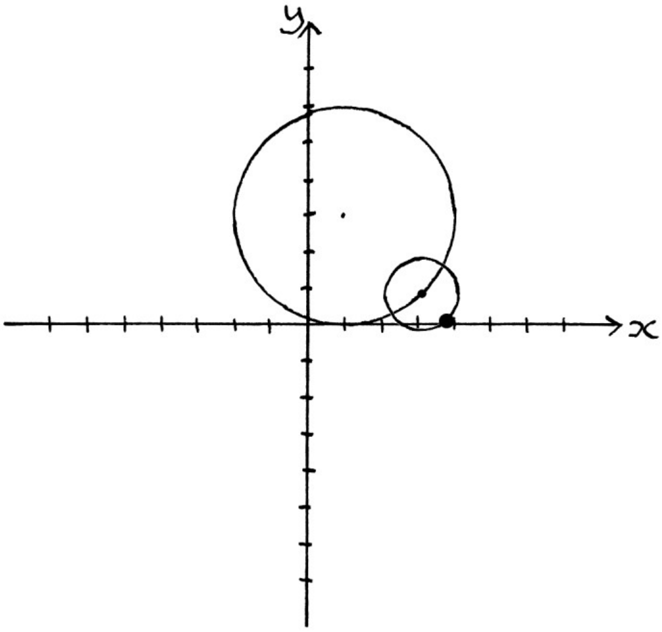
At 1.25 seconds:



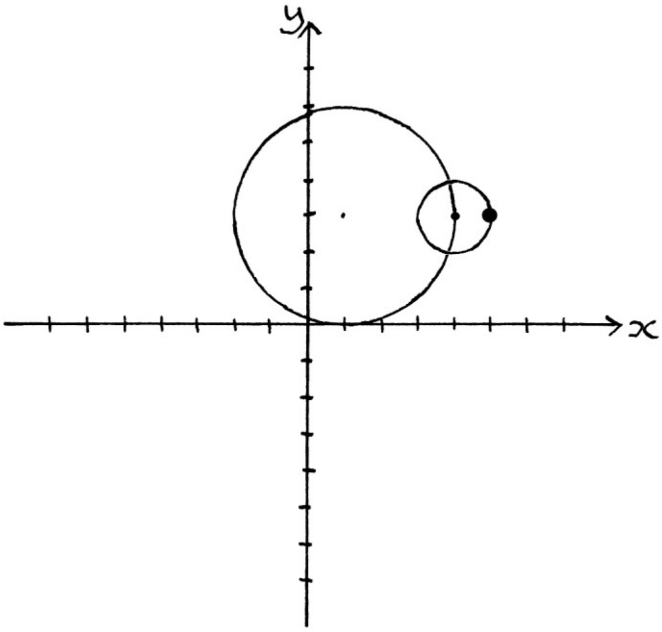
At 1.5 seconds:



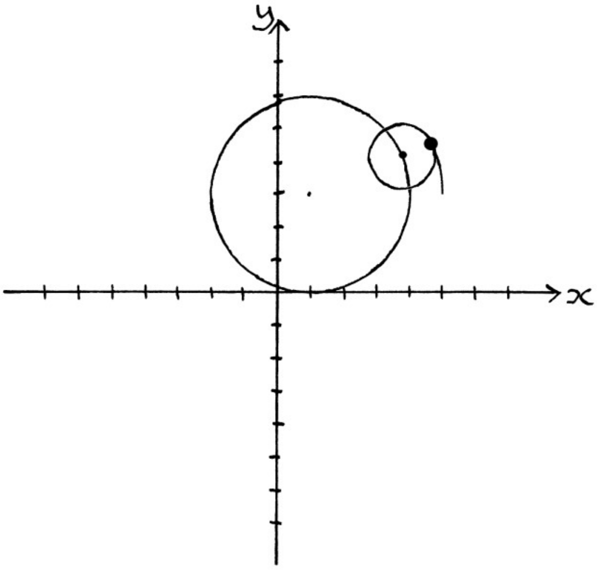
At 1.75 seconds:

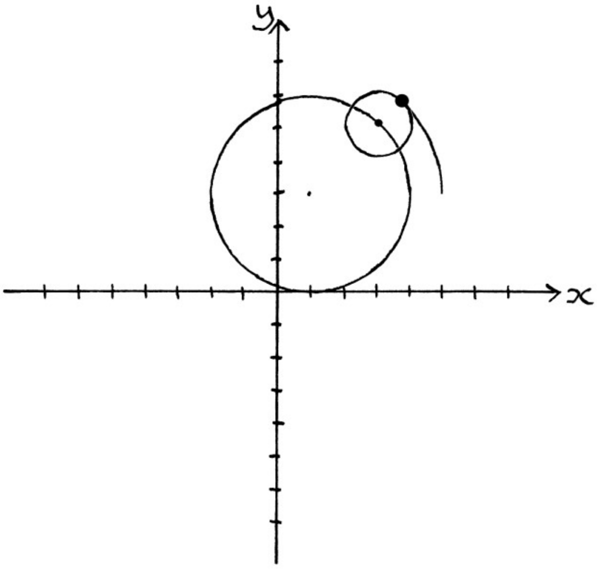


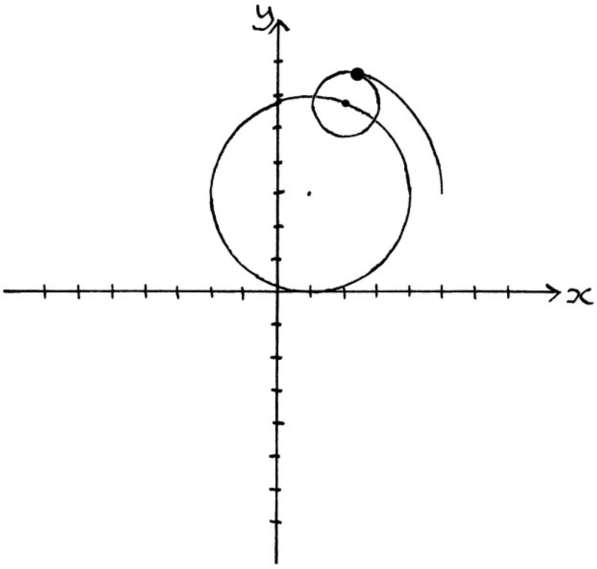
At 2 seconds, the circles are like this:

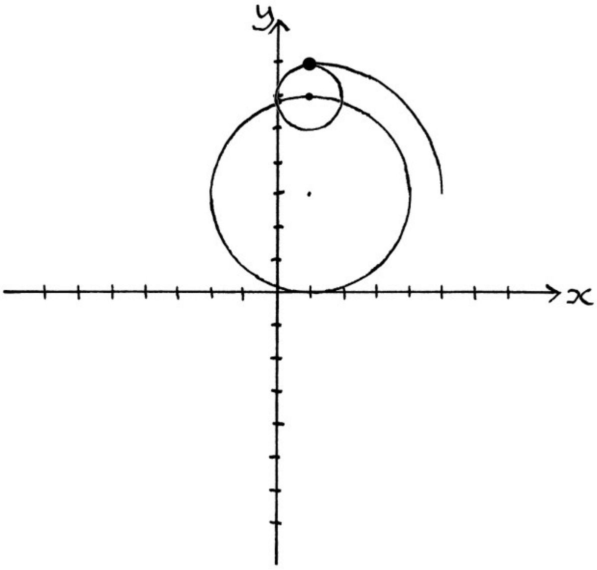


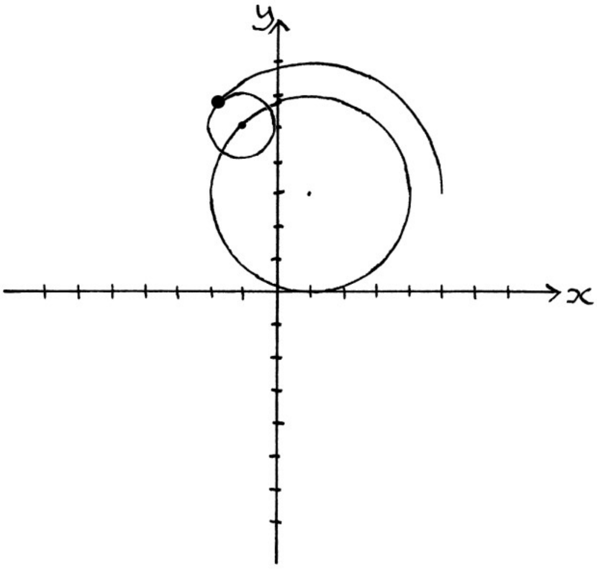
The outer object draws out a circle as it rotates around its circle, and as that circle rotates around the inner circle:

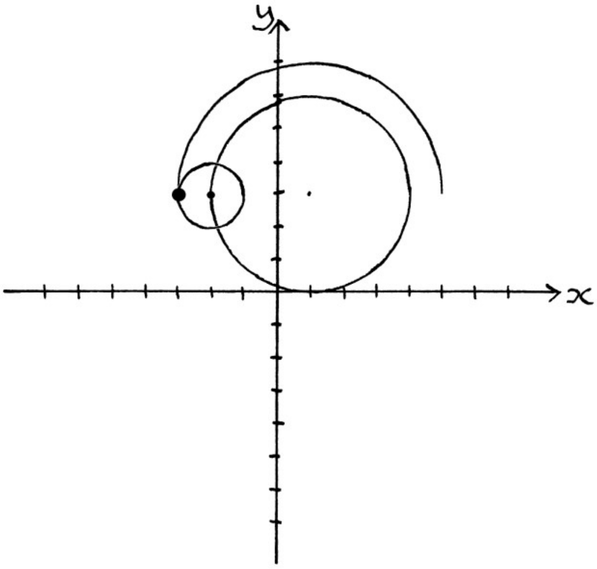


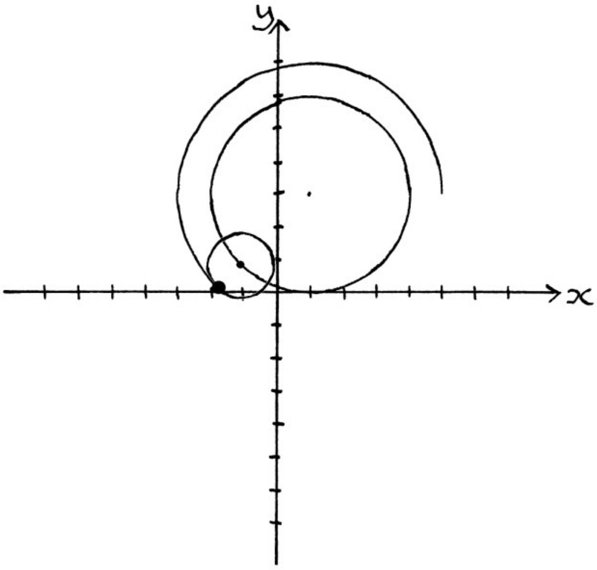

y
x

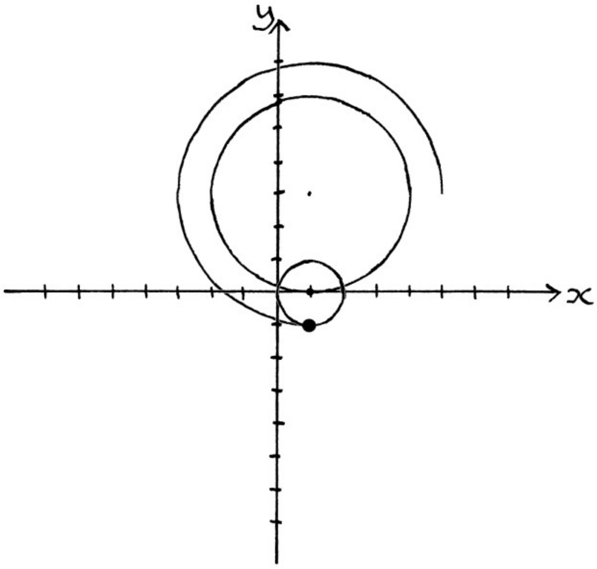

y
x

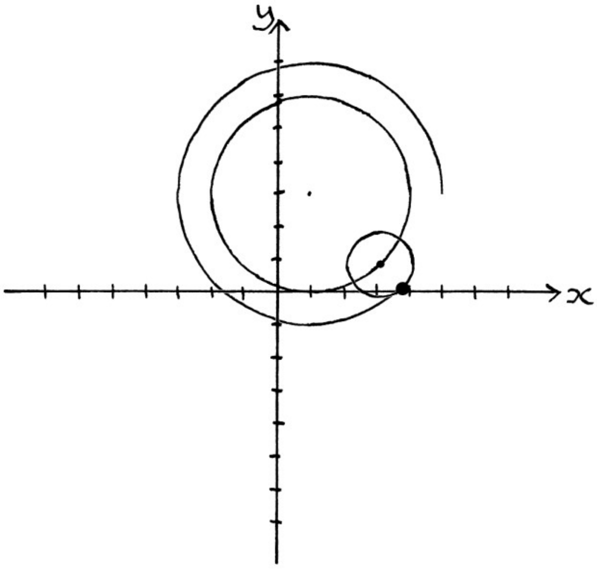

y
x

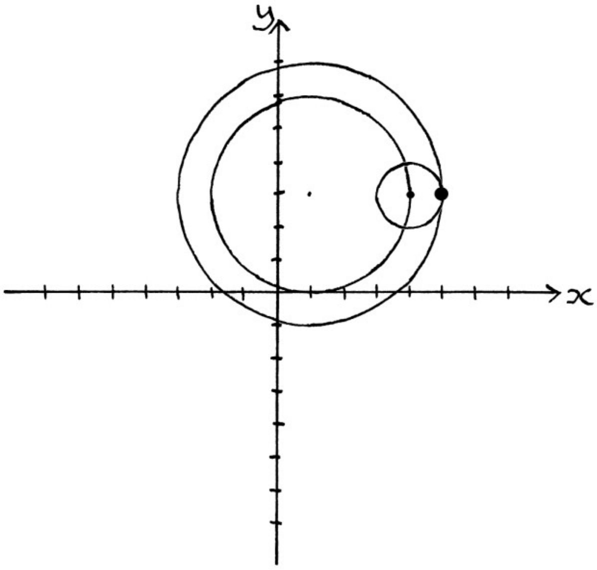

y
x


y
x

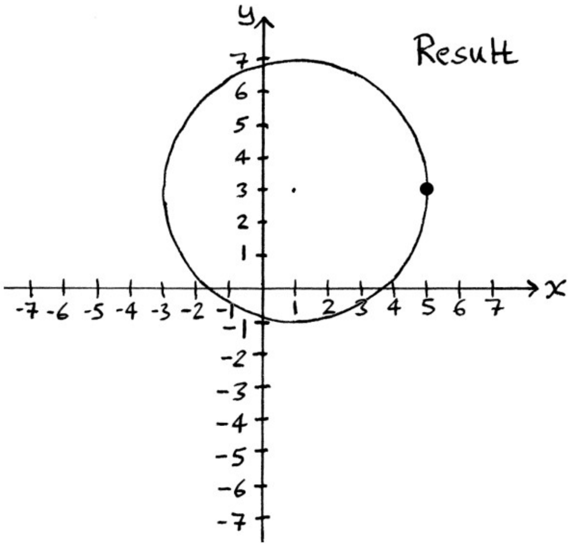

y
x


y
x


y
x


y
x

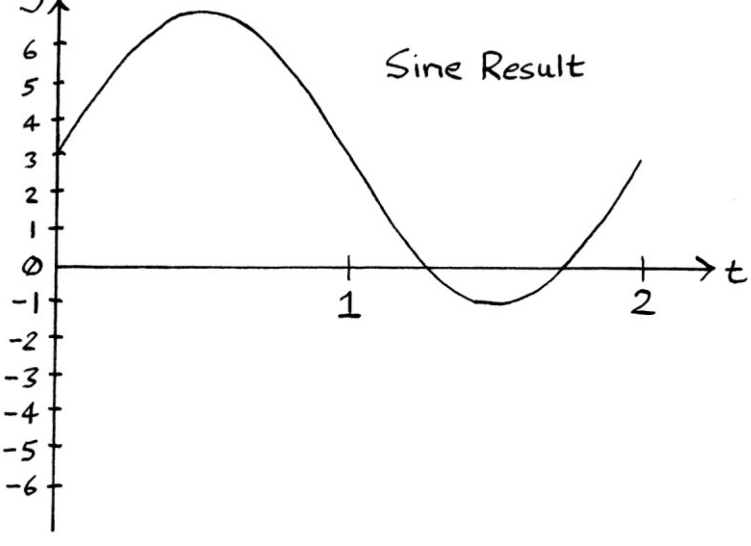
The resulting shape looks like this:



It is a circle with a radius of 4 units, a phase of 0 degrees, and a centre at (1, 3). We know that the phase of the circle is at 0 degrees, because the outer object started at 0 degrees in relation to the centre of the inner circle.

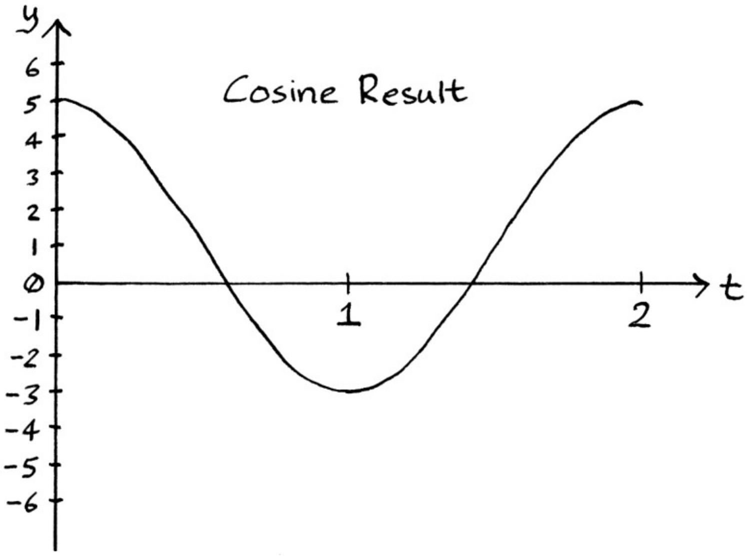
The y-axis values of the object rotating around the outer circle over time when drawn on a graph will produce a wave that is the sum of the two Sine waves derived from each original circle. The resulting wave has the formula:
“y = 3 + 4 sin (360 * 0.5t)”

Note how this wave’s mean level is 3 units, which is the same as the y-axis of the centre of the resulting circle.



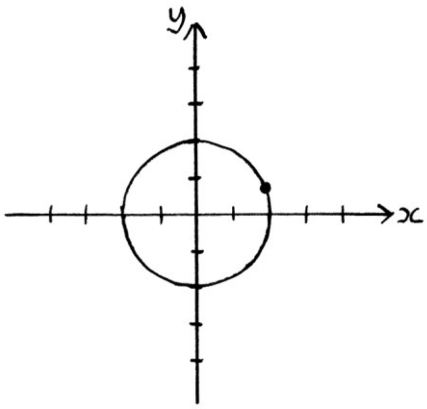
The x-axis values of the object rotating around the outer circle over time will draw out a wave that is the sum of the two Cosine waves derived from each original circle. The resulting wave has the formula:
“y = 1 + 4 cos (360 * 0.5t)”.

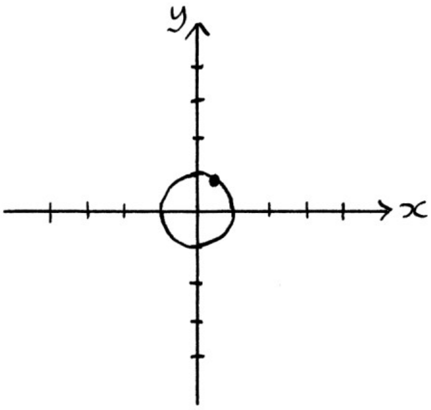
Note how this wave’s mean level is 1 unit, which is the same as the x-axis value of the centre of the resulting circle.



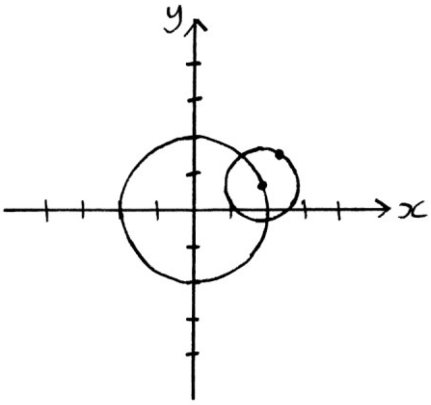
Phase

If the two circles have the same frequency, any mean level, any amplitude, but different phases, then the process of adding circles is still straightforward.

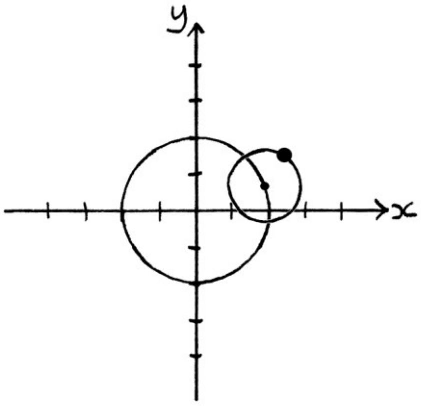


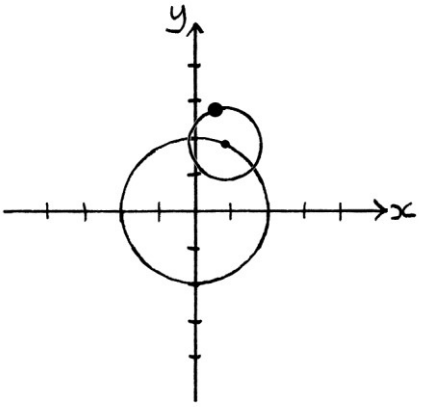


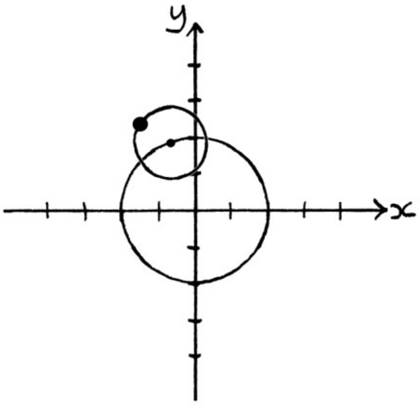
The method for adding circles with non-zero phases is really the same as adding circles with zero phases. The outer circle is centred on the object rotating around the inner circle. This is the same thing as saying that the outer circle is centred on the phase point of the inner circle. We were doing this before in the examples for amplitude and mean level, but in those cases, the phase points of each circle were at zero degrees. Now the phase points of one or more of the circles will not be at zero degrees.

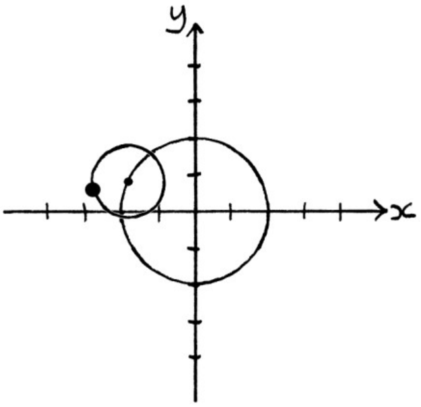


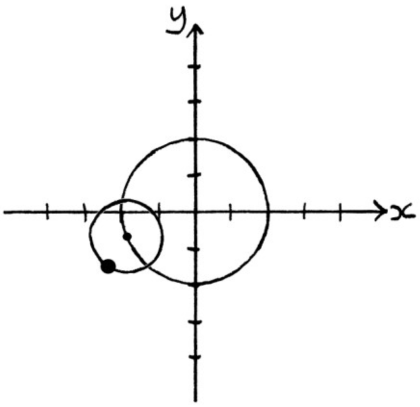
All measurements are still read off the position of the object rotating around the outer circle. However, the starting point of that object will not be at 0 degrees to the centre of the inner circle. The outer object rotates around the outer circle, which in turn rotates around the inner circle.

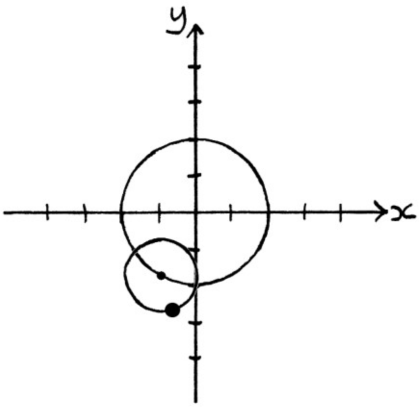


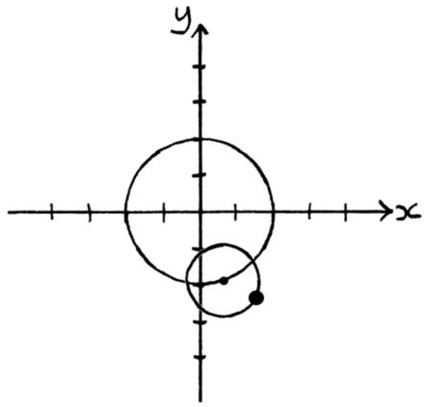


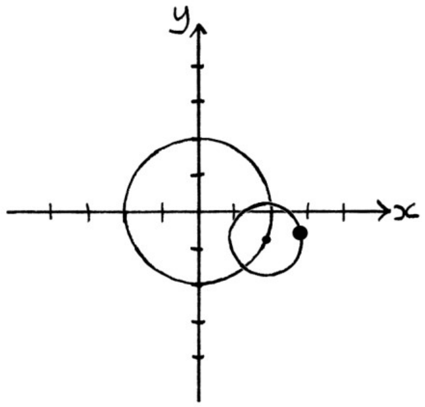

y
x

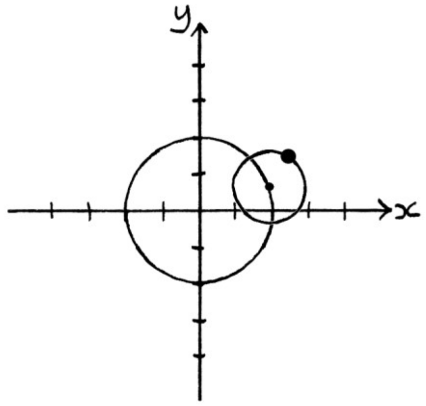

y
x


y
x

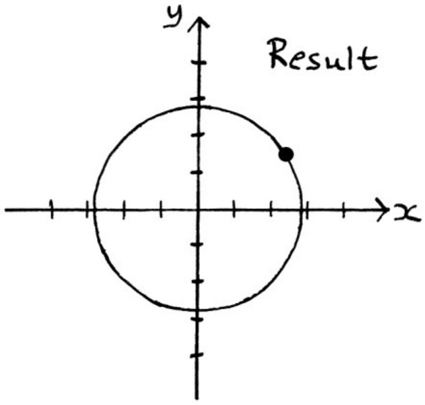

y
x


y
x


y
x


y
x

If we observe the movement of the outer object as it rotates around its circle, and as that circle rotates around the inner circle, we will see it draw out a shape, which will turn out to be this circle:

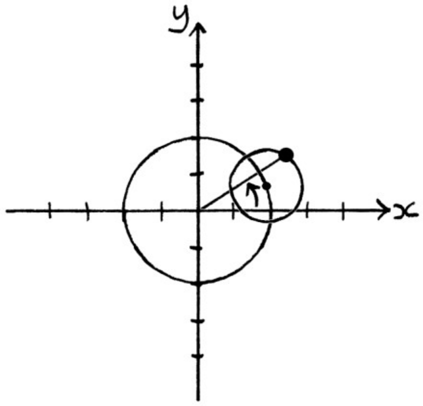


As before, the y-axis values of the outer object over time give the result of the sum of the Sine waves; the x-axis values of the outer object over time give the result of the sum of the corresponding Cosine waves.

If the circles have non-zero mean levels, then it is easiest if the mean levels are added and applied to the first circle.

The sum of two circles with a difference in phase, but the same frequency, still results in the outer object drawing out a circle. This shows that the resulting waves will be pure waves. This is consistent with the last chapter, when we saw that adding waves with the same frequency, but different phases still produced pure waves.

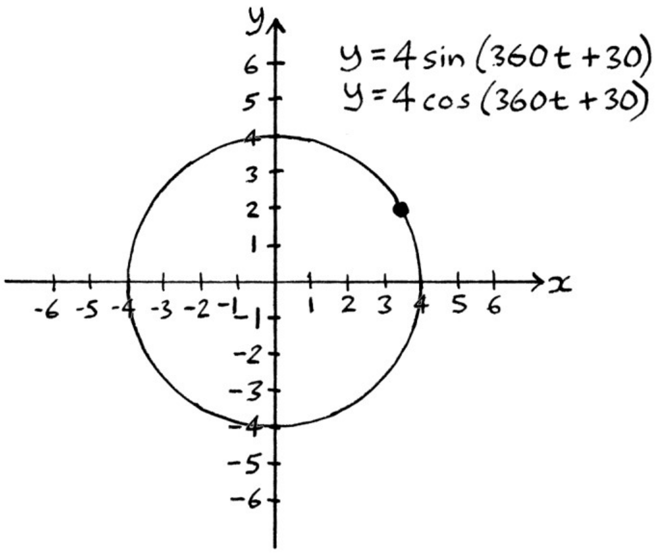
The phase of the resulting pair of waves will be the same as the angle of the outer object with respect to the centre of the inner circle at t = 0:

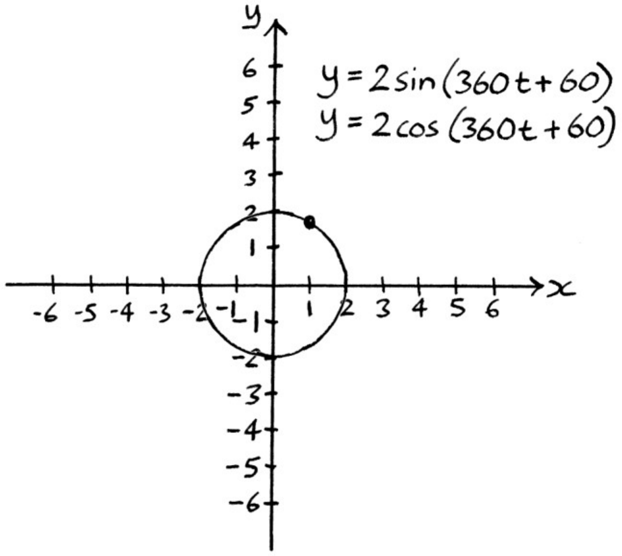


Example

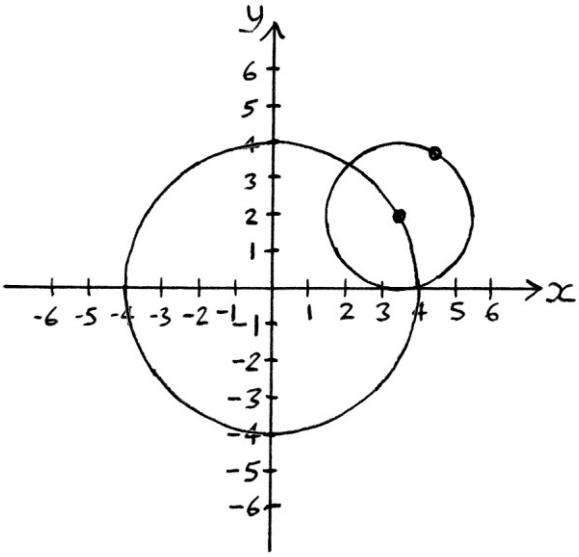
We will add the circles that represent the following pairs of waves:
“y = 4 sin (360t + 30)” and “y = 4 cos (360t + 30)”
... and:
“y = 2 sin (360t + 60)” and “y = 2 cos (360t + 60)”

The circles for these waves are as so:



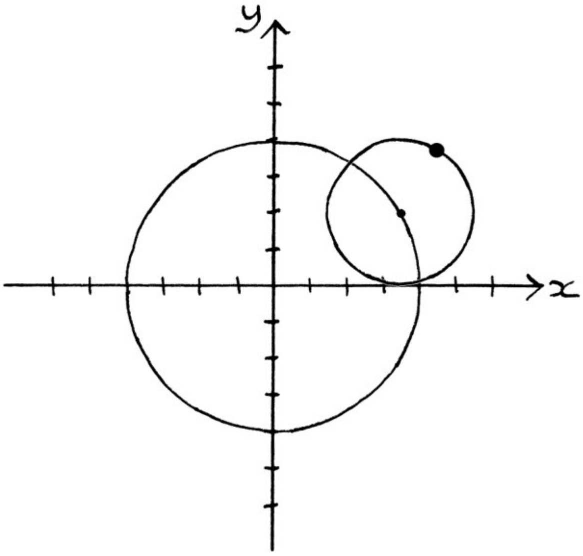


We place the bigger circle over the origin of the axes, and the smaller circle over the phase point of the bigger circle.

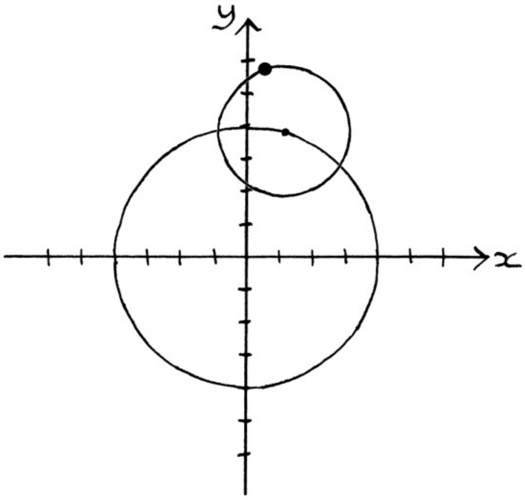


Then we will watch the position of the outer object as it rotates around the outer circle, and while the outer circle rotates around the inner circle.

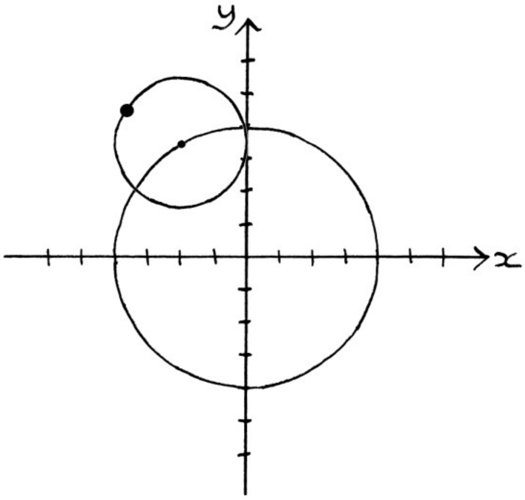
At t = 0, the circles look like this:



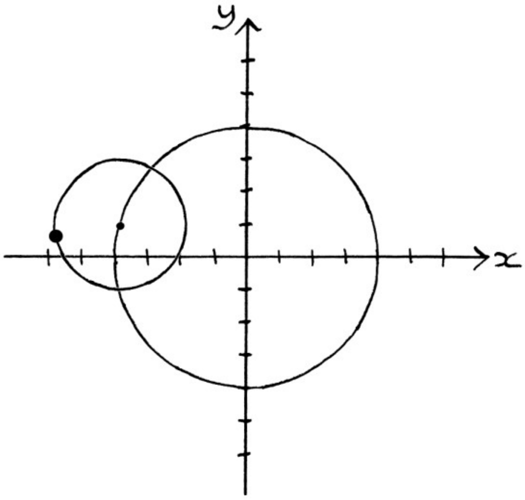
At 0.125 seconds, the circles look like this:



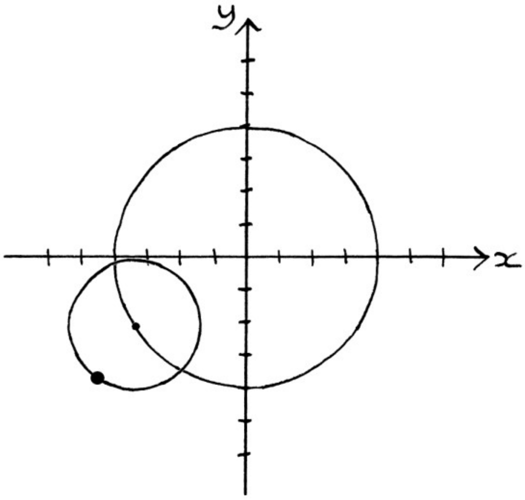
At 0.25 seconds:



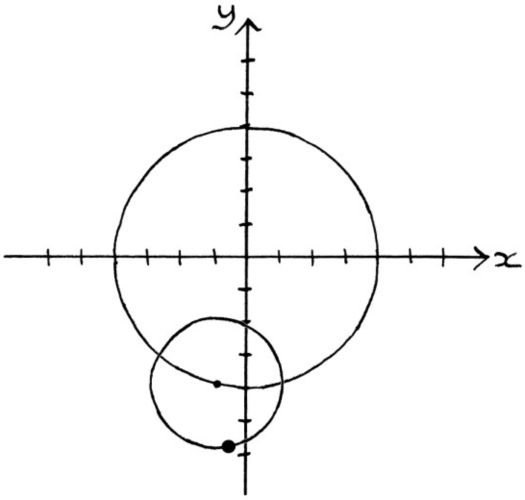
At 0.375 seconds:



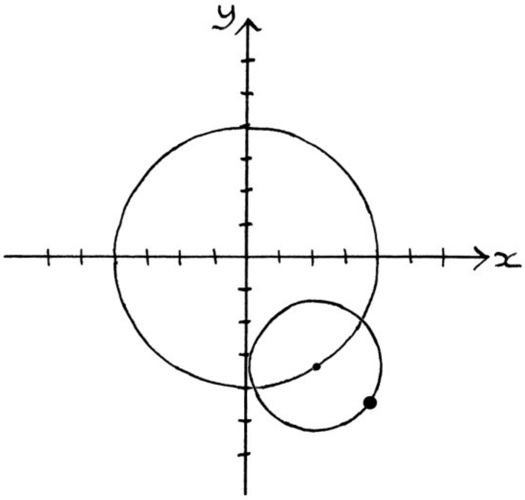
At 0.5 seconds:



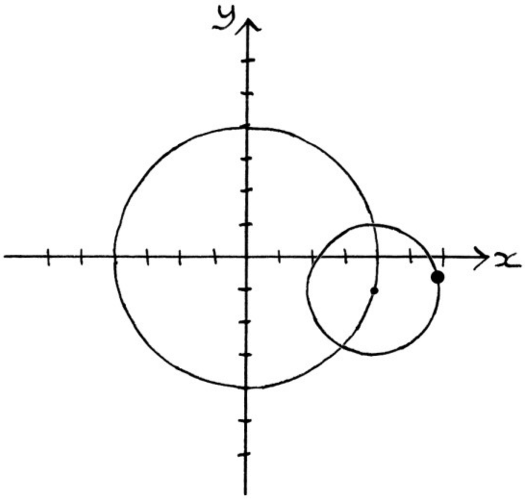
At 0.625 seconds:



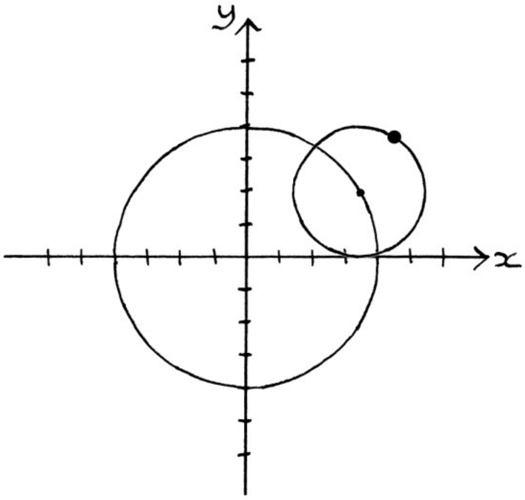
At 0.75 seconds:



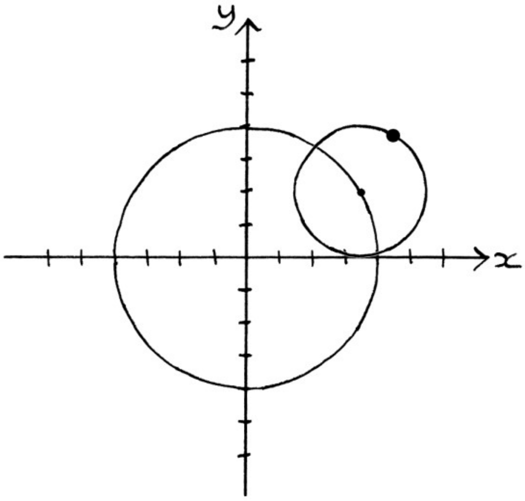
At 0.875 seconds:

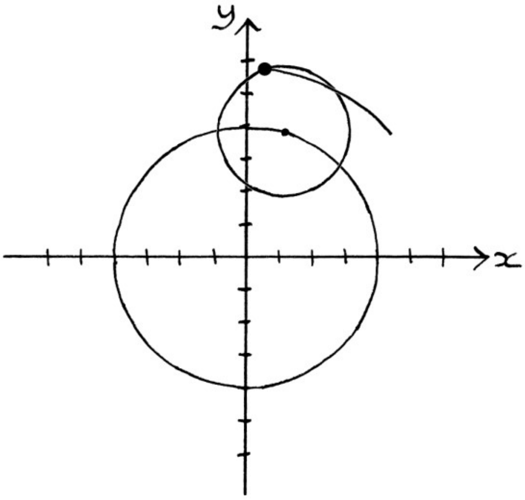


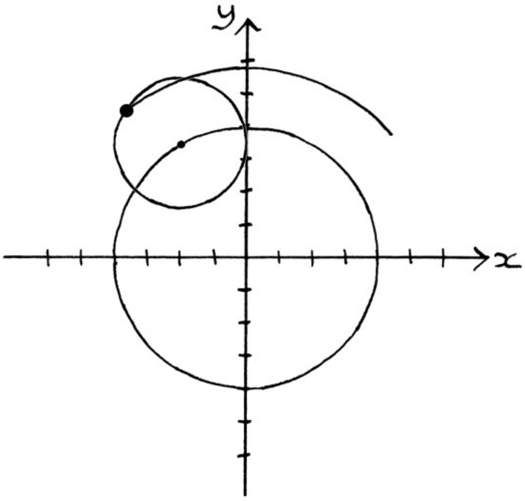
At 1 second:

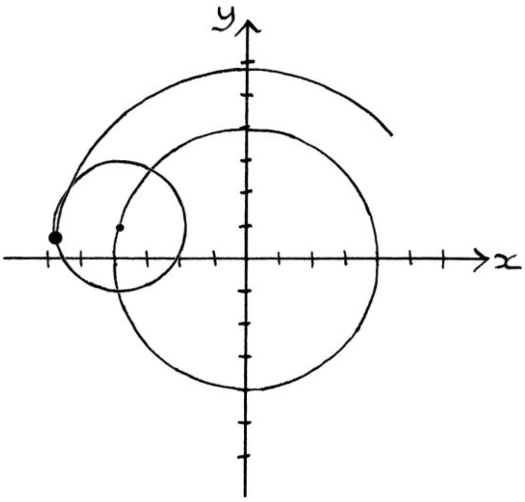


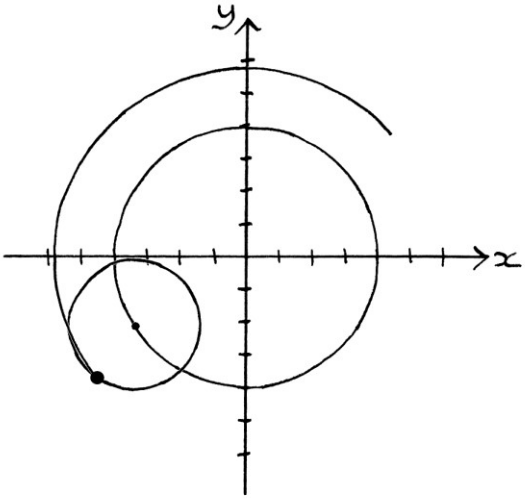
If we follow the position of the outer object as it rotates around its circle, and that circle rotates around the inner circle, it will draw out a shape:

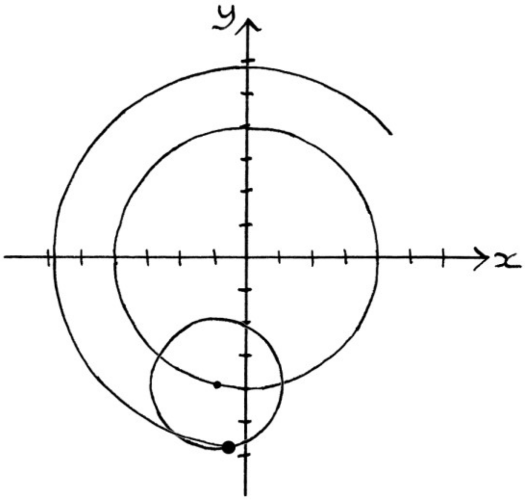


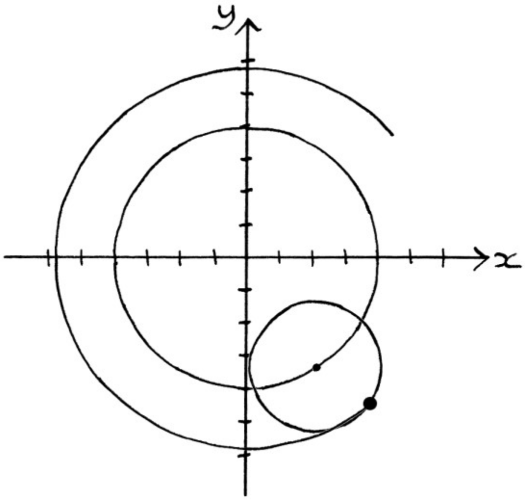

y
x

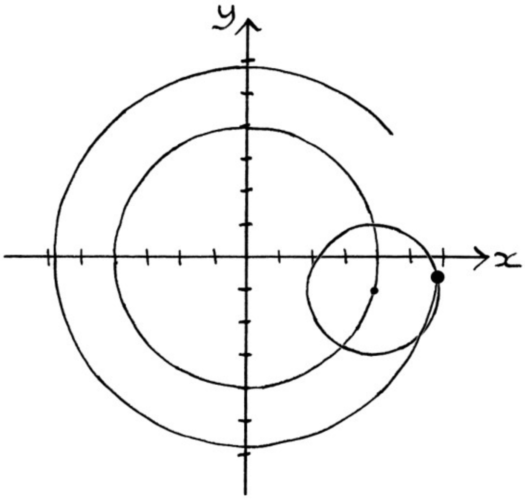

y
x

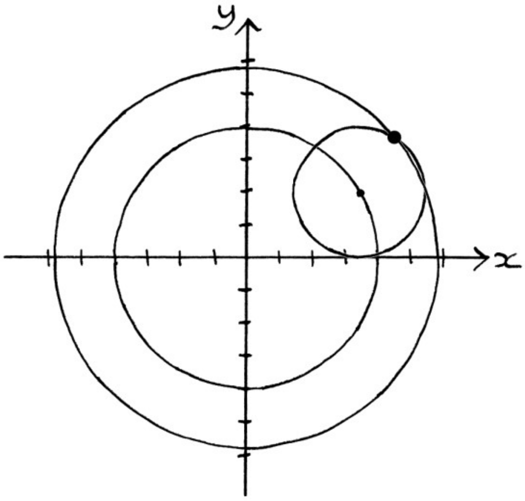

y
x


y
x

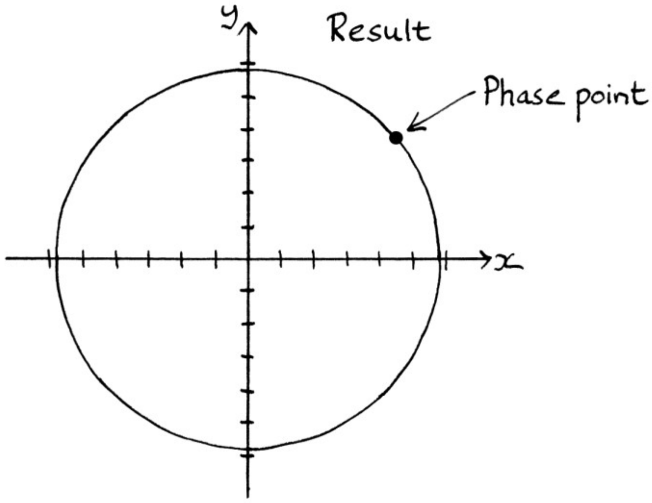

y
x


y
x



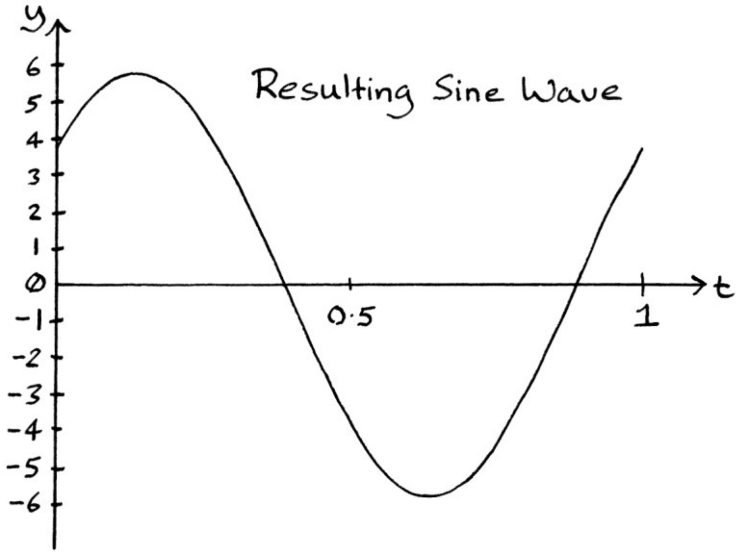


The resulting shape is a circle (as all results will be when the frequencies are the same), and it looks like this:

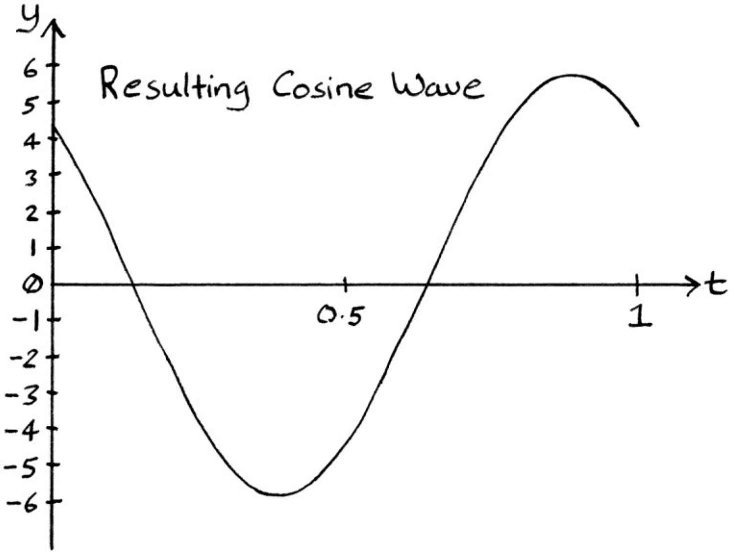


The phase point of the resulting circle is the point where the object on the outer circle was situated at t = 0.

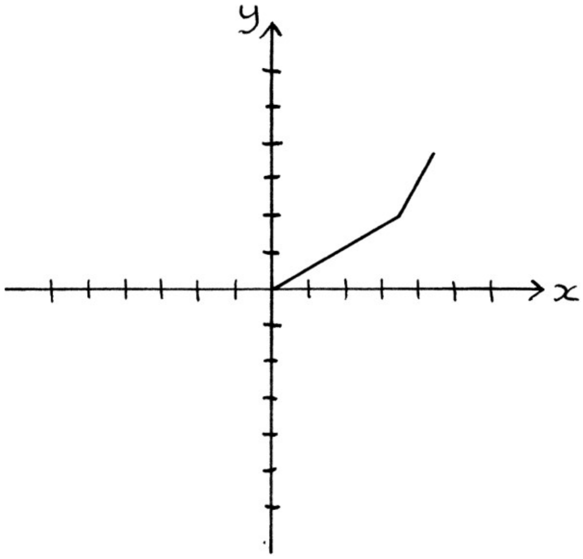
The y-axis values of the position of the outer object over time can be plotted on a graph, and they will give the results of the sum of:
"y = 4 sin (360t + 30)" and "y = 2 sin (360t + 60)"



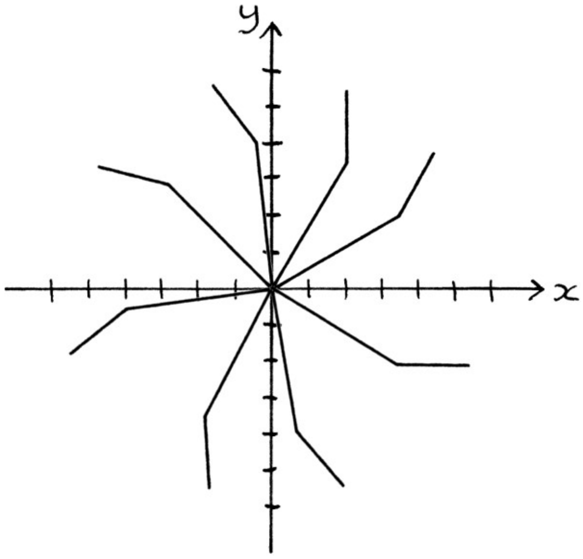
The x-axis values of the position of the outer object over time can be plotted on a graph, and they will give the results of the sum of:
"y = 4 cos (360t + 30)" and "y = 2 cos (360t + 60)"



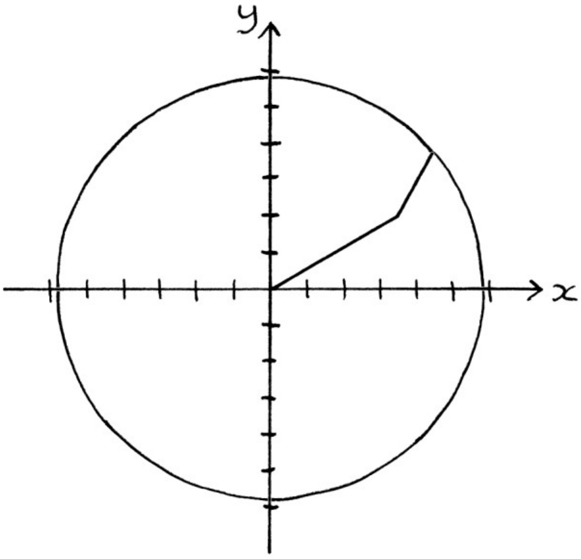
With different phases (and the same frequency), the line from the outer object to the centre of the outer circle remains at a set angle to the line from the inner object to the centre of the inner circle. It is as if the outer object is fixed to a bent arm that revolves around the circle without straightening or contracting. At t = 0, the arm looks like this:



As the outer object moves around on its journey, the arm moves around keeping a fixed bend throughout.



This means that the overall movement of the outer object is a circle. Therefore, the resulting Sine and Cosine waves are pure waves.

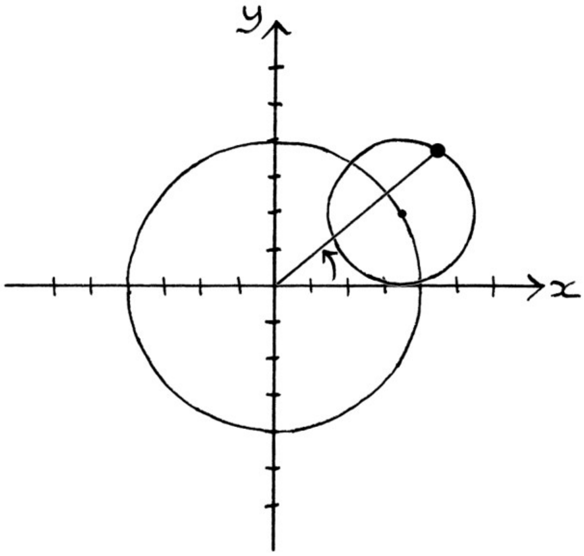


Calculating the phase

As the outer object rotates around its circle, and as that circle rotates around the inner circle, the outer object maintains the same angle in relation to the inner object for its entire journey, as portrayed by the "arm" mentioned above.

Knowing this means that we can calculate the phase for the sum of two circles that differ in phase, maybe differ in amplitude and mean level, but have the same frequency. The phase will be the angle of the outer object at t = 0 with reference to the centre of the inner circle (which might or might not also be the origin of the axes). In other words, the phase point of the outer circle in relation to the centre of the inner circle will be the phase point of the resulting circle, and the angle of that phase point will be the same as the phases in the two resulting waves derived from the resulting circle.

If we have a drawing of the two circles, then the phase can be calculated using a protractor to measure the angle:



Otherwise, we can imagine a drawing of the two circles and think about a triangle. We can calculate the starting point of the outer object by calculating its x-axis distance and y-axis distance from the centre of the inner circle. The x-axis position will be the adjacent side of a right-angled triangle; the y-axis position will be the opposite side of a right-angled triangle.

In the above example, the y-axis distance of the phase point (i.e. the point at t = 0) of the inner, larger, circle to its own centre is at:
4 sin ((360 * 0) + 30) = 4 sin 30 = 2 units.

The x-axis distance from its centre is:
4 cos ((360 * 0) + 30) = 4 cos 30 = 3.4641 units.

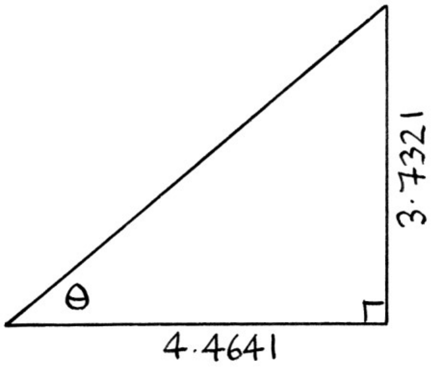
The vertical distance of the phase point of the second, smaller, circle *to its own centre* is:
2 sin ((360 * 0) + 60) = 2 sin 60 = 1.7321 units.

The horizontal distance from the phase point to *its circle's own centre* is:
2 cos ((360 * 0) + 60) = 2 cos 60 = 1 unit.

Therefore, the overall y-axis distance from the outer circle's phase point to the centre of the inner circle is:
2 + 1.7321 = 3.7321 units.

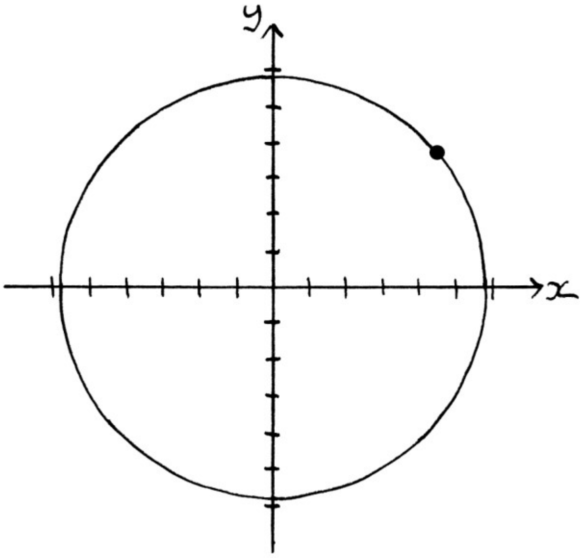
The overall x-axis distance from the outer circle's phase point to the centre of the inner circle is:
3.4641 + 1 = 4.4641 units.

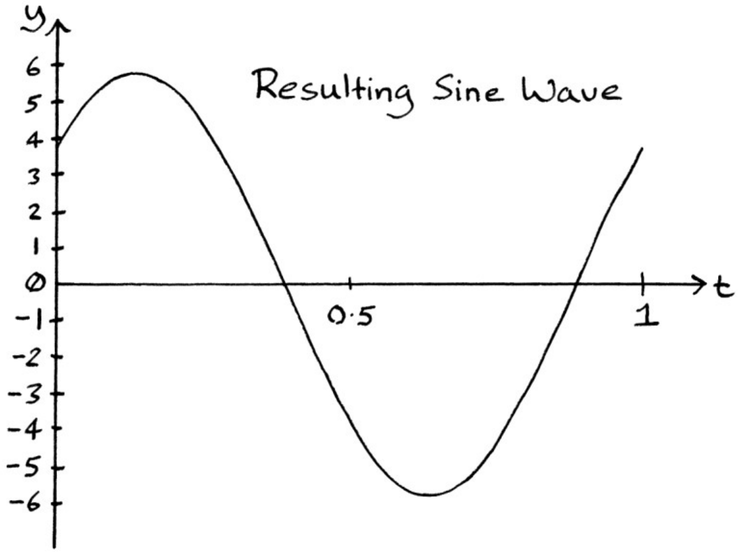
From all of that, we now have two sides of a right-angled triangle:

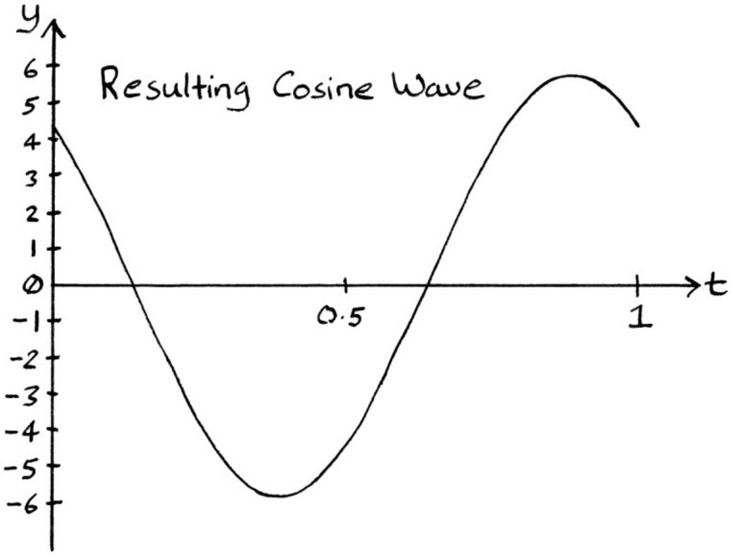


We can calculate the angle using arctan. The angle is arctan (3.7321 ÷ 4.4641), which, if we use the full accuracy of those rounded up numbers, is 39.8961 degrees. Given how two angles produce the same gradient, we need to think about whether this (or 39.8961 + 180 = 219.8961) is the result we need. As the phase point is in the top right quarter of the circle, it means that 39.8961 degrees is the result that we want.

Therefore, the phase point of the resulting circle is 39.8961 degrees, and the phases of the Sine and Cosine waves derived from that circle will also have phases of 39.8961 degrees.

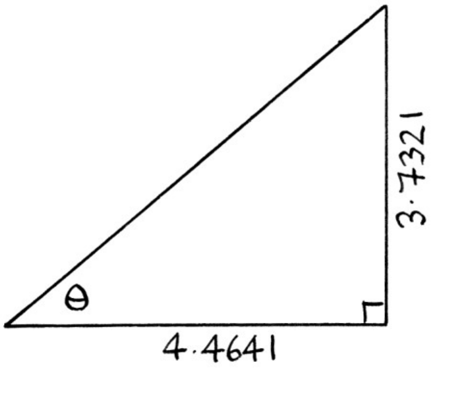






Calculating the amplitude

The amplitude of the resulting circle will be equal to the direct distance of the phase point of the outer object from the centre of the inner circle. Therefore, we can think of the triangle again, and we just need to work out the length of the hypotenuse side:

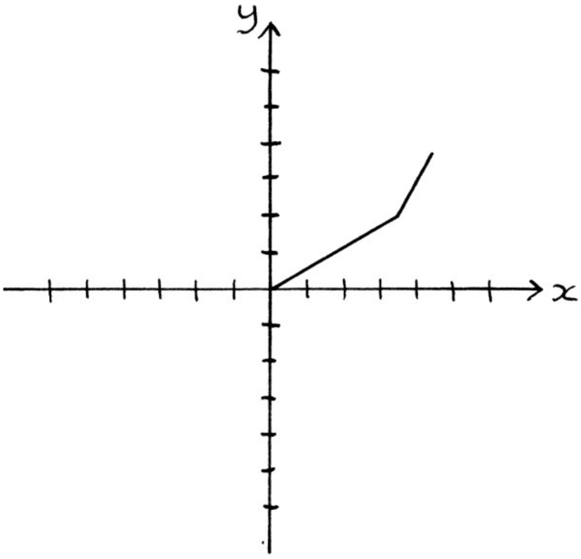


The opposite side is 3.7321 units long; the adjacent side is 4.4641 units long. Using Pythagoras's Theorem (and the full unrounded values), we can calculate the hypotenuse as being 5.8186 units. This means that the resulting circle will have a radius of 5.8186 units, and the waves derived from the circle will have amplitudes of 5.8186 units.

We can now say that our circle has a radius of 5.8186 units, a phase of 39.8961 degrees, and a centre at (0, 0). The formulas of the derived waves are:
"y = 5.8186 sin (360t + 39.8961)"
... and:
"y = 5.8186 cos (360t + 39.8961)"

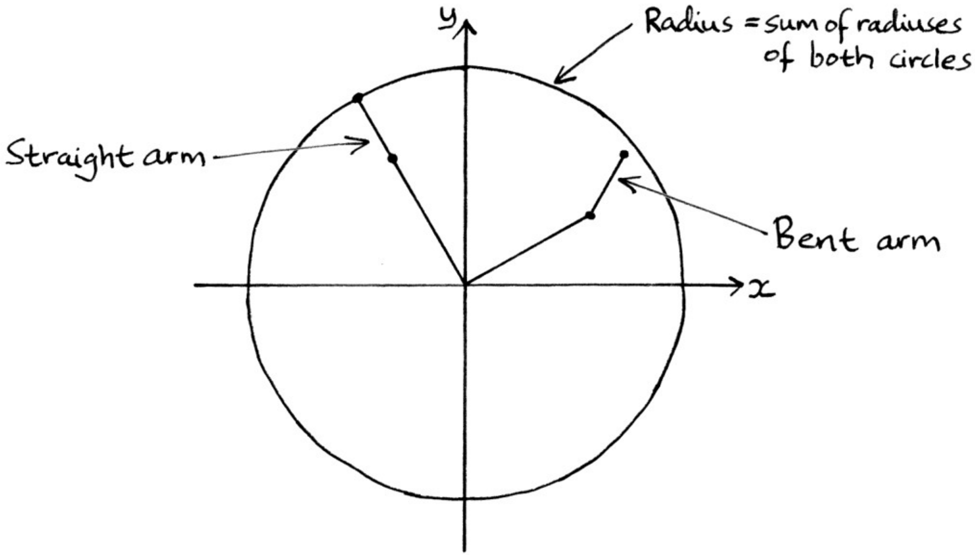
Arms

One interesting thing to notice is that if there is a phase difference between two added circles, the resulting radius will always be less than the sum of the two original radiuses. The resulting radius can never be equal to, or more than, the sum of the two original radiuses. The reason for this becomes clear when we think of the two phase points being connected by an "arm":



If the phases are different, then the arm will not be straight. The only time the resulting amplitude can be equal to the sum of the two added amplitudes is when the arm is straight.

The idea is portrayed in the following picture. The radius of this particular circle is the sum of the radiuses of the two underlying circles. A bent arm will never extend from the centre of the circle to the edge of this circle. Only a straight arm can do this. Two different phases will make a "bent arm" and so can never reach out to the edge of the circle.



If the phases are non-zero but identical, the arm will be straight, but if the phases are different, the arm will not be straight.

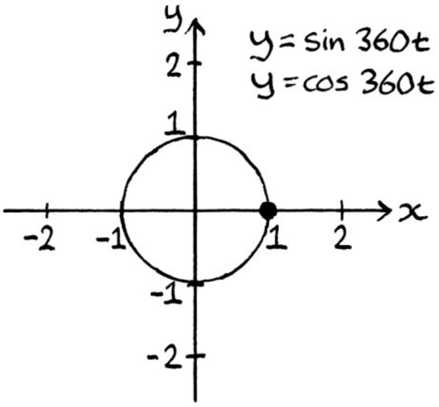
This means that if there is a phase difference between two added *waves*, the resulting *amplitude* will always be less than the sum of the two amplitudes. In such a case, the resulting amplitude can never be equal to the sum of the amplitudes. [It is also worth remembering that the resulting amplitude can never be more than the sum of the amplitudes, no matter what the amplitude, frequency, phase or mean level of the added waves.]

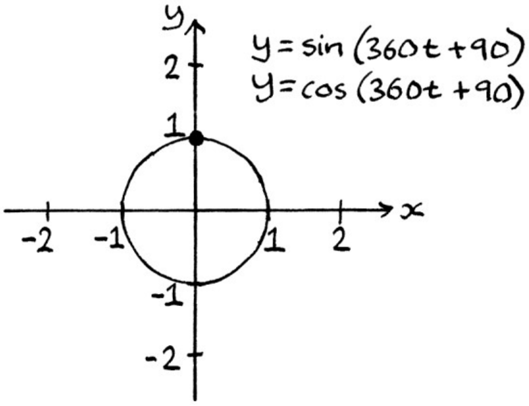
Sine added to Sine + 90

In the last chapter, we added a Sine wave and a Cosine wave, which amounts to adding a Sine wave and a Sine wave with a 90-degree phase in its formula. The exact formulas we added were "y = sin 360t" and "y = cos 360t". Now we will convert these into circles and add them as circles.

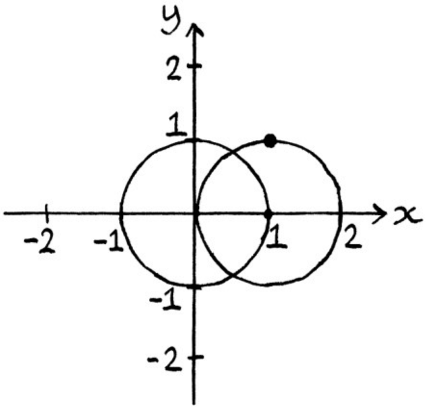
To make them into circles, we have to convert them into the same type of wave. Therefore, we will make them into Sine waves – they become:
“y = sin 360t”
... and:
“y = sin (360t + 90)”.

The two circles that portray these Sine waves are as follows:

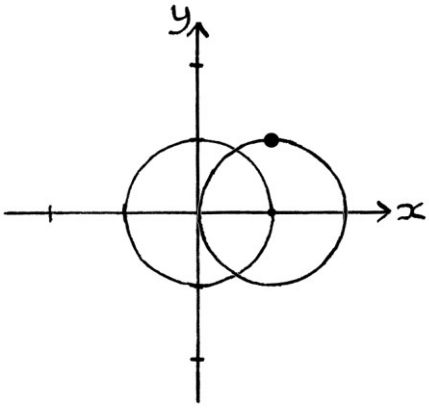




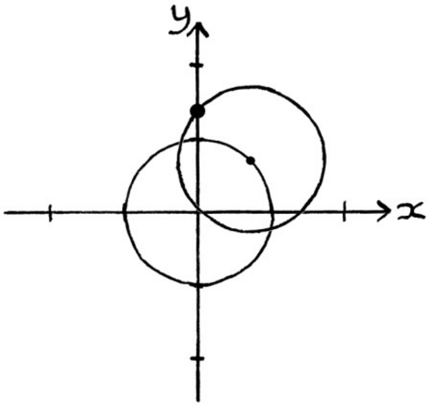
The circles have the same radius, so it is a matter of choice as to which way around we arrange them. We will arrange them with the circle with zero phase on the origin of the axes, and the circle with a 90-degree phase centred over the first circle’s phase point. As always, it does not actually matter which way around they are placed, but sometimes one way looks clearer than the other way.



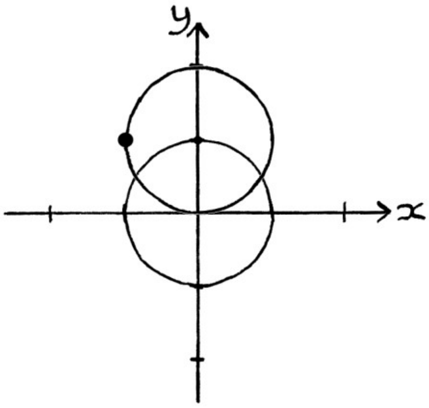
The outer object rotates around its circle, and that circle rotates around the inner circle. At t = 0 seconds, they look like this:



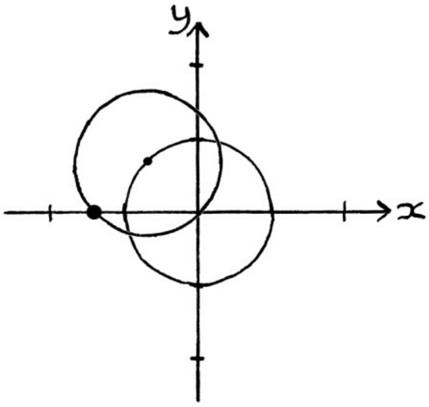
At t = 0.125:



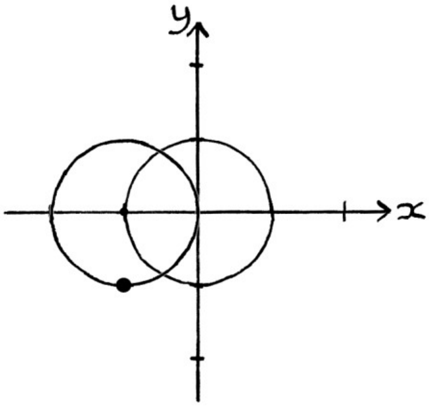
At t = 0.250:



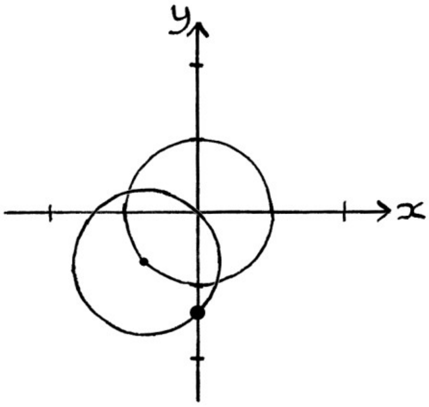
At t = 0.375:



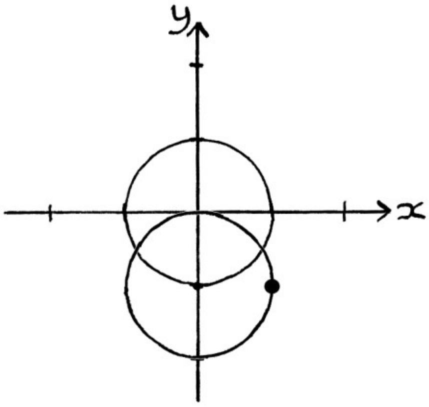
At t = 0.5:



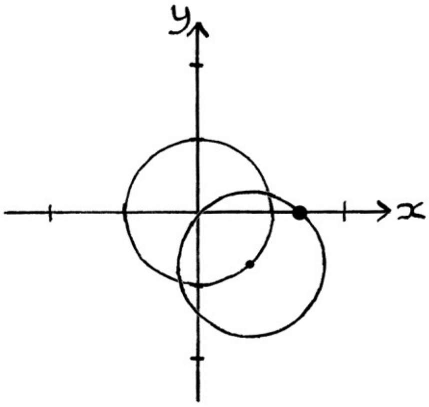
At t = 0.625:



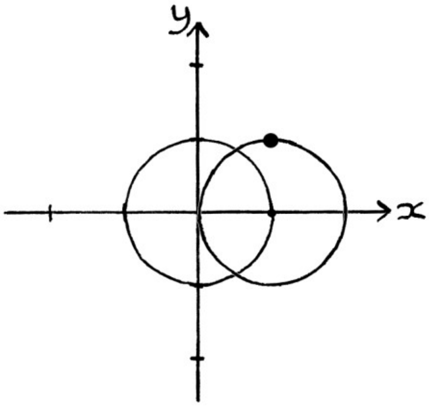
At t = 0.750:



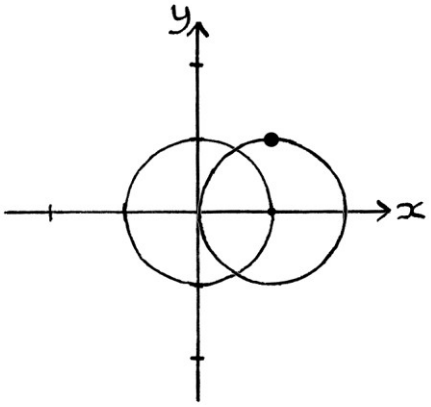
At t = 0.875:

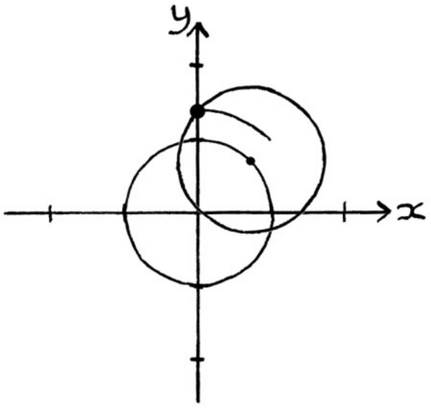


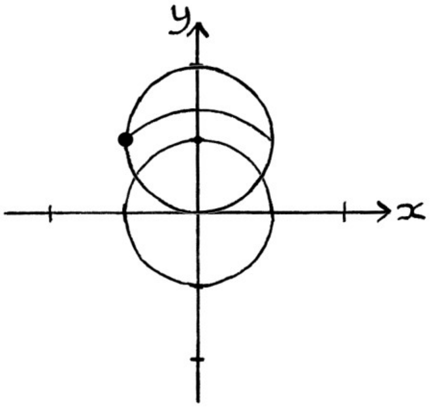
At t = 1 second:

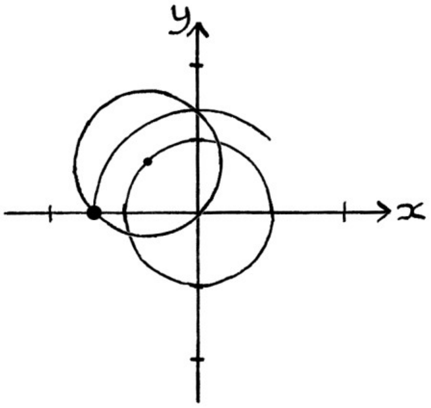


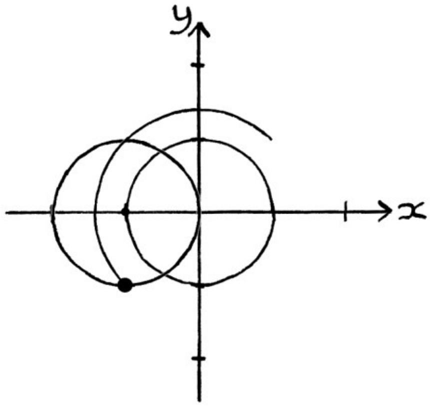
If we observe the outer object on its journey, it will draw out a circle:

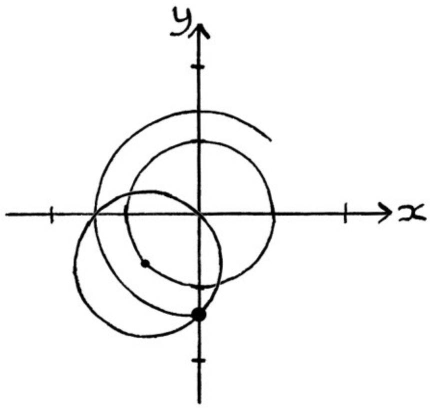


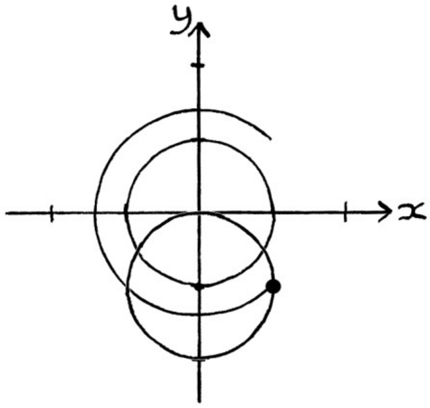


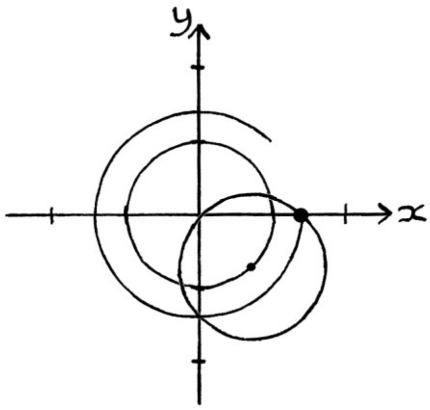


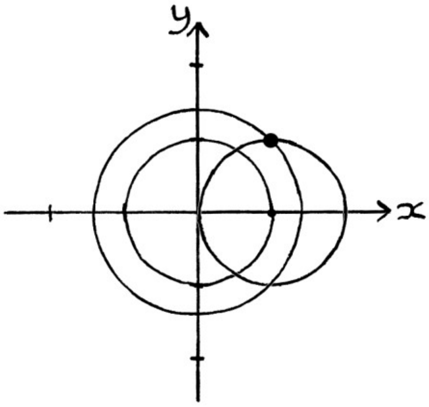



y
x

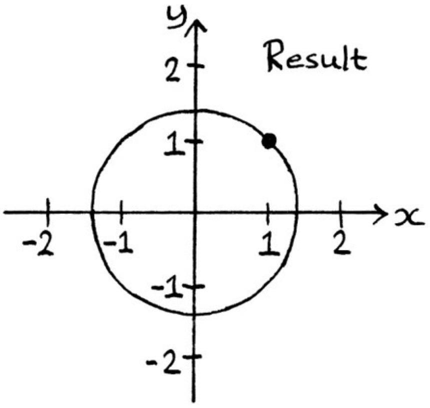

y
x


y
x

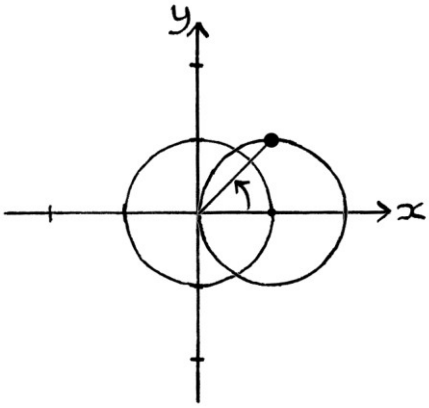

y
x



The resulting circle looks like this:



The phase of the resulting circle can be calculated by finding the angle of the outer circle’s phase point with reference to the centre of the inner circle (which in this case is also the origin of the axes) at t = 0.



To do this we can calculate the x-axis and y-axis distance of the outer circle’s phase point to the centre of the inner circle’s centre, and treat those values as the adjacent and opposite sides of a right-angled triangle.

The x-axis distance from the outer object to the centre of *its own* circle at t = 0 is:
cos (0 + 90) = cos 90 = 0 units.

The x-axis distance from the inner circle's phase point (which is the centre of the outer circle) to the centre of the inner circle is:
cos (0 + 0) = cos 0 = 1 unit.

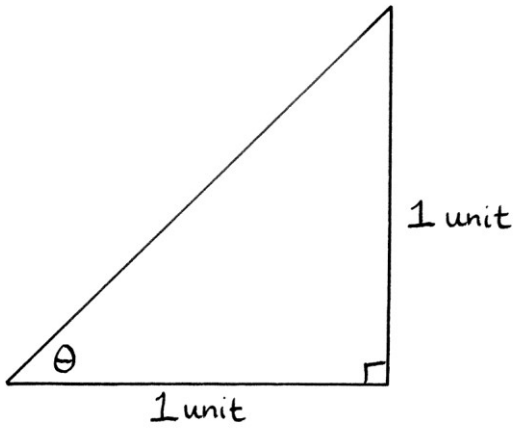
Therefore, the outer object is 1 horizontal unit away from the centre of the inner circle.

The y-axis distance from the outer object to *its own* circle's centre at t = 0 is:
sin (0 + 90) = 1.

The y-axis distance from the inner object (which is the centre of the outer circle) to its circle's centre is:
sin (0 + 0) = 0.

Therefore, the outer object is 1 vertical unit away from the centre of the inner circle.

We can portray these values on a right-angled triangle:



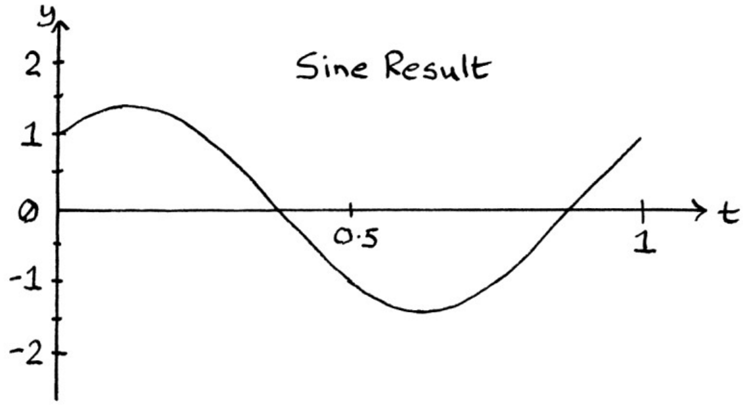
The phase, which is the angle of the triangle, is calculated with:
arctan (1 ÷ 1) = 45 degrees.

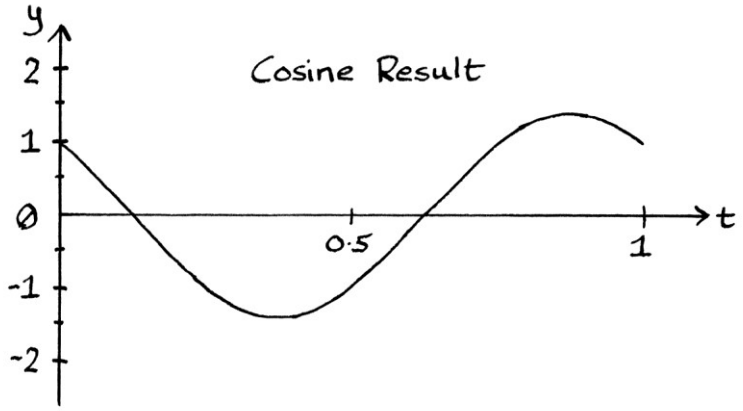
As always when using arctan, we think about this result to see if this is the correct angle for the gradient we gave to arctan (and not 45 + 180 = 225 degrees), and it is, so the phase is definitely 45 degrees.

The radius of the resulting circle will be the hypotenuse of the triangle, which we can calculate using Pythagoras's Theorem. This is the square root of 1 + 1, which is approximately 1.4142.

Therefore, the resulting circle has a radius of 1.4142, a phase point at 45 degrees and is centred at (0, 0). The waves derived from this circle will have the formulas:
"y = 1.4142 sin (360t + 45)"
... and:
"y = 1.4142 cos (360t + 45)"

The waves look like this:



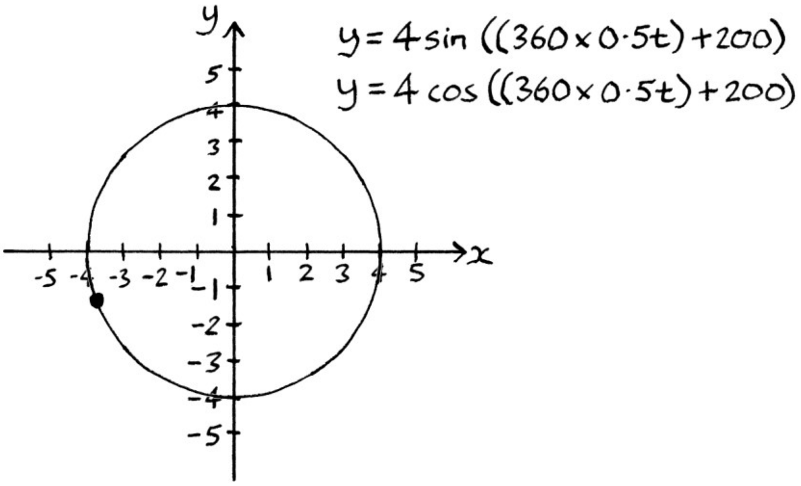


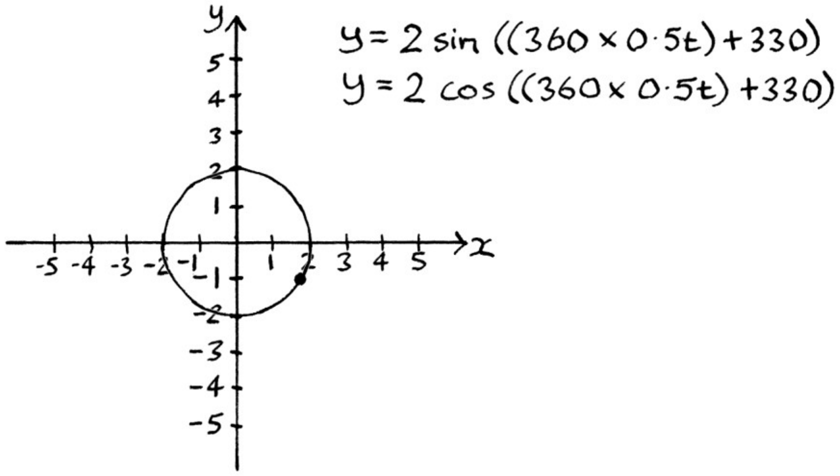
Without thinking of the circle, when adding a Sine wave and a Sine wave with a 90-degree phase in its formula, it would be much harder to understand why the amplitude of the result would be 1.4142 and why the phase of the result would be 45 degrees. Now we are using circles, everything becomes a lot more intuitive.

Another example

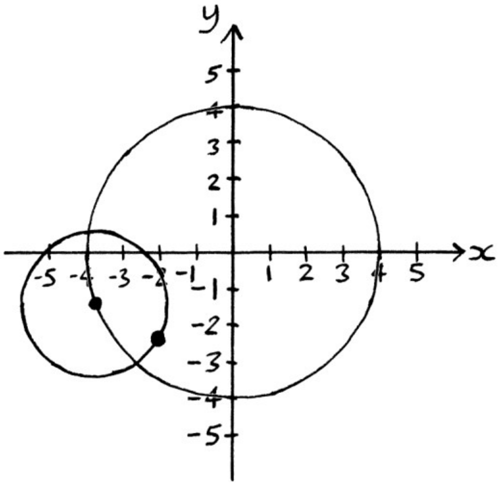
We will add the circles representing:
“y = 4 sin ((360 * 0.5t) + 200)” and “y = 4 cos ((360 * 0.5t) + 200)”
... and:
“y = 2 sin ((360 * 0.5t) + 330)” and “y = 2 cos ((360 * 0.5t) + 330)”

The two circles look like this:

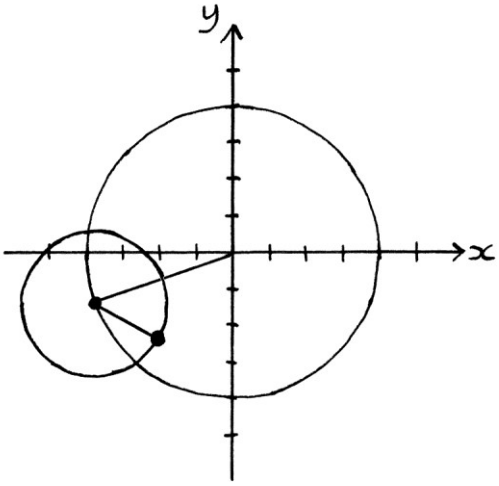

y = 4 sin ((360 × 0·5t) + 200)
y = 4 cos ((360 × 0·5t) + 200)
y
x
5
4
3
2
1
-1
-2
-3
-4
-5
-5 -4 -3 -2 -1 1 2 3 4 5


y = 2 sin ((360 × 0·5t) + 330)
y = 2 cos ((360 × 0·5t) + 330)
y
x
5
4
3
2
1
-1
-2
-3
-4
-5
-5 -4 -3 -2 -1 1 2 3 4 5

When combined, they look like this:



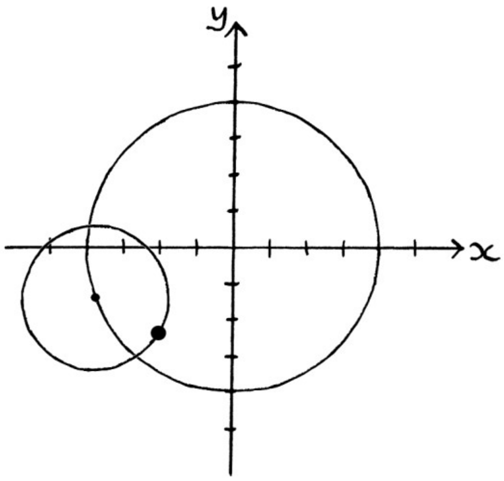
Note how the difference in phases means that the object on the outer circle is actually within the perimeter of the inner circle. The "arm" for this pair of circles ends up pointing back towards the centre of the inner circle:



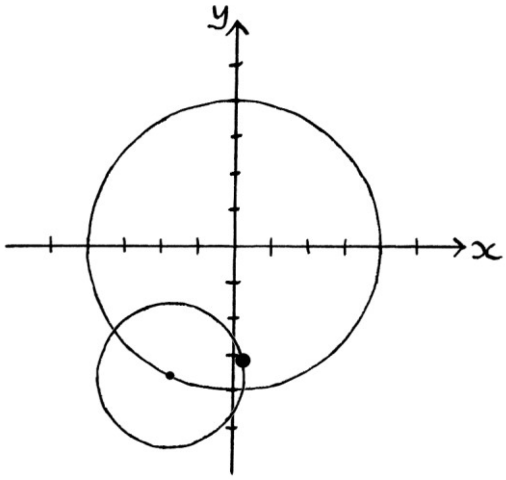
This means that the resulting circle will be smaller than the inner circle.

The following pictures show the movement of the object on the outer circle as it rotates around its circle, and as that circle moves around the inner circle.

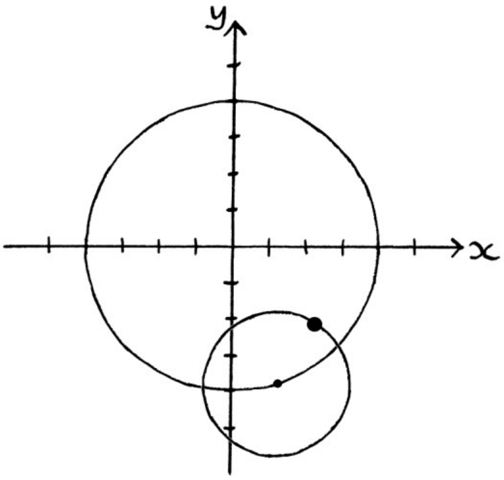
At 0 seconds:



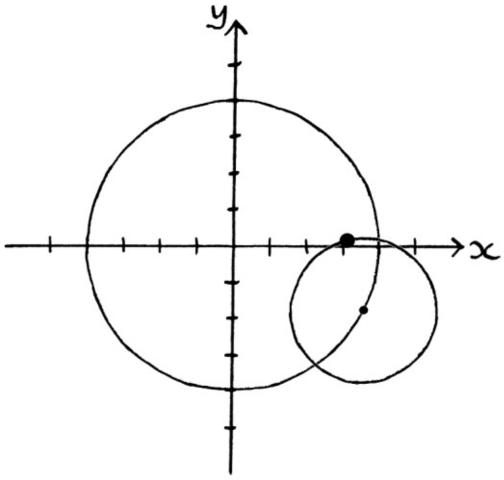
At 0.25 seconds:



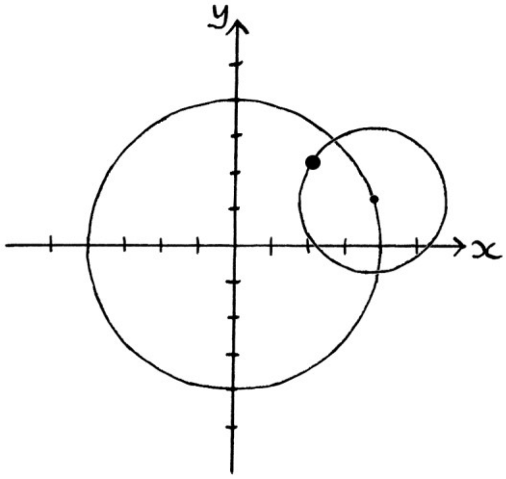
At 0.5 seconds:



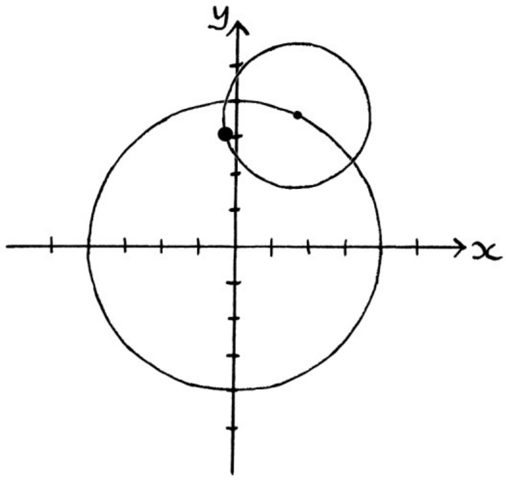
At 0.75 seconds:



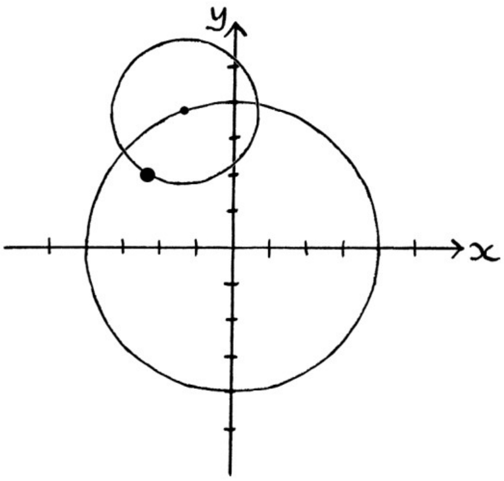
At 1 second:



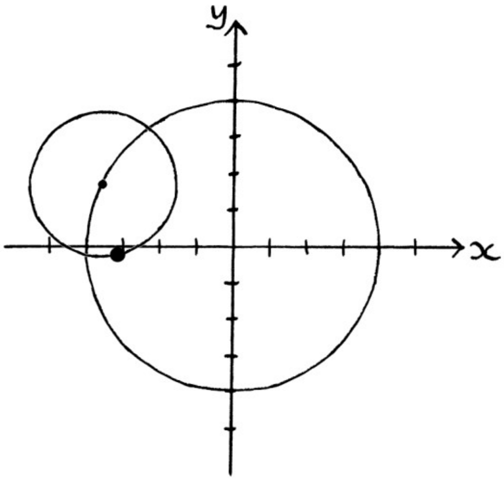
At 1.25 seconds:



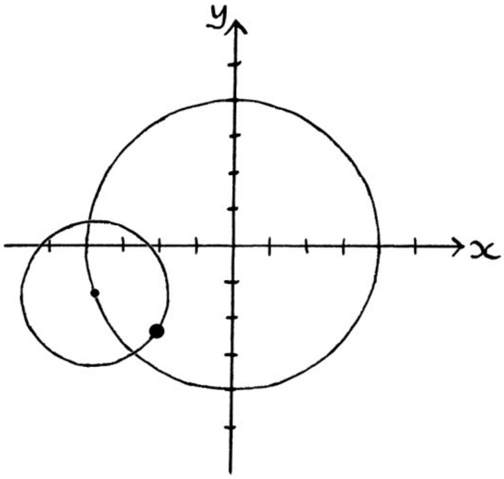
At 1.5 seconds:



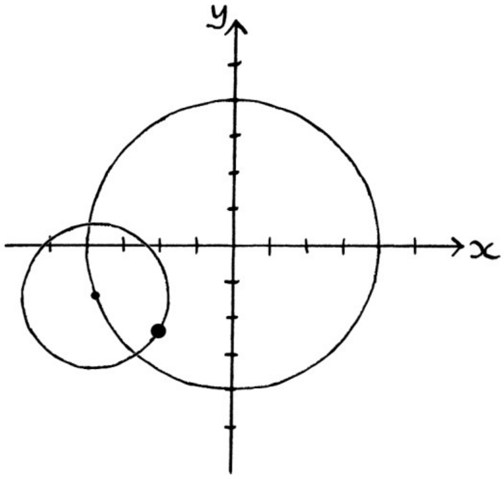
At 1.75 seconds:

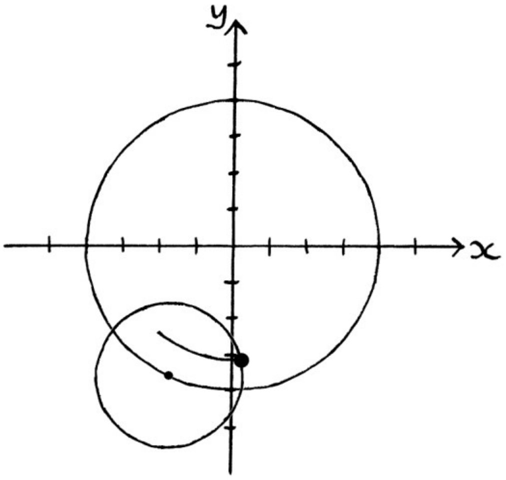


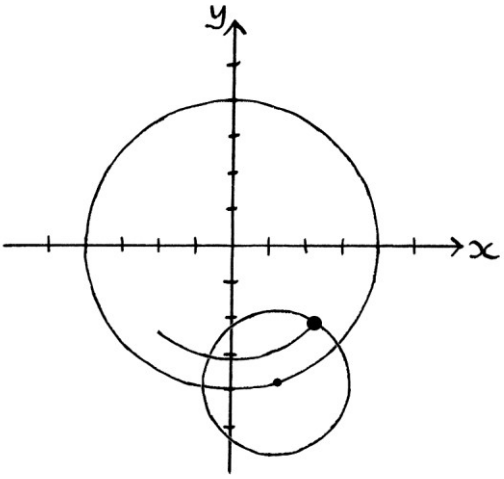
At 2 seconds:

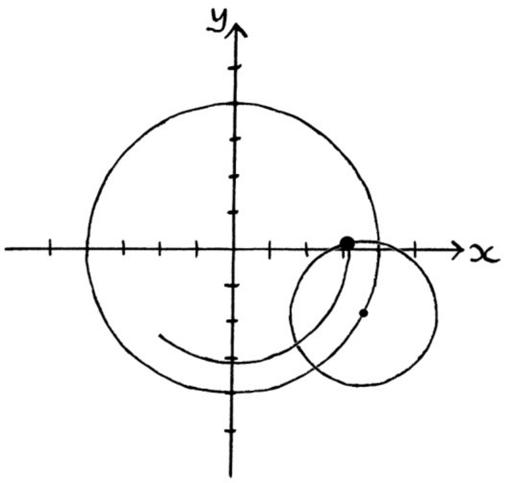


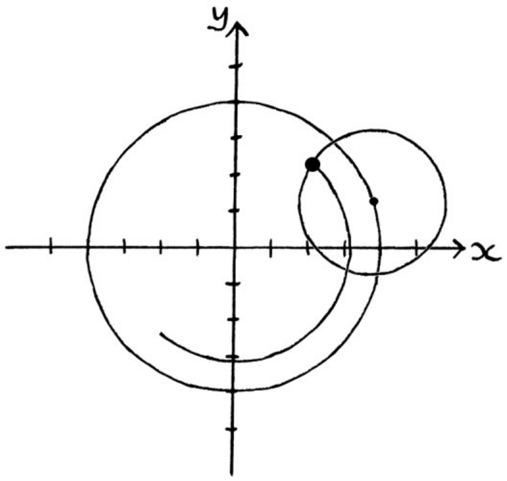
As the outer object moves around its circle, and as that circle moves around the inner circle, the outer object draws out a circle. The following pictures show the state of the two circles at intervals of 0.25 seconds:

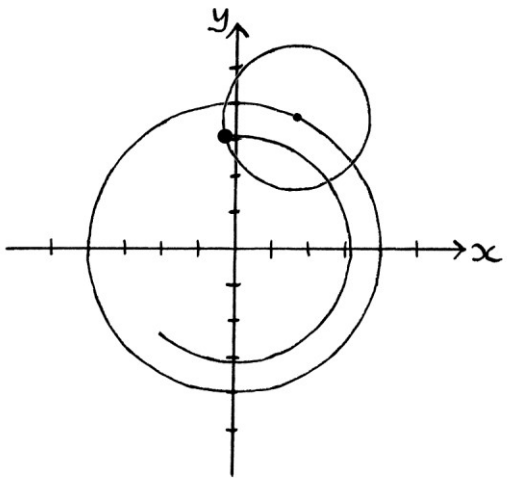


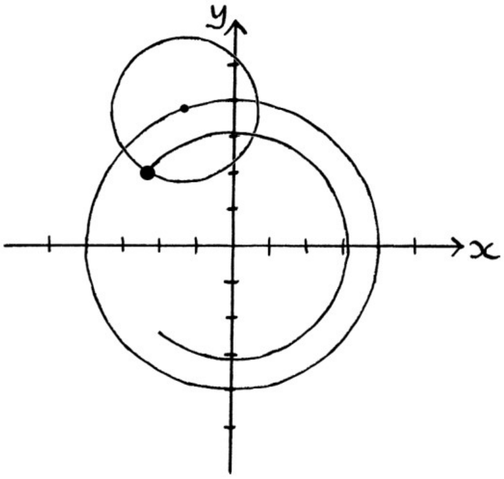


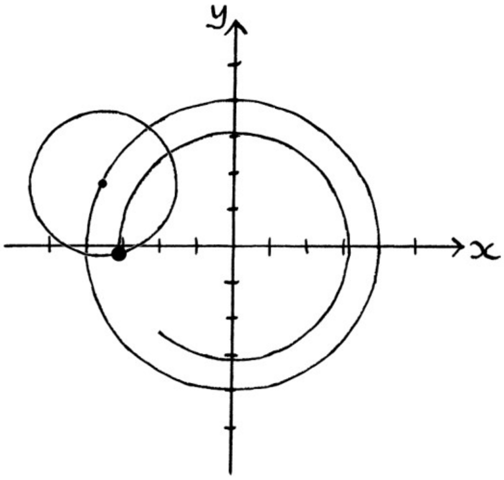


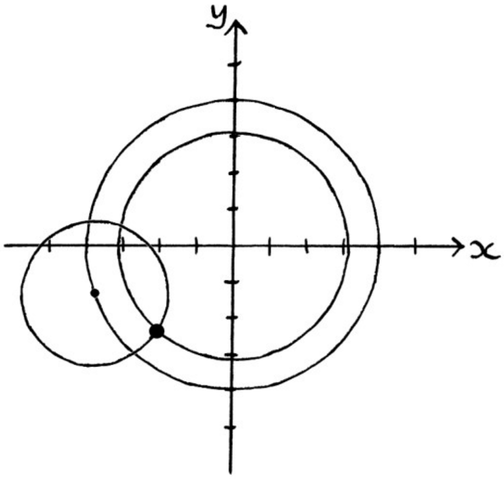

y
x


y
x

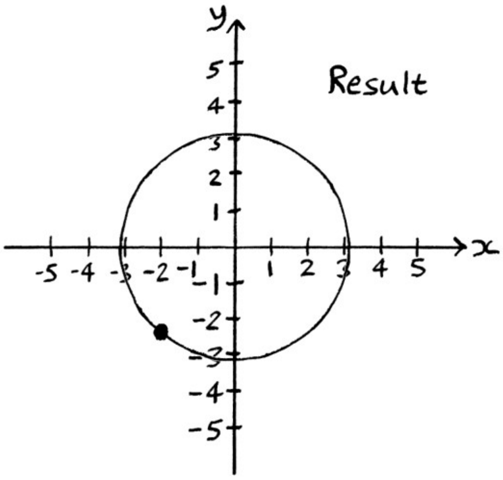

y
x


y
x

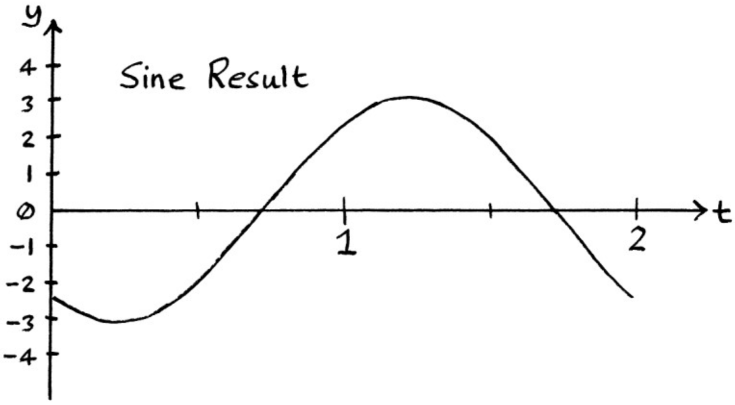

y
x

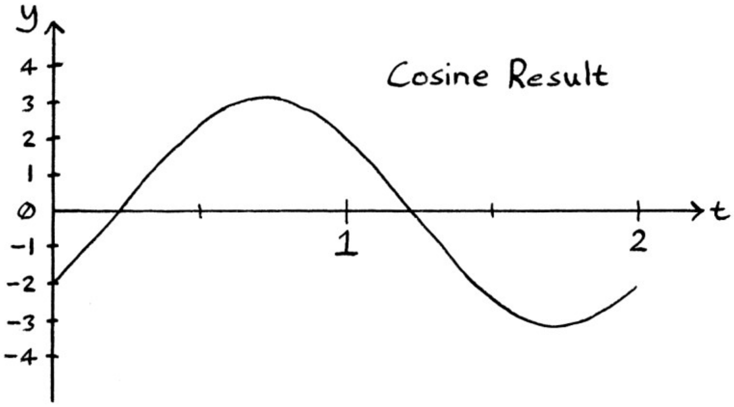

y
x

The resulting circle looks like this:

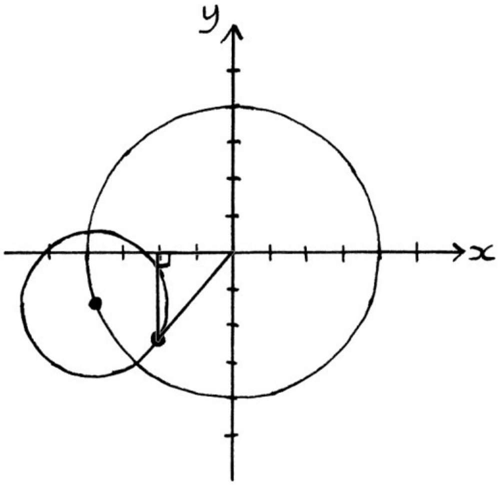


The two waves derived from this resulting circle look like this:





We can calculate the phase of the resulting circle by thinking of the two original circles at t = 0, and by using a right-angled triangle.



At t = 0, the horizontal distance of the phase point of the outer circle to *its own* centre is:
2 cos ((360 * 0) + 330) = 2 cos 330 = 1.7321 units.

The horizontal distance of the centre of the outer circle to the centre of the inner circle is:
4 cos ((360 * 0) + 200) = 4 cos 200 = −3.7588 units.

Therefore, the phase point of the outer circle is:
−3.7588 + 1.7321 = −2.02672 horizontal units away from the centre of the inner circle.

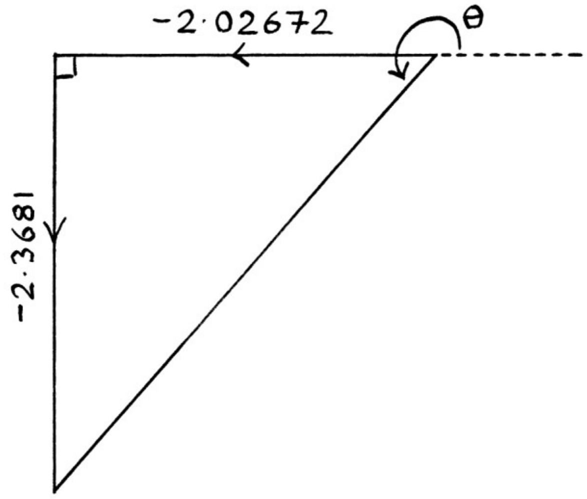
At t = 0, the vertical distance of the phase point of the outer circle to *its own* centre is:
2 sin ((360 * 0) + 330) = 2 sin 330 = −1 units.

The vertical distance of the centre of the outer circle to the centre of the inner circle is:
4 sin ((360 * 0) + 200) = 4 sin 200 = −1.3681 units.

Therefore, the phase point of the outer circle is:
−1 – 1.3681 = −2.3681 vertical units away from the centre of the inner circle.

Therefore, our right-angled triangle has an adjacent side of −2.02672 units and an opposite side of −2.3681 units. The triangle is upside down and the wrong way around.

The triangle looks like this, with the phase angle that we want to calculate marked:



The angle can be calculated with:
arctan (−2.3681 ÷ −2.02672) = 49.4415 degrees.

We think about whether we want this angle or the other angle that would produce the same gradient (49.4415 + 180 = 229.4415 degrees). As the phase point of the outer circle is in the bottom-left quarter of the circle, it means we want 229.4415 degrees as our result.

The radius of the circle will be the hypotenuse of the triangle. We can use Pythagoras's theorem to calculate this from the opposite and adjacent sides. The result is 3.1170 units.

Therefore, the resulting circle has a radius of 3.1170 units, a phase of 229.4415 degrees, and is centred at (0, 0). The derived waves will have amplitudes of 3.1170 units, phases of 229.4415 degrees and mean levels of zero units. They are:
"y = 3.1170 sin ((360t *0.5t) + 229.4415)"
... and:
"y = 3.1170 cos ((360t *0.5t) + 229.4415)"

Phases and mean levels

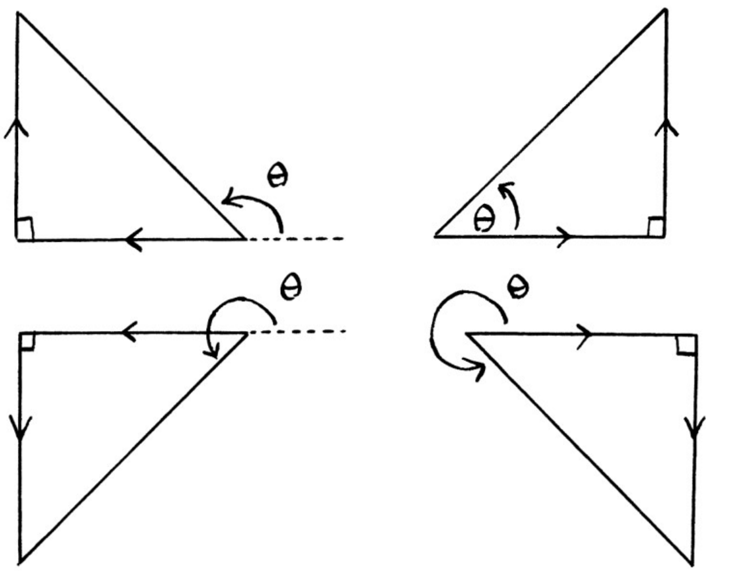
If there are no mean levels in the added circles, then we can calculate the radius and phase point of the circle as before. However, if there are non-zero mean levels, we need to group the mean levels together, and then calculate the angle and distance from the outer circle's phase point to *the centre of the inner circle*. We were doing this before, but then it just so happened that the centre of the inner circle was also the origin of the axes. If there is a non-zero mean level, the centre of the inner circle will not be on the origin of the axes. The mean levels are independent of the phase and radius, and would need to be added on to any graph formulas afterwards.

Phase addition summary

A summary of how to calculate the phase and radius of a circle made from adding two other circles is as follows:

Draw (or think of) a right-angled triangle that shows how the outer object is connected to the centre of the inner circle at t = 0.

The opposite side is the vertical distance of the phase point of the outer circle from the centre of the inner circle. The adjacent side is the horizontal distance of the phase point of the outer circle from the centre of the inner circle. The phase of the resulting circle will be the angle of the hypotenuse with respect to 0 degrees on the circle. This means if the triangle is the wrong way round or upside down, we measure the angle outside the triangle:



The amplitude of the resulting circle will be the hypotenuse of the right-angled triangle.

All of this can be summarised more mathematically as:

If we have the two waves:
$h_1 + A_1 \sin (360ft + \phi_1)$
... and:
$h_2 + A_2 \sin (360ft + \phi_2)$
... where f is the same for each wave, then:

The resulting phase will be:
$\arctan ((A_1 \sin \phi_1 + A_2 \sin \phi_2) \div (A_1 \cos \phi_1 + A_2 \cos \phi_2))$

[As always with arctan, remember to think about the circle to make sure that the result is the correct one of the two possible results. Also, if either value in the division is zero, we can know the angle from thinking about a circle, so we do not need to bother using arctan.]

The resulting amplitude will be:
The square root of: $(A_1 \sin \phi_1 + A_2 \sin \phi_2)^2 + (A_1 \cos \phi_1 + A_2 \cos \phi_2)^2$

The resulting mean level will be:
$h_1 + h_2$

The resulting frequency will be the same as in the two original waves:
f.

Potential sources of confusion so far

A circle does not represent a second

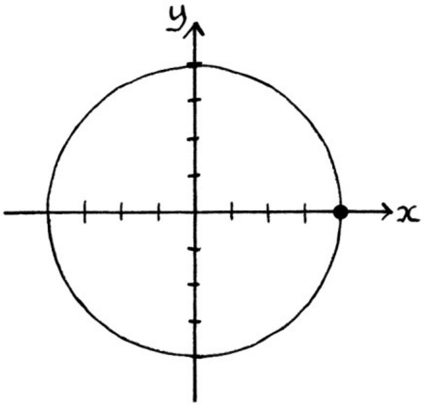
When adding circles, it pays to remember that the circles represent a rotation of 360 degrees. They *do not* represent a time. In some of the previous examples, where objects were rotating at one cycle per second, 360 degrees would have been completed in one second, and that can trick people into thinking that the circle represents time. Any confusion can be diminished when we think about what happens if the frequency is 2 cycles per second – the object rotating around the outer circle will complete two full circles every second, in the same way that it will complete two full cycles on the wave graphs.

If there is a circle, it always represents one cycle. This might seem obvious, but sometimes it is easy to become confused.

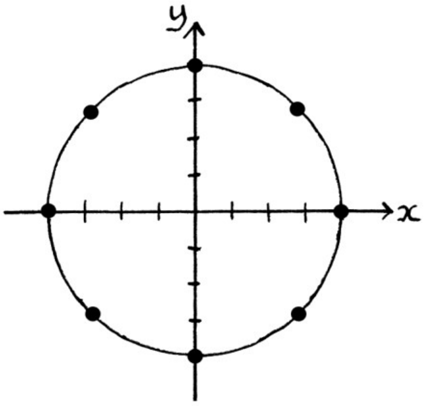
We cannot deduce the waves from a static circle

If we are given a picture of a completed circle and no other information, it is impossible to know the frequency or what the derived Sine and Cosine waves look like. This is because, without other information, we cannot know the time when the outer object was at any particular place on the circle's edge. [We can, however, know the phase.]

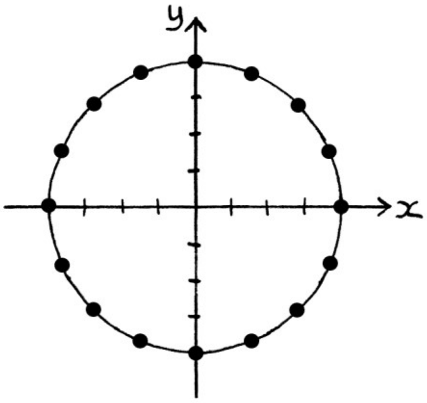
For example, this is a picture of a resulting circle:



If an object were travelling at 0.5 cycles per second, its position along the circumference would be as so at every 0.25 seconds:



If an object were travelling at 0.25 cycles per second, its position along the circumference would be as so at every 0.25 seconds:



These are different frequencies, but the same resulting circle. The derived waves would have different frequencies. We need to know where the object was at any particular time to create the derived Sine and Cosine waves.

Different frequencies

Using circles to add two sets of waves that have different frequencies is slightly more complicated than if they have the same frequency, but there are much more interesting results.

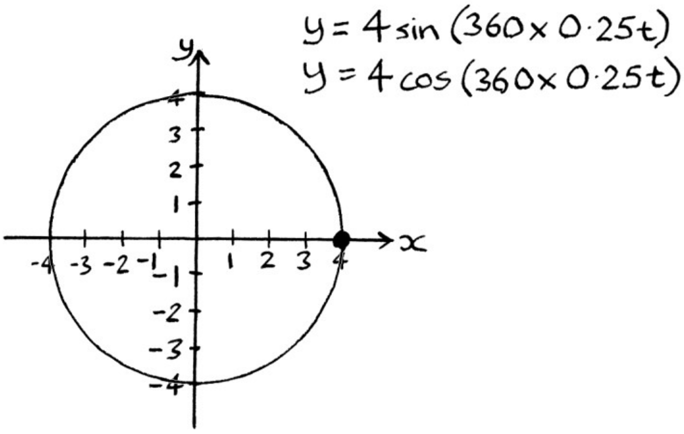
The main thing to know is that the object rotating around the inner circle rotates at a different speed to the object rotating around the outer circle. Therefore, the outer object might complete several revolutions in the time it takes the inner object to complete just one revolution, or conversely the inner object might complete several revolutions in the time it takes the outer object to complete just one revolution. The outer object might move backwards and forwards across the perimeter of the inner circle as it moves. This movement creates much more complicated resulting shapes than when the frequencies are the same. It is important to know that if the objects have different frequencies, the resulting shape cannot be a circle.

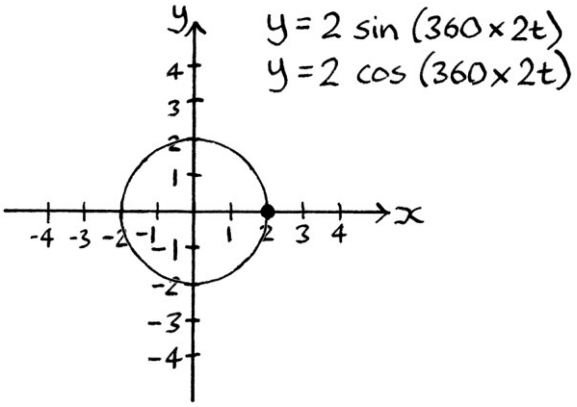
This is easiest to explain with examples.

Example 1

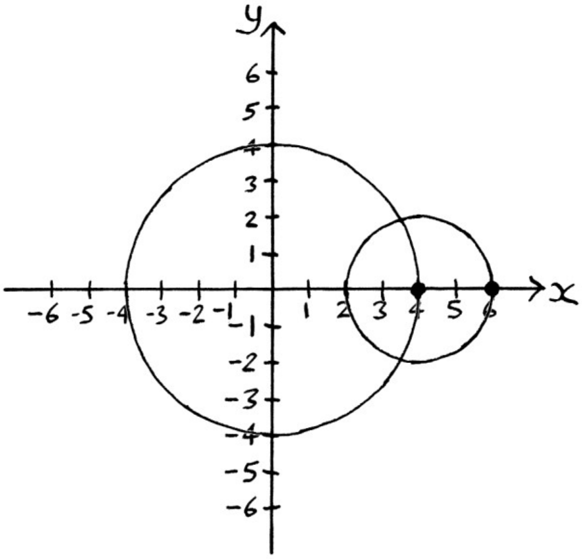
We will add the circles that represent the waves:
“y = 4 sin (360 * 0.25t)” and “y = 4 cos (360 * 0.25t)”
... and:
“y = 2 sin (360 * 2t)” and “y = 2 cos (360 * 2t)”

The circles look like this:



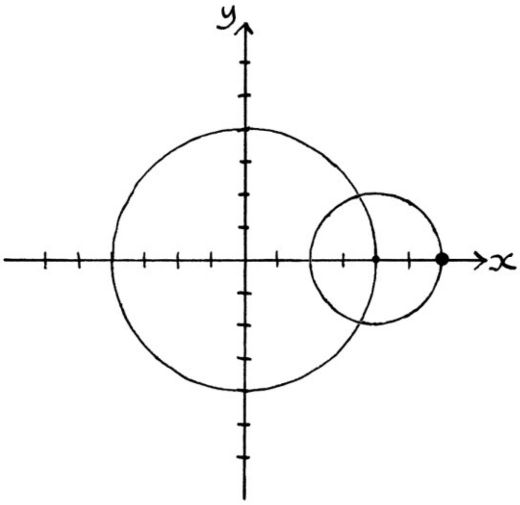


We will put the larger circle (which has the slower frequency) on the origin of the axes, and the smaller circle (which has the faster frequency) over the object rotating around the inner circle.

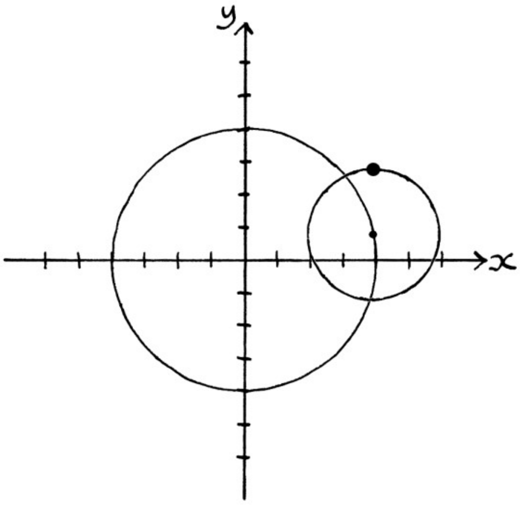


As the outer object rotates around its circle, and as that circle rotates around the inner circle, the two circles appear as in the following pictures. Notice how the outer object rotates at a faster rate than the inner object.

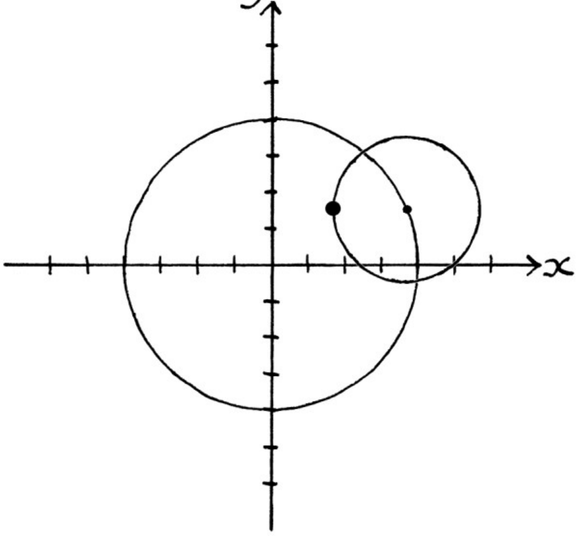
At t = 0:



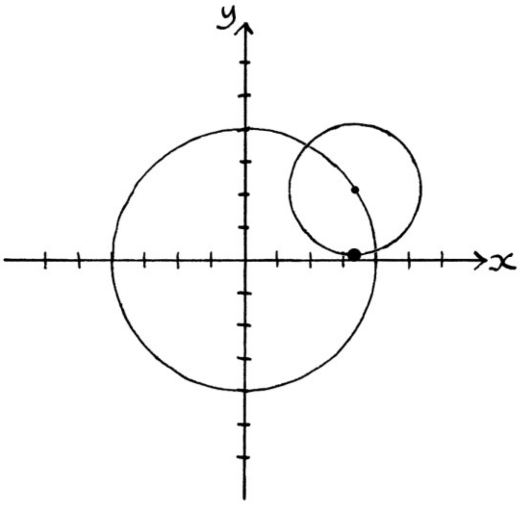
At 0.125 seconds:



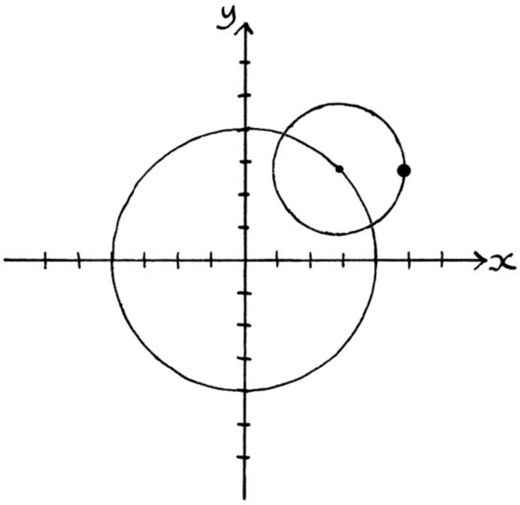
At 0.25 seconds:



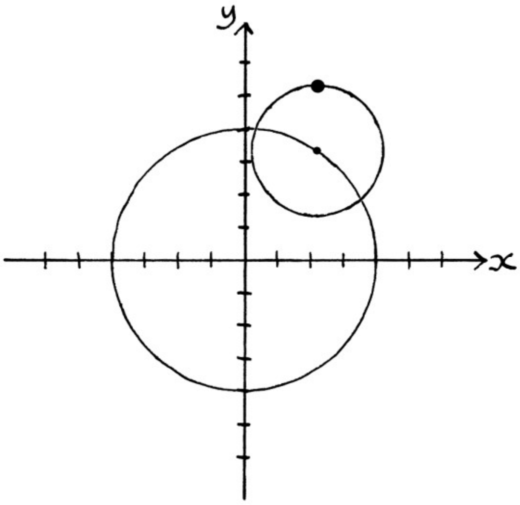
At 0.375 seconds:



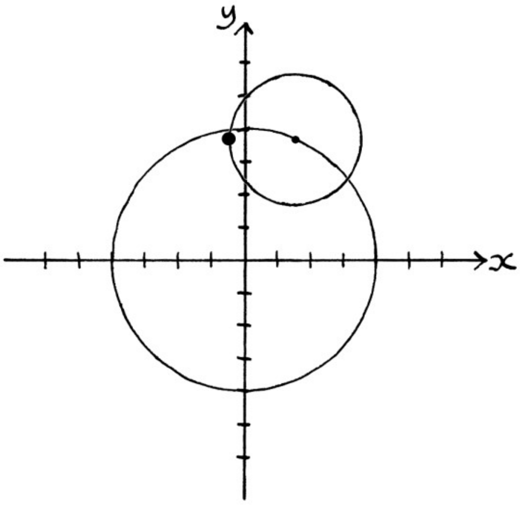
At 0.5 seconds. Notice how the outer object has completed one full revolution, while the inner object has only completed an eighth of a revolution.



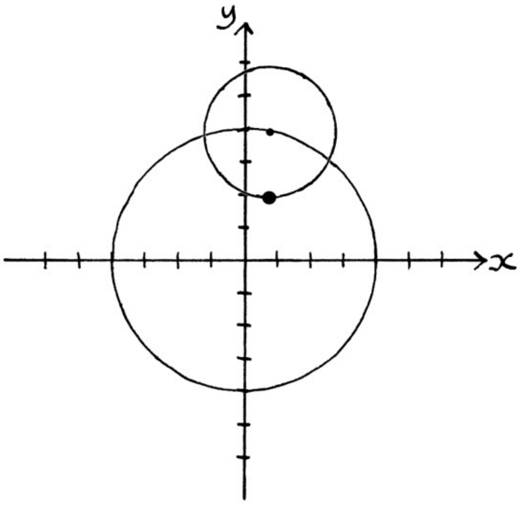
At 0.625 seconds:



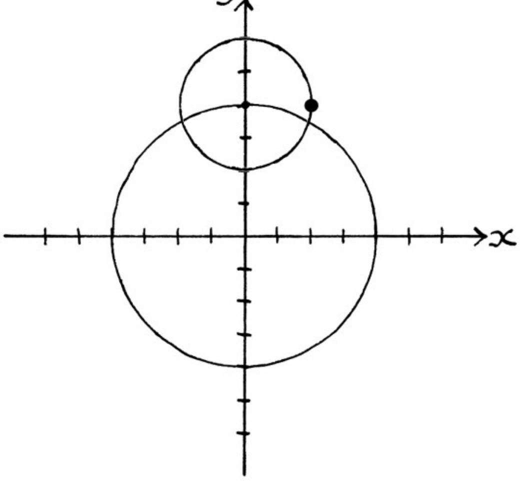
At 0.750 seconds:



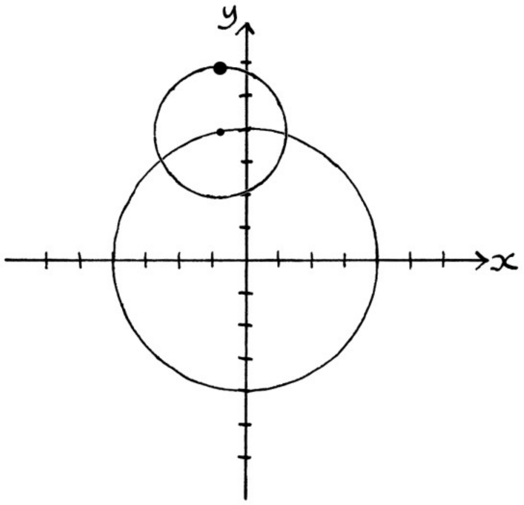
At 0.875 seconds:



At 1 second. Notice how the outer object has now completed two revolutions, while the inner circle has only completed a quarter of a revolution.



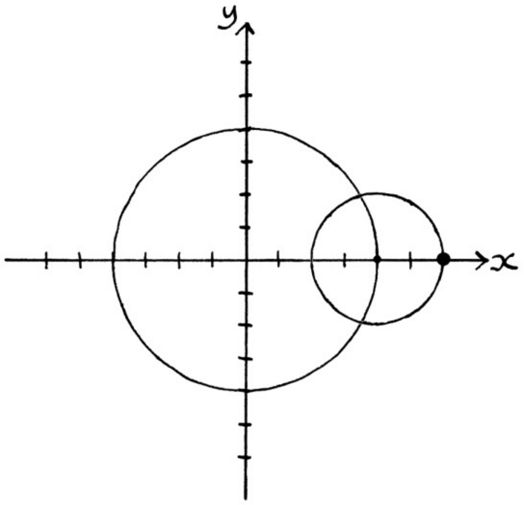
At 1.125 seconds:



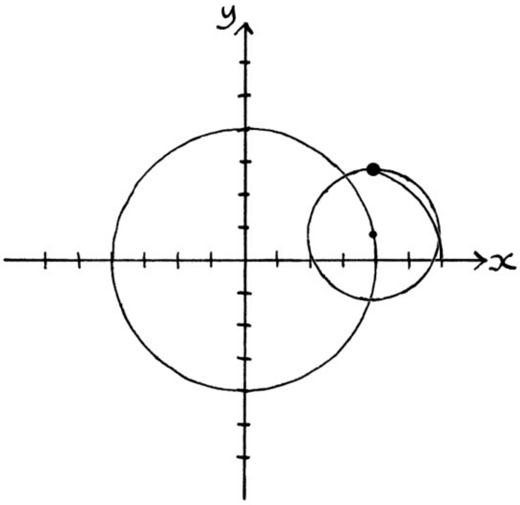
The circles continue to move around in this way until 4 seconds have passed, at which time, the outer object will have completed 8 revolutions around its own circle, and the inner object will have completed 1 revolution around its circle. Overall, the outer object will have completed 1 revolution around the centre of the inner circle, but it does this in a convoluted way.

As the outer object moves around its circle, and as that circle moves around the inner circle, the outer object draws out a shape, which, because the frequencies are different, is not a circle:

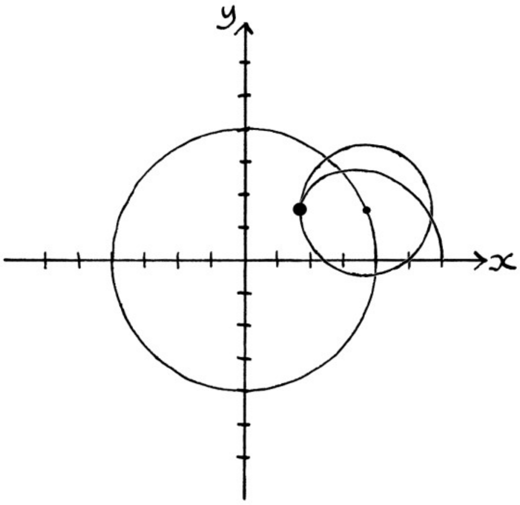
At t = 0:



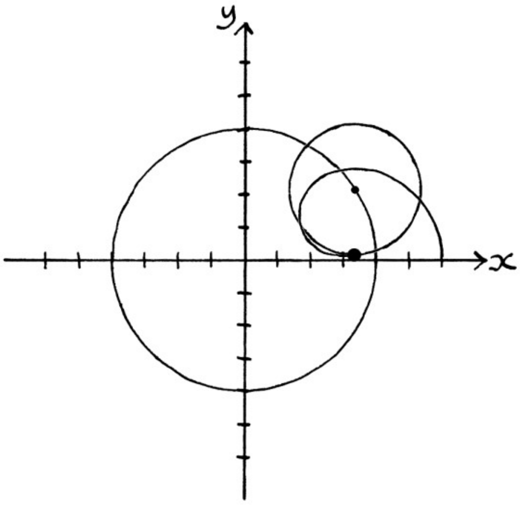
At t = 0.125 seconds:



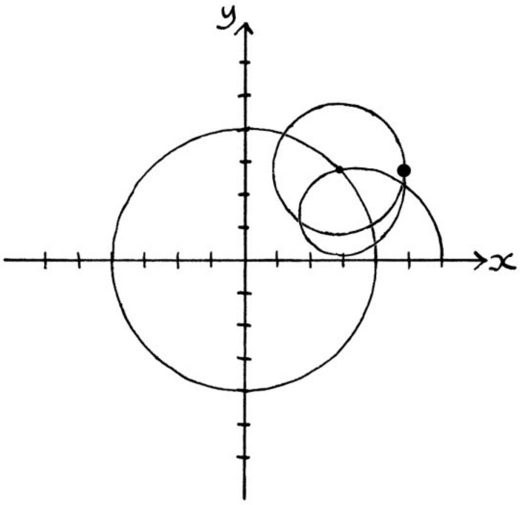
At 0.25 seconds:



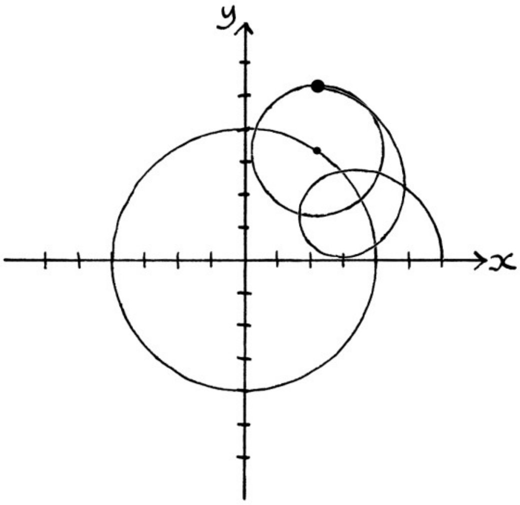
At 0.375 seconds:



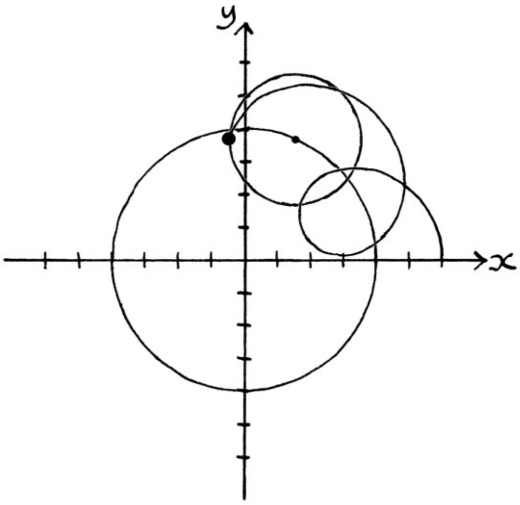
At 0.5 seconds:



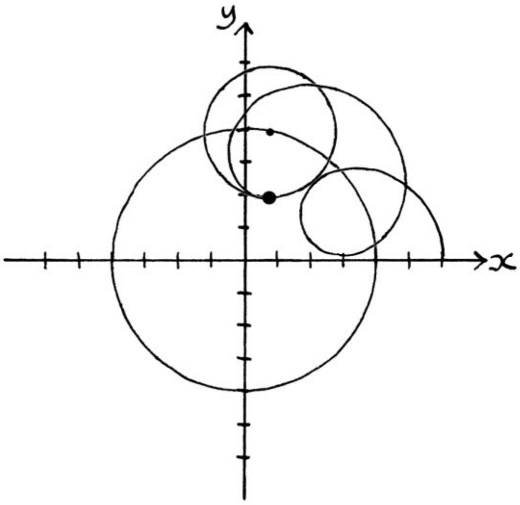
At 0.625 seconds:



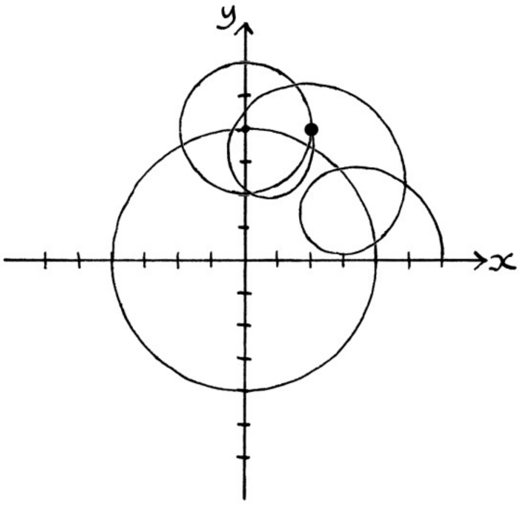
At 0.75 seconds:



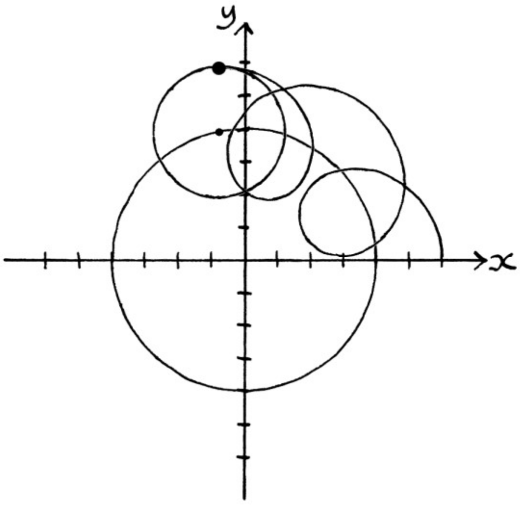
At 0.825 seconds:



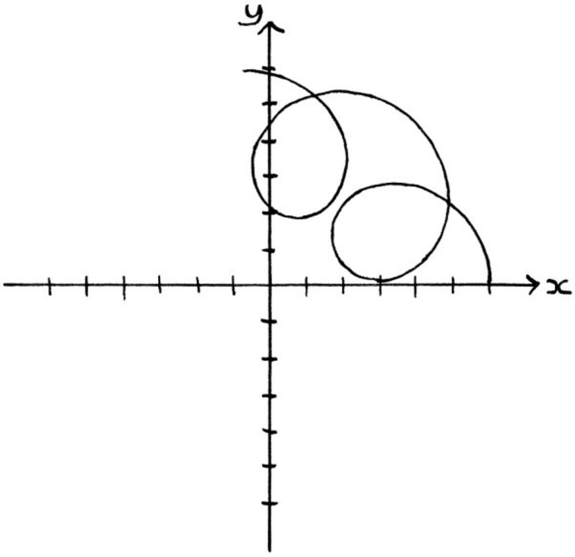
At 1 second:



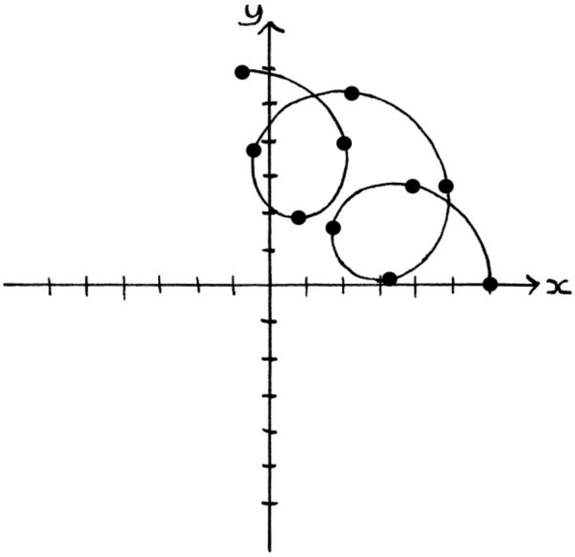
At 1.125 seconds:



The outer object continues to draw out the shape until 4 seconds have passed. Before we look at the final shape, we will look at the shape so far. At 1.125 seconds, the shape looks like this (drawn without the underlying circles):



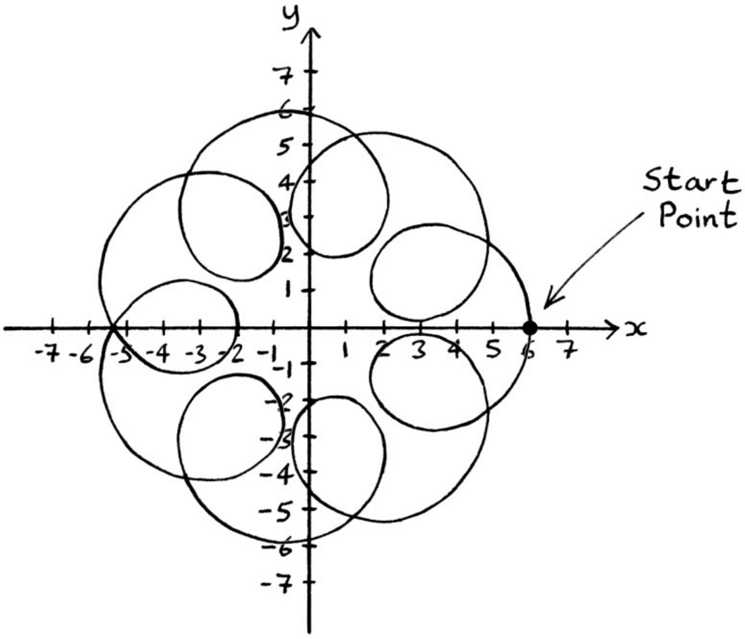
If we mark the position of the outer object on the shape at every 0.125 seconds, we will have the following picture:



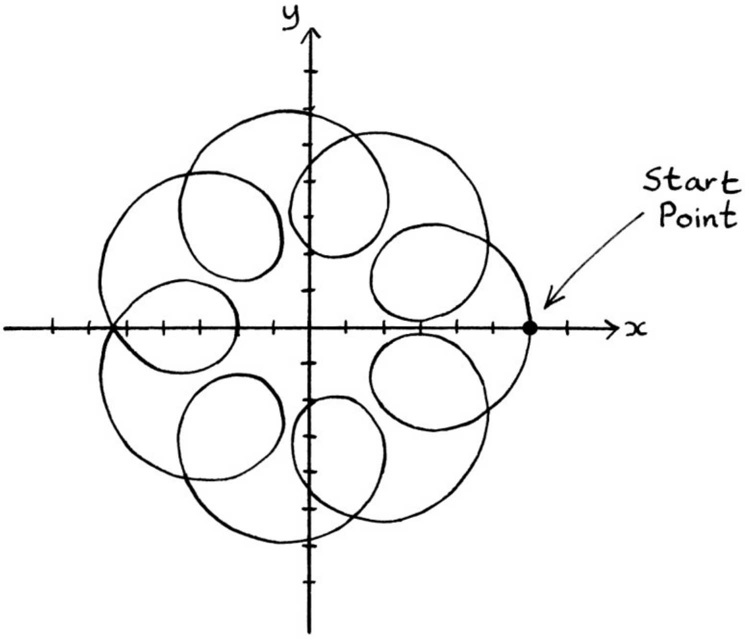
This demonstrates how, when we add two different frequencies, the resulting shape can be much harder to read. The angle of the position of the outer object with respect to the centre of the inner circle at particular times is not proportional to the time. The angle of the outer object from the centre of the inner circle increases *and decreases* as time progresses depending on the position of the two circles over time.

Example 1: resulting shape

If we let the outer object continue for the full 4 seconds, the resulting shape will look like this:



Or, with the axes numbering removed to make it clearer:



If we let the outer object continue for longer, it would follow exactly the same path. The cycles of the two objects align at 4 seconds, and are in the same place as when they started. Therefore, any more movement will be identical to before. Such behaviour is reflected in the derived signals where the cycles repeat identically every 4 seconds.

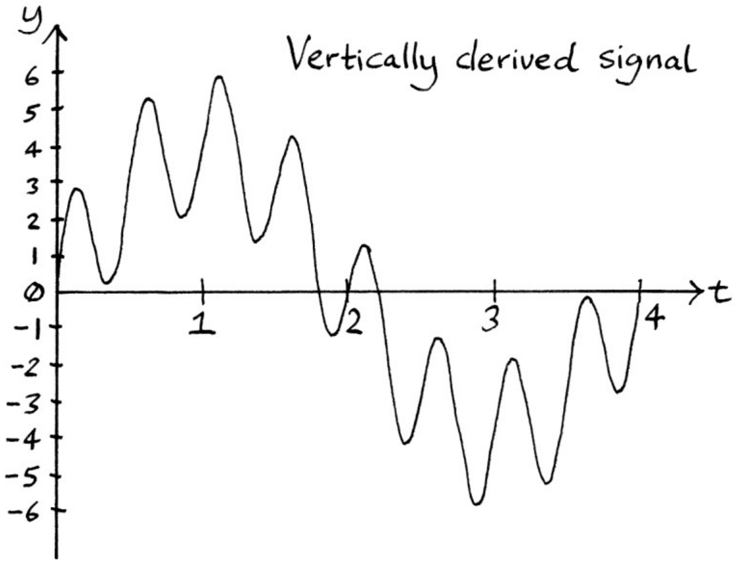
Note that if we were adding two circles of the *same* frequency and phase, the resulting shape would be larger than either of the added circles. When we are adding two circles with *different* frequencies, the size of the resulting shape is much harder to predict.

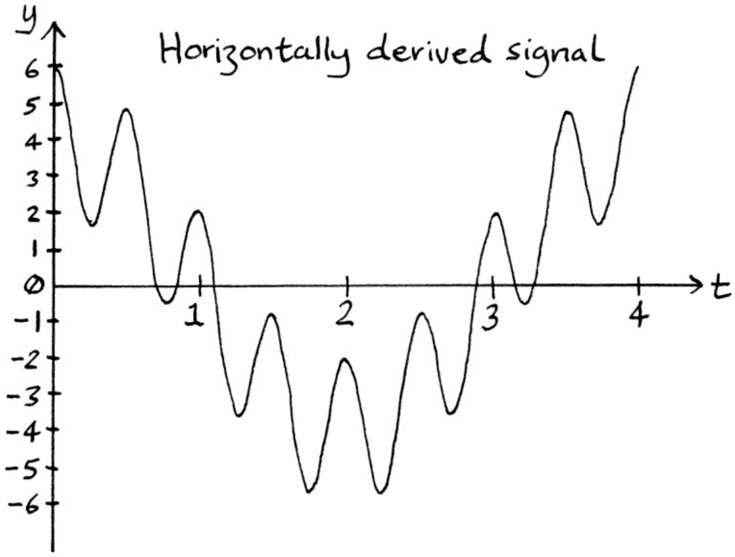
Example 1: resulting signals

As the resulting shape is not a circle, we cannot derive a resulting Sine wave or Cosine wave from it – we can only have a Sine wave or Cosine wave if we have a circle. One of the resulting *signals* will be the sum of two Sine waves, and the other resulting signal will be the sum of two Cosine waves, but the signals will not be pure waves. We can however, derive the resulting *signals* from the shape. The y-axis points of the shape at evenly spaced *times* will produce the signal that is equivalent to the sum of the two original Sine waves; the x-axis points of the shape at evenly spaced *times* will produce the signal that is equivalent to the sum of the two original Cosine waves.

To emphasise the distinction between Sine and Cosine waves, and the nature of the result here, I will call the resulting signals, the "vertically derived signal" and the "horizontally derived signal". A vertically derived signal is the signal derived from a shape by reading the vertical axis (the y-axis) values of the outer object at evenly spaced times. A horizontally derived signal is the signal derived from a shape by reading the horizontal axis (the x-axis) values of the outer object at evenly spaced times. A vertically derived signal might be a Sine wave or the sum of two or more Sine waves; a horizontally derived signal might be a Cosine wave or the sum of two or more Cosine waves. I am using these terms because they are much more precise in their meaning.

The two resulting signals look like this:





The vertically derived signal shows the y-axis values of the outer object on the shape at particular moments in *time*. Likewise, the horizontally derived signal shows the x-axis values of the outer object on the shape at particular moments in *time*. Note that these have nothing to do with the y-axis or x-axis values of the object on the shape at particular *angles*. It is important to understand this distinction, which is much more obvious in this example than it would have been when we were looking at adding circles that still produced proper circles.

If we were to calculate the derived waves by measuring the x-axis and y-axis values of points on the shape at particular *angles*, we would not be representing the shape. For one thing, we would have the problem of having up to three values for the same angle in the loops, which would not make sense. As with a time related circle, we need to calculate the derived signals by measuring the x-axis and y-axis values of the *outer object* at particular, evenly spaced *times*. Sometimes the outer object moves quickly; sometimes it moves slowly. Sometimes, it changes direction and moves backwards (for example in the loops). The outer object moves around its *own* circle at a set number of angles per second, but it does not move around the centre of the inner circle at a set number of angles per second. It is important to understand the difference between the angles and time.

If we took corresponding y-axis values from each resulting signal at regular intervals in time, and used them as coordinates (with the x-axis value of the shape coming from the horizontally derived signal, and the y-axis value of the shape coming from the vertically derived signal), we would recreate the resulting shape.

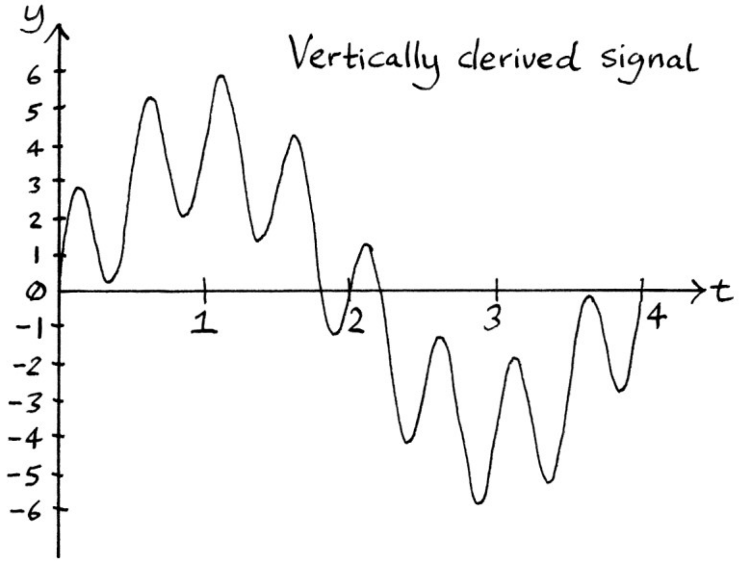
Example 1: resulting frequency

At t = 4 seconds, the cycles of the two pairs of underlying waves align. On the circles, the two objects are back in their original positions. This shows that the period of the resulting signal is 4 seconds (because it takes 4 seconds for the cycles to align again). The frequency of the resulting signal is, therefore, 1 ÷ 4 = 0.25 cycles per second. We could also have worked this out from seeing how the faster frequency (0.5 cycles per second) is an integer multiple of the slower frequency (0.25 cycles per second), so therefore, the resulting frequency will be the same as the slower frequency. This was explained in the previous chapter.

Example 1: the idea of phase

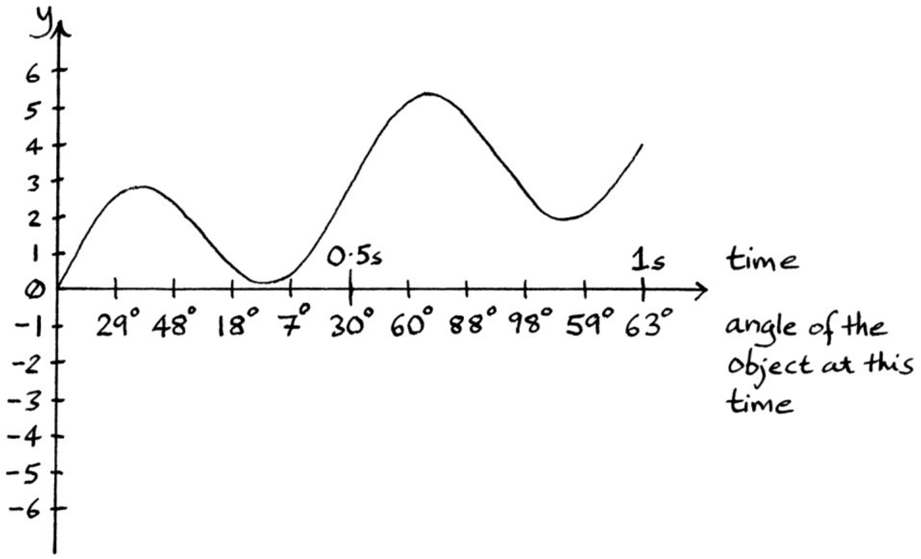
The resulting shape is a good example of how the concept of phase becomes much more abstract when adding circles (or waves) of different frequencies. In this example, the outer object starts at 0 degrees from the centre of the inner circle at t = 0. However, supposing the object had started at 40 degrees, knowing the angle would not have helped us know at which exact point it had started as there are 3 different times where the object is at 40 degrees. Saying the "phase is 40 degrees" because the object started at 40 degrees would not be helpful. I do not think it is sensible to use the term "phase" when dealing with the results of adding two or more circles or waves with different frequencies. [Other people might disagree, though.] The outer object still has a starting point on the resulting shape, but it cannot be distinguished by its angle. It can, however be distinguished by its coordinates. The starting point is still an important attribute of the resulting shape. For shapes that are not circles, I will call the point where the outer object starts, the "start point". The start point on a shape that is not a circle is equivalent to the phase point on a circle.

The possible meaninglessness of the term "phase" with signals that are not pure waves leads on to how the term "phase shift" is also essentially meaningless in such situations. To explain why, we will look at the vertically derived signal again:



If we were dealing with pure waves, and we wrote on the time axis both the time and the angle of the object on the circle at that time, we would see how there was a consistent relationship with the time and the angle. With a pure wave, the angle increases at the same rate as the time. When it comes to impure signals, as with the vertically derived signal here, it is still possible to write the object's angle on the t-axis, but doing so reveals that the time and the angle have a much more complicated relationship. There is not a consistent relationship between the time and the angle.

The following graph has both time and angle written on it, and is only drawn up to the first second to make it clearer:

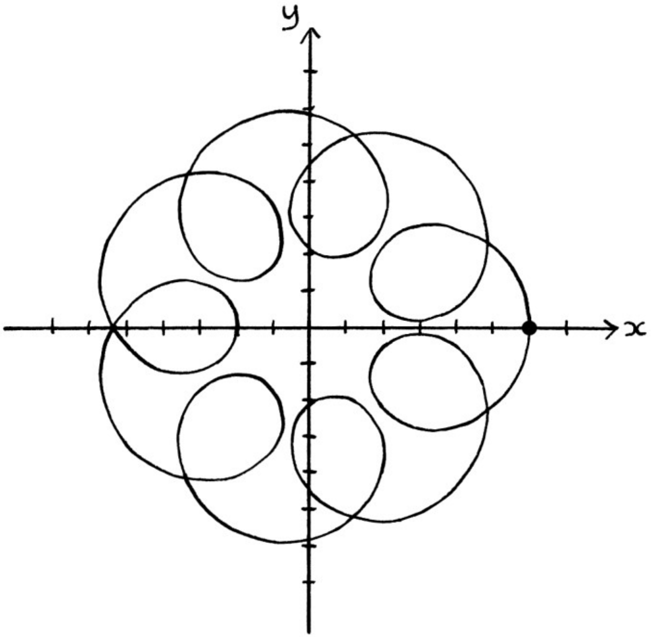


As we can see from the graph, the angle of the outer object over time does not increase at an even rate. Sometimes, the object's angle increases; sometimes, the object's angle decreases. Sometimes, it increases quickly; sometimes, it increases slowly. There will still be a pattern in the angles that will repeat every time the signal repeats its shape, but that pattern is not particularly useful to us. Writing the angle on the time axis here has little purpose, except to indicate that the relationship between the two is not straightforward. With a pure wave, we can easily work out the angle of the object at particular times by looking at the curve. With the graph of a signal such as that shown here, we can have a vague idea of the angle at certain times, but it is impossible to know it accurately most of the time.

Given how the angle is not consistently related to the time, if we shifted the signal left or right along the time axis, we could not say that we were shifting it by a particular *angle*. Angles are less meaningful now. If we wanted to shift the signal left or right by say 10 degrees, we would have to work out what a 10-degree shift would mean for that graph, while bearing in mind that degrees and time are no longer linked in the same way as for a pure wave. Shifts left or right for impure signals are much better described in terms of time. Therefore, we could describe a signal as having been shifted, say, half a second to the left, and it would be easy to understand what this meant and what was happening.

Example 1: resulting heights and widths

Although it is hard to tell from the drawings of the resulting shape, the highest y-axis value on the shape is +5.9253 units. The lowest y-axis value is −5.9253 units. The highest x-axis value on the shape is +6 units (when t = 0); the lowest x-axis value is −5.7044 units. The reason that the highest and lowest values are not all +6 and −6 might be clearer when we look at the way the outer loops of the shape do not all match up with the axes. The outer loops of the shape are the outermost points, and the only place where an outer loop coincides with an axis is at t = 0 when the x-axis value is 6.



The highest and lowest y-axis values of the shape are the negative of each other because the shape, in this particular case, is symmetrical vertically. The highest and lowest *x-axis* values are not the negative of each other because the shape, in this particular case, is not symmetrical horizontally.

Example 1: resulting mean levels

If we were given one cycle of a *Sine wave* or a *Cosine wave*, we would instantly know that the mean level was halfway between the maximum point and the minimum point. However, when we are given signals that are more complicated, such as we have here, we cannot deduce the mean level so easily, and using that method would give a wrong answer. We have to remember that the mean level of a signal or a wave is literally the average y-axis value for one cycle [or an integer number of cycles, or for all time]. Given the vertically derived signal, ideally, we would measure an infinite number of y-axis values off the graph, and then take the average. As this is impossible, we could take readings every 0.05 seconds or so, and take the average of those. This would give an approximate average, which might be good enough. If we wanted to have a very accurate result, and we knew the formulas that were added together, we could use calculus, or more specifically, Integration, to calculate the area under the curve, and from that we could calculate the mean level. In this example, the mean level of both waves is zero, which also means the shape is centred at the origin of the axes.

The shape's elaborate pattern makes it difficult to tell exactly where the centre of the shape is by looking at it. The coordinates of the centre of a shape are the average x-axis value of its perimeter, followed by the average y-axis value of its perimeter. In this case, these are (0, 0). [Note that the centre is not the average of the maximum and minimum, but the average of every point.]

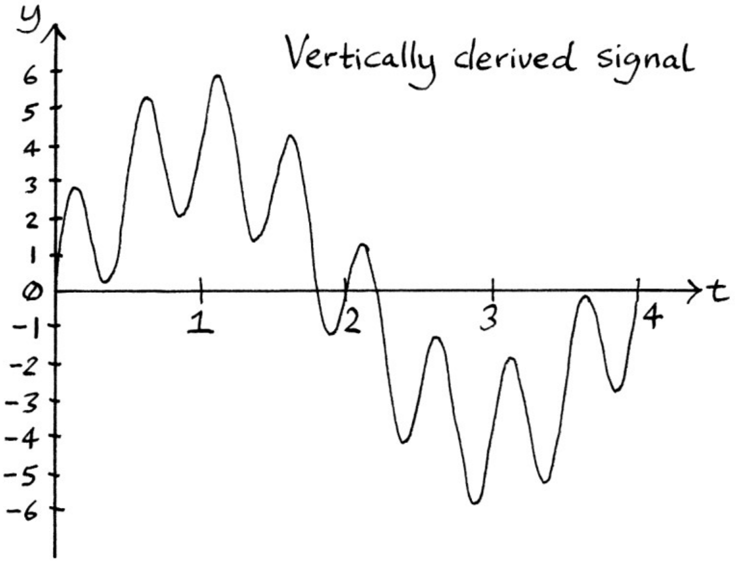
It is helpful to remember the rule that adding two waves with any amplitude, frequency or phase, and zero mean level, will always produce a resulting signal with zero mean level [as long as neither has zero frequency]. If a shape or circle is made up of coordinates from two signals with zero mean levels, then it will always be centred at (0, 0). The rule for circles is that if each of the added circles was centred on the origin of the axes (before they were arranged for adding), then the resulting shape will be centred on the origin of the axes.

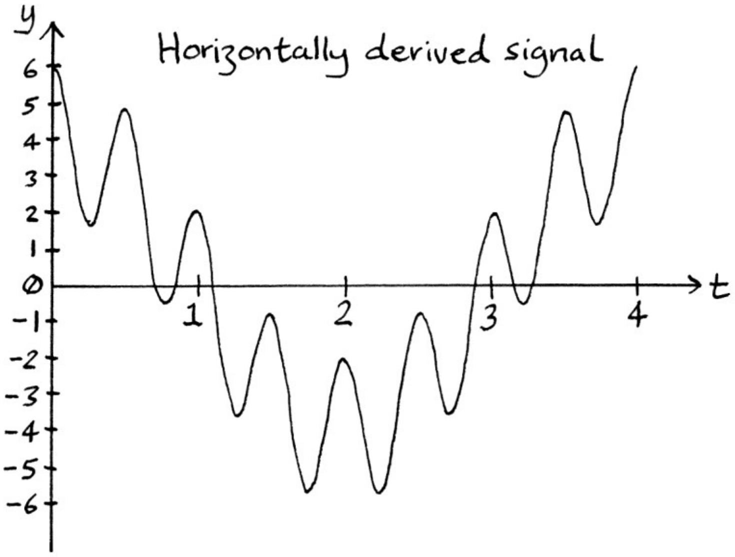
Example 1: the idea of amplitude

As I said in the previous chapter, to say that the resulting signals of an addition of two waves with different frequencies have "an amplitude" is possibly meaningless. It would be similar to saying that the resulting shape has a radius, when in reality the distance from the centre to the perimeter of such a complicated shape cannot be defined by just one value. On the resulting shape in this example, the highest y-axis value is +5.9253 and the lowest y-axis value is −5.9253, and the centre is at y = 0. If we said that amplitude could apply to such a shape, we might say it was 5.9253. However, this is shown to be wrong when we see that the highest x-axis value is 6, and the lowest x-axis value is −5.7044 units. These are not even equal distances from the centre of the shape, and neither is equal to 5.9253 units. In my view, it is better to avoid using the term "amplitude" when discussing signals that are not pure waves. It is better to use the terms "maximum" and "minimum" instead.

Example 1: the signals are not shifted versions of each other

If we look at the resulting signals again...





... we can see that the horizontally derived signal is not the same as the vertically derived signal shifted along the time axis, as it would be if they were both pure waves. The two signals have different shapes. This is most obvious in how the two lowest dips in the first signal are different from each other, but the two lowest dips in the second signal are the same as each other. If we have the sum of two waves, we cannot recreate the sum of the two corresponding waves by shifting the signal along the time axis. Shifting along the time axis only works for individual pure waves. If we were given the vertically derived signal and wanted to find the corresponding horizontally derived signal, we would have to figure out which pure waves were added together to make the vertically derived signal, then shift those pure waves 90 degrees to the left, and then add those shifted waves together. [Again, it is important to remember that the idea of shifting either of these derived signals by 90 degrees is almost meaningless anyway, given how the angle of the object rotating around the shape is not proportional to the time].

Example 1: formula

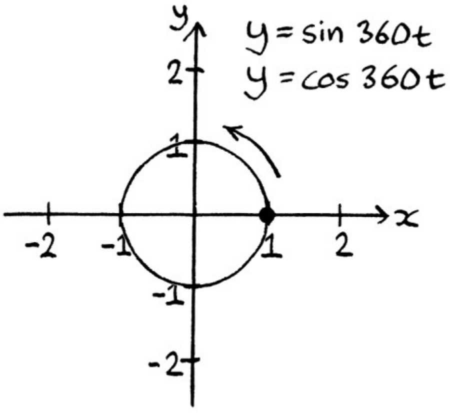
If we wanted to describe the resulting shape mathematically, we could do so by calling it "the shape from which these two signals are derived":
"y = 4 sin (360 * 0.25t) + 2 sin (360 * 2t)"
... and:
"y = 4 cos (360 * 0.25t) + 2 cos (360 * 2t)"

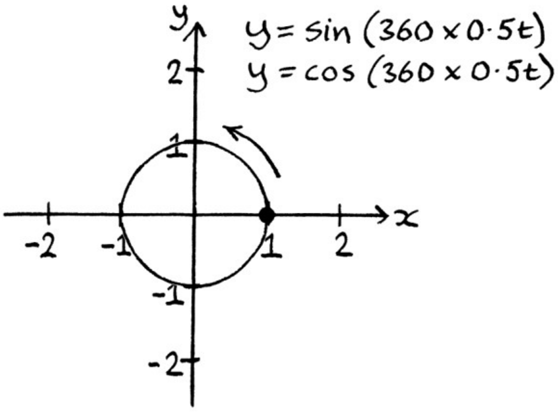
... or, to put it another way, it would be the shape whose coordinates are given by:
(4 cos (360 * 0.25t) + 2 cos (360 * 2t)), (4 cos (360 * 0.25t) + 2 cos (360 * 2t))
... for values of "t" from 0 to 4 seconds or more.

Example 2

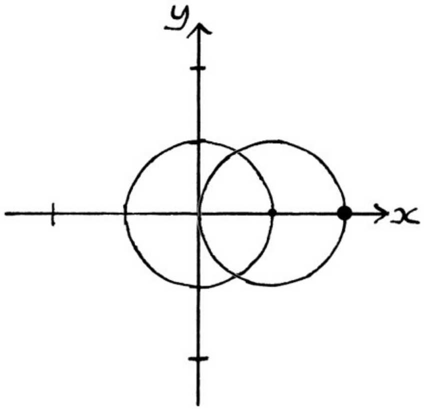
As a simpler example, we will add the circles that are based on the following waves:
"y = sin 360t" and "y = cos 360t"
... and:
"y = sin (360 * 0.5t)" and "y = cos (360t * 0.5t)"

The two circles look the same, but have different frequencies:

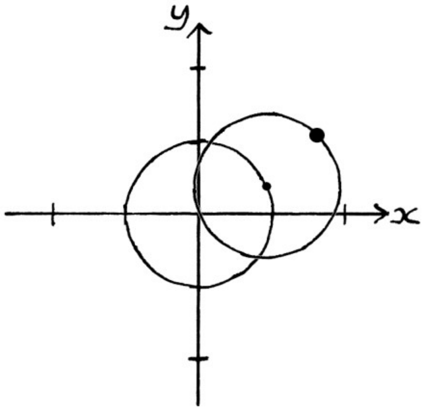




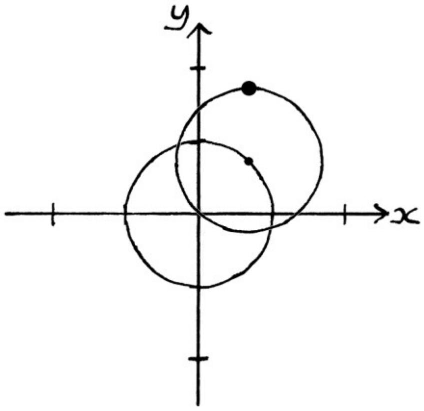
When placed together, with the faster frequency on the outside, the circles look like this at t = 0.



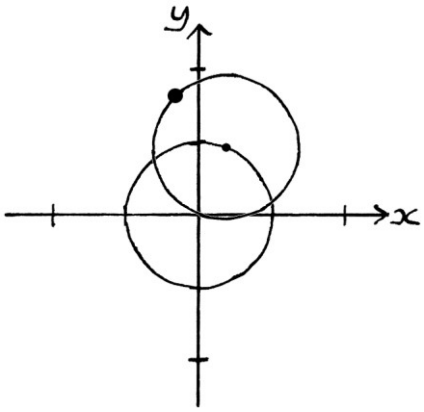
The outer object moves around its circle, and that circle moves around the inner circle. At t = 0.125 seconds, they look like this:



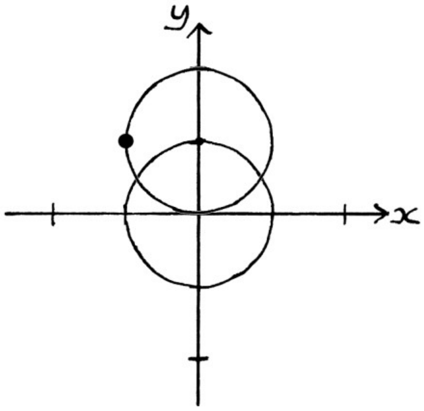
At t = 0.25 seconds:



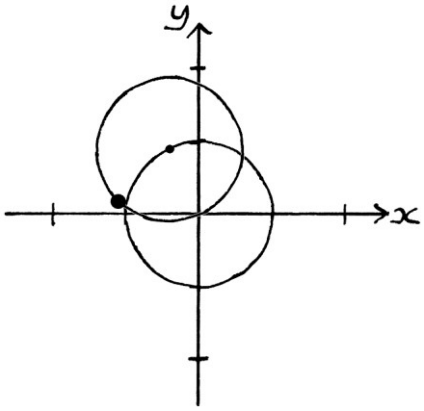
At t = 0.375 seconds:



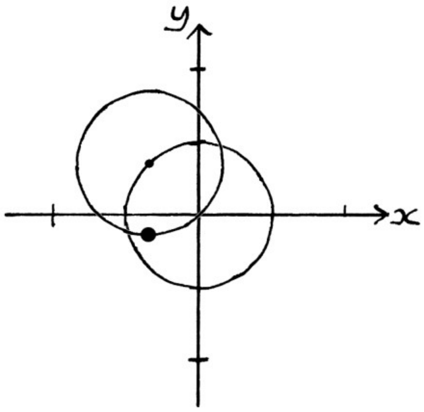
At 0.5 seconds, the outer object has completed half a revolution; the inner object has completed quarter of a revolution:



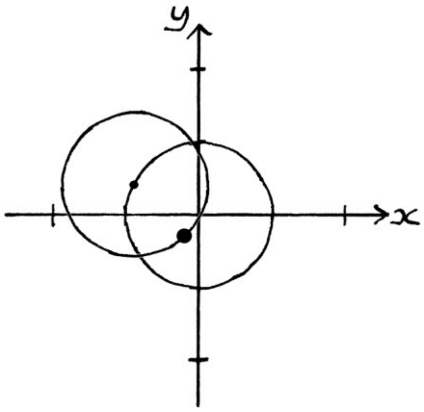
At 0.625 seconds:



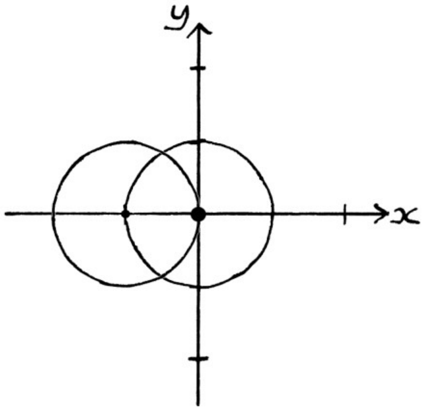
At 0.75 seconds:



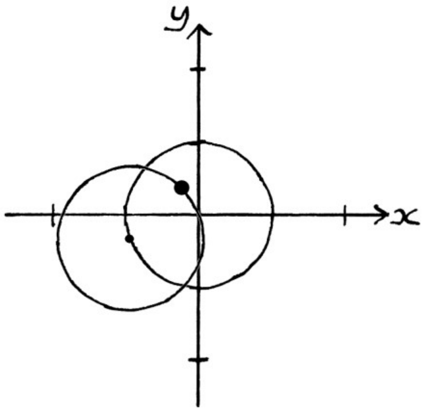
At 0.875 seconds:



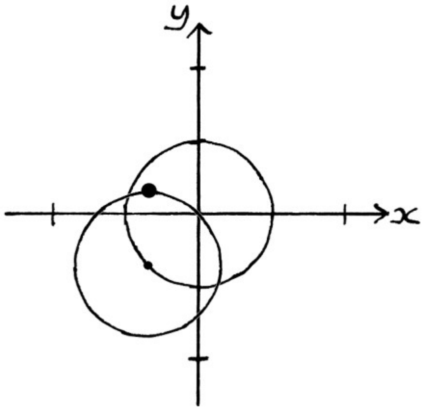
At 1 second, the outer object has completed a full revolution; the inner object has completed half a revolution:



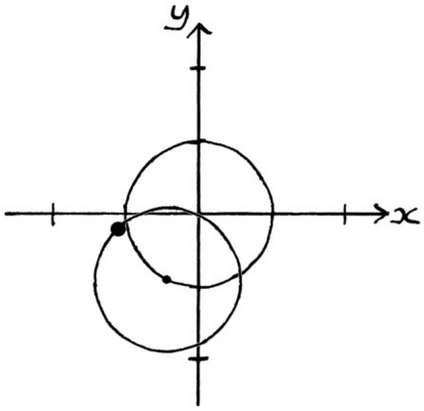
At 1.125 seconds:



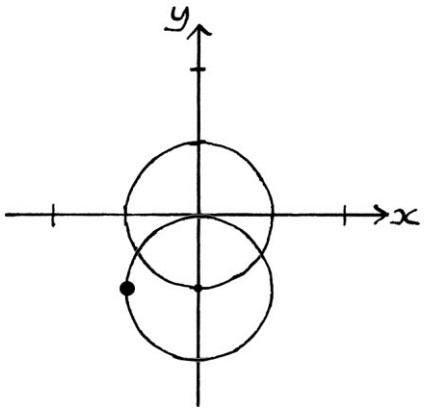
At 1.25 seconds:



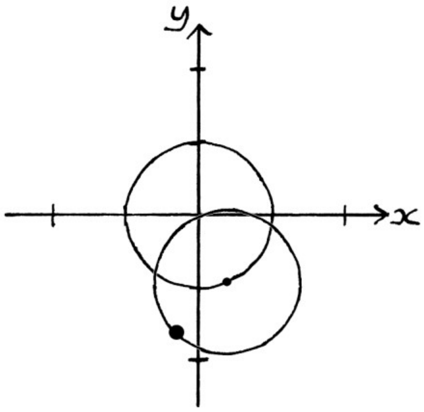
At 1.375 seconds:



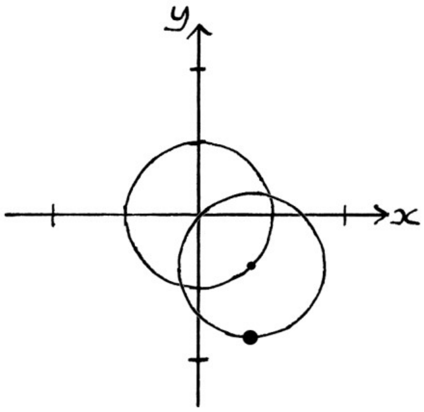
At 1.5 seconds:



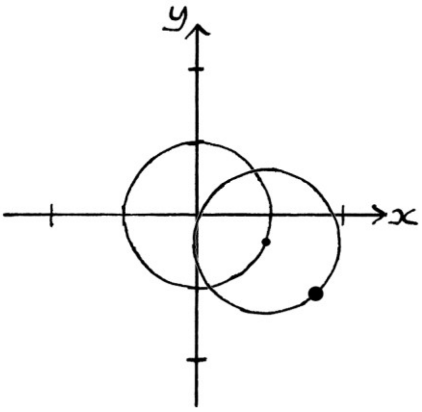
At 1.625 seconds:



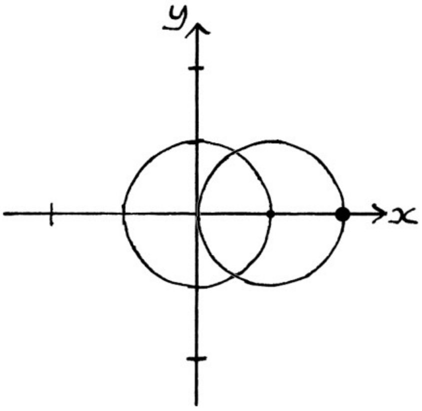
At 1.75 seconds:



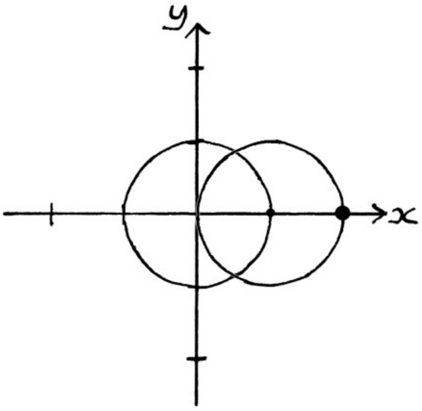
At 1.875 seconds:

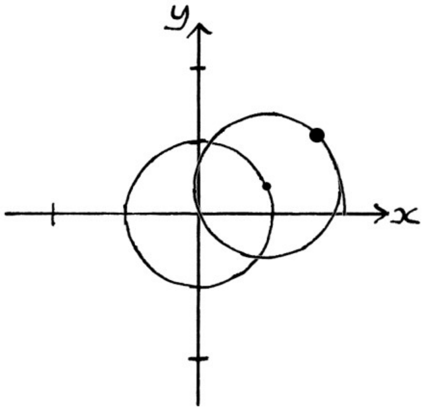


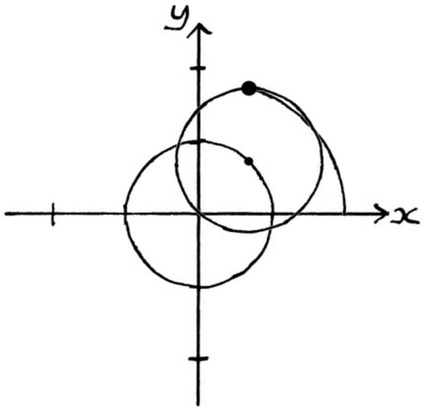
At 2 seconds, the outer object has completed two full revolutions, and the inner object has completed one full revolution:

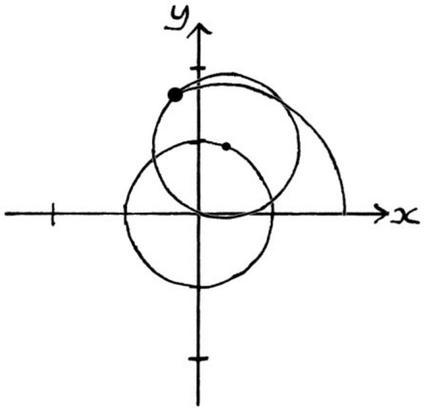


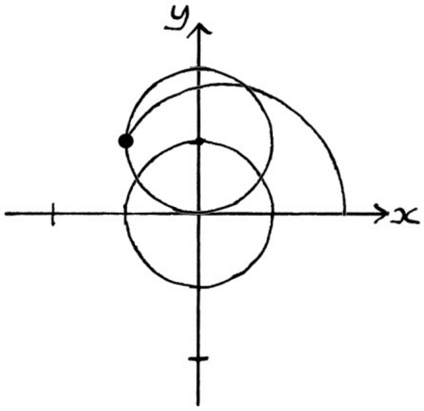
The outer object draws out a shape as it rotates around its own circle, and as that circle rotates around the inner circle:

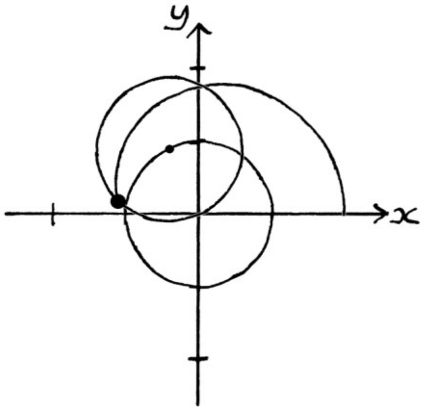


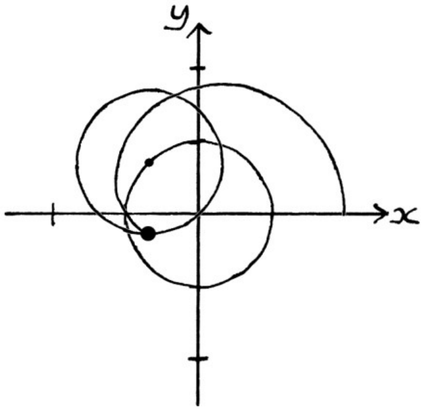

y
x

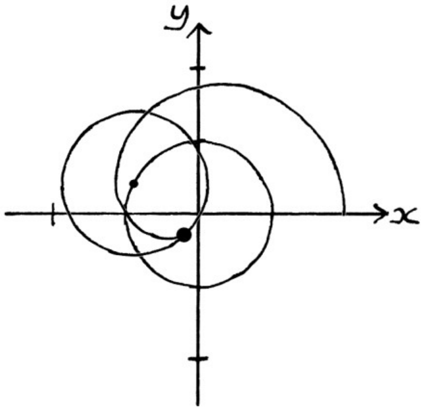

y
x

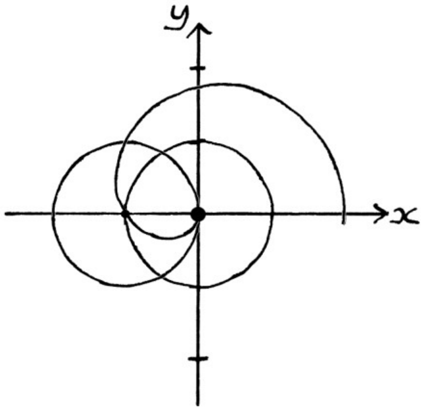

y
x

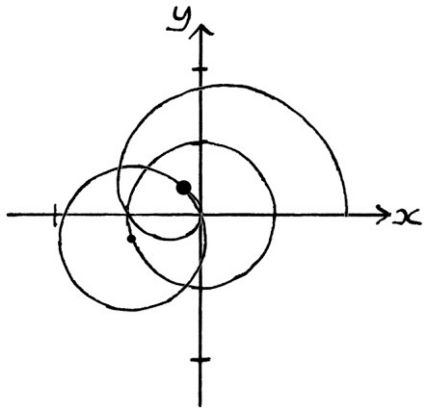

y
x

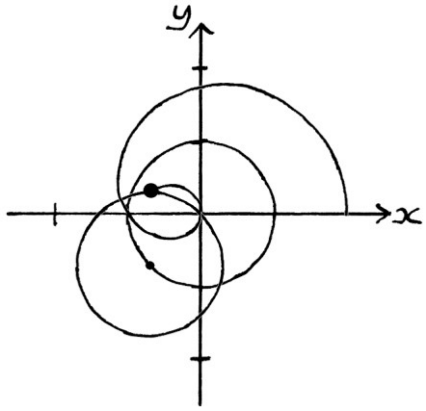

y
x

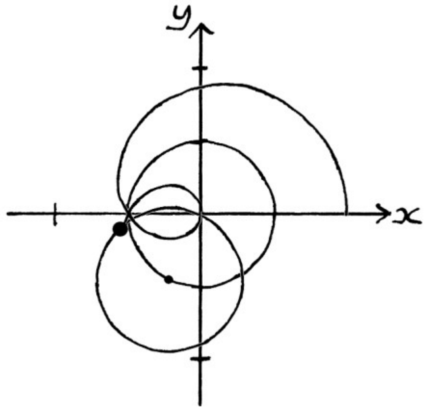

y
x

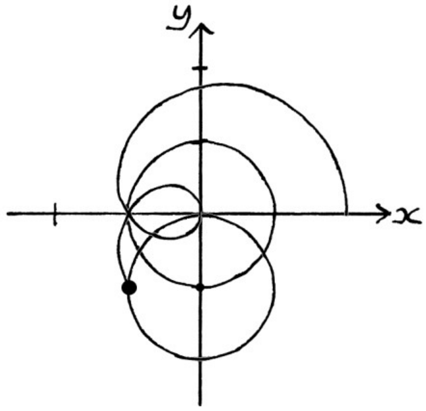

y
x

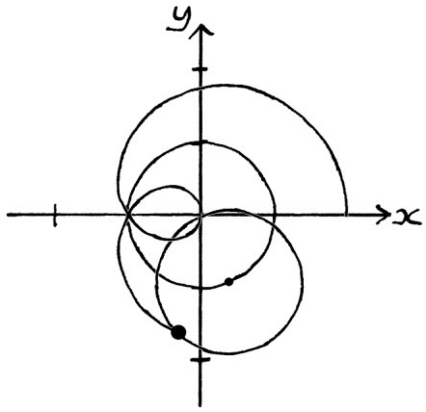

y
x

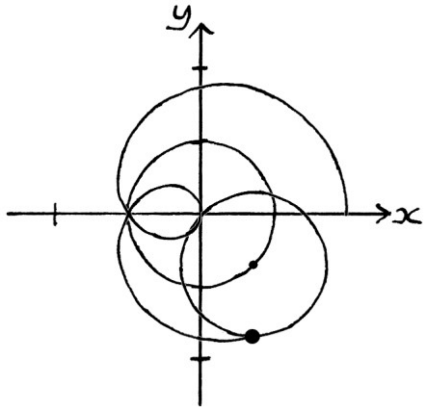

y
x

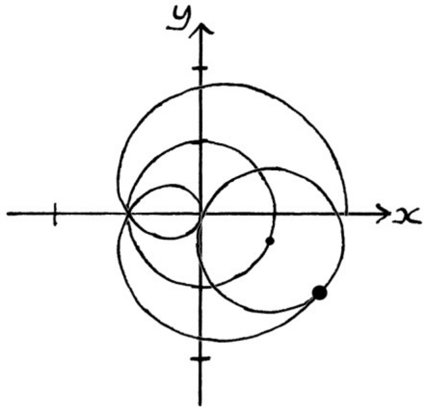

y
x

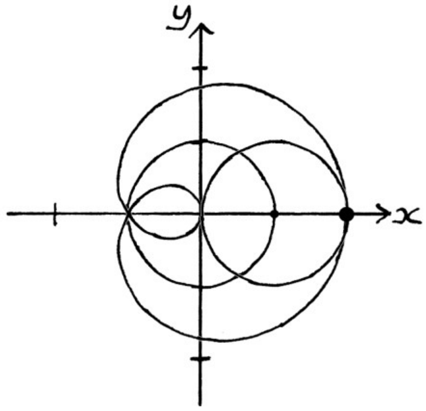

y
x


y
x

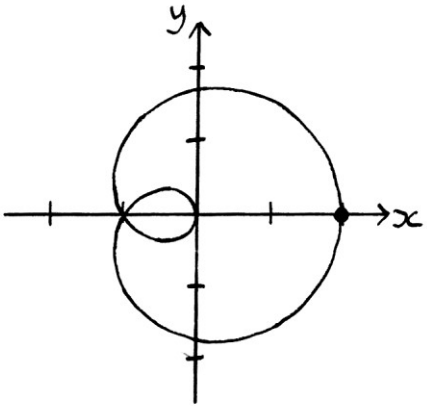

y
x


y
x

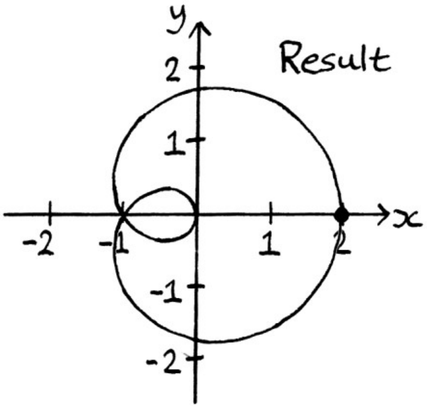

y
x


y
x

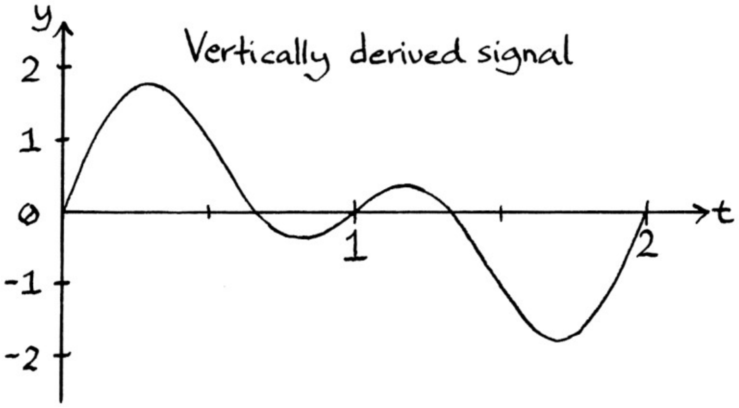
The resulting shape, with the circles removed, looks like this:



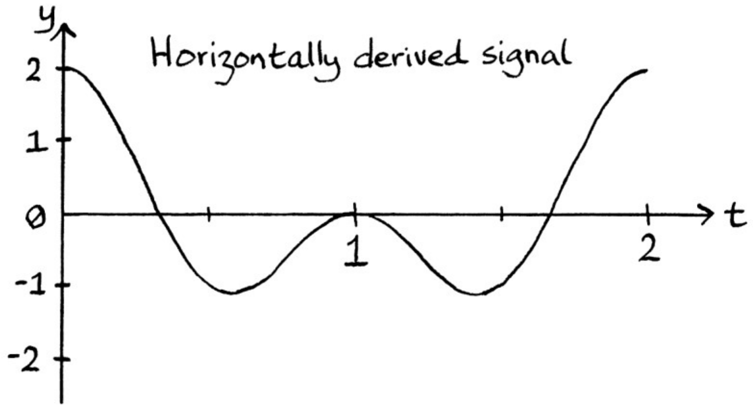
...or drawn with the axes numbered:



The vertically derived signal from this shape looks like this:



The horizontally derived signal looks like this:



Some thoughts about the resulting shape and signals are:

- The two derived signals have mean levels of zero units, which means that the shape's centre must be at (0, 0) on the circle chart.

- The two derived signals repeat every 2 seconds, which means that they both have frequencies of 0.5 seconds. The outer object on the shape takes two seconds to complete one revolution around the origin of the axes.

- The vertically derived signal has a maximum value of 1.7602 units, and a minimum value of −1.7602 units.

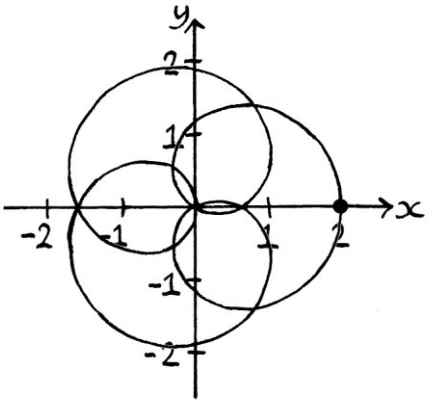
- The horizontally derived signal has a maximum value of 2 units, and a minimum value of −1.125 units. Despite this difference in the maximum and minimum, the mean level is still zero – in this case in each cycle, there are two smaller dips below y = 0, and one bigger dip above y = 0, and the average y-axis value for all points in a cycle is zero.

- As with the previous example, we can see that we could not recreate either signal from the other by shifting it left or right by 90 degrees. To recreate one signal from the other, it would be necessary to guess or calculate the individual waves that made up the signal, shift each individual one, and then add the shifted waves together.

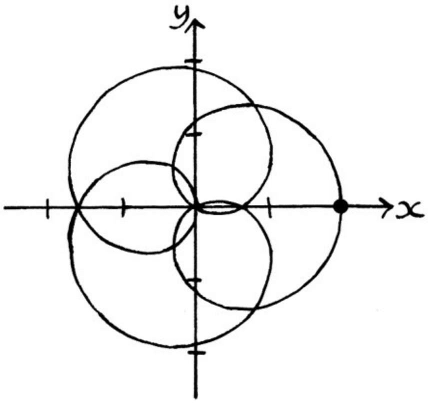
Example 3

Now we will add the circles that represent:
“y = sin 360t” and “y = cos 360t”
... and:
“y = sin (360 * 4t)” and “y = cos (360 * 4t)”

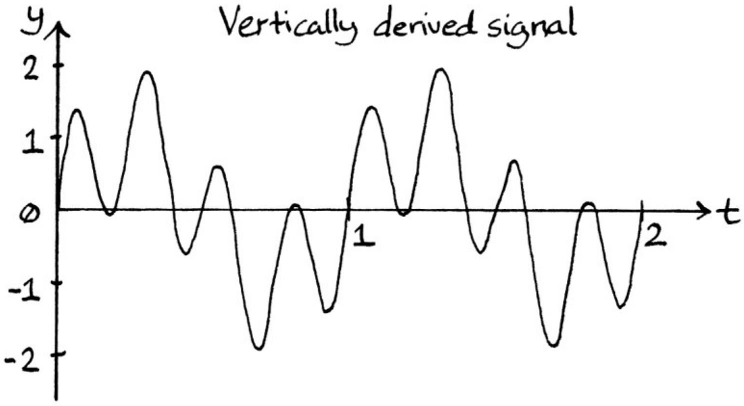
We will skip the pictures of circles and go straight to the result, which looks like this:

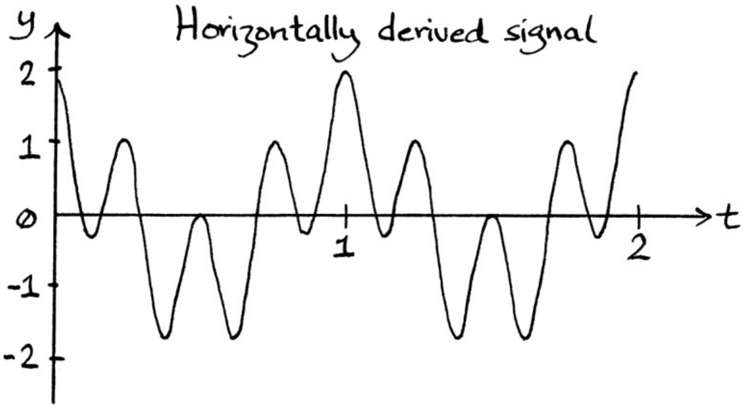


... or drawn without axis numbering:



The two derived signals look like this:





- Both derived signals repeat every second, so have a frequency of 1 cycle per second.

- The mean level of both derived signals is zero.

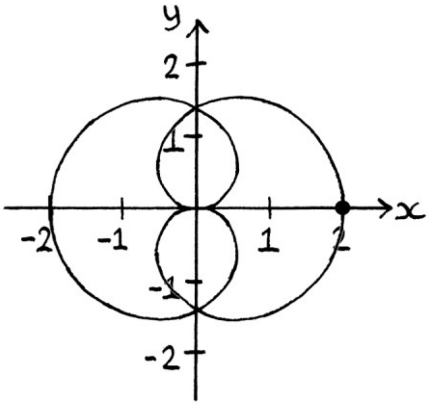
- For the vertically derived signal, the maximum y-axis value is +1.9282; the minimum y-axis value is −1.9282. We can tell that the maximum and minimum in this case would be the negative of each other by how the top half of the shape on the circle chart is the mirror image of the bottom half.

- For the horizontally derived signal, the maximum y-axis value is 2 units; the minimum y-axis value is −1.7221 units.

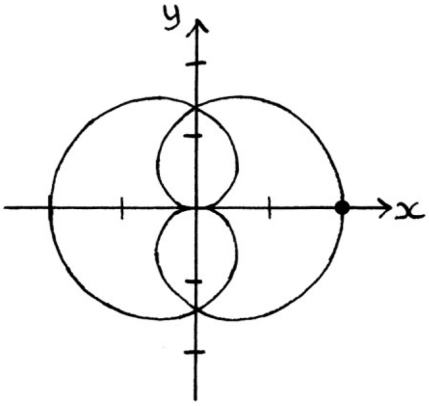
Example 4

We will add the two circles that represent the following waves:
“y = sin (360 * 1.5t)” and “y = cos (360 * 1.5t)”
... and:
“y = sin (360 * 4.5t)” and “y = cos (360 * 4.5t)”

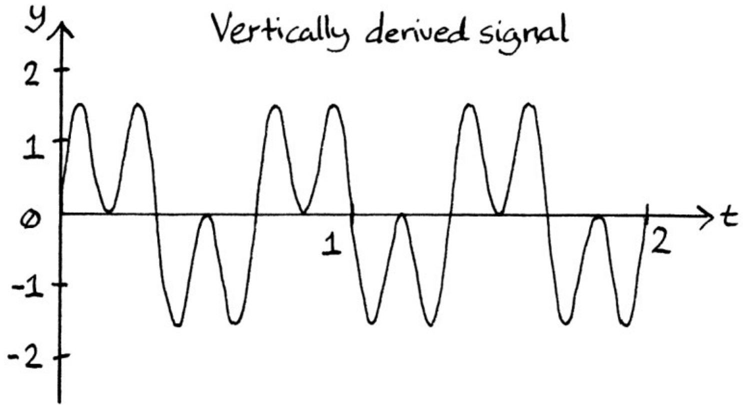
The resulting shape looks like this:

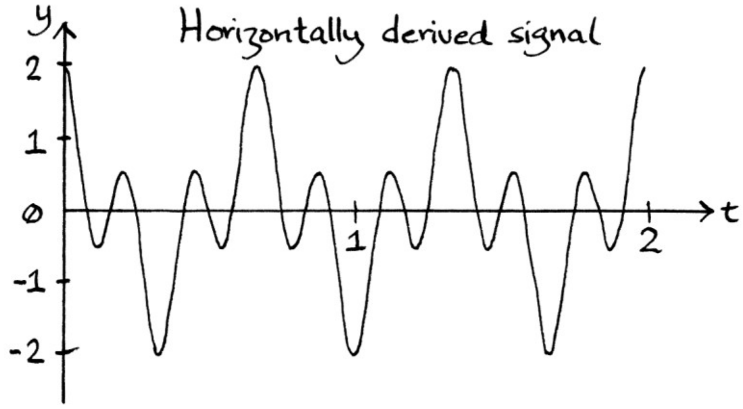


... or with the axes numbers removed:



The two derived signals look like this:





- The vertically derived signal has the standard pattern of a signal created from adding two Sine waves that have a frequency ratio of 3 : 1. [It has the typical MW pattern.] The horizontally derived wave has the standard pattern of a signal created from adding two Cosine waves that have a frequency ratio of 3 : 1 [This is a bit harder to recognise.]

- Both signals repeat every 0.6667 seconds, so have a frequency of 1 ÷ 0.6667 = 1.5 cycles per second. Their mean levels are both zero.

- For the vertically derived signal, the maximum y-axis value is +1.5396 units; the minimum y-axis value is −1.5396 units.

- For the horizontally derived signal, the maximum y-axis value is +2 units; the minimum y-axis value is −2 units.

Patterns For frequency ratios

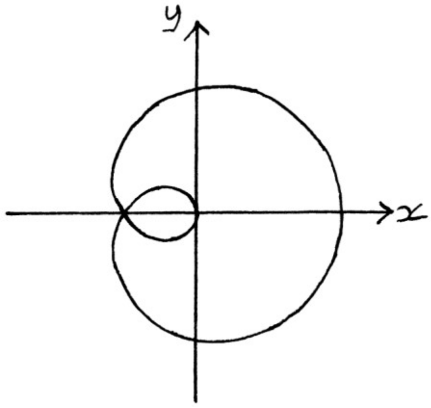
For circles with matching amplitudes, the same phase and the same ratio of frequencies (for example, 2 : 1), the resulting shape will *always* look the same. It might be bigger or smaller, depending on the amplitude, but it will always have the same shape. It does not matter what the actual frequencies are – it only matters what their ratio is. This is because the outer object travels along the path of a shape based on how the inner and outer objects align with each other, and how they align with each other is based on the ratio. The ratio of the frequencies dictates what the path will be; the actual frequencies dictate how quickly the object travels along that path. [Note that non-zero mean levels will still create the same shape, but it will not be centred on the origin of the axes].

If, for example, we add two circles with radiuses of 5 units, zero phases, zero mean levels, and frequencies of 10 cycles per second and 5 cycles per second, then the frequency ratio will be 2 : 1. The resulting shape will be *identical* to adding two circles with radiuses of 5 units, zero phases, zero mean levels, and frequencies of 30 cycles per second and 15 cycles per second. The second pair of circles will result in derived waves that complete their cycles more quickly, but the actual shapes on the circle chart will be identical.

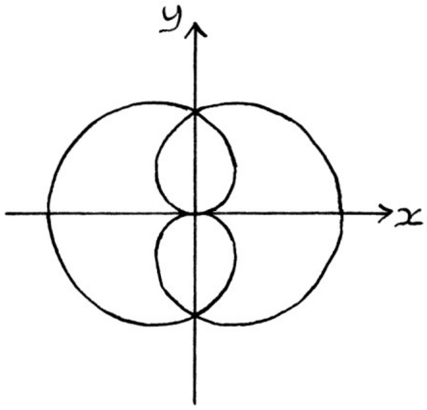
One way of thinking about this is by considering gears in the real world. Imagine we connect a large gear to a small gear on a piece of paper, and stick a pencil in a hole in the small gear. As we rotate the pencil, the small gear will rotate and move around the big gear. The pencil will draw out a shape in a similar way that the outer object in previous frequency examples draws out a shape. The significant thing to realise here is that it does not matter how fast we move the pencil around because we will *always* draw out the same shape. The overall speed at which the gears move is irrelevant to the resulting drawing – the main thing that matters is the *ratio* of one gear's speed to the other (and of course, the radius, starting point, and where the two gears are placed). With circles and frequency, the only things that matter for the resulting shape are the ratio of one frequency to another, and the amplitude, phase and mean levels. [Gears moving around each other, and circles moving around each other have slight differences, but are similar enough to make a suitable analogy].

If we are adding two circles with matching amplitudes, zero phases, zero mean levels, and different frequencies, there are noticeable patterns in the resulting shapes. These patterns depend on the ratio between the frequencies of the two circles.

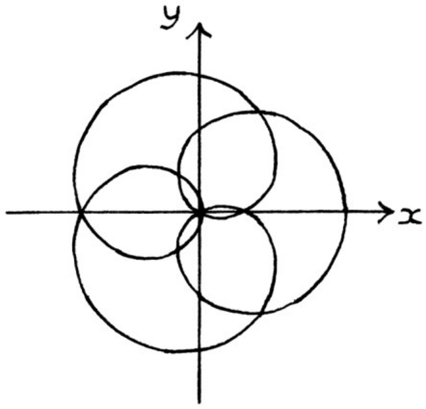
If the ratio is 2:1, the resulting shape will look like this:



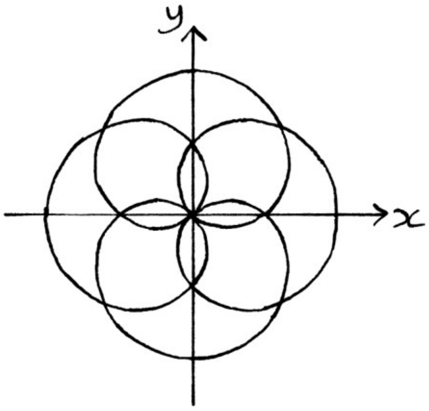
If the ratio is 3:1, the shape will look like this:



If the ratio is 4:1, the shape will look like this:



If the ratio is 5:1, the shape will look like this:



Fixed proportions

The fact that the shapes depend on the frequency ratio means that the maximum and minimum y-axis and x-axis values of the shape's edge will be proportional for any two added circles with matching amplitudes, the same phase, and the same frequency ratio.

For any pair of circles with matching radiuses of any value, the same phase and the same frequency ratio, there will be:

- A particular fraction of the summed radiuses that will give the maximum y-axis value.
- A particular fraction for the minimum y-axis value.
- A particular fraction for the maximum x-axis value.
- A particular fraction for the minimum x-axis value.

In some cases, two or more of these fractions will be the same or the negative of each other.

For example, if we have two circles with the same radiuses, zero phase, zero mean level, and frequencies of 2 and 1, then:

- The fraction for the maximum y-axis value will be 0.8801
- The fraction for the minimum y-axis value will be −0.8801
- The fraction for the maximum x-axis value will be 1
- The fraction for the minimum x-axis value will be −0.5625

This means that if the radiuses are both 1 unit then:

- The maximum y-axis value = 0.8801 * (1 + 1) = +1.7602
- The minimum y-axis value = −0.8801 * (1 + 1) = −1.7602
- The maximum x-axis value = 1 * (1 + 1) = +2
- The minimum x-axis value = −0.5625 * (1 + 1) = −1.125

If the radiuses are both 7 units then:

- The maximum y-axis value = 0.8801 * (7 + 7) = +12.3214
- The minimum y-axis value = −0.8801 * (7 + 7) = −12.3214
- The maximum x-axis value = 1 * (7 + 7) = +14
- The minimum x-axis value = −0.5625 * (7 + 7) = −7.875

The derived signals will have maximums and minimums that are the same as those taken from the resulting shape.

If the frequency ratios are different, or the amplitudes are different, then the fractions will be different.

There are fixed fractions because the shapes will be the same, but larger or smaller, and scaled in proportion.

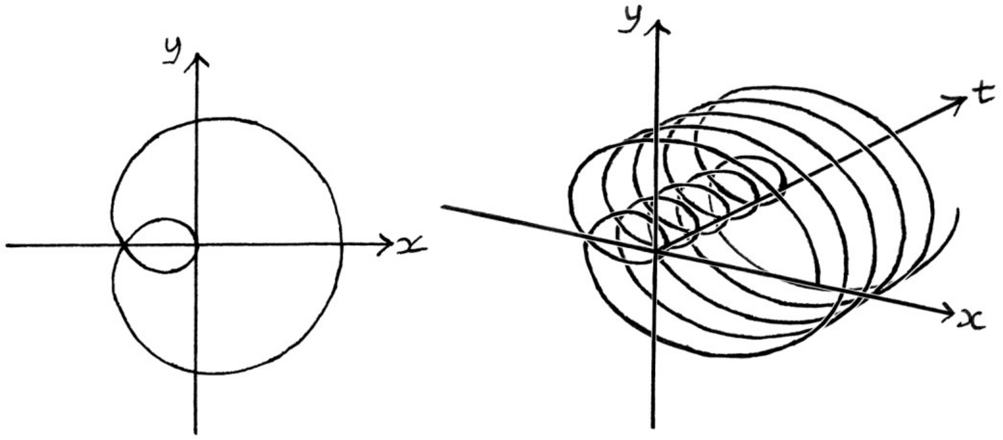
The easiest way to calculate the fractions (that I am aware of) is by plotting the derived signals on a graphing calculator and measuring the maximum and minimum points.

I introduced this idea in the last chapter on adding waves of different frequencies, but the reason why there is a fixed value to multiply by the sum of the radiuses makes more sense now that we are looking at circles. It still does not make it any easier to work out the fractions, but it is clearer why the fractions are the same.

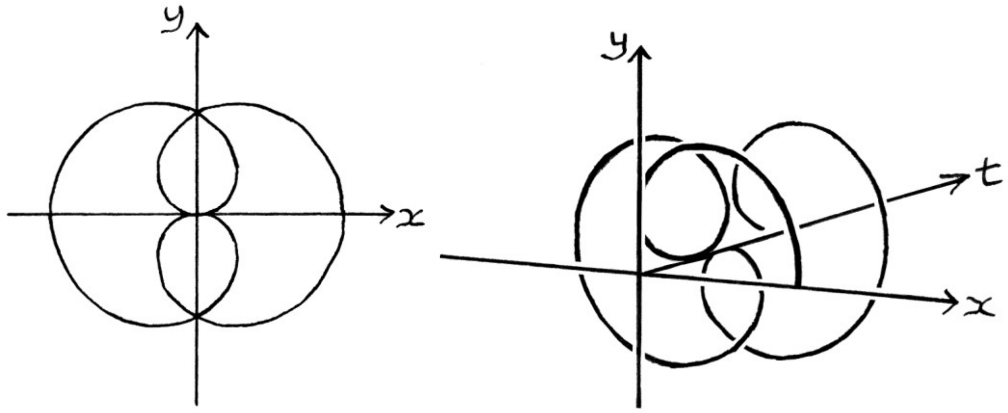
Helices

The resulting shapes can also be drawn on the helix chart. They can be drawn in the same way as helices, but they will not actually be helices. Drawing them in this way makes the position of the outer object over time much clearer. From the helix, if it is drawn well, it is much easier to understand why there are loops in the shapes.

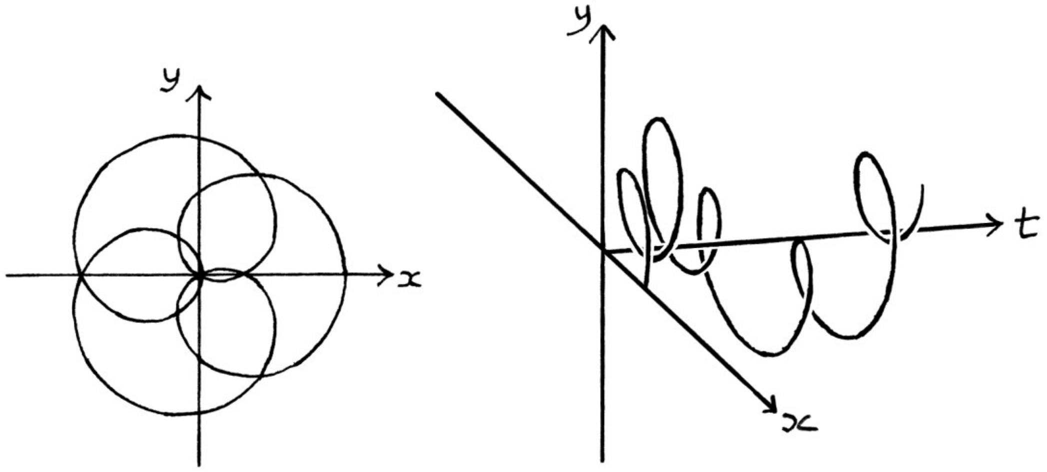
Here is the helix chart for the sum of two circles, with the same amplitude, zero phase, zero mean level, and a frequency ratio of 2 : 1:



Here is the helix for the same circles, but with a frequency ratio of 3 : 1:



Here is the helix for the same circles, but with a frequency ratio of 4 : 1:



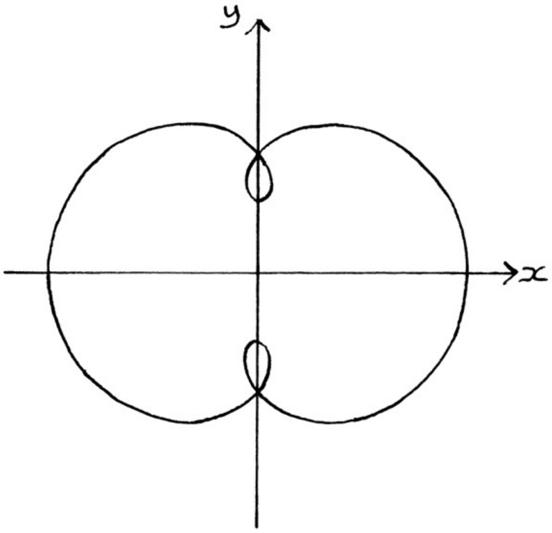
When we dealt with helices for pure waves, we could see the Sine wave by looking at the helix side on (with the time axis to the right, and the x-axis pointing directly at us), and we could see the Cosine wave by looking at helix from underneath (with the time axis to the right, and the y-axis pointing directly away from us). The same idea still works with the helix representing the sum of two circles. The vertically derived signal can be seen by looking at the helix side on, and the horizontally derived signal can be seen by looking at the helix from underneath.

Different frequencies and amplitudes

When adding two circles with different frequencies, the smaller one radius is in relation to the other, the less effect it will have in the resulting sum. This idea is the same as how a wave with a much smaller amplitude in a pair of added waves has less effect on the resulting signal.

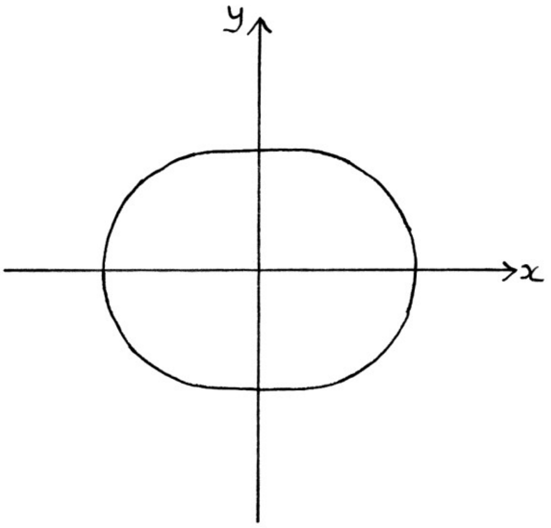
When adding waves, the larger the difference between the amplitudes of the waves, the more the resulting signal will look like the wave with the higher amplitude, and therefore, the more the result will look like a pure wave. When adding circles, the larger the difference between the radiuses of the circles, the more the resulting shape will look like the larger circle. The idea can be phrased more succinctly as, “the higher the ratio between the radiuses of two added circles, the more circle-like the result will be.”

This might be apparent from how an object rotating around a tiny circle that rotates around a larger circle has much less in and out movement. There is much less possible variation in the extremes of its movement. The following picture shows the result of adding the two circles that represent the pairs of waves:
“y = 4 sin 360t” and “y = 4 cos 360t”
... and:
“y = 2 sin (360 * 3t)” and “y = 2 cos (360 * 3t)”

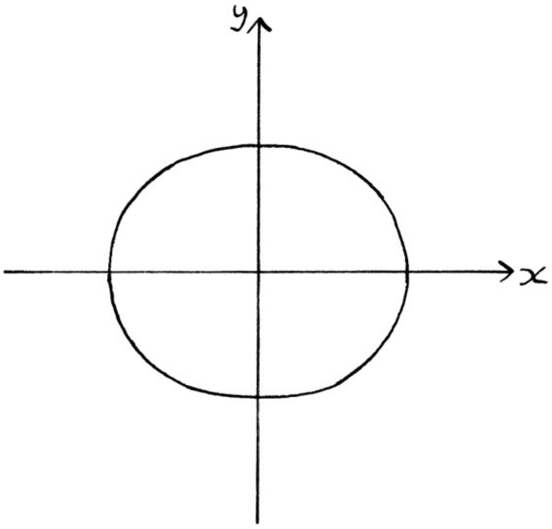


This next picture shows the result of adding the two circles that represent the pairs of waves:
“y = 4 sin 360t” and “y = 4 cos 360t”
... and:
“y = 0.5 sin (360 * 3t)” and “y = 0.5 cos (360 * 3t)”.

The amplitude of the second circle is smaller, and has less of an effect on the result:



This next picture shows the result of adding the two circles that represent the pairs of waves:
"y = 4 sin 360t" and "y = 4 cos 360t"
... and:
"y = 0.3 sin (360 * 3t)" and "y = 0.3 cos (360 * 3t)".



As the second circle has increasingly smaller radiuses, the result becomes more circle-like.

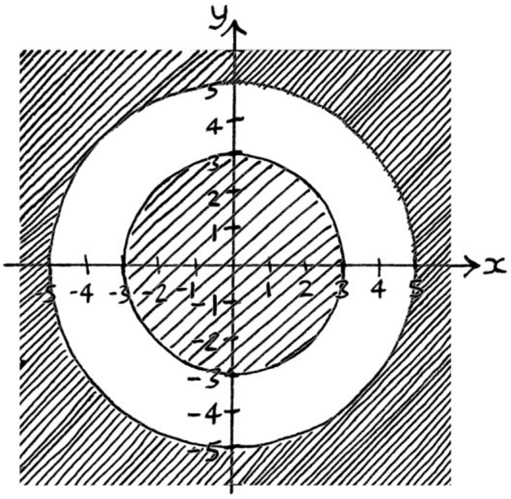
To summarise this section, regardless of the values or ratios of the frequencies, the greater the difference between the two amplitudes, the rounder the resulting shape. The derived signals of a rounder shape will more closely resemble pure waves, and furthermore, they will more closely resemble the constituent wave with the highest amplitude.

Estimating results

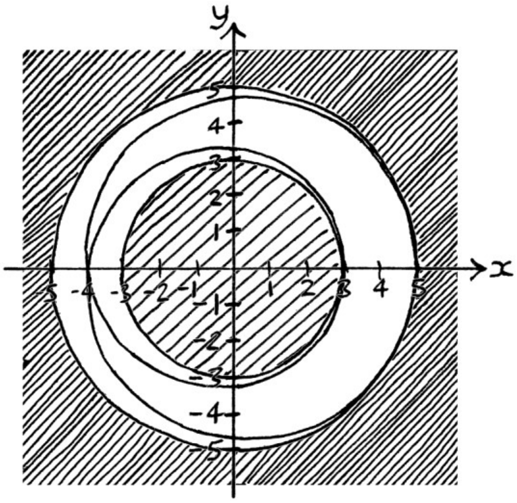
Even without knowing the frequencies of two circles, we can tell that the amplitude of the resulting shape will never be more or less than certain values.

We will say that we have been given two circles with radiuses of 4 and 1 units, phases of zero degrees, and mean levels of zero units, but we do not know their frequencies. As the outer object moves around its circle, which in turn moves around the inner circle, we know that the outer object can never be further away from the perimeter of the inner circle than the radius of the outer circle. The object's maximum distance from the centre of the inner circle can never be more than the sum of the two radiuses. The outer object can never be nearer to the centre of the inner circle than the value of the radius of the inner circle minus the radius of the outer circle.

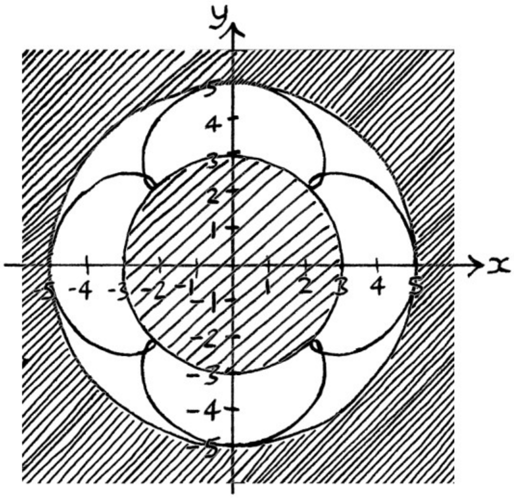
In this example, the outer object can never be further out than 4 + 1 = 5 units, and can never be nearer to the origin of the axes than 4 – 1 = 3 units. Therefore, we can draw two "boundary circles" that mark the limits of the outer object's movement, and also mark the limits of the resulting shape. The outer object will always be in the non-shaded part of the following picture:



No matter what the frequencies of the two original added circles, the resulting shape will always fit within the boundaries. For example, if the frequencies are 4 cycles per second and 2 cycles per second, we end up with this shape:



If the frequencies are 1 cycle per second and 5 cycles per second, we end up with this shape:



Guessing the added circles

Given a shape made from adding two zero-phase circles together, it can sometimes be possible to work out the radiuses of the two original circles using the previous idea. First, draw a circle centred at the middle of the shape and with a circumference that touches the outermost point of the shape. Then, draw a smaller circle, centred in the same place, that touches the innermost point of the shape. We will call these circles the outer boundary and the inner boundary.

The radiuses of the two circles that were added to make up the shape, when added together will be equal to the radius of the outer boundary; the radius of the smaller original circle subtracted from the radius of the larger original circle will be equal to the radius of the inner boundary. From that, it is possible to work out the amplitudes.

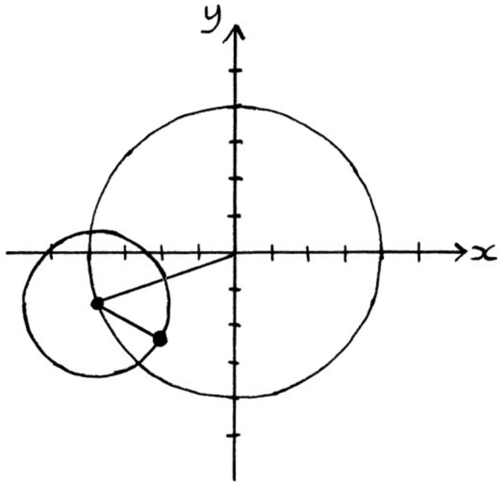
If we call the radius of the original larger circle that was added, “Radius A”, and we call the radius of the original smaller circle that was added, “Radius B”, then:
A + B = the radius of the outer boundary.
A – B = the radius of the inner boundary.

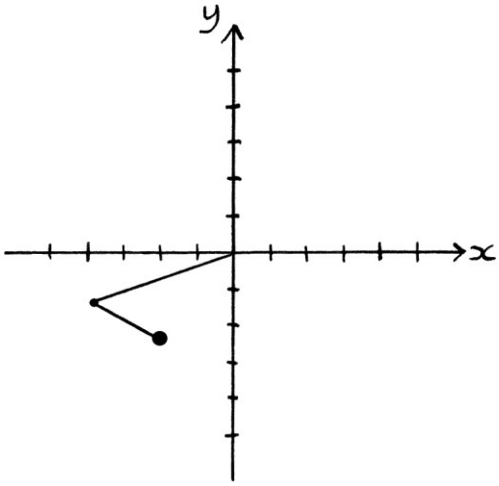
We just need to work out A and B from knowing this.

This will not work all the time. For example, if the phases of the two circles are different, it will not work. However, it might still have some uses.

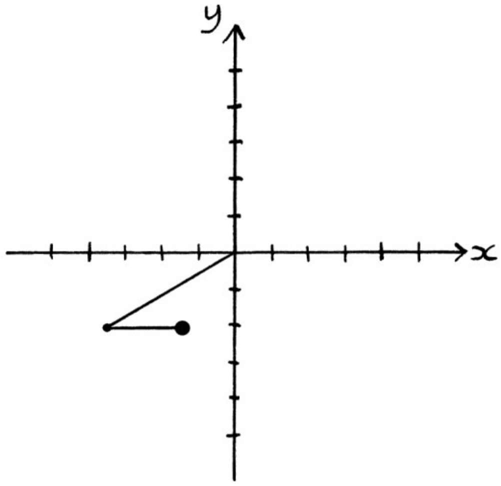
Arms

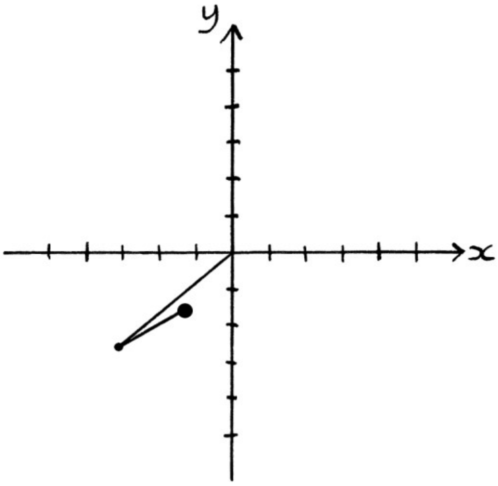
When we looked at phase earlier in this chapter, we saw how the two added circles could be portrayed using an arm to indicate the angle of the inner object to the centre of the inner circle, and the angle of the outer object to the centre of the outer circle. For two circles with the same frequency, the bend in the arm remains fixed as the outer object moves around:

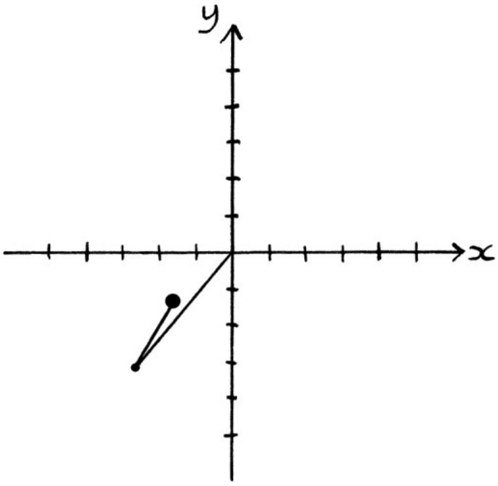


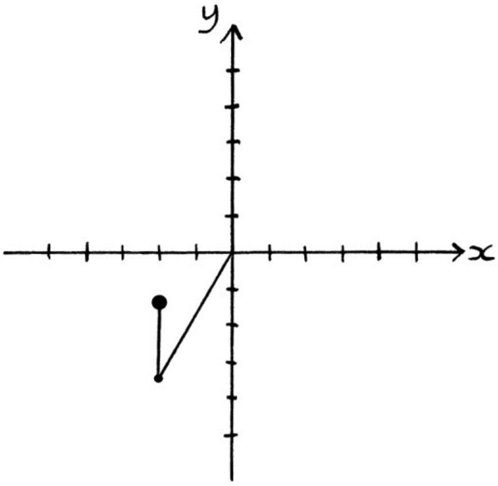


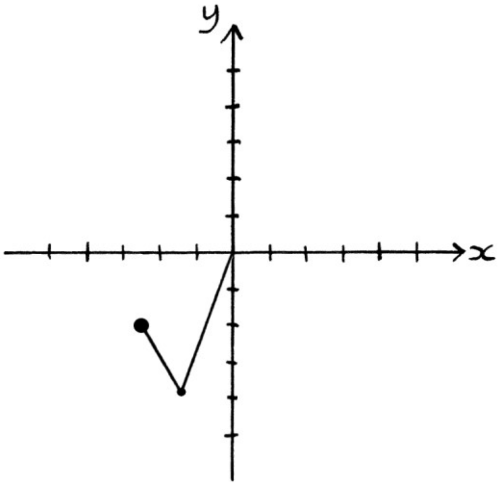
For two circles with different frequencies, the bend in the arm is constantly changing:

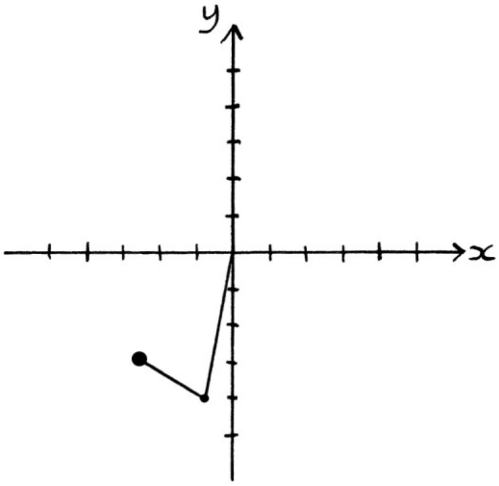



y
x


y
x


y
x





Different frequencies and mean levels

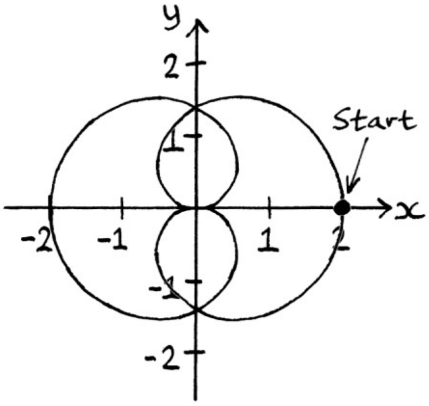
Adding circles with different frequencies and non-zero mean levels is not much different from adding circles with zero mean levels. First, we group the mean levels together and apply them to the inner circle. Then the outer object rotates around as usual. The resulting shape will be identical to that created when the mean levels are zero, but just placed in a different position on the circle chart.

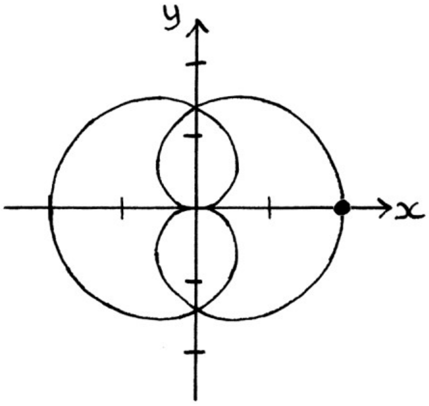
Different frequencies and phases

Adding circles with different frequencies and non-zero phases causes the shape on the circle chart to rotate or distort depending on the ratio between the two phases, and the ratio between the two frequencies.

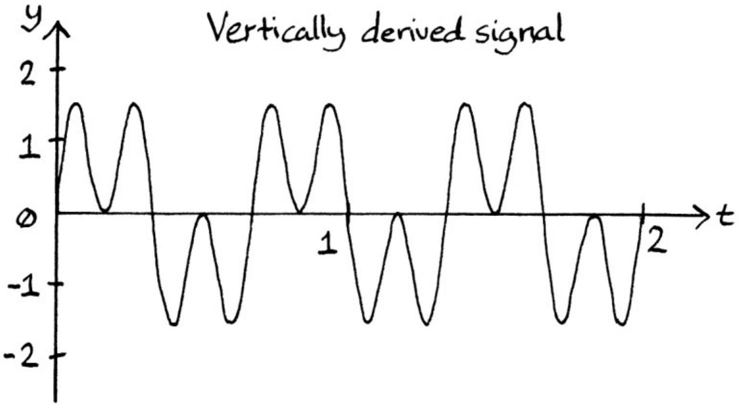
In Example 4 in the first section of adding circles with different frequencies, from earlier in this chapter, we added the circles that represent the waves:
“y = sin (360 * 1.5t)” and “y = cos (360 * 1.5t)
... and:
“y = sin (360 * 4.5t)” and “y = cos (360 * 4.5t)

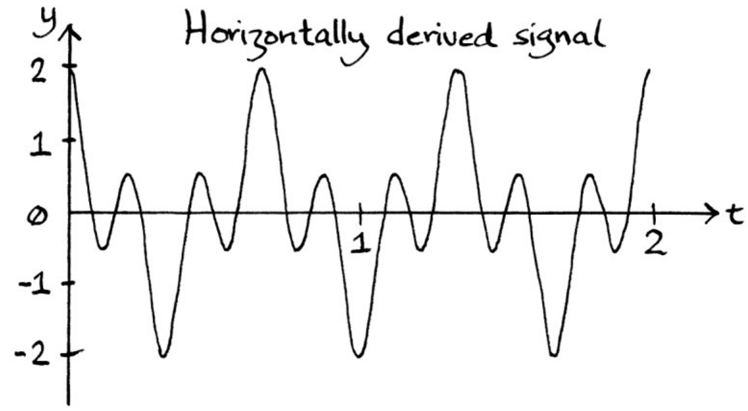
We ended up with the following shape, which I have drawn first with the axes numbers, and then without the axes numbers, as it makes a clearer picture.





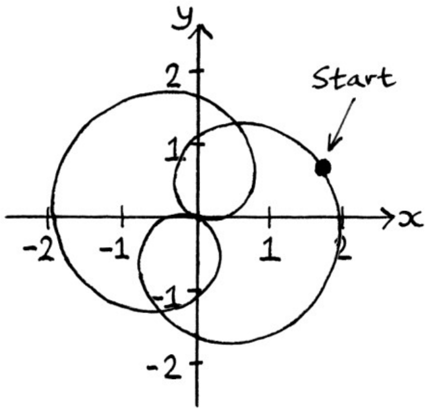
The shape produces these two derived signals:

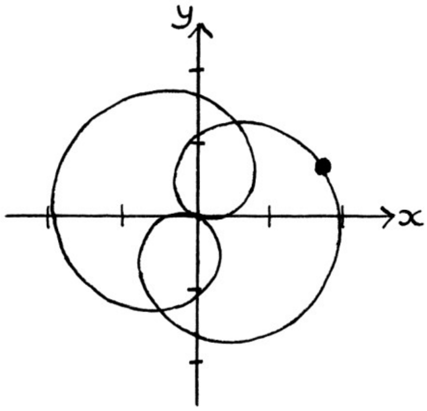




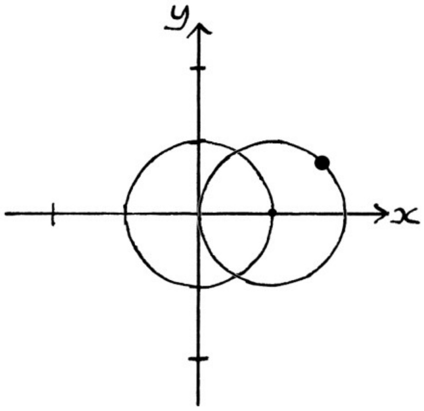
If we have a phase of 45 degrees in the formula of the second circle, we will be adding the circles that represent these pairs of waves:
“y = sin (360 * 1.5t)” and “y = cos (360 * 1.5t)”
... and:
“y = sin ((360 * 4.5t) + 45)” and “y = cos ((360 * 4.5t) + 45)”.

The resulting shape will look like this (shown with axes numbers and without):





Note how the entire shape has been rotated by 22.5 degrees clockwise. The point where the outer object starts is no longer at 0 degrees, but instead at 22.5 degrees. That the outer object starts here can be calculated by imagining the initial position of the original circles:

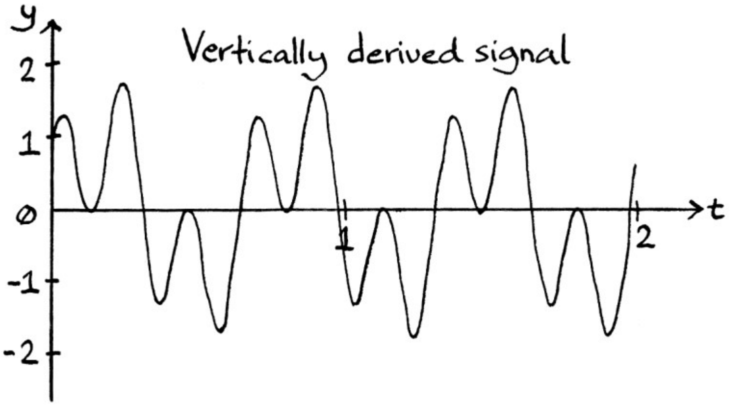


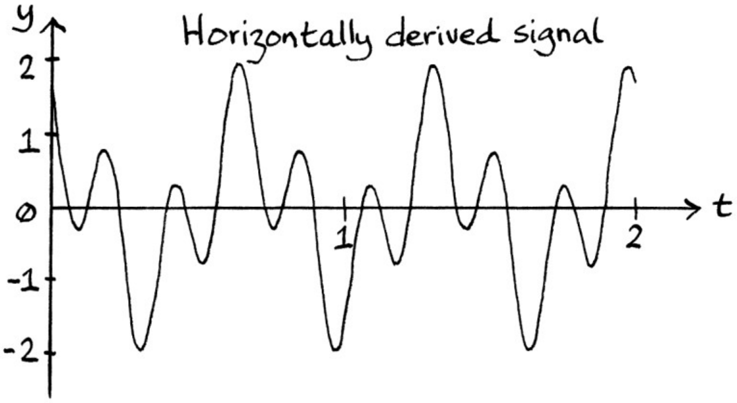
The outer object's horizontal distance from the centre of the inner circle is cos 0 + cos 45 = 1.7071 units. Its vertical distance is sin 0 + sin 45 = 0.7071. Its angle from the centre of the inner circle is arctan (0.7071 ÷ 1.7071) = 22.5 degrees (and a check to see if this is the result we want from the two possible ones will confirm that it is).

The reason the whole shape has been rotated 22.5 degrees clockwise is made clearer with some thought. The frequencies are the same as before, so the cycles of the two circles will match up as often as they did before. However, because the second circle has a head start of 22.5 degrees, the entire resulting shape will reach its resulting points 22.5 degrees sooner than it did before. Therefore, it becomes rotated by 22.5 degrees.

Although the shape is the same as before but rotated, the derived signals will not be shifted versions of the shape made from circles with zero phase. The derived signals are made up of the y-axis and x-axis values of the shape's edge, so rotating the shape distorts the signals.

The two derived signals look like this:

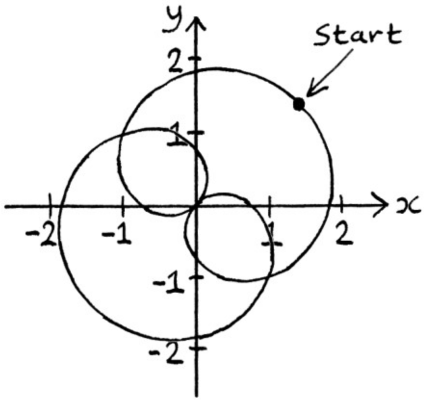


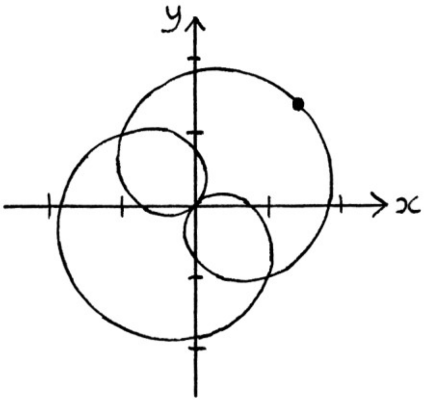


The shapes are similar to before, but are distorted.

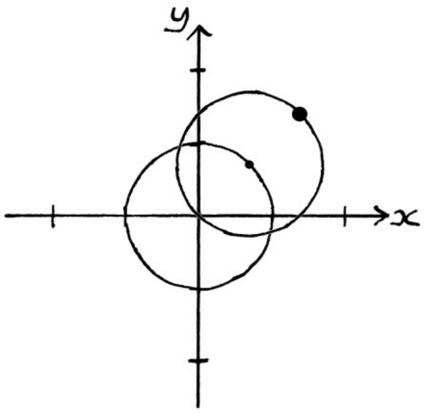
If we have a phase of 45 degrees in the formulas for *both* circles, we will be adding the circles that represent these pairs of waves:
“y = sin ((360 * 1.5t) + 45)” and “y = cos ((360 * 1.5t) + 45)”
... and:
“y = sin ((360 * 4.5t) + 45)” and “y = cos ((360 * 4.5t) + 45)”.

The resulting shape looks like this (shown with axes numbers and without axes numbers):





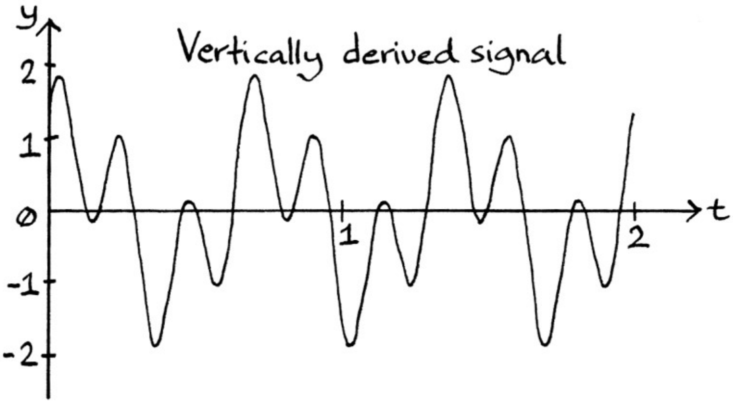
Note how the shape is the same as one created from circles with no phase, but it has been rotated by 45 degrees. The outer object starts at 45 degrees in relation to the centre of the inner circle. The reason for this is made clearer by looking at the position of the original circles at t = 0:

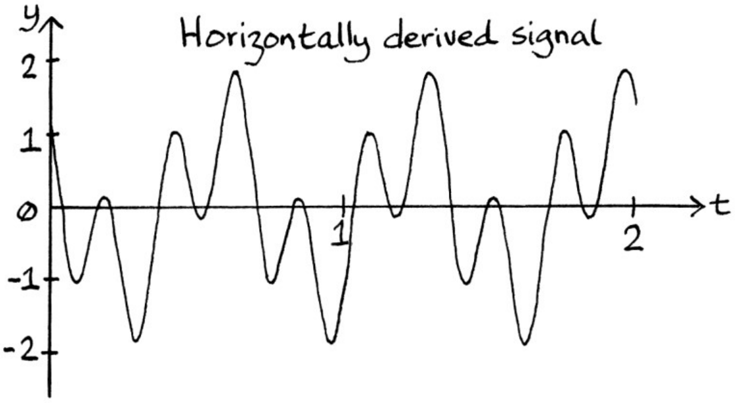


The outer circle is centred at 45 degrees to the inner circle, and the outer object is at 45 degrees to the centre of the outer circle. Therefore, the outer object is at 45 degrees to the centre of the inner circle.

The resulting shape is rotated by 45 degrees because the frequencies of the circles are the same as before, so the cycles align in the same way as before, but the points of the shape are reached 45 degrees earlier.

The derived signals still look like distorted versions of those from the shape with zero phase:

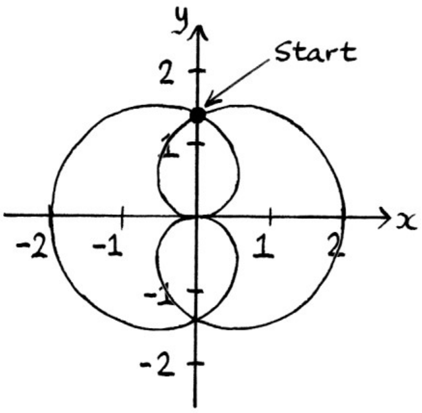


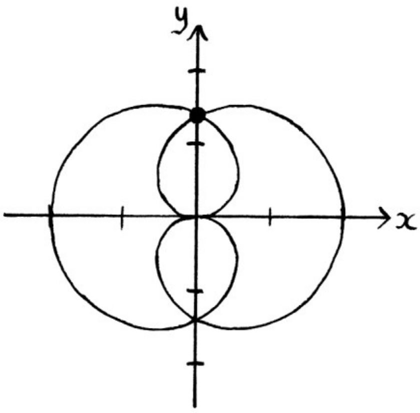


Next, we will give the circles phases that have the same ratio to each other as the frequencies of each circle have to each other. In other words, because the frequency ratio of the two circles is 1 : 3, we will give the circles phases that have the same ratio. Therefore, we will give the first circle a phase of 45 degrees, and the second circle a phase of 45 * 3 = 135 degrees. We will be adding the circles that represent these pairs of waves:

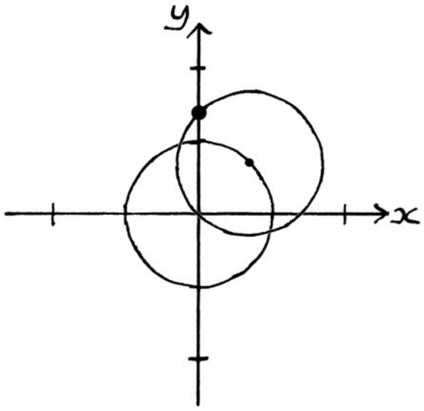
“y = sin ((360 * 1.5t) + 45)” and “y = cos ((360 * 1.5t) + 45)”
... and:
“y = sin ((360 * 4.5t) + 135)” and “y = cos ((360 * 4.5t) + 135)”.

The resulting shape looks like this:





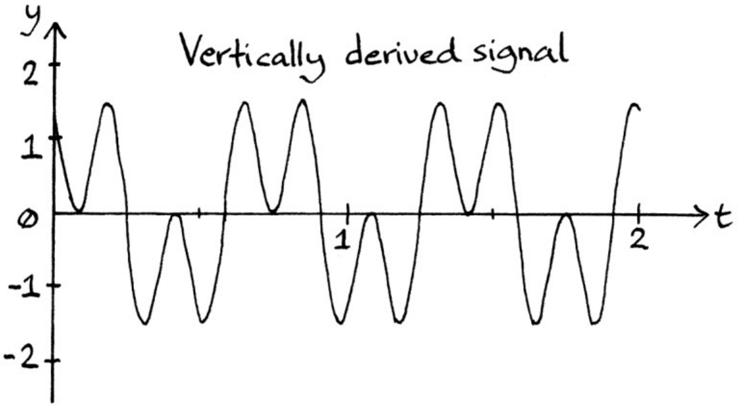
This shape is identical in look and rotation to the shape made from circles with no phase. The difference is that the outer object starts at 90 degrees instead of at 0 degrees. The reason it starts at 90 degrees is made clearer by looking at the circles at t = 0:

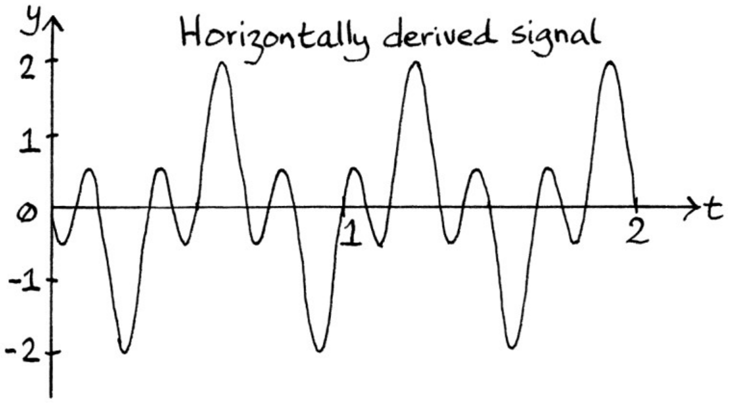


The shape is not rotated because the circles combine in an identical way to when there was no phase, but they start at 90 degrees instead of at 0 degrees. Having phases in the same ratio as the frequencies draws an identical shape with an identical rotation to having zero phases, but the object starts in a different place.

The fact that the shape has the same rotation, but the object starts in a different place means that the derived signals will have an identical curve to those derived from adding the same circles with no phase, but they will be shifted. In this particular case, they will be shifted by 0.08333 seconds to the left (a twelfth of a second).

The derived signals look like this:





Seeing how the shape becomes rotated, and how the starting point of the outer object is changed depending on the phases, explains how adding pure waves with different phases can have skewed results. We looked at this in the last chapter, but now we have seen the behaviour of the circles, it should make more sense.

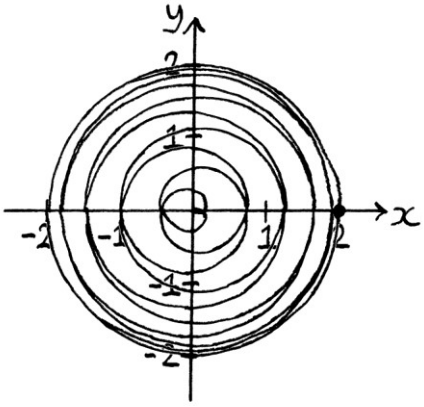
Oddities of adding frequencies

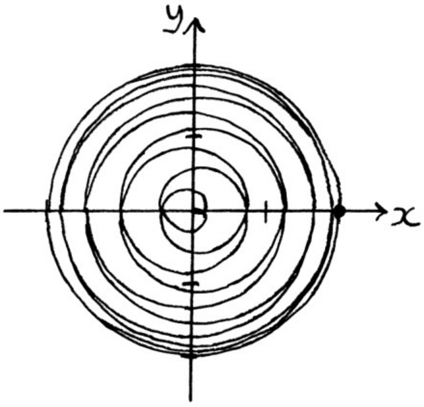
Close frequencies

If the frequencies of the added circles are close to each other, then they will take longer to align, and therefore, the resulting shape will have more lines going around it. The reason being that the circles nearly align at the end of one revolution around the axes, but are slightly off. It can take several revolutions for the cycles to align.

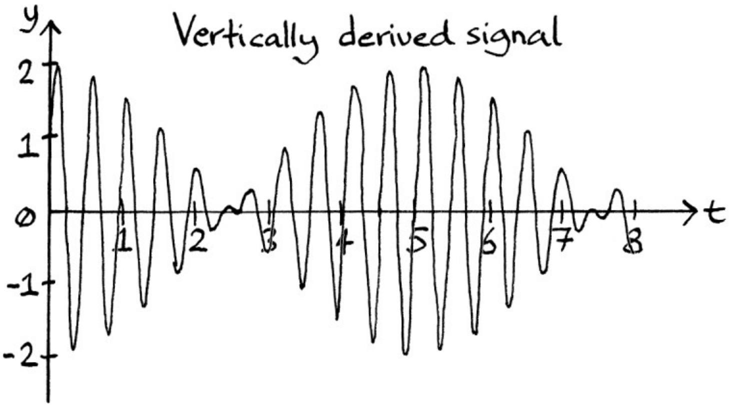
For example, we will add the two circles that represent the pairs of waves:
“y = sin (360 * 2t)” and “y = cos (360 * 2t)”
... and:
“y = sin (360 * 2.2t)” and “y = cos (360 * 2.2t)”

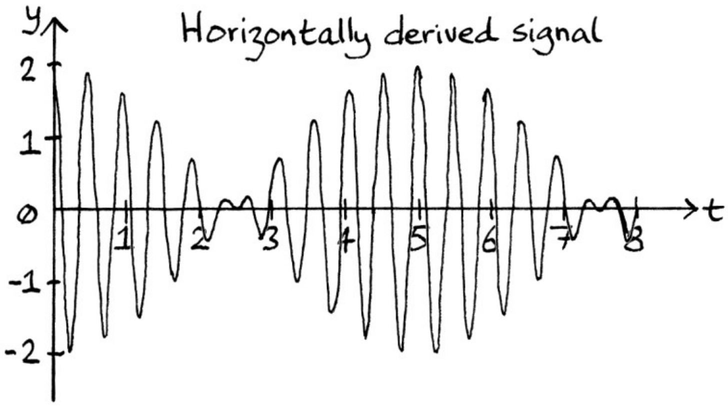
The result is shown with axes numbering and without axes numbering to make the picture clearer.





The derived signals from this shape have cycles that take longer to repeat:



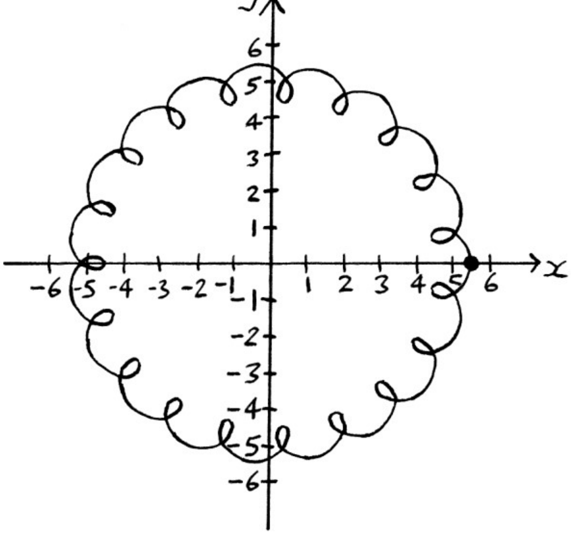


Significantly different amplitudes and frequencies

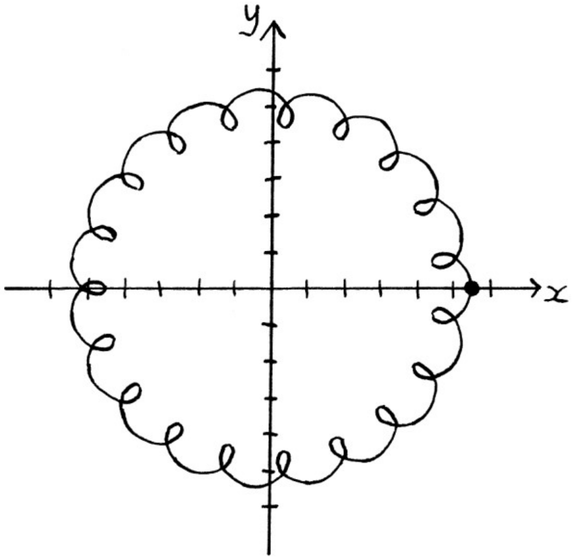
If the frequencies of two circles are significantly different from each other, then the faster object will rotate many times in the time it takes the slower object to move just slightly. If the amplitudes are also significantly different, with the faster circle having the smaller amplitude, then this will result in lots of loops in the resulting shape.

As an example, we will add the circles that represent the two pairs of waves:
“y = 5 sin 360t” and “y = 5 cos 360t”
... and:
“y = 0.5 sin (360 * 20t)” and “y = 0.5 cos (360 * 20t)”

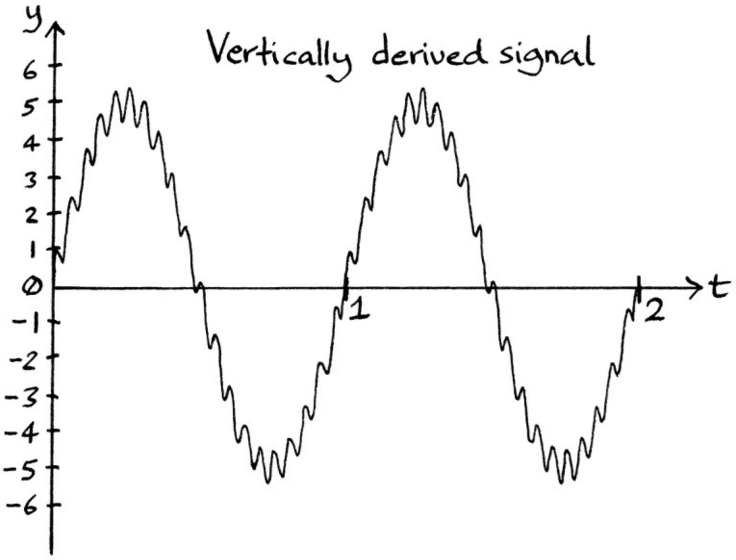
The resulting shape looks like this:

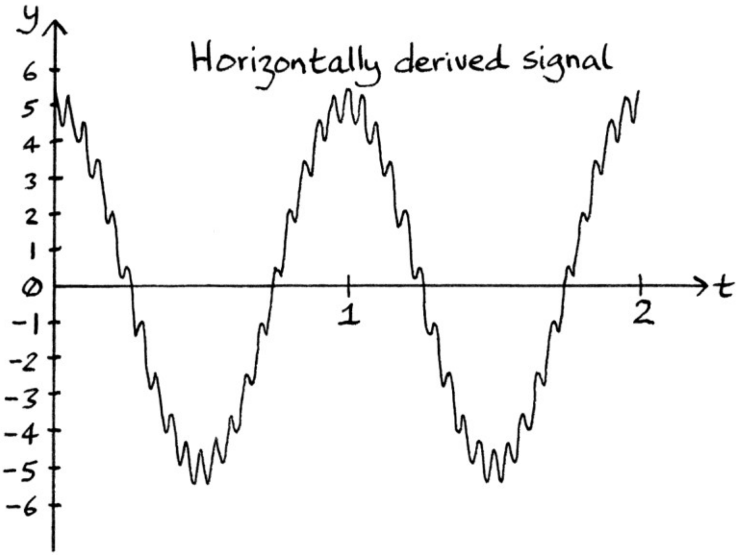


Or, if we remove the axis numbering to make the picture clearer, it looks like this:



The derived signals from this shape have the overall curve of the slower frequency wave, but with many small fluctuations:





Adding more than two circles

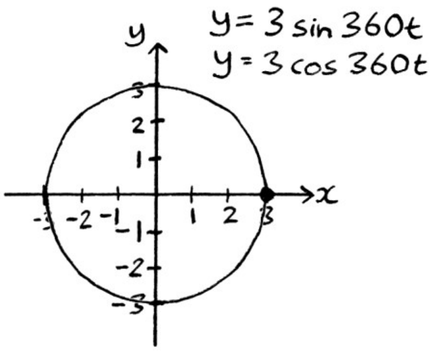
The concept of adding more than two circles is just as straightforward as the concept of adding two circles. We arrange the circles using the same system as for two circles. Each circle is centred on the phase point of the previous circle, and the outer object gives the path of the resulting shape.

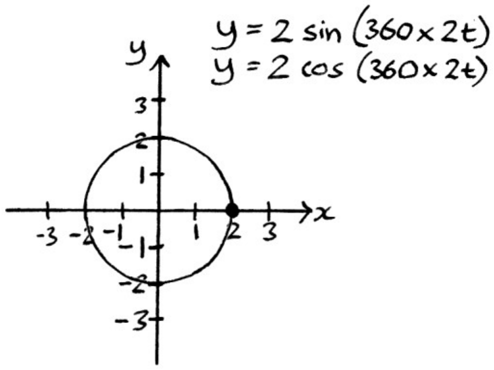
With every extra circle, the overall movement of the outer object becomes more complicated. There is no limit to the number of circles we can add together.

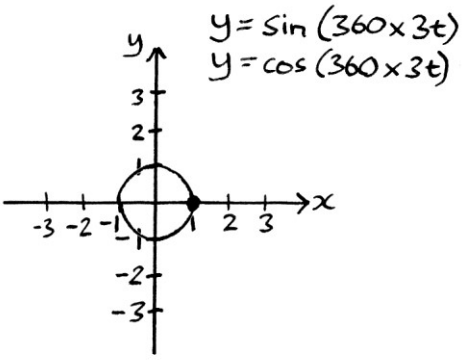
The actual order of the circles does not make any difference to the movement of the outer object. You might have been able to guess that the order makes no difference, because if we were adding the individual waves derived from each circle, it would not make any difference in which order we added them.

For the purposes of illustrating how adding several circles works, it is clearer to have the larger radius circles nearer the origin of the axes than the smaller radius circles.

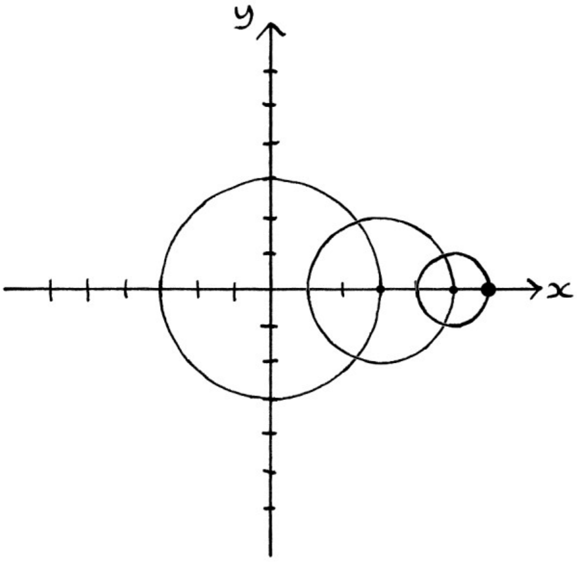
As an example, we will add the following three circles:





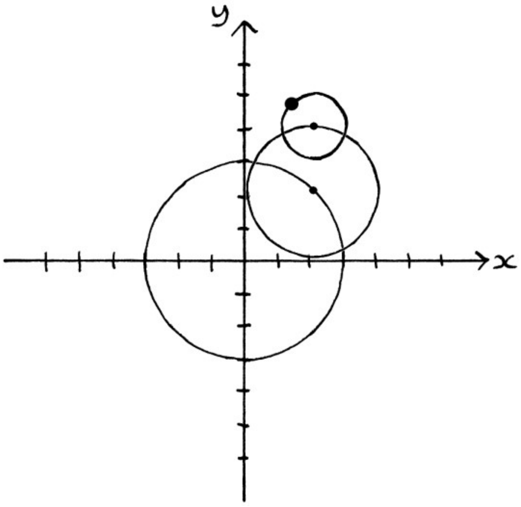


We will arrange them as so, but the order does not matter:

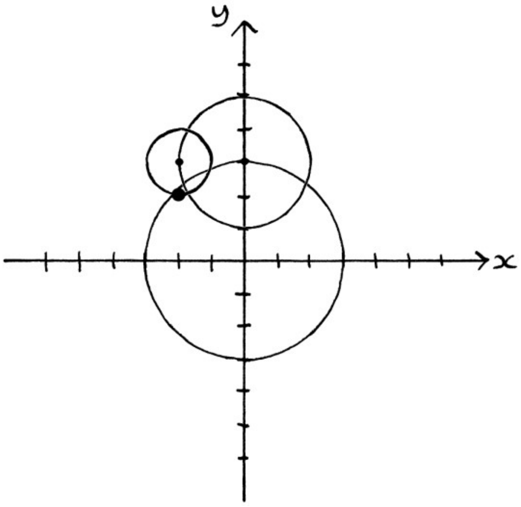


The outer object rotates around its circle, and that circle rotates around the middle circle, and the middle circle rotates around the inner circle.

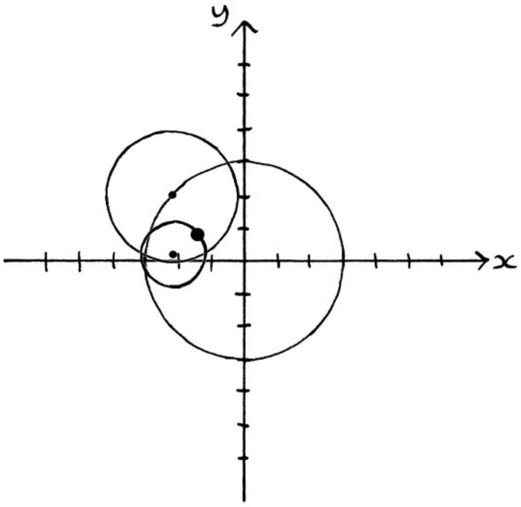
At t = 0.125 seconds, they look like this:



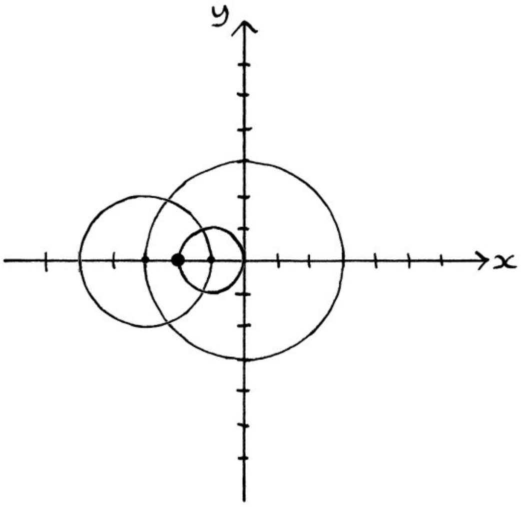
At t = 0.25 seconds:



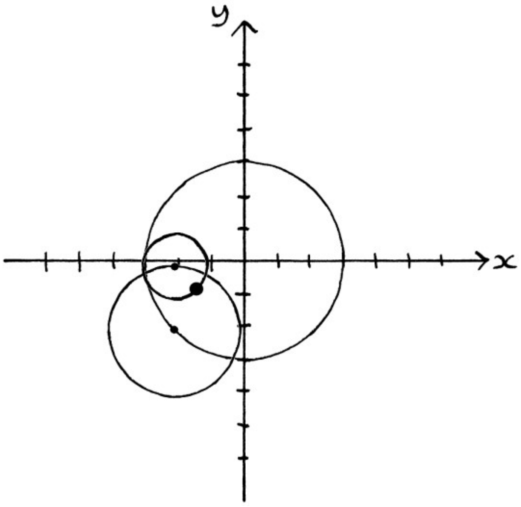
At t = 0.375 seconds:



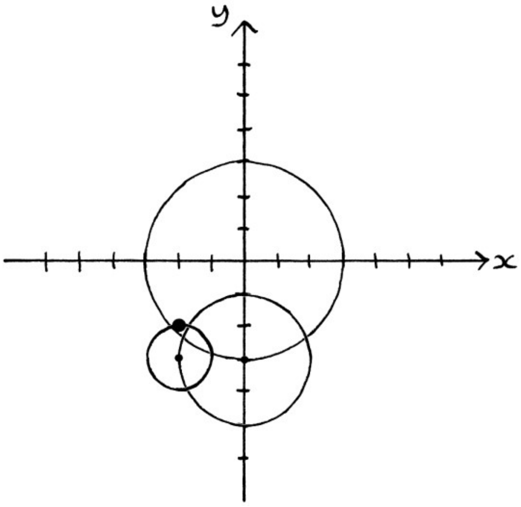
At t = 0.5 seconds:



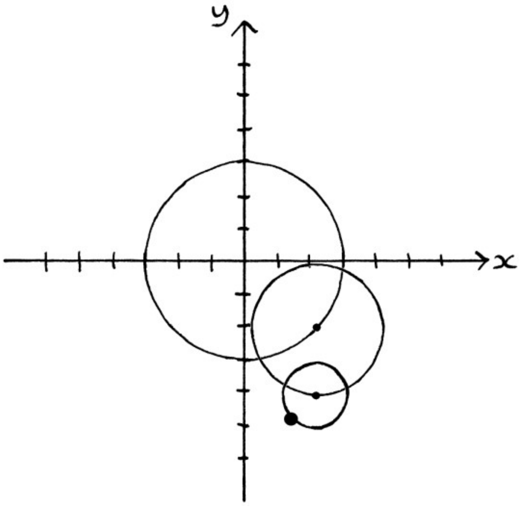
At t = 0.625 seconds:



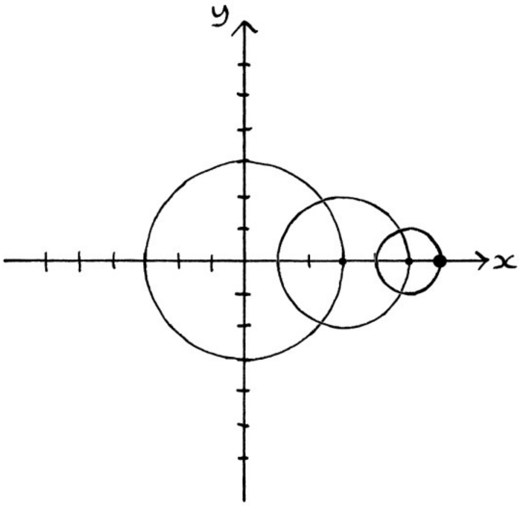
At t = 0.75 seconds:



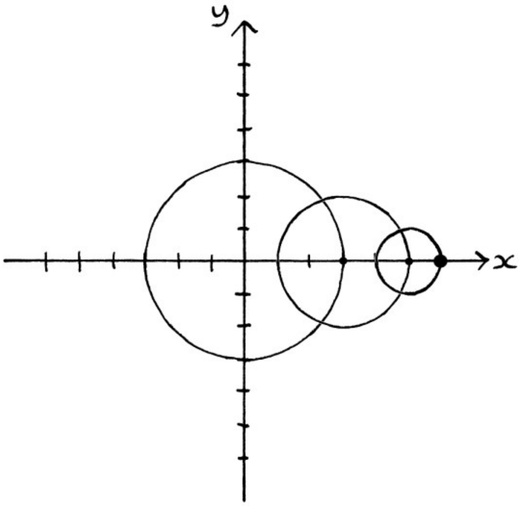
At t = 0.875 seconds:

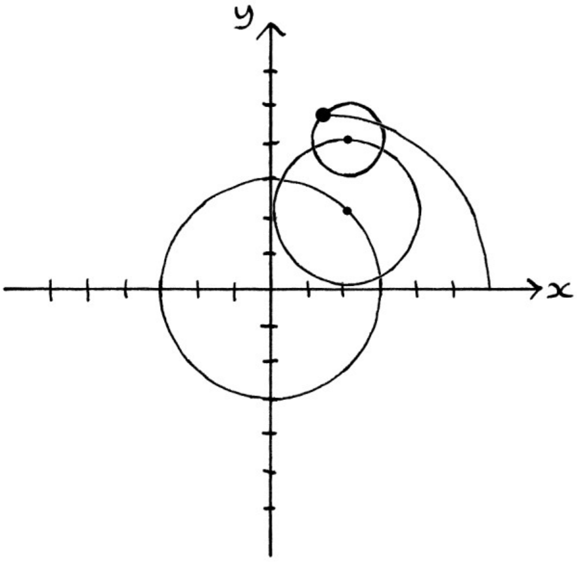


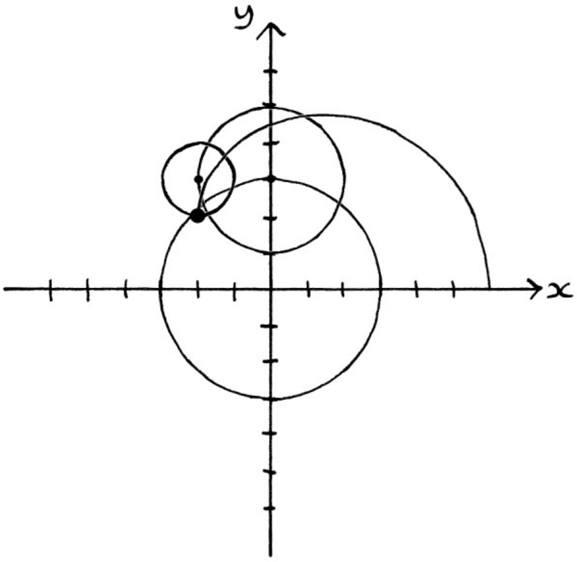
At t = 1 second:

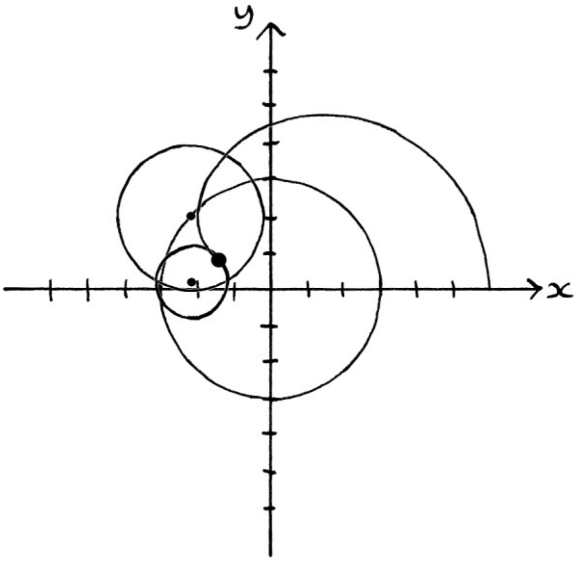


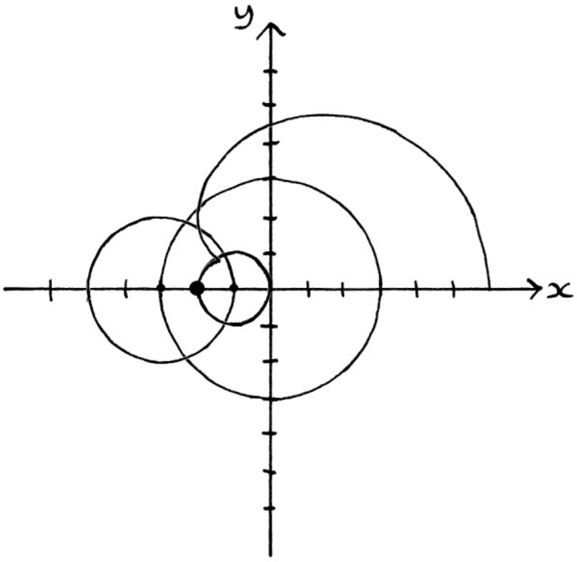
The outer object draws out a shape as it moves:

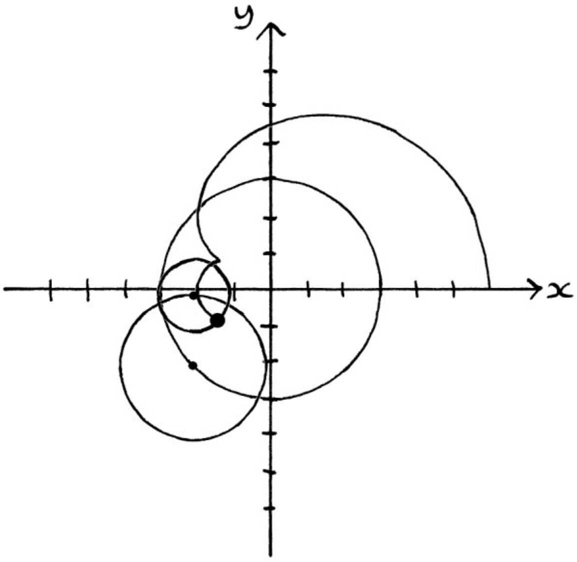


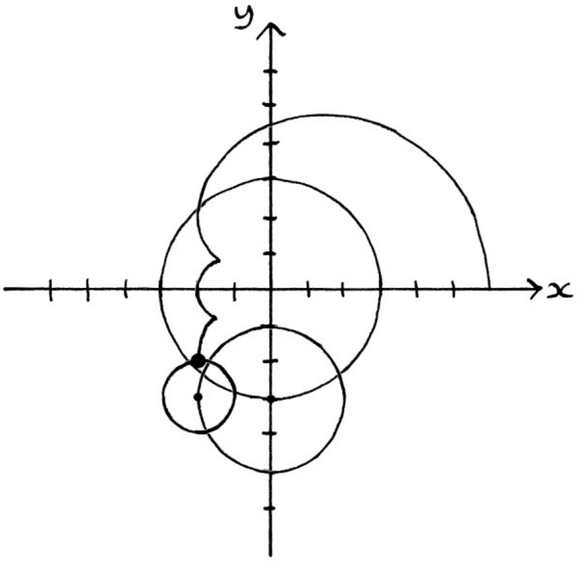

y
x

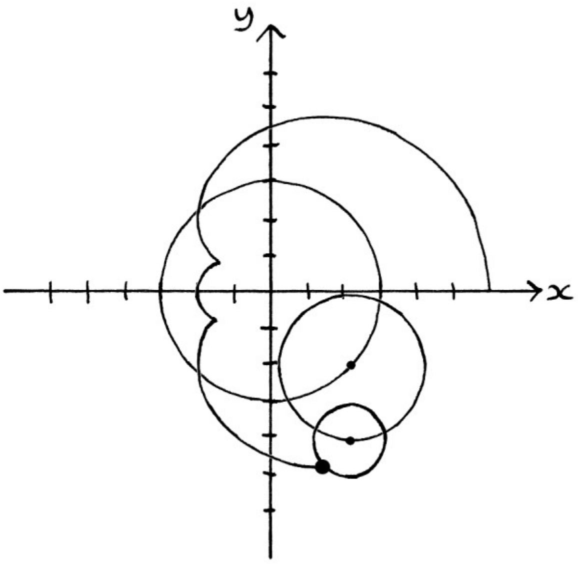

y
x

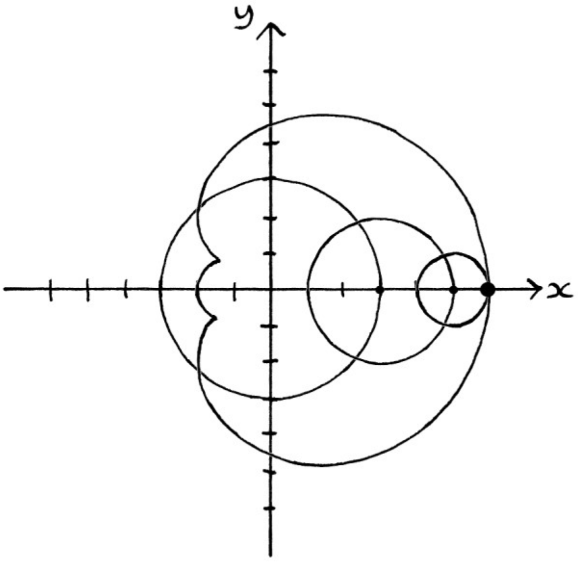

y
x


y
x

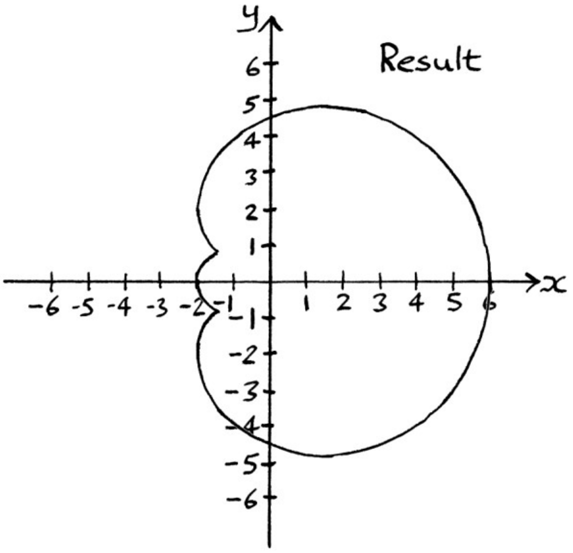

y
x


y
x

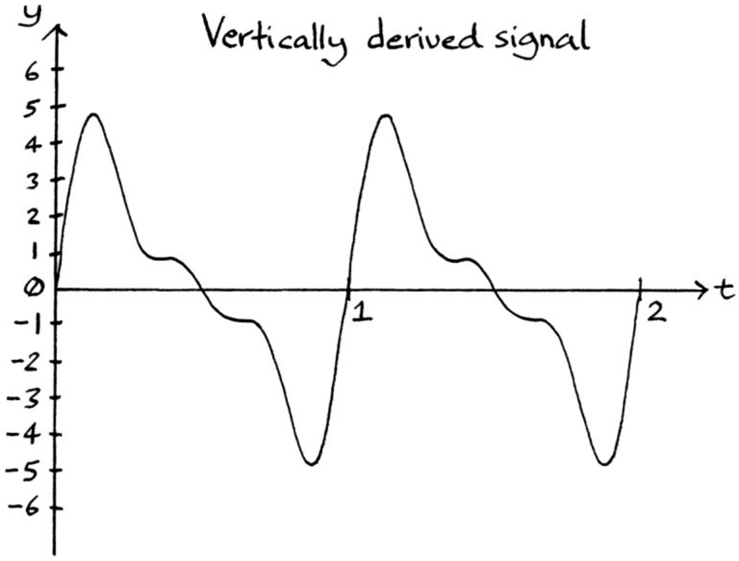


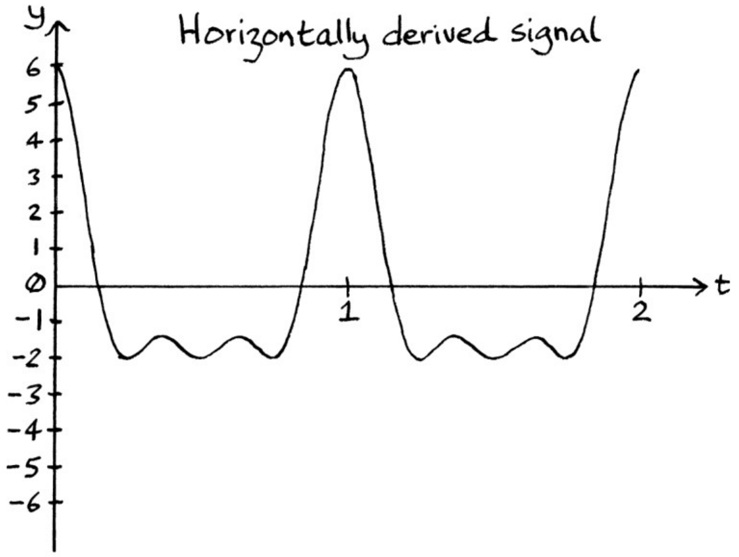


The resulting shape looks like this:



The two derived signals look like this:





Subtracting circles

Subtracting one circle from another is as straightforward as adding two circles, and it can be done in exactly the same way as adding two circles.

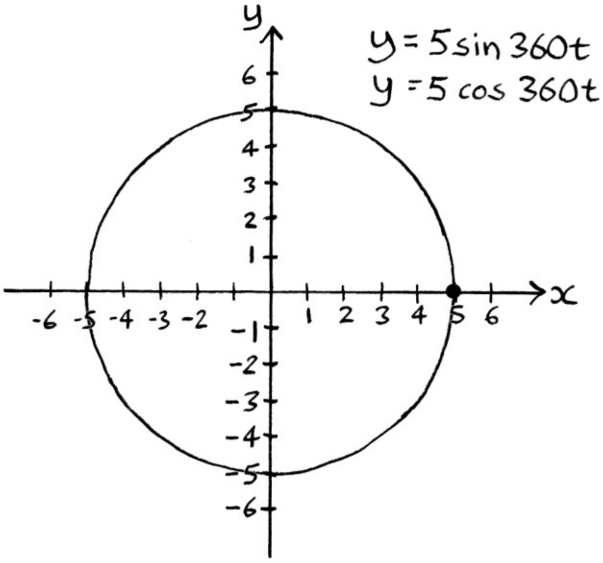
If we had the calculation:
Circle A − Circle B
... we would know that the x-axis points of the resulting shape would be made up of the Cosine wave from Circle A minus the Cosine wave from Circle B for corresponding moments in time. The y-axis points of the resulting shape would be made up of the Sine wave from Circle A minus the Sine wave from Circle B for corresponding moments in time.

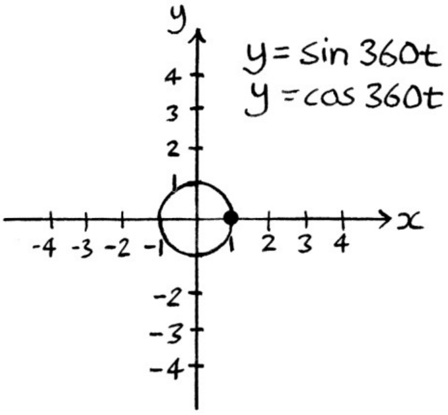
A wave minus another wave is the same as a wave added to a wave with a 180-degree phase in its formula. Therefore, we can rephrase the subtraction of circles as:
Circle A + (Circle B with a 180-degree phase).

Subtraction of one circle from another is just addition of the same circles with the subtracted one rephrased to have a +180 degree phase. [If the subtracted circle had a non-zero phase to start with, then we would add or subtract 180 degrees to or from that phase to achieve the same effect].

We will use one of the very first examples in this chapter, where the phase and mean level are zero, and the frequencies are the same, but do a subtraction instead of an addition. We will start with the circle that represents this pair of waves:
“y = 5 sin 360t” and “y = 5 cos 360t”
... and we will subtract from that, the circle that represents the following pair of waves:
“y = sin 360t” and “y = cos 360t”.

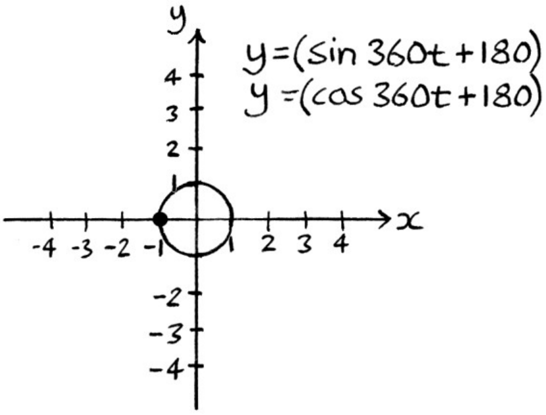
The circles for these waves look like this:



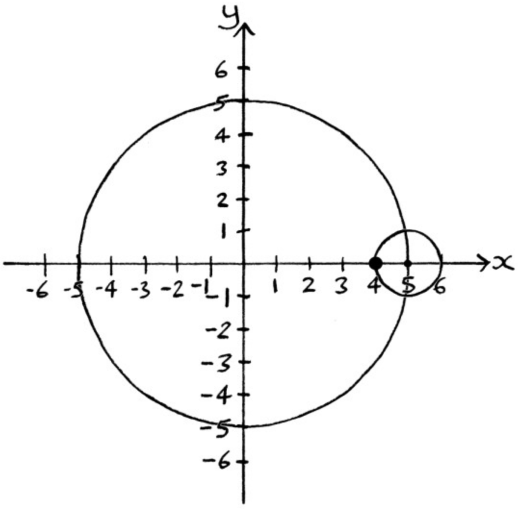


As we are doing a subtraction, we can rephrase the calculation to be an addition with the subtracted circle given a phase of +180 degrees. Therefore, we end up adding the two circles that represent the following pairs of waves:
“y = 5 sin 360t” and “y = 5 cos 360t”
... and:
“y = sin (360t + 180)” and “y = cos (360t + 180)”

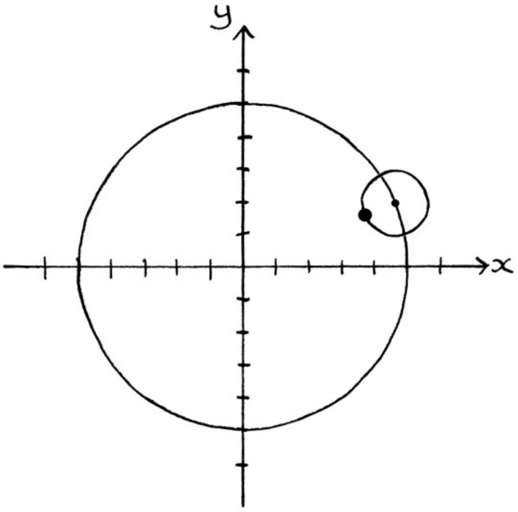
The new little circle looks like this. The phase point has moved to 180 degrees:



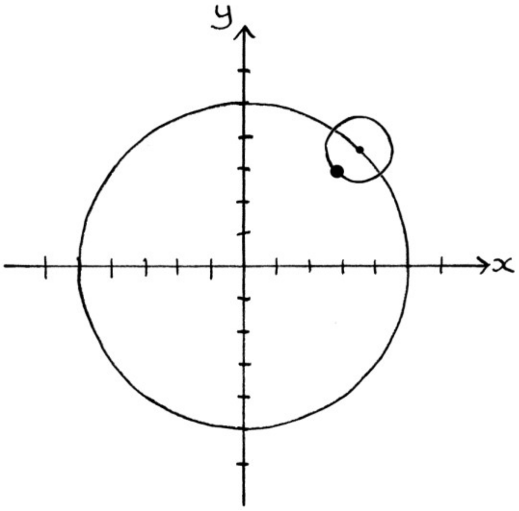
We arrange the circles, so at t = 0, they look like this:



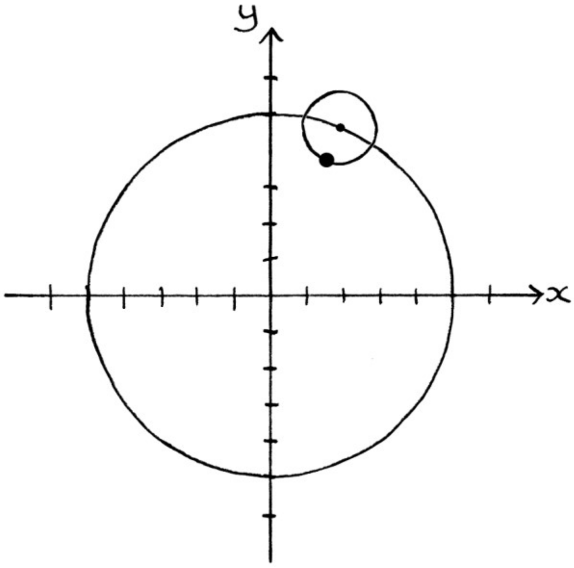
At t = 0.0625 seconds, the circles look like this:



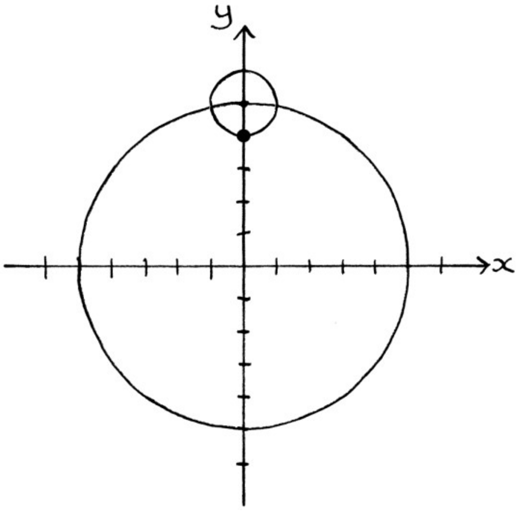
At t = 0.125 seconds:



At t = 0.1875 seconds:



At 0.25 seconds:



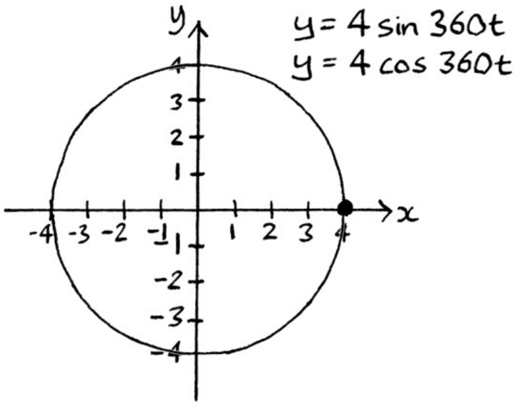
... and so on. As we can see, the object rotating around the outer circle will draw out a perfect circle that has a radius of 5 – 1 = 4 units. This is the result we would expect from knowing how the addition and subtraction of waves works.

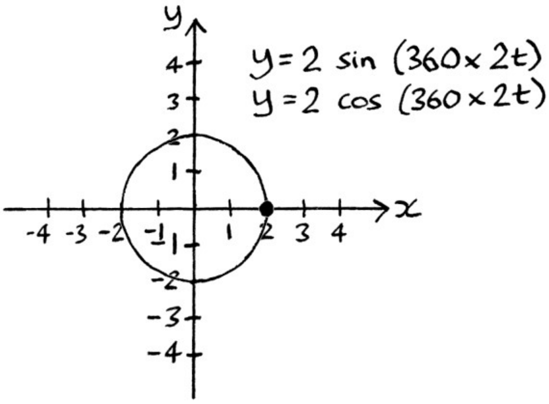
Subtraction with circles of different frequencies

If the frequencies are the same, then subtraction and its results are straightforward. If the frequencies are different, then subtraction is still straightforward, but it is harder to know how the result will look.

As an example, we will start with the circle that represents these waves:
“y = 4 sin 360t” and “y = 4 cos 360t”
... and subtract from it the circle that represents these waves:
“y = 2 sin (360 *2t)” and “y = 2 cos (360 * 2t)”.

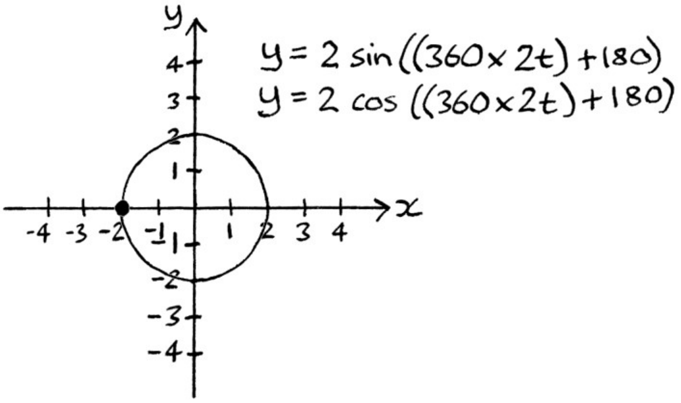
The circles look like this:



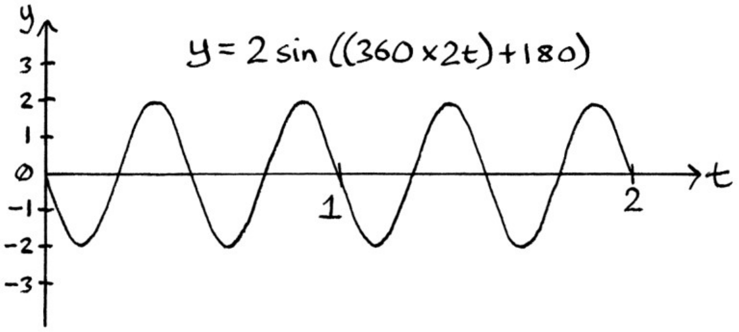


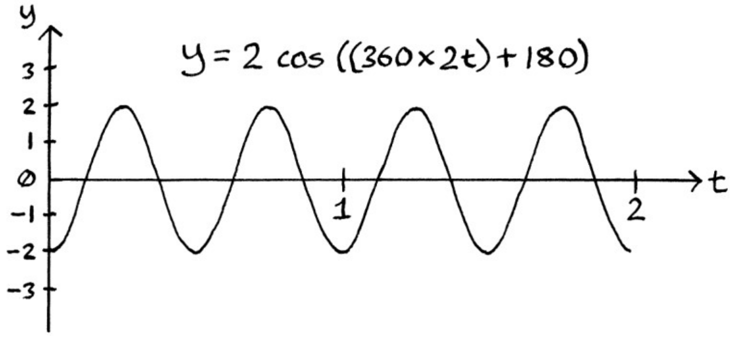
As we are doing a subtraction, we can rephrase it as an addition with the subtracted circle rewritten to have a phase of +180 degrees. Therefore, we can say we are adding the two circles that represent the following two pairs of waves:
“y = 4 sin 360t” and “y = 4 cos 360t”
... and:
“y = 2 sin ((360 *2t) + 180)” and “y = 2 cos ((360 * 2t) + 180)”.

The smaller circle now looks like this (the phase point has moved from 0 degrees to 180 degrees):

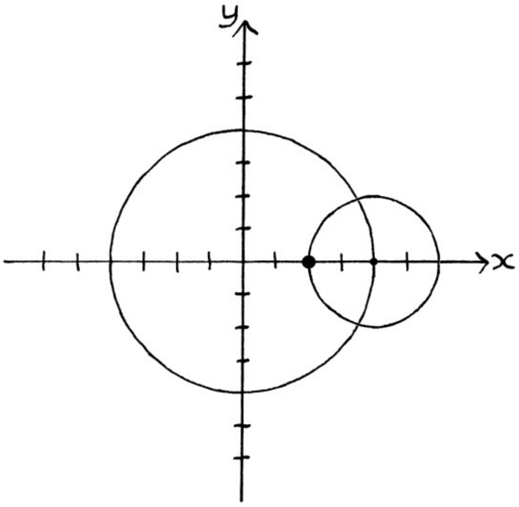


The waves that this new circle represents look like this:

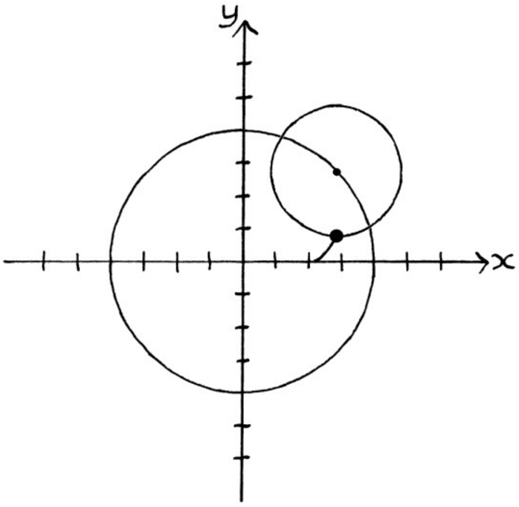




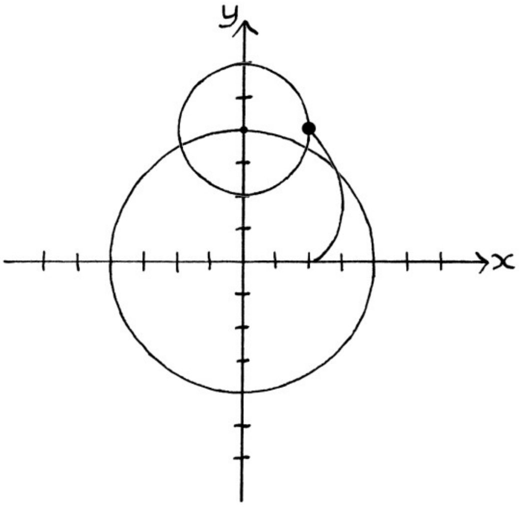
We put the circles together, and at t = 0, they look like this:



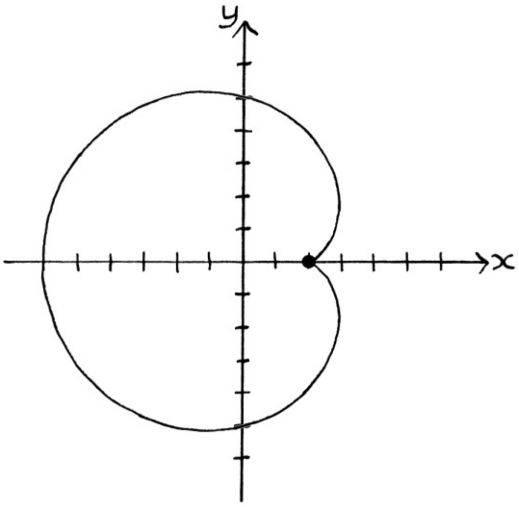
At t = 0.125 seconds, the circles look like this, drawn with the path that the outer object makes:



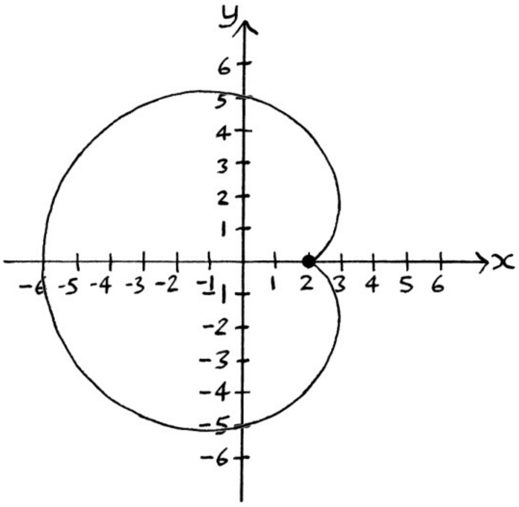
At t = 0.25 seconds, they look like this:



... and skipping ahead to 1 second, the shape that the outer object will have drawn will look like this:

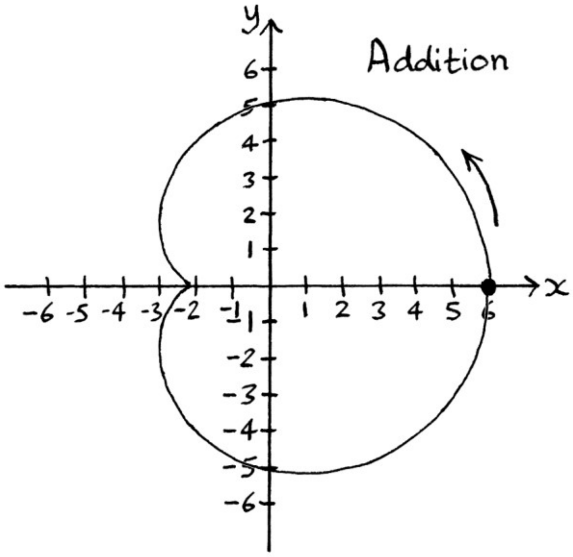


... and that same shape with numbered axes looks like this:



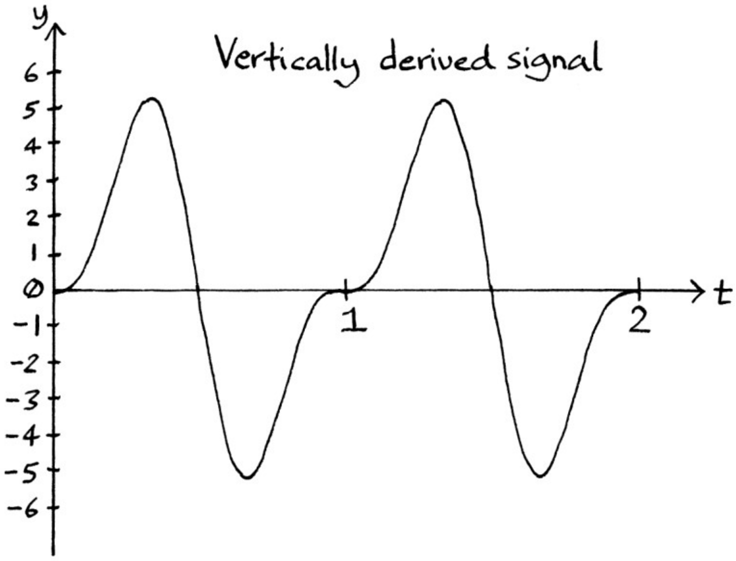
What is interesting about this shape is that it is the mirror image of the result of *adding* the two circles that represent the pairs of waves:
“y = 4 sin 360t” and “y = 4 cos 360t”
... and:
“y = 2 sin (360 *2t)” and “y = 2 cos (360 * 2t)”

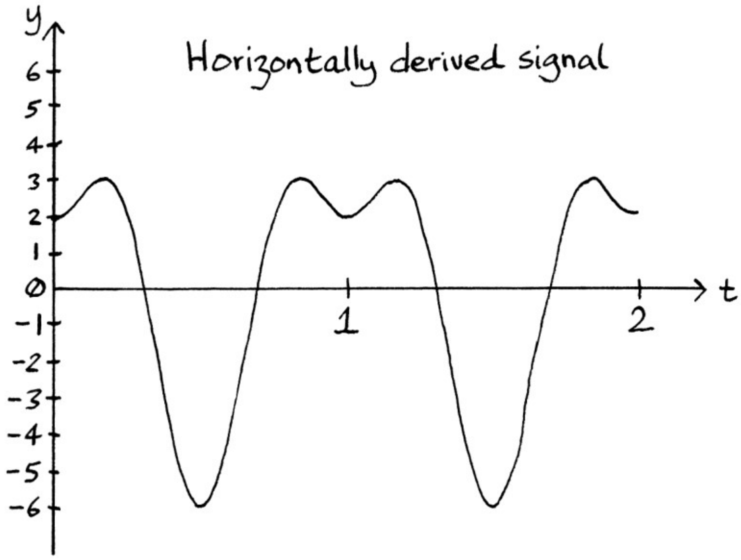
That addition results in this:



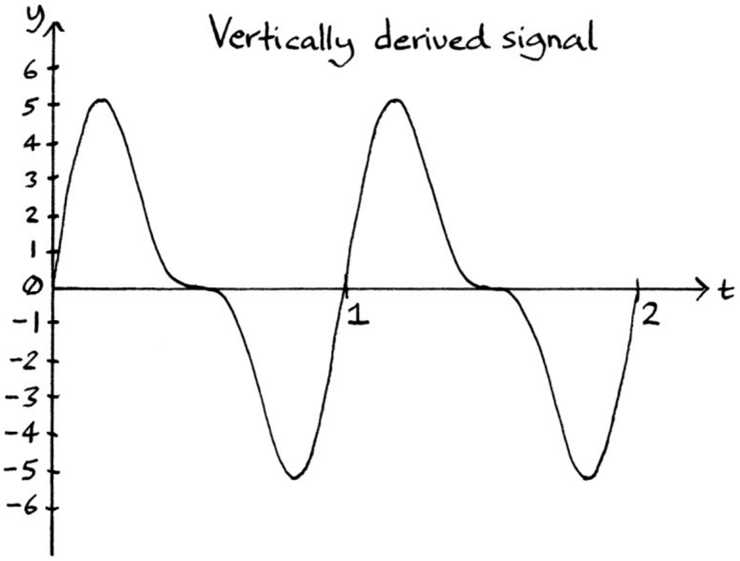
The area inside the shape is the same whether we add or subtract the second circle. This is a good example of how seemingly unpredictable the results of addition or subtraction can be when dealing with circles with different frequencies.

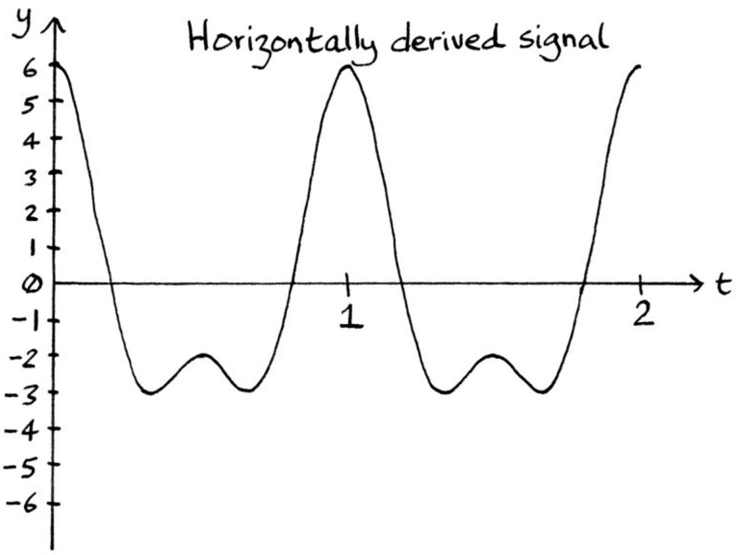
The derived signals from the *subtraction* look like this:





...whereas the derived waves from the *addition* would have looked like this:





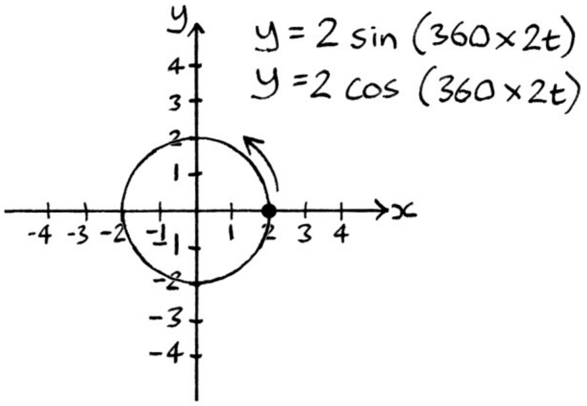
The derived signals from the addition are the same as those from the subtraction, but flipped vertically and shifted to the left (or right) by half a second.

Subtraction of identical circles

The subtraction of one circle from an identical circle will result in nothing on the circle chart. Subtracting one Sine wave from an identical one, or subtracting one Cosine wave from an identical one, will result in a flat line at y = 0. Therefore, subtracting a circle (from which a Sine wave and Cosine wave derive) from an identical circle will result in a shape where all the x-axis coordinates are zero and all the y-axis coordinates are zero for all time. Although this might be obvious, we can see the process working by using rotating circles.

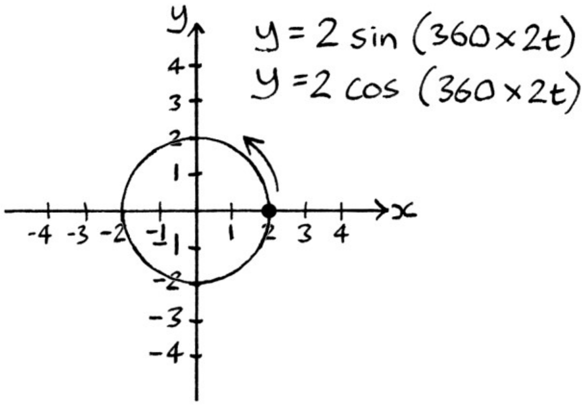
We will start with the circle that represents the following two waves:
"y = 2 sin (360 * 2t)" and "y = 2 cos (360 * 2t)"
... and we will subtract it from a copy of itself.

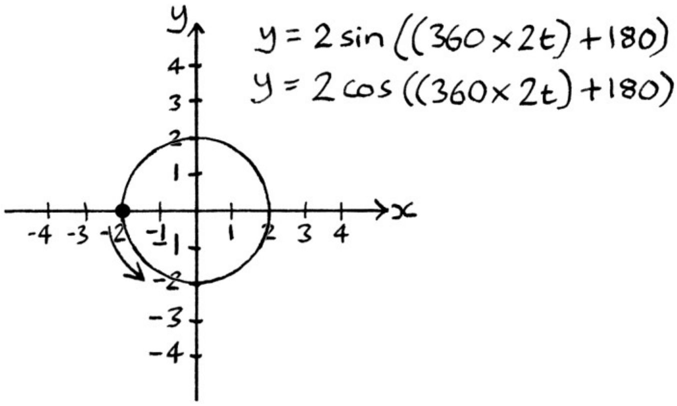
Therefore, both our circles will look like this one:



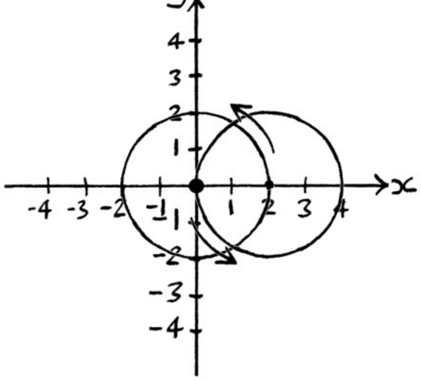
We will turn the calculation into an addition, where the subtracted circle becomes a positive circle with a phase of 180 degrees. The second circle therefore represents these waves: "y = 2 sin ((360 * 2t) + 180)" and "y = 2 cos ((360 * 2t) + 180)".

The circles that we are now adding look like this:

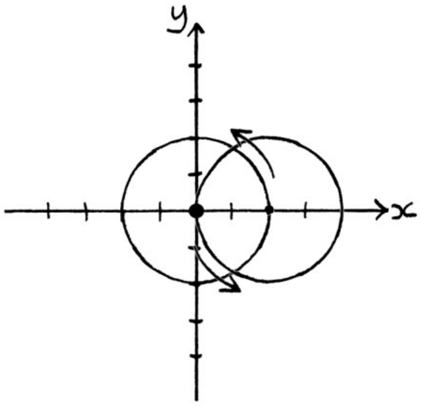



$y = 2\sin((360 \times 2t) + 180)$
$y = 2\cos((360 \times 2t) + 180)$
y
x
4
3
2
1
-1
-2
-3
-4
-4
-3
-2
1
2
3
4

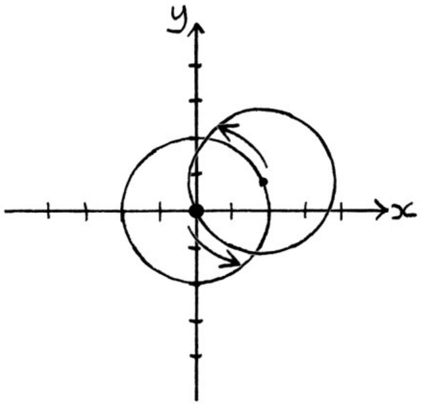
We will arrange the circles as in the following picture, with the circle with the 180-degree phase on the outside (although we could arrange them either way around, as it does not make a difference).

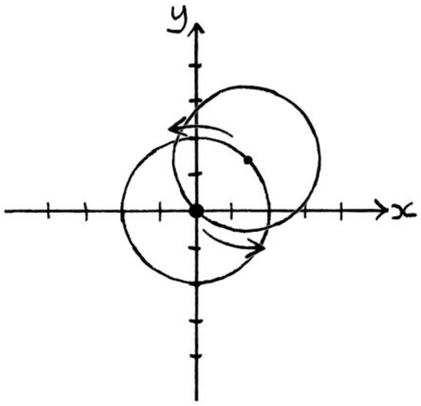


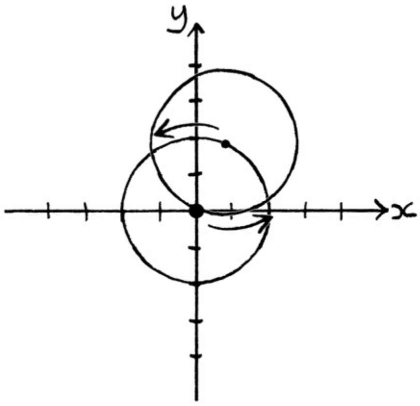
At t = 0 the circles look like this:

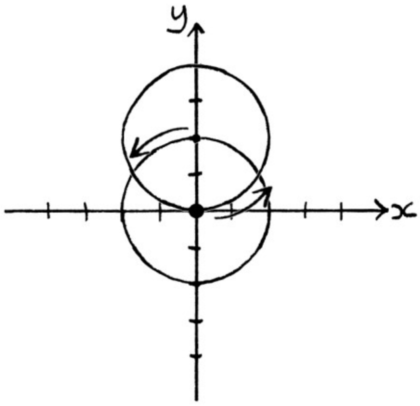


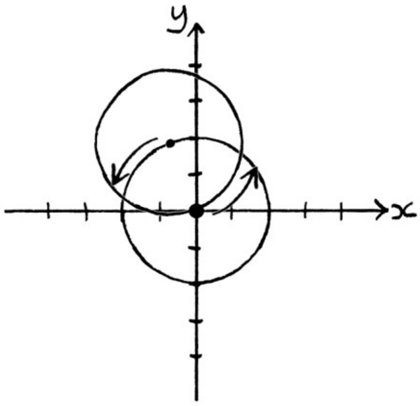
... and then the circles appear as in the following pictures for the next 0.5 seconds, at intervals of 0.03125 seconds (1/32 of a second):

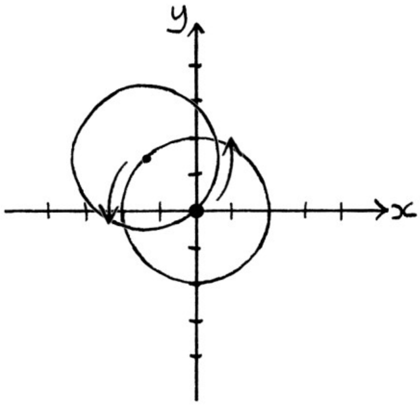


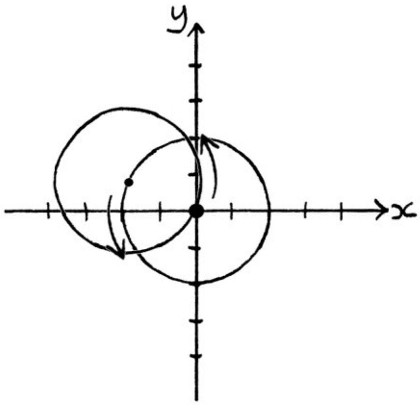


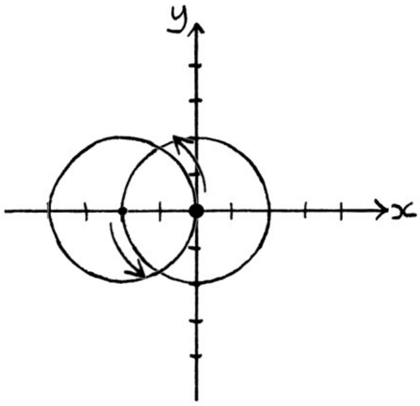

y
x

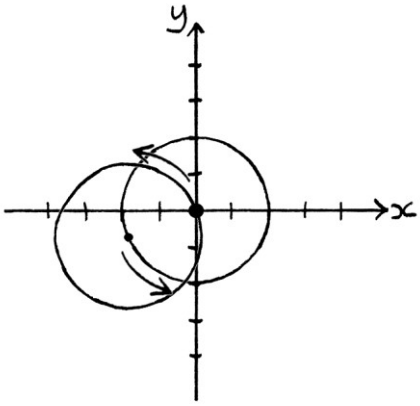

y
x

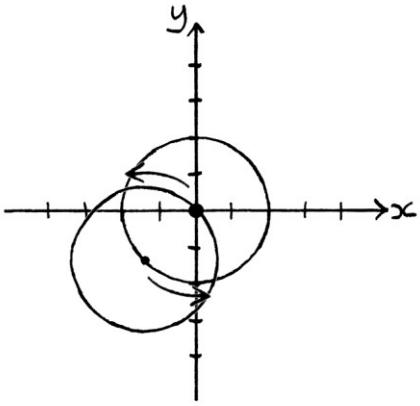

y
x

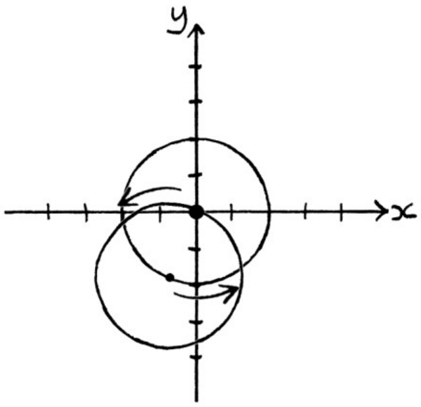

y
x

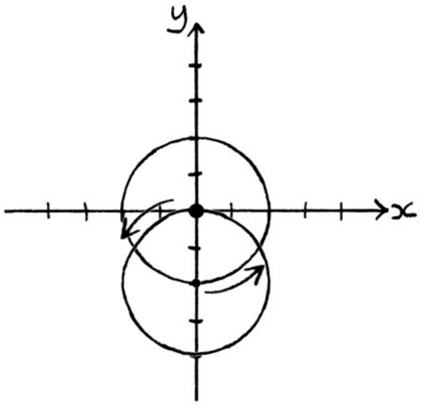

y
x

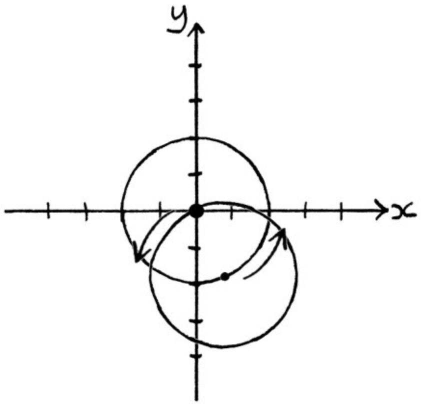

y
x

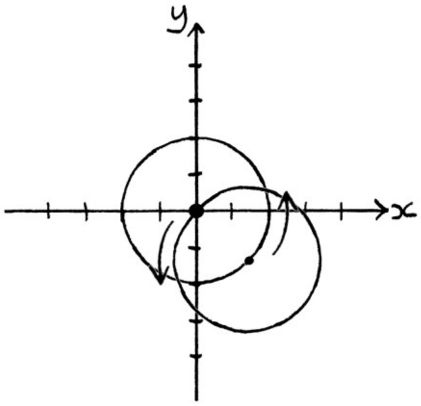

y
x

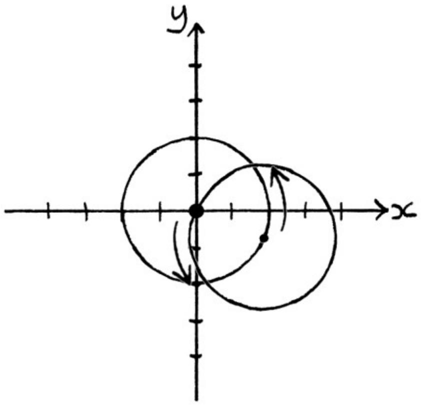

y
x

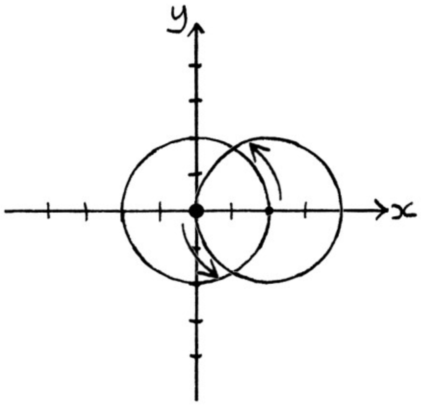

y
x


y
x


y
x

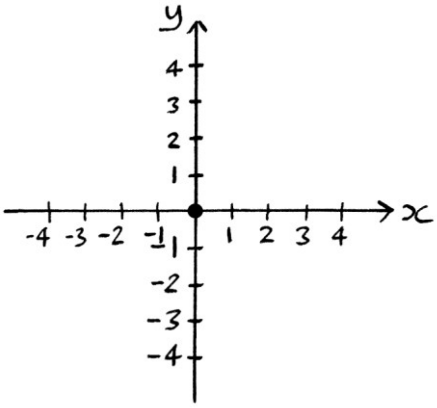

y
x



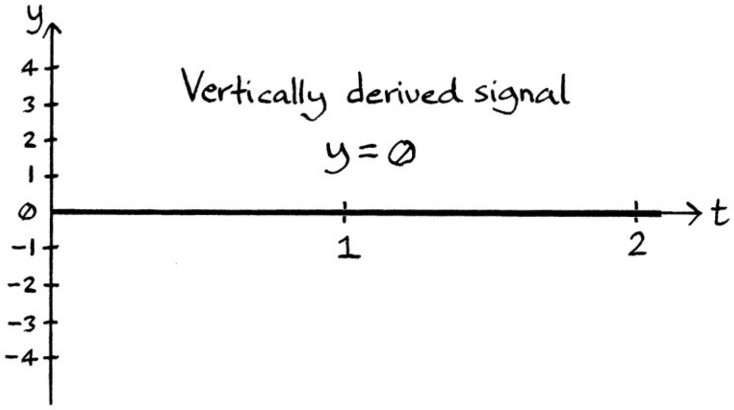


As we can see, the outer object remains at the origin of the axes, regardless of the rotation of the two circles. This is because the frequencies are the same. There is no resulting shape to this calculation.

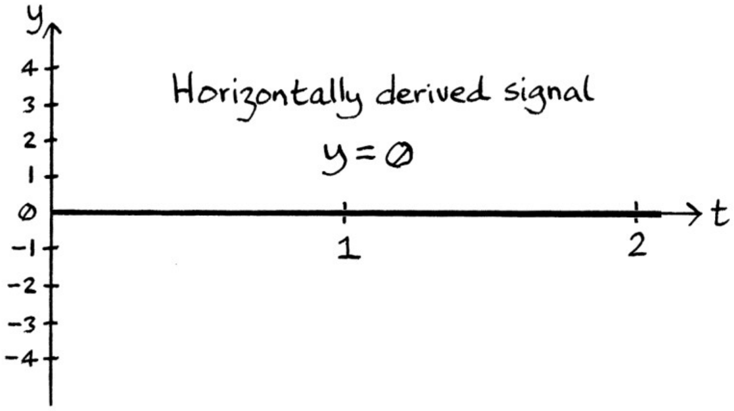
The outer object is at the coordinates (0, 0) for all time:



This means that the vertically derived signal will just be a horizontal line at y = 0:



... and the horizontally derived signal will also just be a horizontal line at y = 0.



Negative frequencies

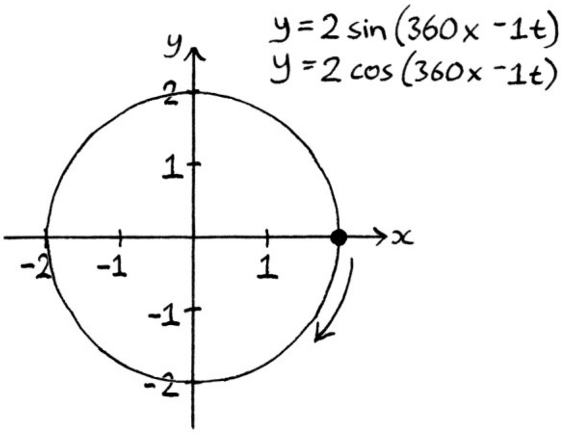
Two negative frequencies

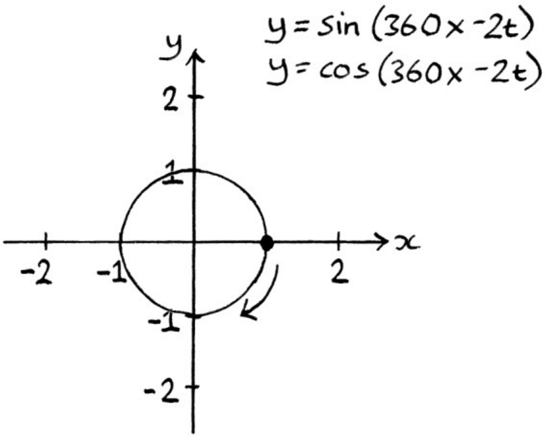
If we add two circles, both with negative frequencies, the outer object will travel the other way around the shape, which means that the shape will be drawn in the opposite direction to if the frequencies had been both positive. If the phases are both zero, the finished resulting shape will look the same as if the two circles had had positive frequencies.

If the phases are both zero, the vertically derived signal will be an upside-down version of the vertically derived signal that we would have if both frequencies had been positive. The horizontally derived signal will be identical to the one we would have if both frequencies were positive. The reason for this is clearer if we consider the individual circles. An object rotating the "wrong way" around a circle produces an upside-down Sine wave and a normal Cosine wave (if there is zero phase). Adding two upside-down Sine waves will produce the upside-down version of adding two correct-way-up Sine waves. Adding two normal Cosine waves results in the addition of two normal Cosine waves (obviously). Therefore, the vertically derived signal is upside down, and the horizontally derived signal is the correct way up.

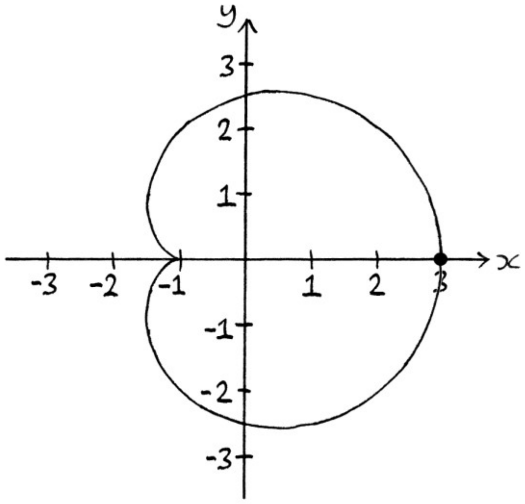
As an example, we will add the two circles that represent the following waves:
"y = 2 sin (360 * −1t)" and "y = 2 cos (360 * −1t)"
... and:
"y = sin (360 * −2t)" and "y = cos (360 * −2t)"

The two circles look like this:

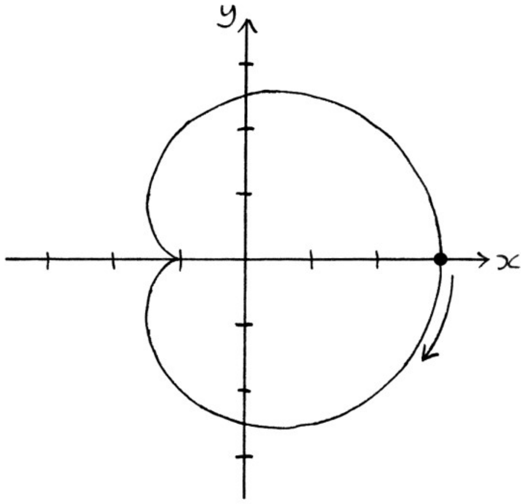


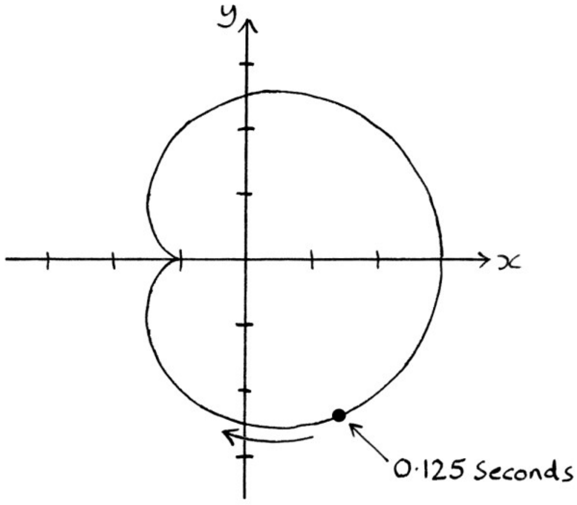


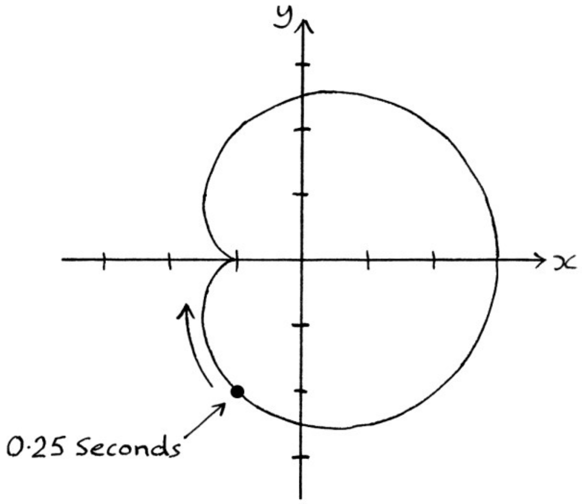
The resulting shape is shown in the following picture. It is identical to the shape that would have been created with positive frequencies...



... however, the outer object moves around the outer shape *clockwise* as so:

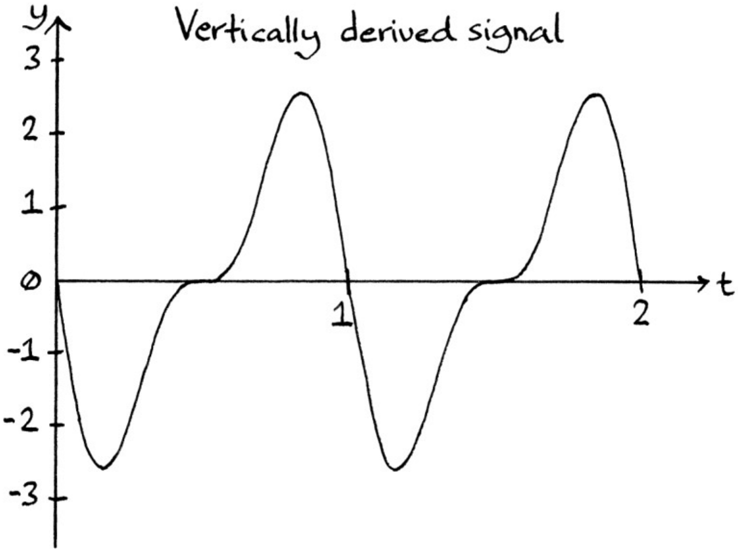






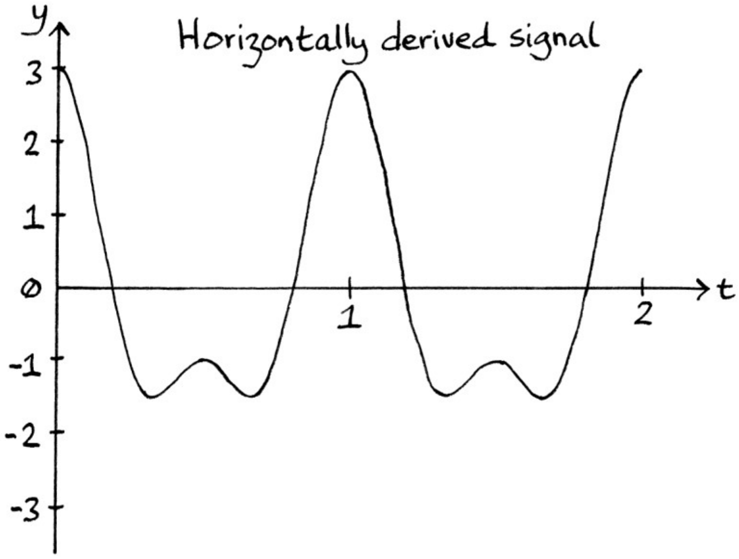
... and so on.

The resulting vertically derived signal looks like this:



It is the upside-down version of the vertically derived signal that would be obtained if the frequencies had both been positive.

The horizontally derived signal looks like this:



This is identical to the horizontally derived signal that would be obtained if both frequencies had been positive.

One negative frequency; one positive frequency

The idea of adding one circle with a positive frequency to another circle with a negative frequency sounds complicated, but is straightforward if we think about what is actually happening. The first thing to realise is that the object on one circle will be moving anticlockwise, while the object on the other circle will be moving clockwise.

It also pays to think about the underlying waves from the two original circles. The positive-frequency circle will have a normal Cosine wave and a normal Sine wave. If the phase is zero, the negative-frequency circle will have a normal Cosine wave and an *upside-down* Sine wave. Therefore, the resulting *x-axis* coordinates of the perimeter of the resulting shape will be the sum of the two Cosine waves for corresponding moments in time. They will be identical to the x-axis coordinates if the two circles both had positive frequencies. The resulting *y-axis* coordinates of the perimeter of the resulting shape will be one Sine wave added to an upside-down Sine wave for corresponding moments in time. As the y-axis points of an upside-down Sine wave are the negative of a correct-way-up Sine wave, this all means that the y-axis coordinates of the resulting shape over time will be one Sine wave minus the other Sine wave.

This will be clearer with an example.

We will add the two circles that represent the following waves:
“y = 4 sin (360 * 0.25t)” and “y = 4 cos (360 * 0.25t)”
... and:
“y = 2 sin (360 * −2t)” and “y = 2 cos (360 * −2t)”

These are almost the same as the very first frequency example in this chapter, but the second circle has a negative frequency.

We know that the *x-axis* coordinates of the resulting shape for every moment in time will be:
4 cos (360 * 0.25t) + 2 cos (360 * −2t)
... which, because the phases are zero, is identical to:
4 cos (360 * 0.25t) + 2 cos (360 * 2t)

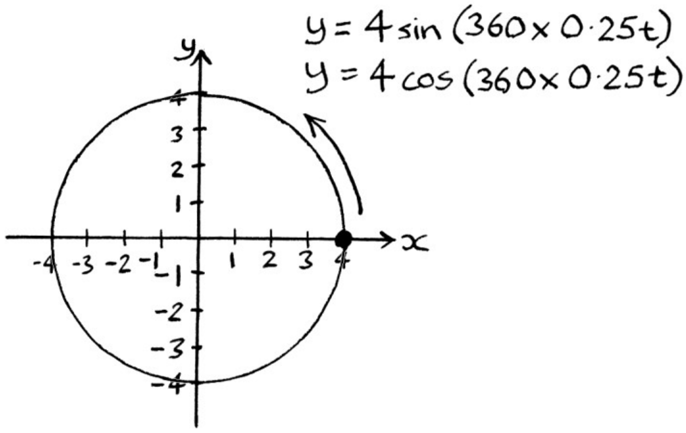
This is because a negative-frequency Cosine wave *with zero phase* has the same shape as, and is therefore identical to, a positive-frequency Cosine wave with zero phase.

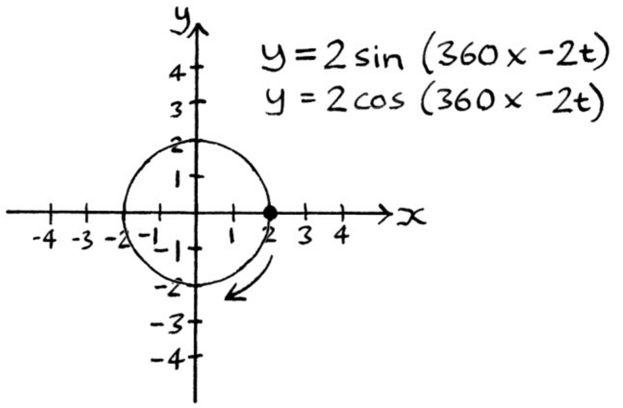
We also know that the *y-axis* coordinates of the resulting shape for every moment in time will be:
4 sin (360 * 0.25t) + 2 sin (360 * −2t)
... and this is identical to:
4 sin (360 * 0.25t) − 2 sin (360 * 2t)

This is because a negative-frequency Sine wave *with zero phase* has the same shape as, and is therefore identical to, an upside-down positive-frequency Sine wave with zero phase. An upside-down positive-frequency Sine wave with zero phase is the same as a positive-frequency Sine wave with zero phase and a negative amplitude. Therefore, adding the negative-frequency Sine wave is identical to adding that same Sine wave with a positive frequency and a negative amplitude, and that is the same as subtracting that same Sine wave with a positive frequency and a positive amplitude.

Another thing we know is that subtracting a wave is the same as adding that same wave with a +180 degree phase. Therefore, we could rephrase the sum to be:
4 sin (360 * 0.25t) + 2 sin ((360 * 2t) + 180)

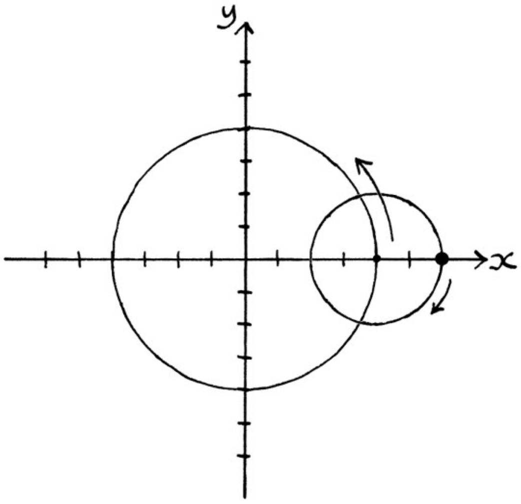
Knowing all this does not particularly help with predicting how the resulting shape will look though, so we will draw the shape using the two circles joined together. The two circles look like this:



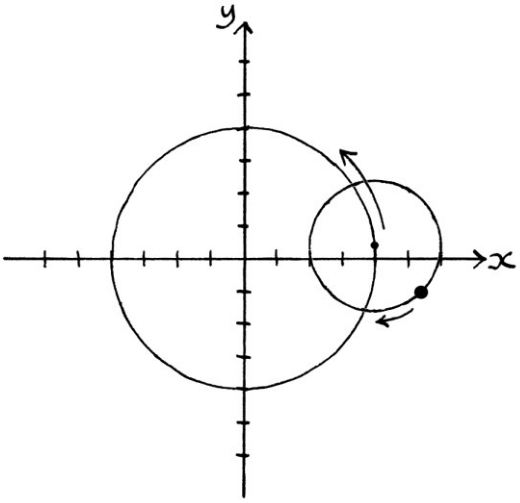


The outer object will rotate *clockwise* around its own circle, and that circle will rotate *anticlockwise* around the inner circle.

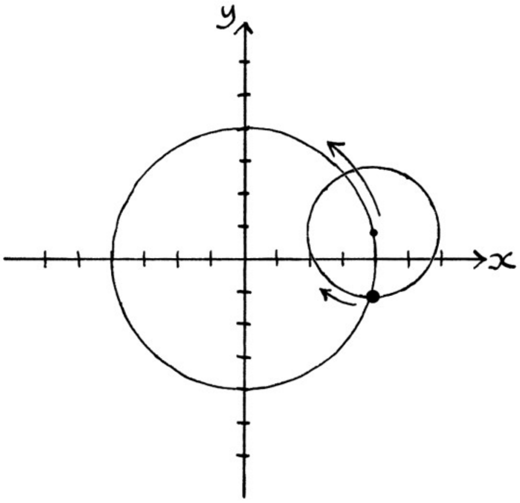
At t = 0, the circles look like this:



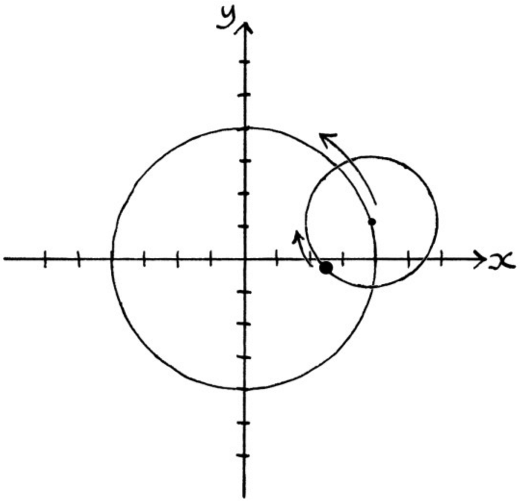
At t = 0.0625 seconds, the circles are as so:



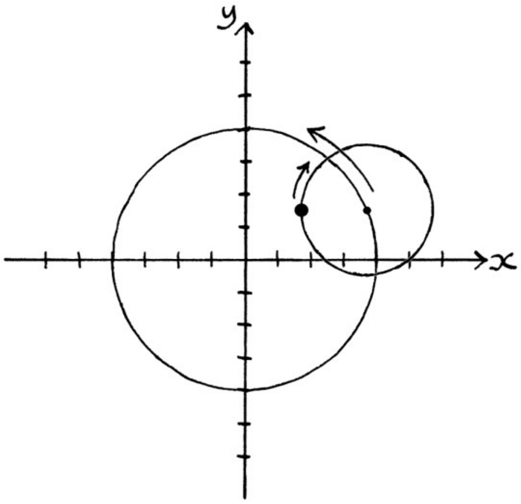
At t = 0.125 seconds, the circles are as so:



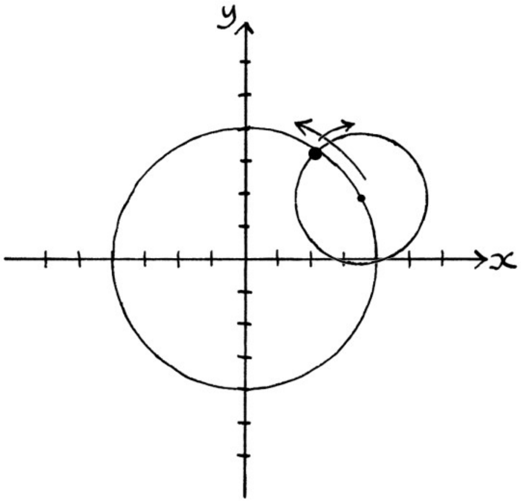
At t = 0.1875 seconds:



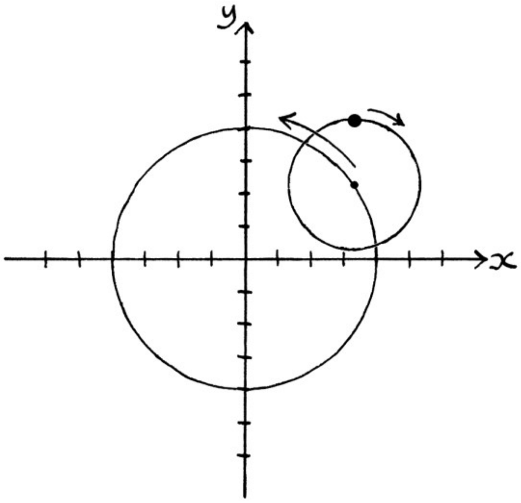
At t = 0.25 seconds, the circles appear as in the following picture. Note that the positions of the inner object and the outer object at this particular moment in time are the same as if the outer object were rotating anticlockwise. They are in the same position, but they took a different route to get there.



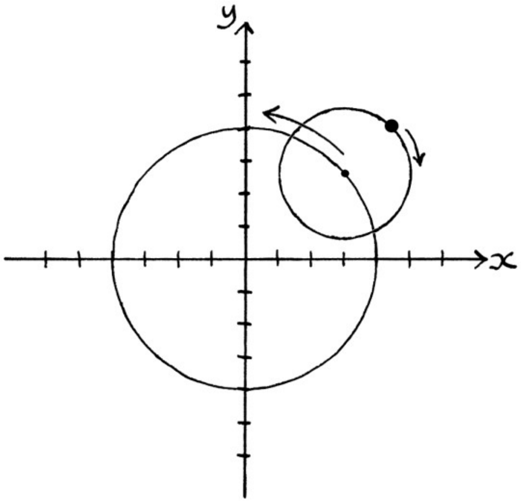
At t = 0.3125 seconds:



At t = 0.375 seconds:

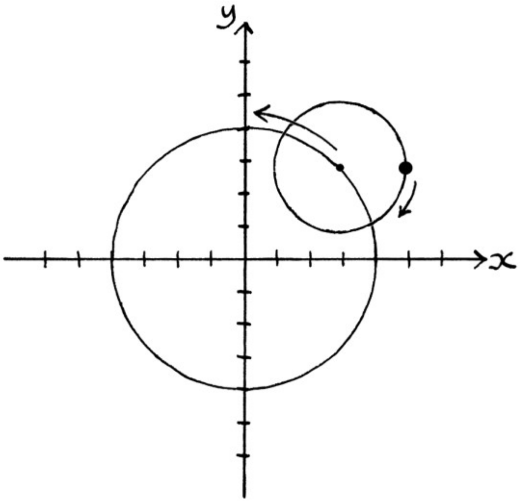


At t = 0.4375 seconds:

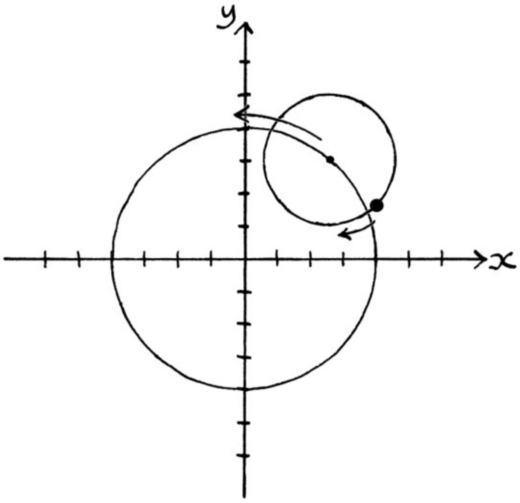


At t = 0.5 seconds, the outer object has completed one full revolution. The inner object has completed an eighth of a revolution. The objects at this particular moment in time are aligned in the same way as if both frequencies had been positive, although they took a different route to reach this position.

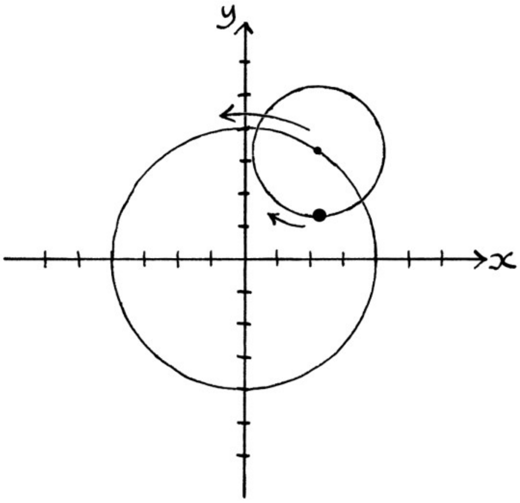
At t = 0.5 seconds:



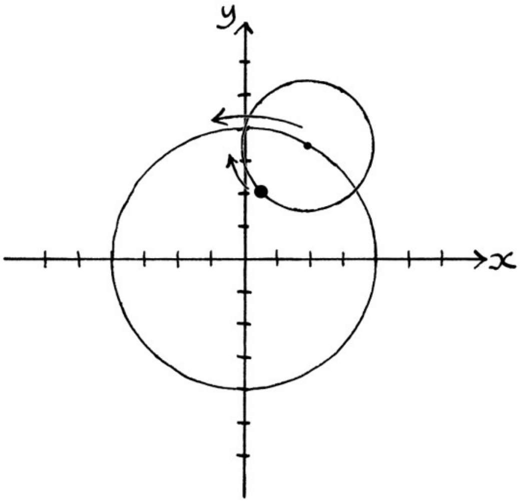
At t = 0.5625 seconds:



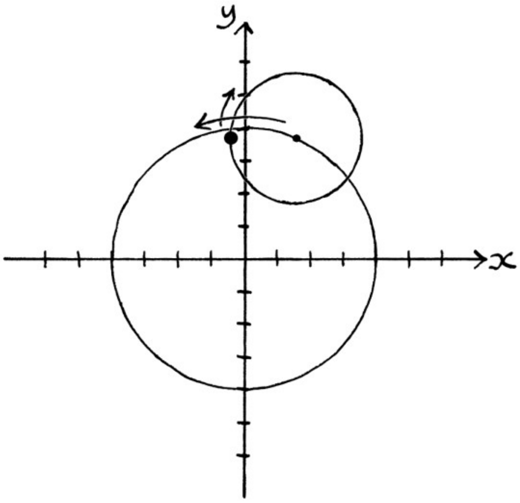
At t = 0.625 seconds:



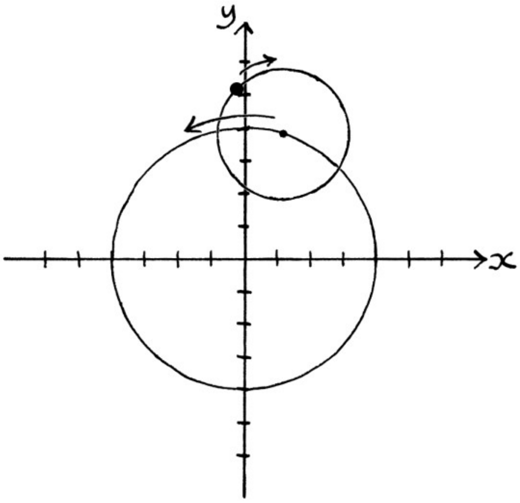
At t = 0.6875 seconds:



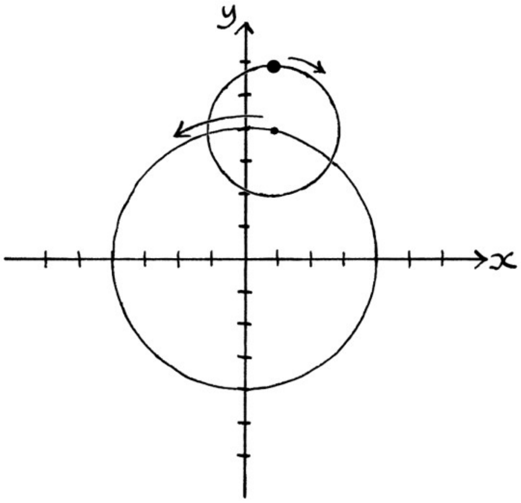
At t = 0.75 seconds, the inner and outer objects are in the same position as if both frequencies had been positive.



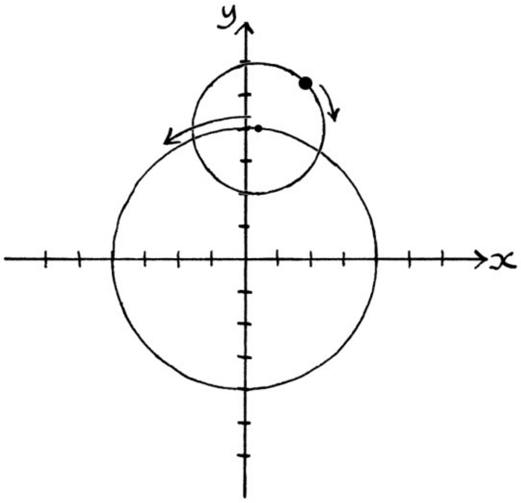
At t = 0.8125 seconds:



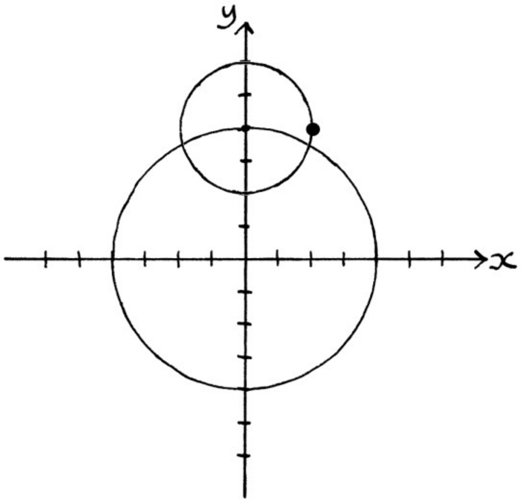
At t = 0.875 seconds:



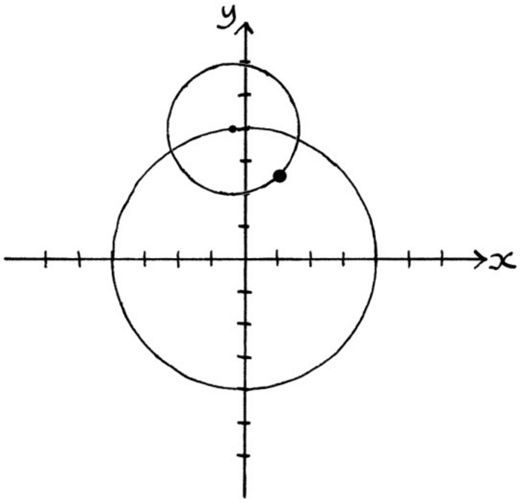
At t = 0.9375 seconds:



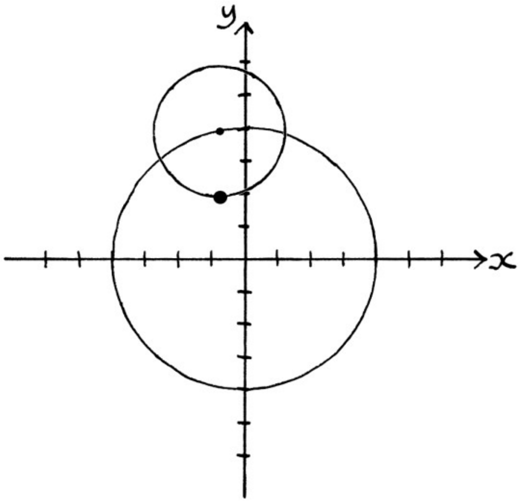
At t = 1 second, the outer object has completed 2 revolutions, and the inner object has completed a quarter of a revolution.



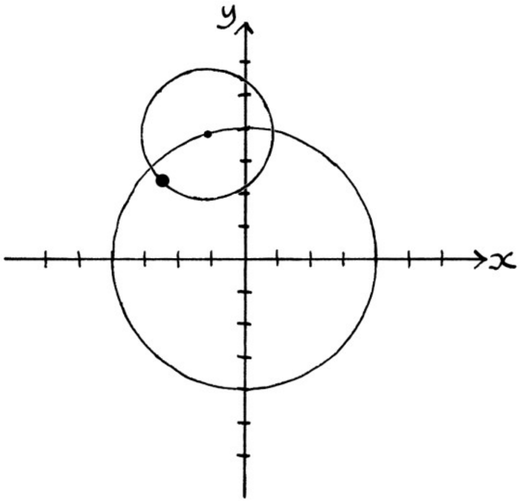
At t = 1.0625 seconds:



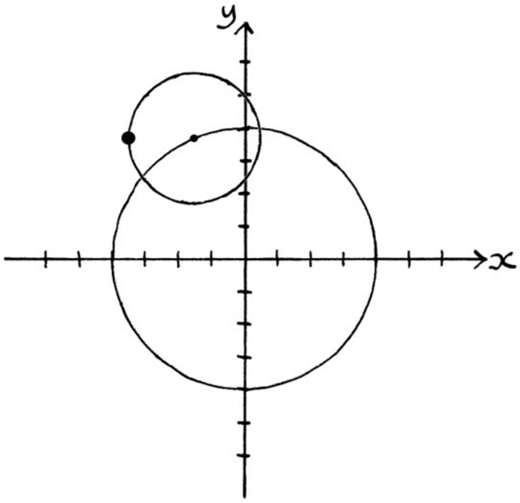
At t = 1.125 seconds:



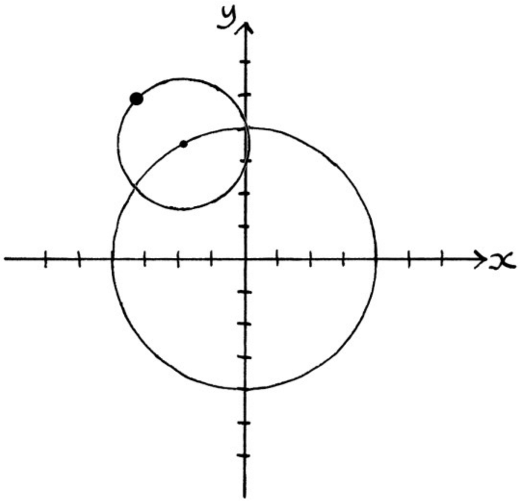
At t = 1.1875 seconds:



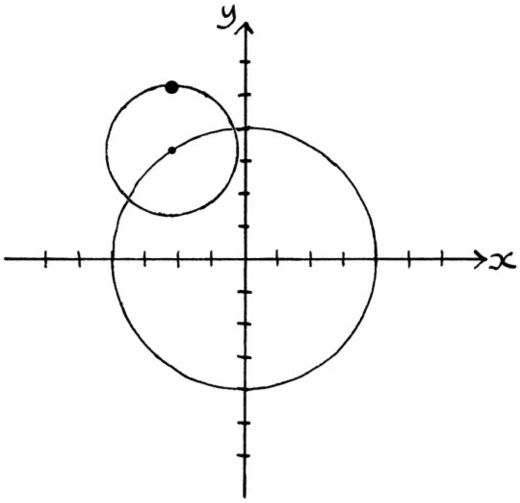
At t = 1.25 seconds:



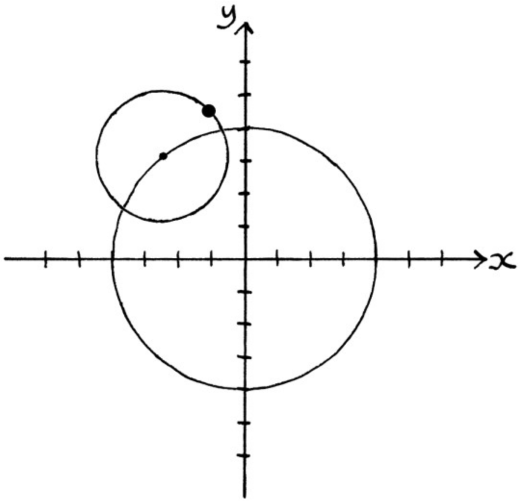
At t = 1.3125 seconds:



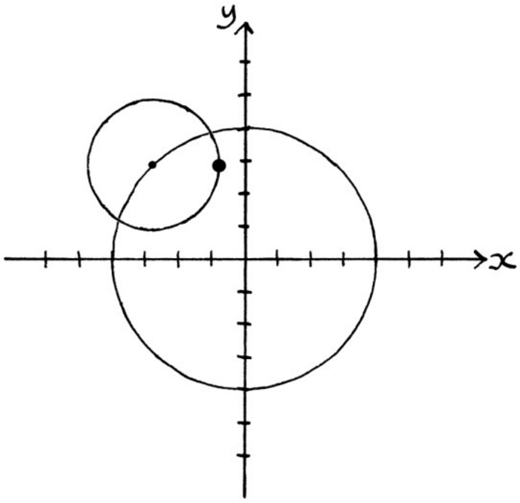
At t = 1.375 seconds:



At t = 1.4375 seconds:

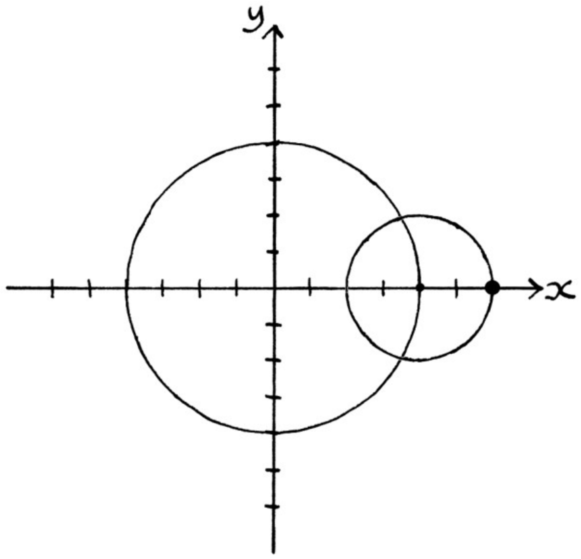


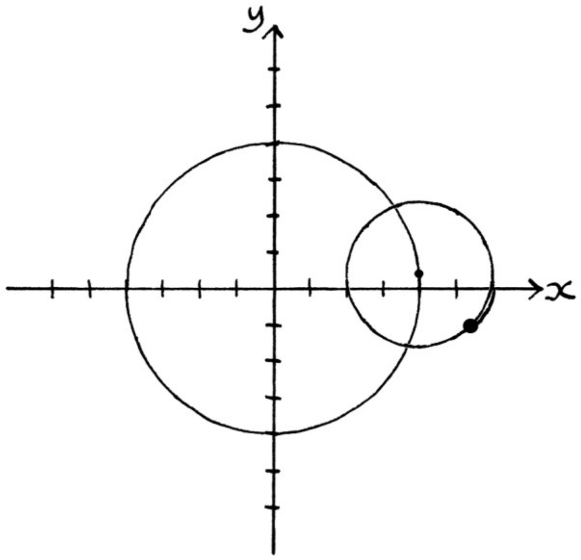
At t = 1.5 seconds:

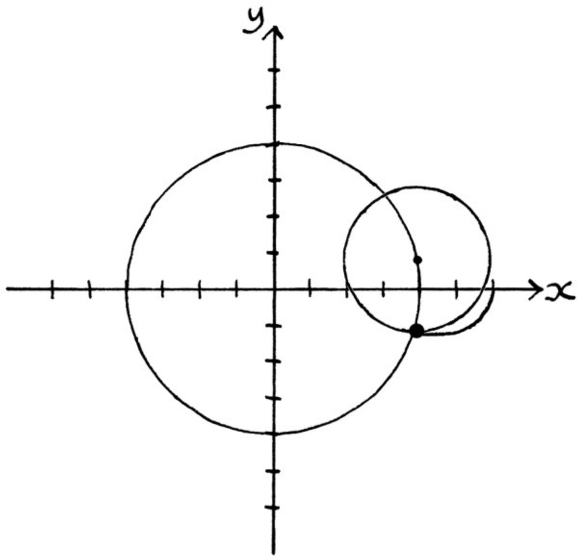


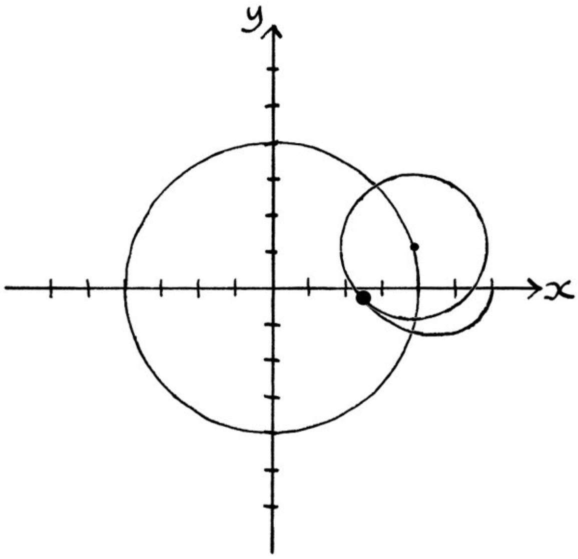
... and so on until 4 seconds have passed, at which time the inner object will have completed one full cycle, and the outer object will have completed 8 cycles.

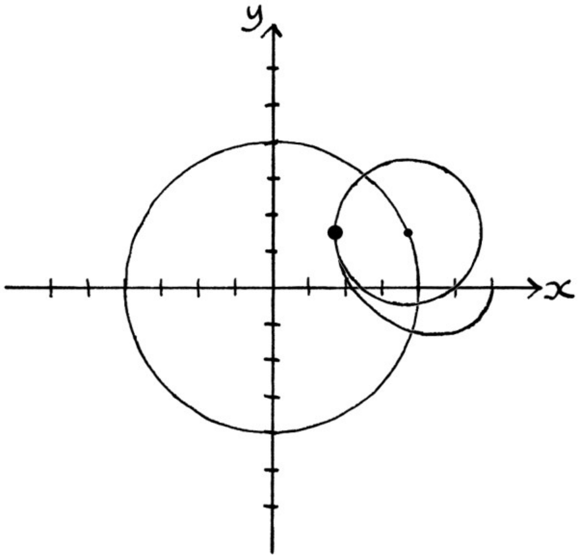
The outer object draws out a shape as it rotates. The following pictures show the circles at intervals of 0.0625 seconds.

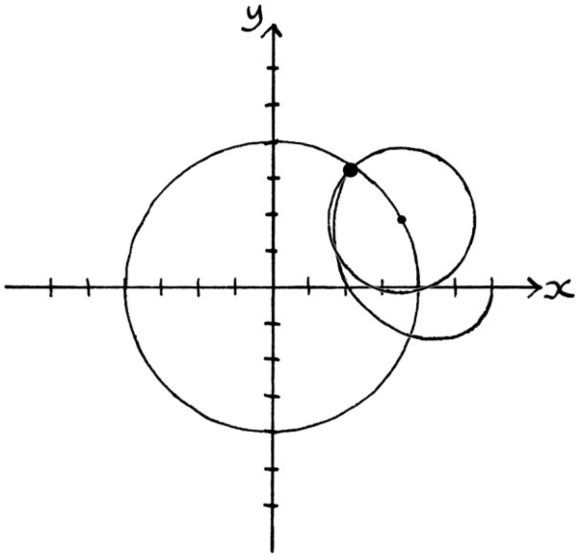

y
x

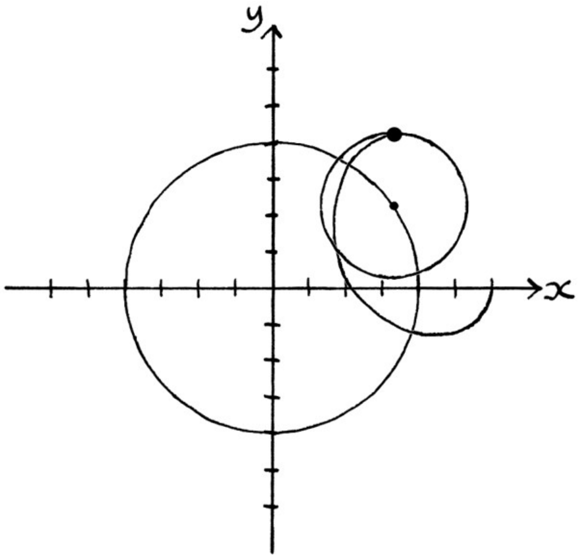

y
x

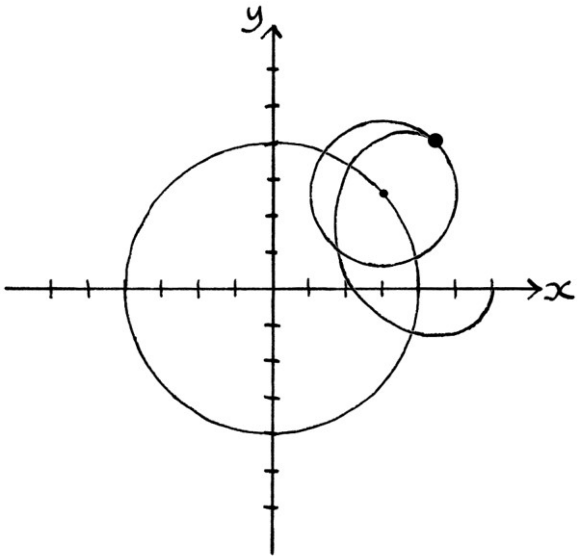

y
x

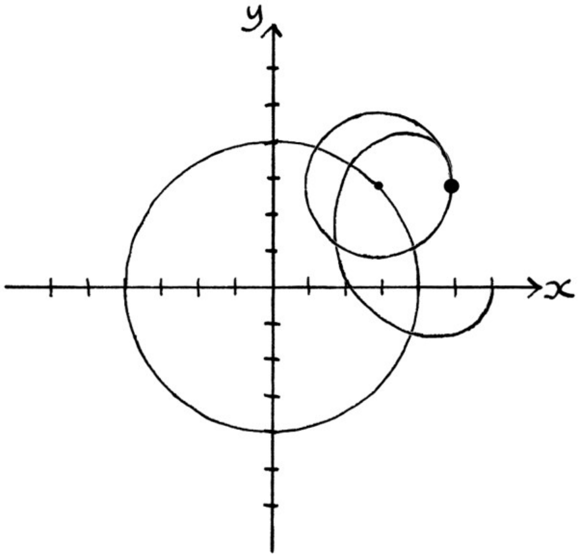

y
x

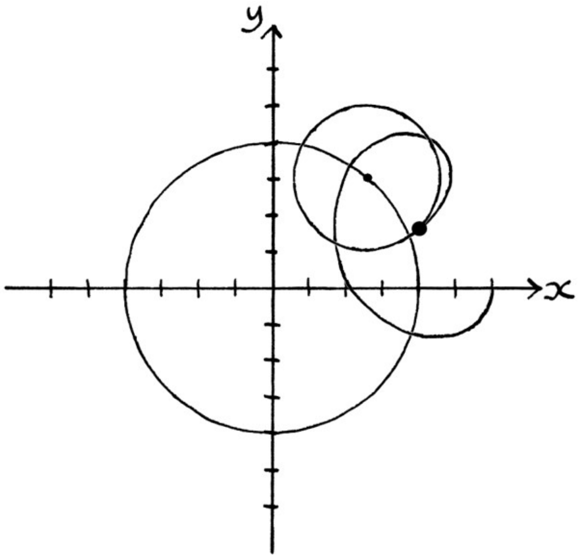

y
x

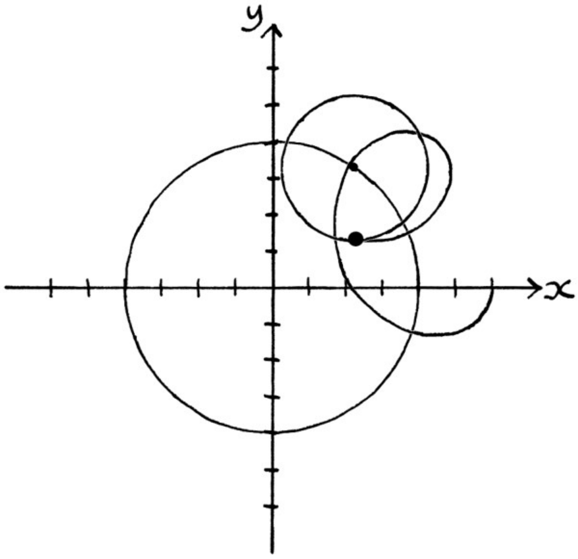

y
x

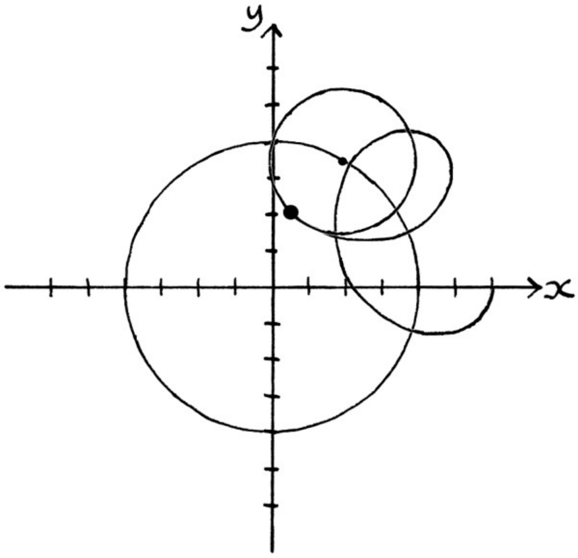

y
x

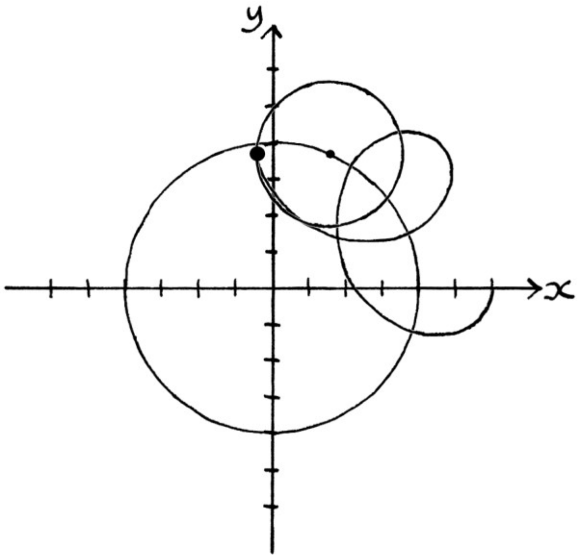

y
x

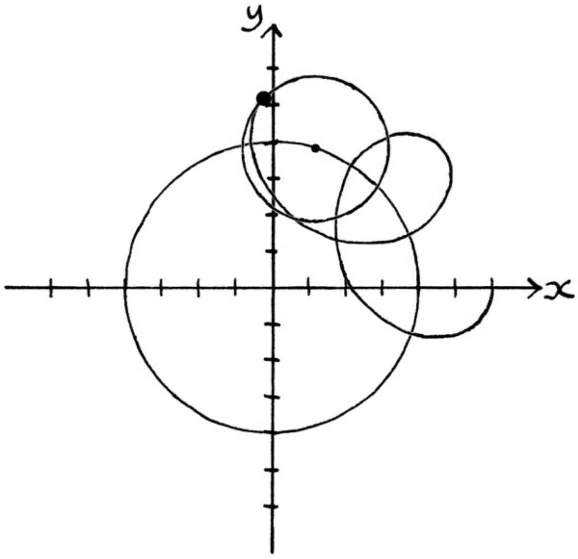

y
x

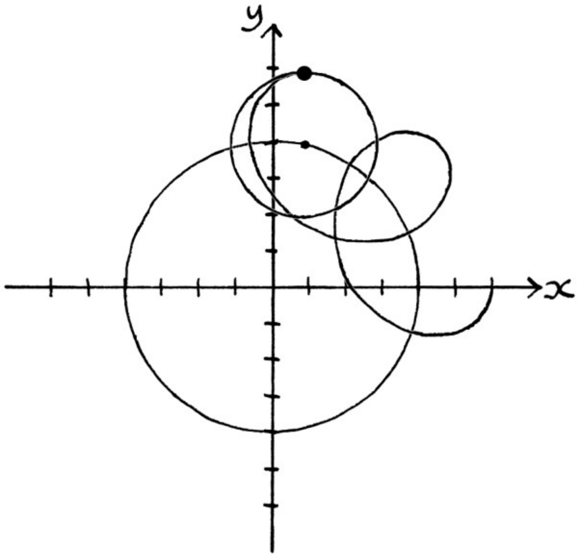

y
x

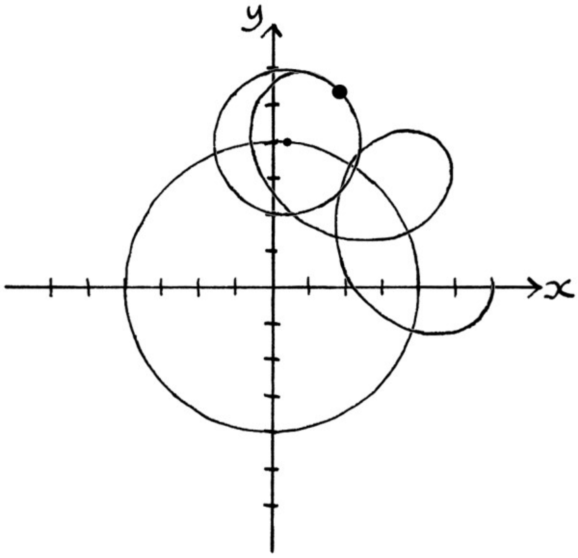

y
x

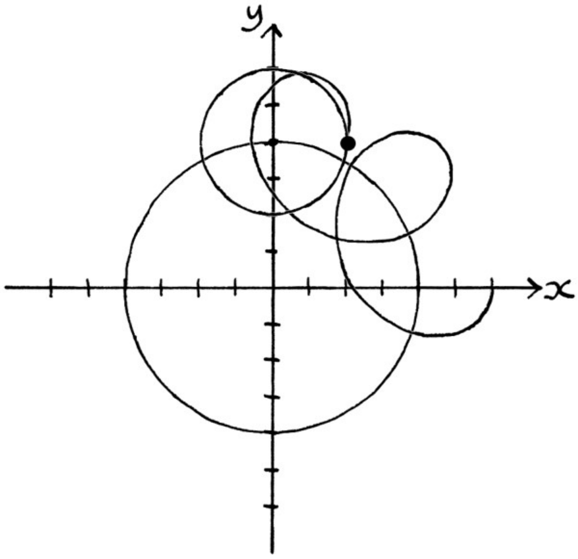

y
x

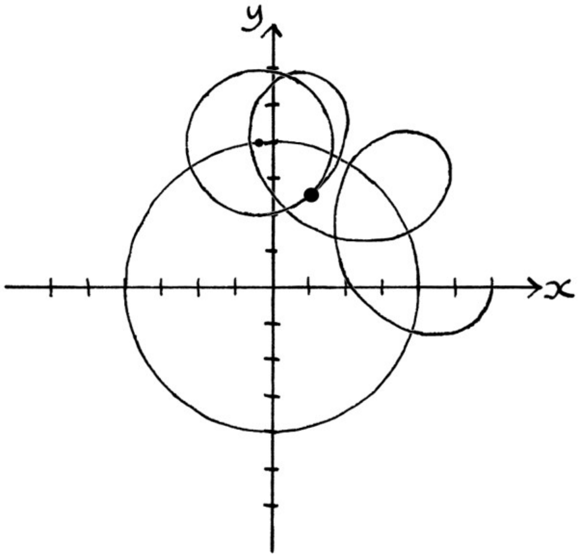

y
x

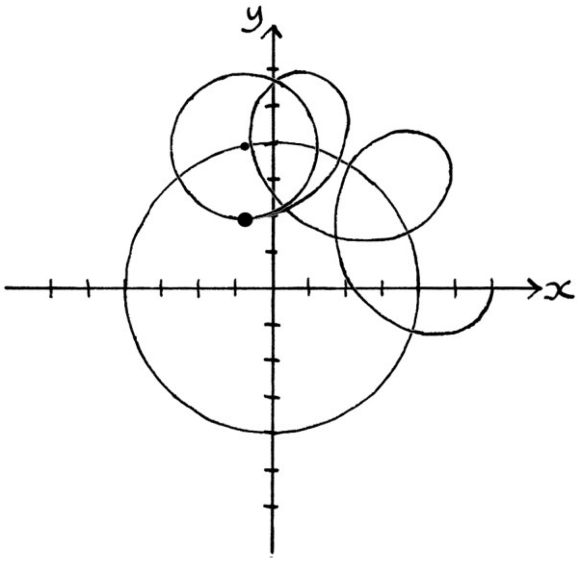

y
x

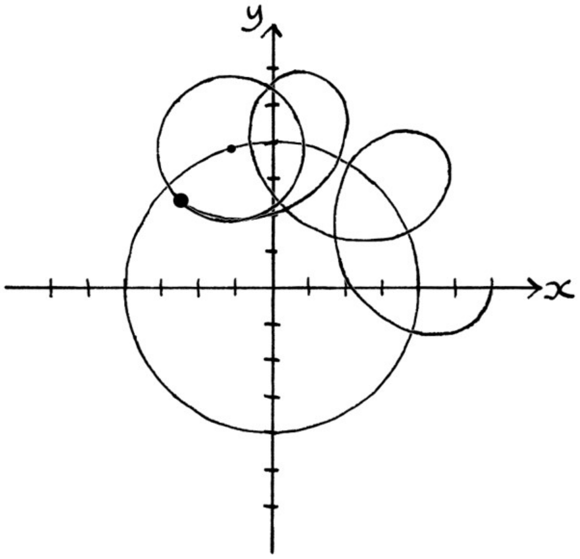

y
x


y
x

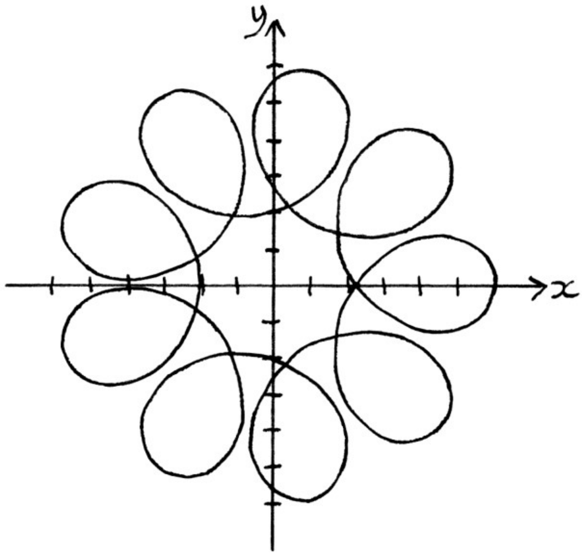

y
x


y
x

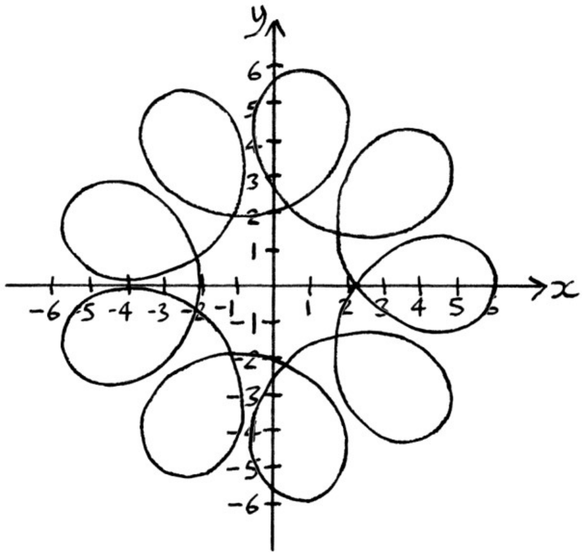




... and so on. The finished resulting shape looks like this:

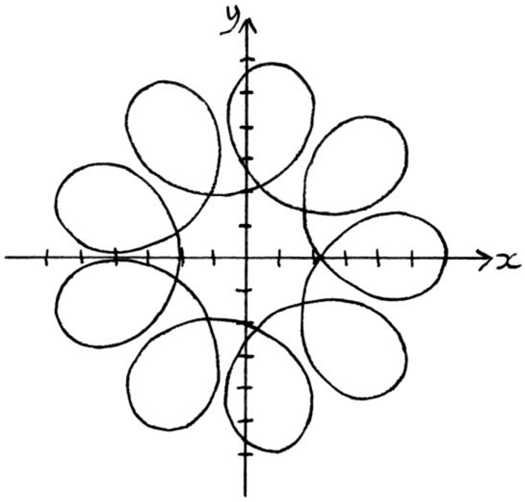


... or shown with the axes numbered:

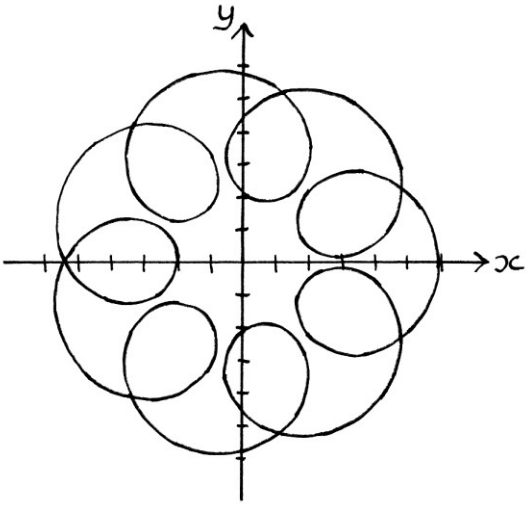


If we compare this with the resulting shape from the first frequency example in this chapter (which had the same circles, but both with positive frequencies), we can see that our new shape is an inside-out version of the old one.

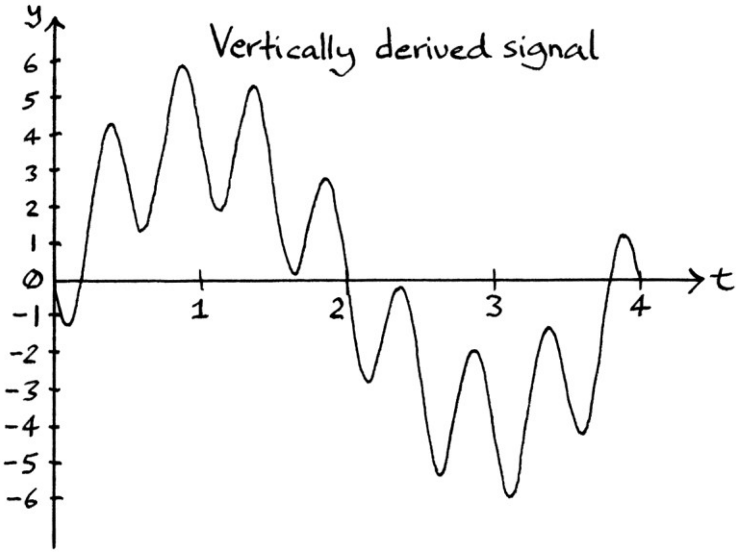
Positive and negative frequencies:

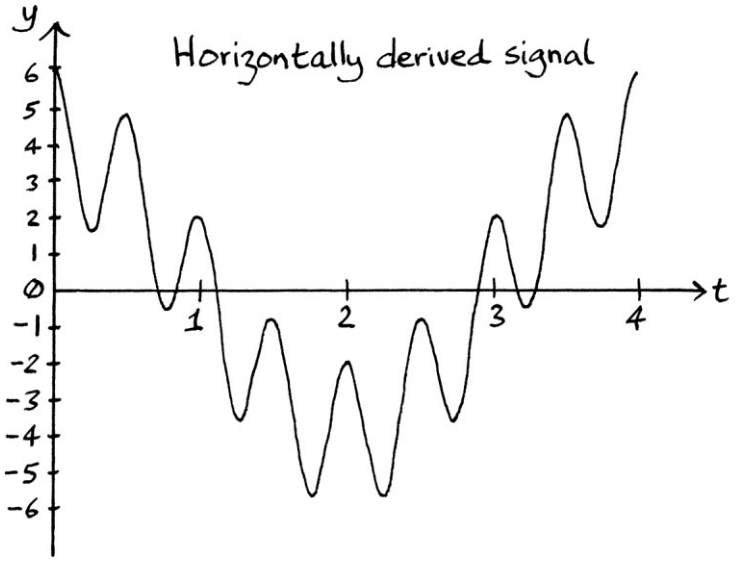


Two positive frequencies:



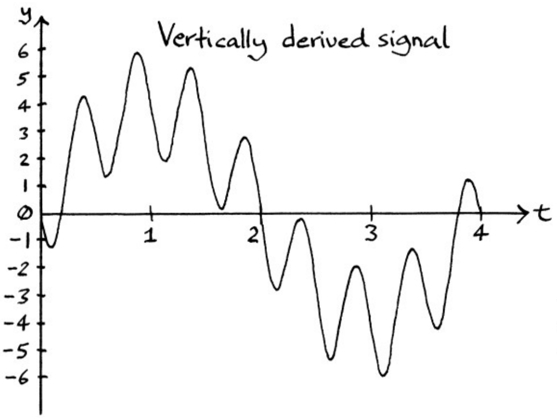
The derived signals from our resulting shape look like this:



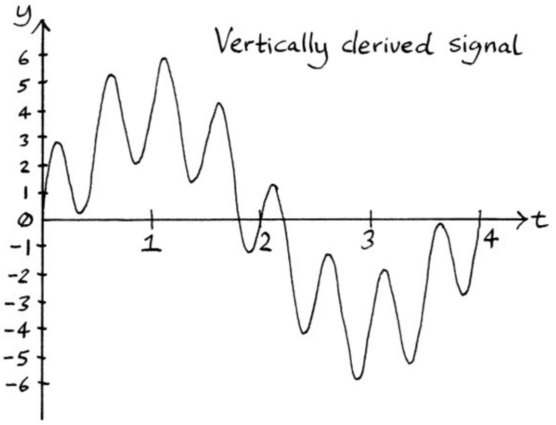


If we compare the vertically derived signals from this example and the first frequency example in this book, we can see how the one from this example is a flipped version of the other, which has also been shifted by half a cycle:

Positive and negative frequencies:

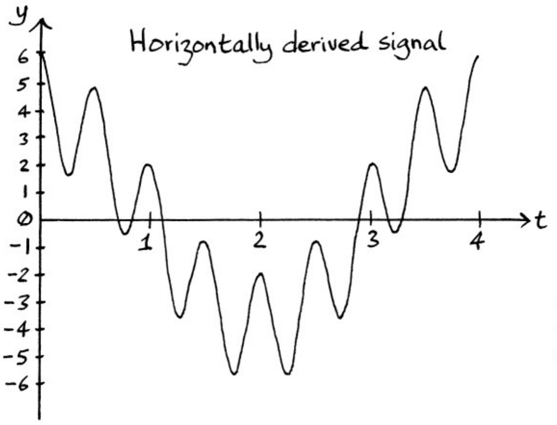


Two positive frequencies:

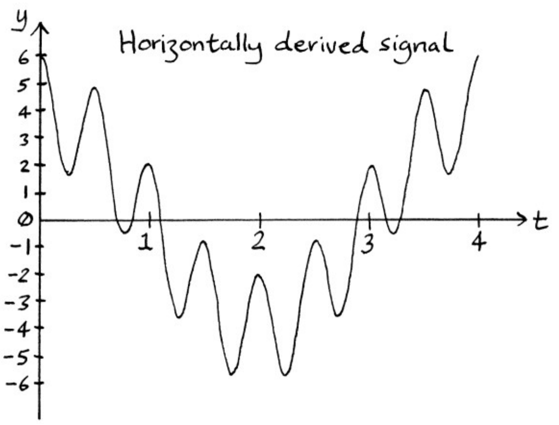


... whereas the two horizontally derived signals are identical:

Positive and negative frequencies:



Two positive frequencies:



Identical circles with opposing frequencies

One useful addition to know about is the addition of one circle with a positive frequency to another circle that is identical in every way, except it has the negative frequency of the first circle.

As an example, we will add the two circles that represent the pairs of waves:
“y = 2 sin (360 * 2t)” and “y = 2 cos (360 * 2t)”
... and:
“y = 2 sin (360 * −2t)” and “y = 2 cos (360 * −2t)”

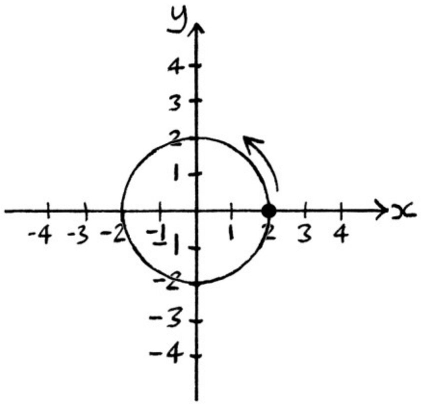
Before we look at the circles, we will think about the addition of the underlying waves. The derived signals of the result of the addition will be:
“y = 2 sin (360 * 2t) + 2 sin (360 * −2t)”
“y = 2 cos (360 * 2t) + 2 cos (360 * −2t)”

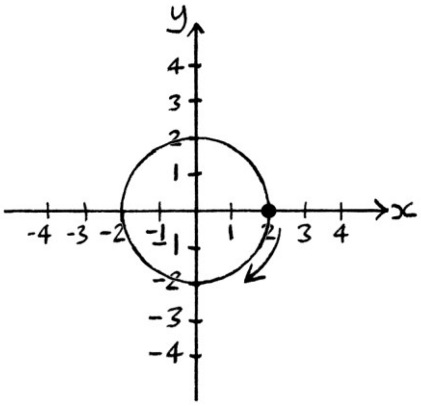
The wave “y = 2 sin (360 * −2t)” is the same as “y = 2 sin ((360 * 2t) + 180)”, which is the same as “y = −2 sin (360 * 2t)”. Therefore, the resulting vertically derived will be:
y = 2 sin (360 * 2t) – 2 sin (360 * 2t)
... which results in “y = 0”. This means that the resulting vertically derived signal will be a horizontal line at y = 0 for all time.

The wave “y = 2 cos (360 * −2t)” is the same as “y = 2 cos (360 * 2t)”, because a negative-frequency Cosine wave *with zero phase* is the same as a positive-frequency Cosine wave *with zero phase*. Therefore, the second resulting signal can be rewritten as:
y = 2 cos (360 * 2t) + 2 cos (360 * 2t)
... which results in “y = 4 cos (360 * 2t).” This means that the resulting horizontally derived signal will be the pure wave “y = 4 cos (360 * 2t)”.

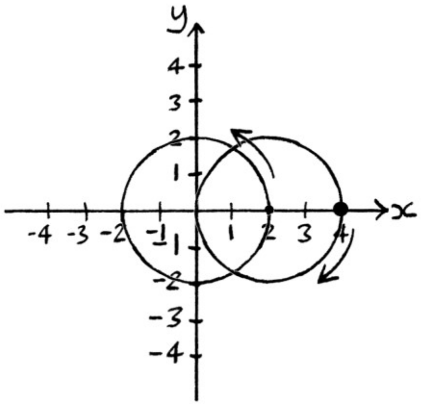
All this means that the resulting shape will have y-axis values that are all zero, and x-axis values that will conform to “y = 4 cos (360 * 2t)”. The shape will be flat but fluctuate up and down the x-axis.

This will be more understandable if we look at the circles. The two circles look like this:

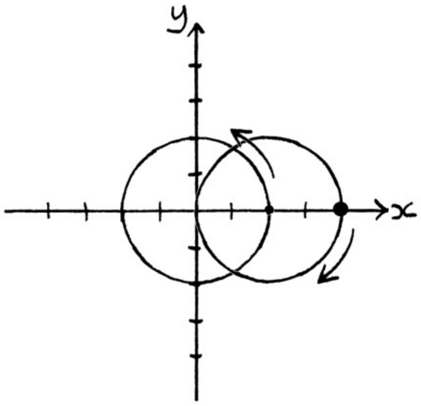




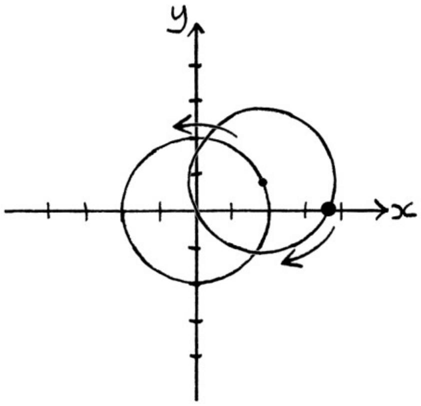
We will arrange them with the positive-frequency circle on the inside, and the negative-frequency circle on the outside, although as always, their order does not matter.



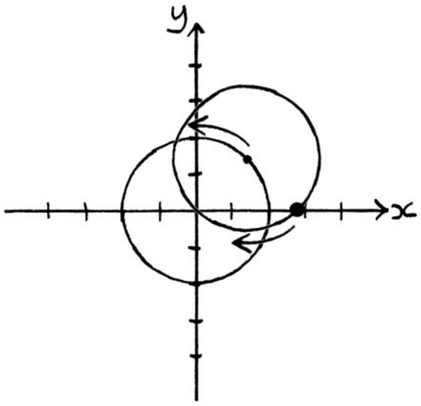
At t = 0, they look like this:

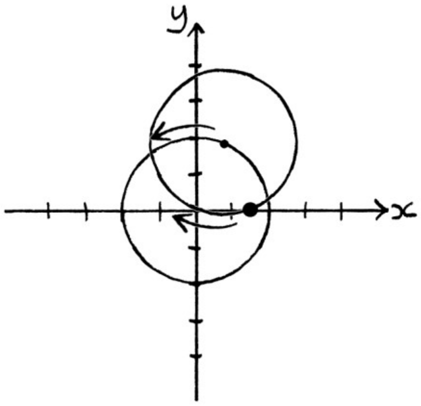


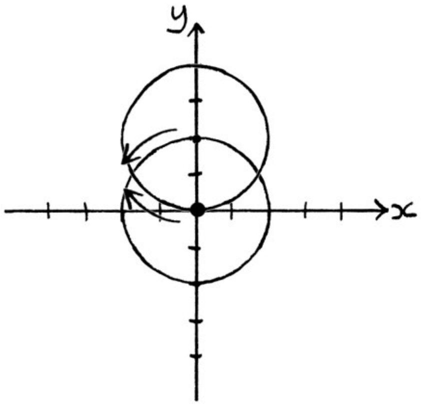
At t = 0.03125 seconds:

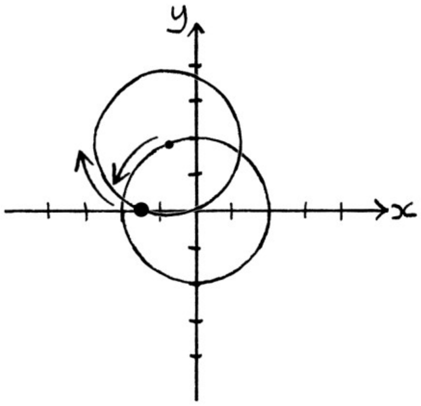


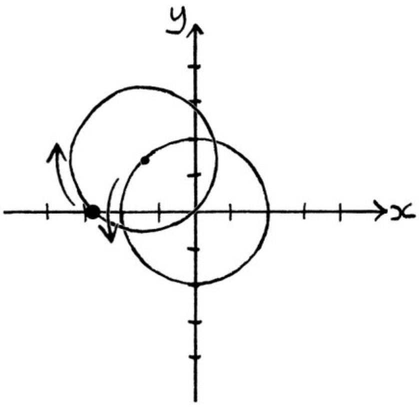
Then up until t = 0.5, the circles appear as in the following pictures (at intervals of 0.03125 seconds):

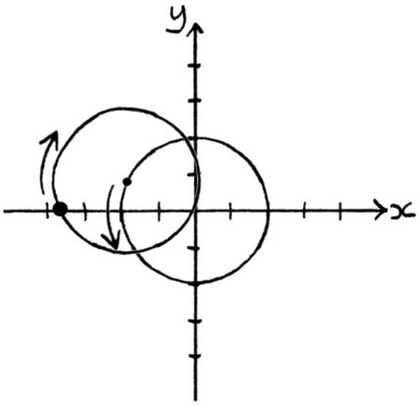


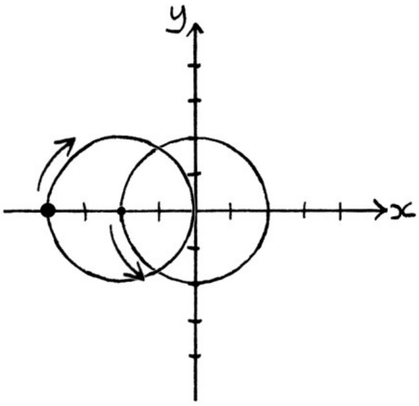


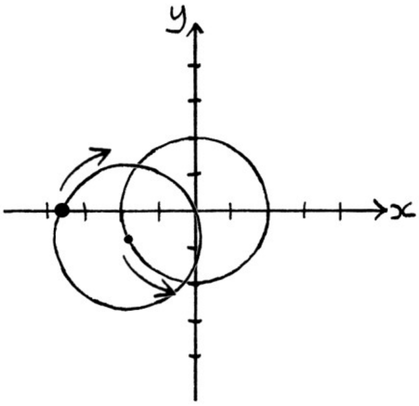


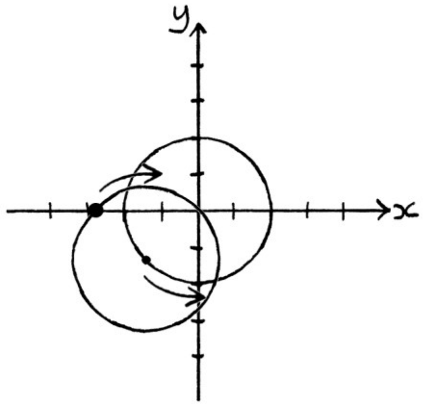


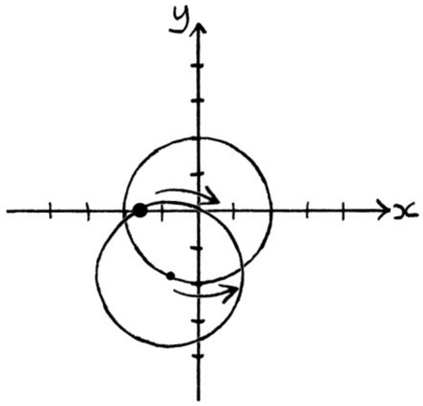

y
x

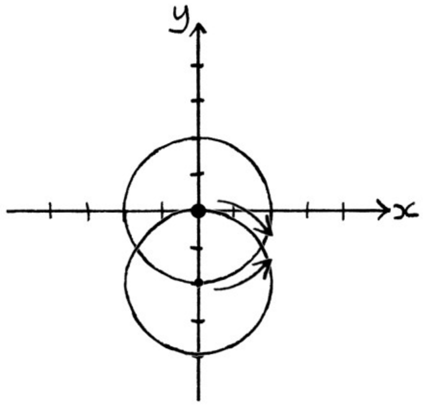

y
x

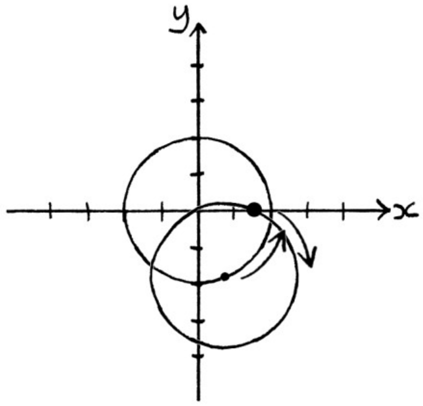

y
x

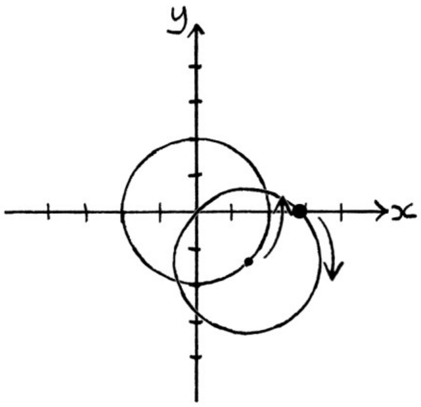

y
x

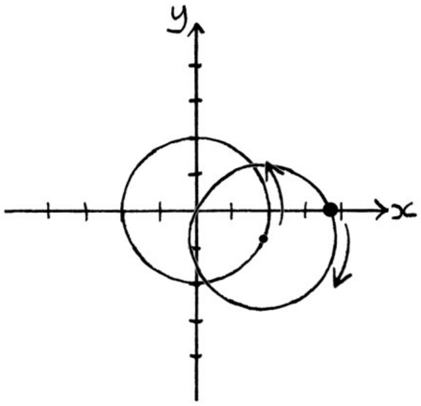

y
x

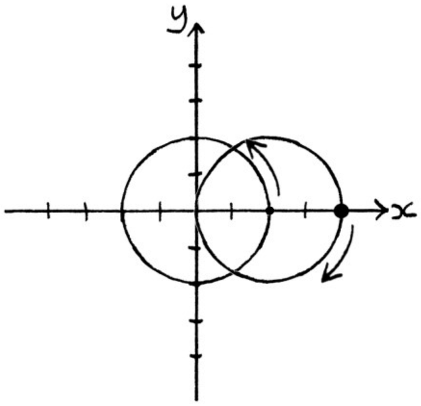

y
x


y
x

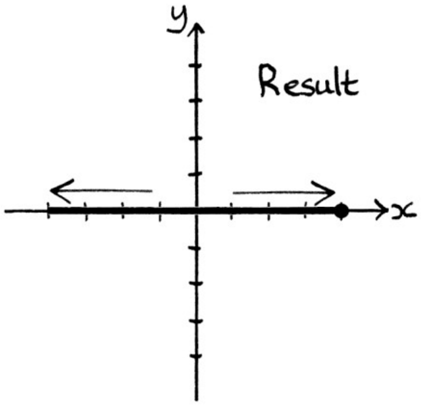

y
x



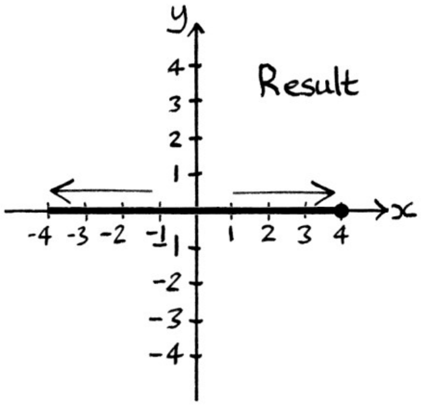




As we can see, the outer object moves up and down the x-axis, but stays fixed on y = 0. The resulting “shape” in this case is just a straight line:



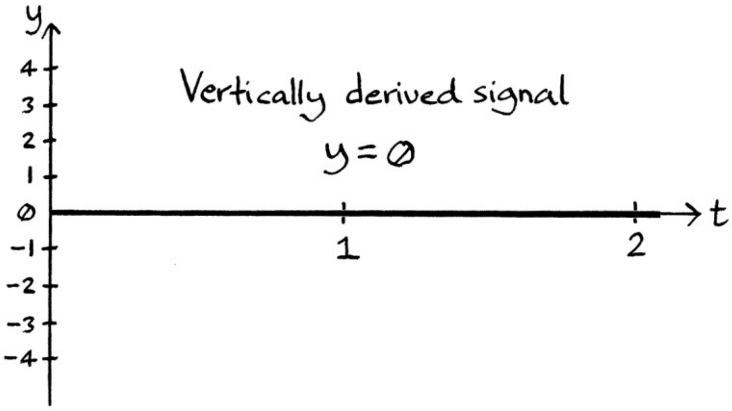
Or, shown with the axes numbered:

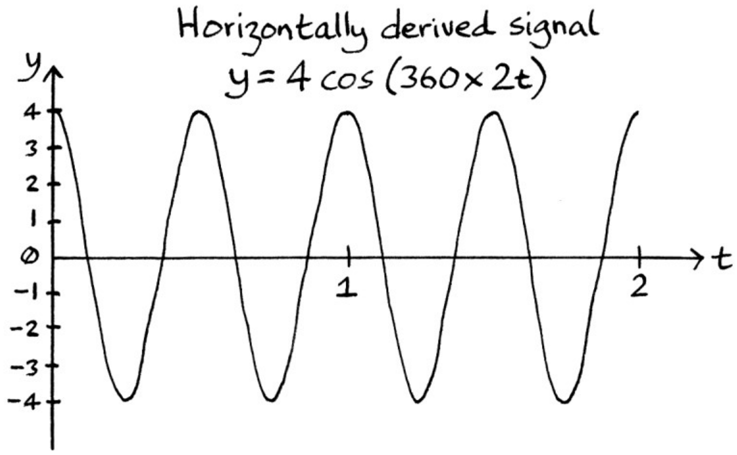


The outer object moves from one end of the line to the other end of the line and back again. It starts at x = +4, moves to x = −4, which it reaches at t = 0.25 seconds, and then it moves back to x = +4, which it reaches at t = 0.5 seconds. If left to continue, it would just move backwards and forwards along the line forever, never moving from the y = 0 line.

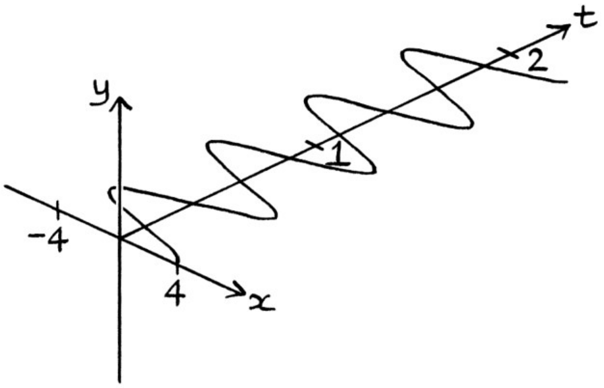
The outer object does not move at an even rate from x = +4 to x = −4 and back. Instead it moves according to the formula of its underlying horizontally derived signal, "y = 4 cos (360 * 2t)". On the circle chart, its horizontal position can be given by "x = 4 cos (360 * 2t)". Its y-axis position can be given by the formula "y = 0t". Its coordinates at any one moment in time on the circle chart are (4 cos (360 * 2t), 0).

The two derived signals are as so:





Everything becomes much clearer when we look at the shape on the helix chart:



If we looked at the helix chart from underneath, we would see a Cosine wave.

From all of this, we can see that we have really created a pure wave in the helix chart. If we think of helices instead of circles, we can say that the addition of two identical helices, with identical but opposing frequencies, creates a pure wave *within the helix chart.* The pure wave is not in a normal two-dimensional "y-axis and "time axis" graph. Instead, it is in the three-dimensional helix chart. The addition of these two helices has created a two-dimensional shape in a three-dimensional chart. This is an idea that will become important when we look at Imaginary powers of "e" in Chapters 27 and 28.

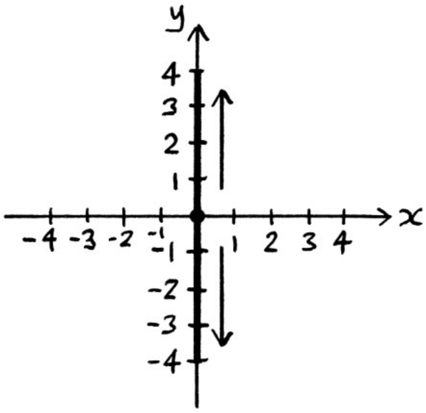
This is one of many situations where it can be useful to think of helices instead of circles. The movement of an object rotating around a circle can be thought of as a circle or it can be thought of as a helix. Which one is best depends on the situation.

As an example of what happens when we *subtract* one circle from a second circle, that is identical except for having an opposing frequency, we will start with this circle:
“y = 2 sin (360 * 2t)” and “y = 2 cos (360 * 2t)”
... and subtract from it this circle:
“y = 2 sin (360 * −2t)” and “y = 2 cos (360 * −2t)”

We can work out the characteristics of the derived signals. The resulting vertically derived signal will be:
“y = 2 sin (360 * 2t) − 2 sin (360 * −2t)”
... which, because a negative-frequency Sine wave with zero phase is the same as that wave with a positive frequency and a phase of 180 degrees, is:
“y = 2 sin (360 * 2t) − 2 sin ((360 * 2t) + 180)”
... which, because a negative-amplitude Sine wave with a phase of 180 degrees is the same as that wave with a positive amplitude and a phase of zero degrees, is:
“y = 2 sin (360 * 2t) + 2 sin (360 * 2t)”
... which is:
“y = 4 sin (360 * 2t)”

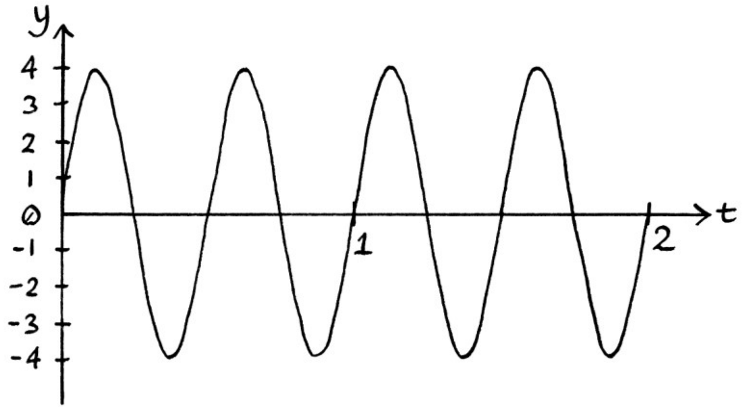
The resulting horizontally derived signal will end up as:
“y = 2 cos (360 * 2t) − 2 cos (360 * −2t)”
... which is:
“y = 2 cos (360 * 2t) − 2 cos (360 * 2t)”
... which is:
“y = 0”
... for all time, or “y = 0t”, if we wanted to phrase it in that way.

The result, as viewed in the circle chart, looks like this:

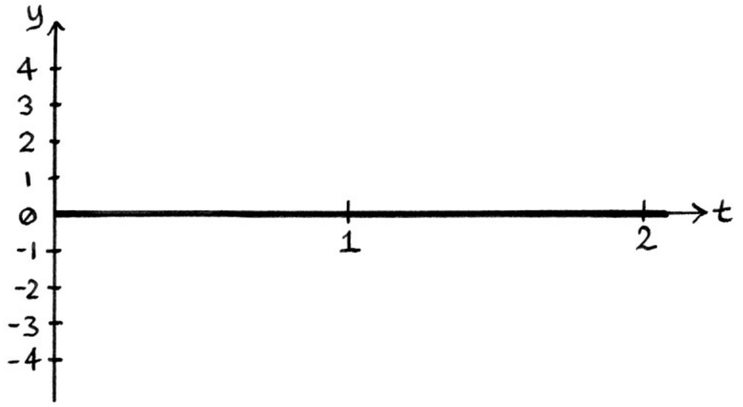


It is just a vertical line. The object moves up and down this line.

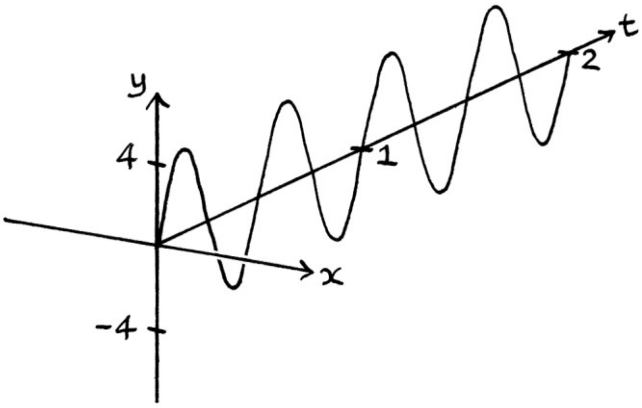
The vertically derived signal is a Sine wave:



The horizontally derived signal is just a line at y = 0:



The situation is made clearer when we view the result on the helix chart:

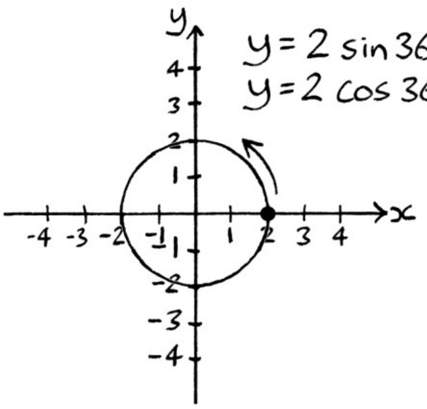


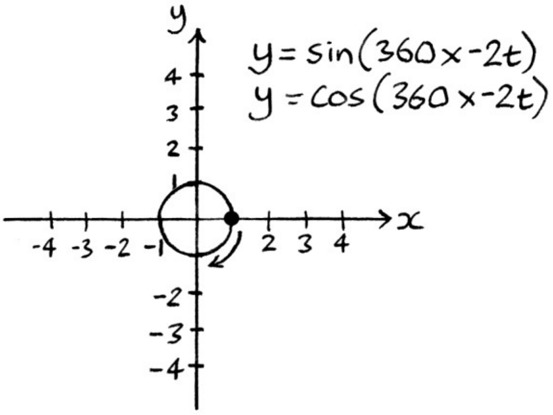
The result is a two-dimensional Sine wave sitting in the three-dimensional helix chart.

Interesting shapes

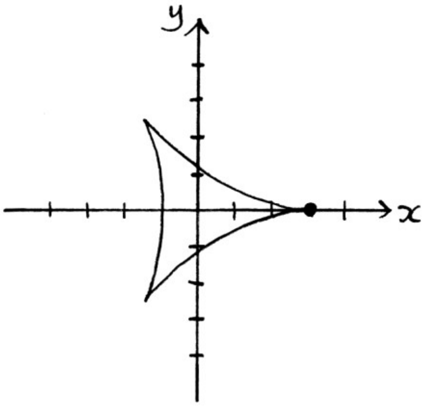
It is possible to make much more interesting shapes by adding circles with opposing frequencies, than if the frequencies are the same. For example, we will add the circles that represent the pairs of waves:
"y = 2 sin 360t" and "y = 2 cos 360t"
... and:
"y = sin (360 * −2t)" and "y = cos (360 * −2t)"

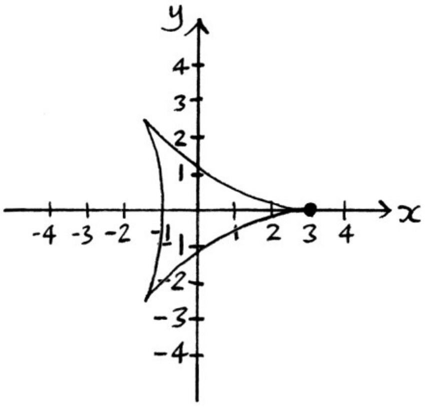
The circles look like this:



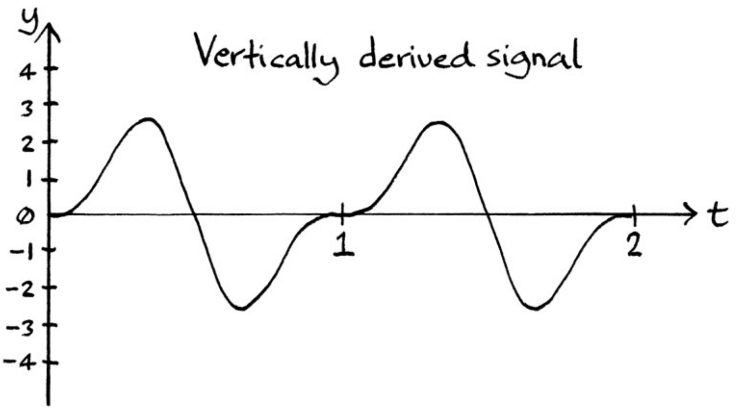


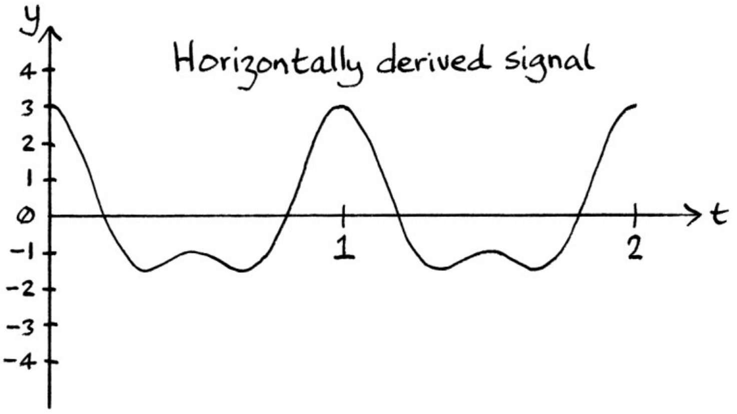
The resulting shape looks like this, shown with the axes unnumbered and numbered:



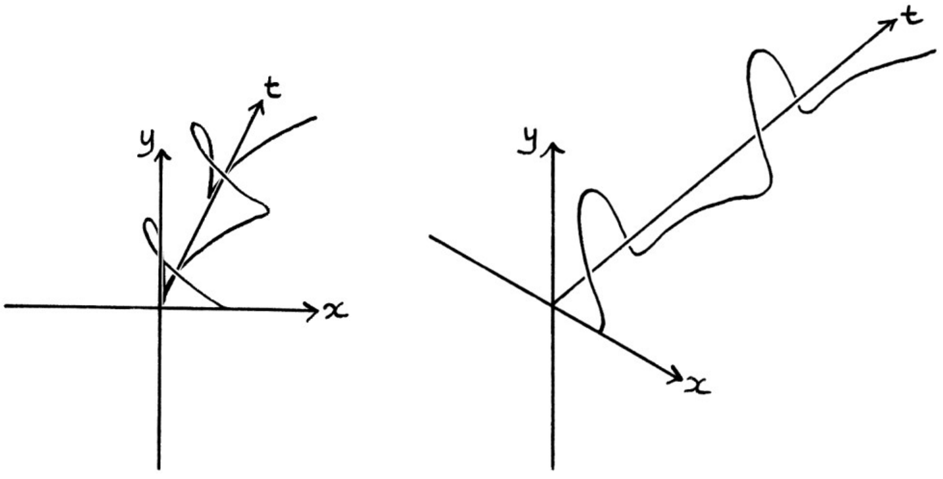


The derived signals look like this, and give no obvious clue as to how they are derived from what is nearly a triangle:





The "helix" looks like this, as viewed from two different angles:



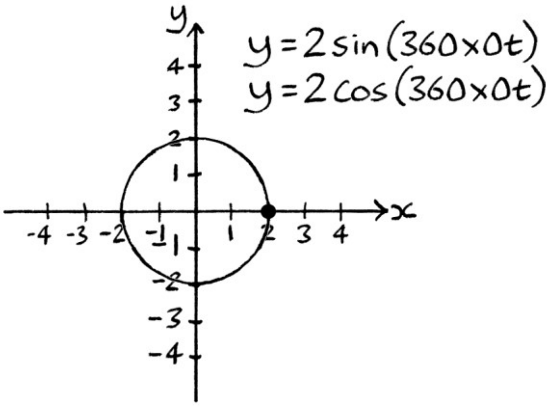
Drawing a triangle in this way gives a clue as to the range of shapes that it is possible to create by adding circles of different frequencies.

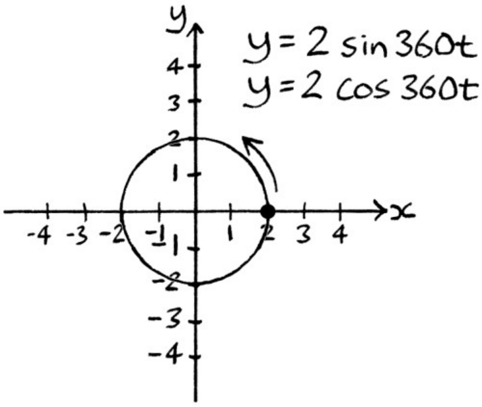
Zero frequencies

If we add two circles where one circle has a frequency of zero cycles per second, then the stationary object "moving" around that circle will act as one or two mean levels for the other circle.

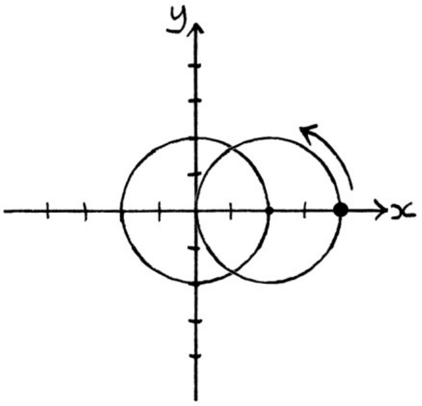
For example, we will add the circles that represent the pairs of waves:
"y = 2 sin (360 * 0t)" and "y = 2 cos (360 * 0t)"
... and:
"y = 2 sin 360t" and "y = 2 cos 360t"

The circles look like this:



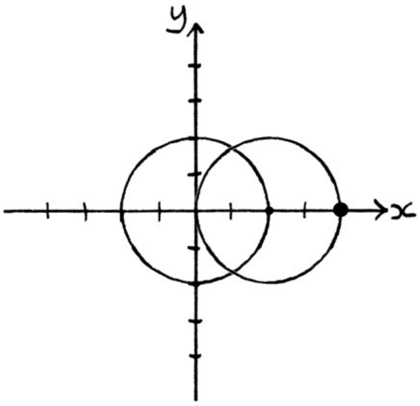


The circles are arranged as so, with the zero-frequency circle over the origin of the axes:

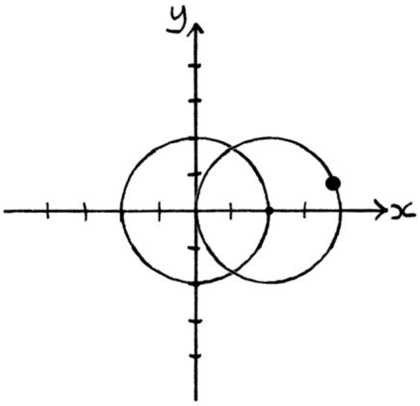


The outer object rotates around the outer circle as normal. The outer circle is centred on the inner object, which, because the inner circle has zero frequency, remains fixed in the same position all the time. Therefore, this has the same effect as the outer object rotating around the point at (2, 0).

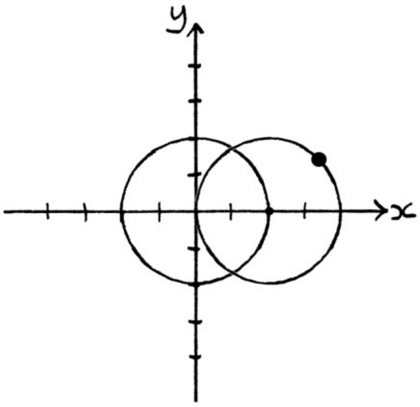
At t = 0 seconds, the circles look like this:



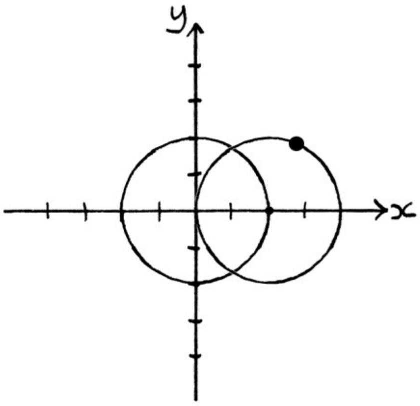
At t = 0.125 seconds:



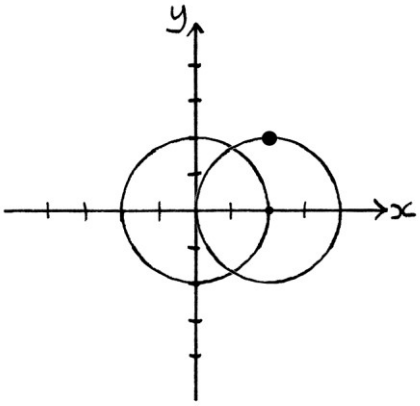
At t = 0.25 seconds:



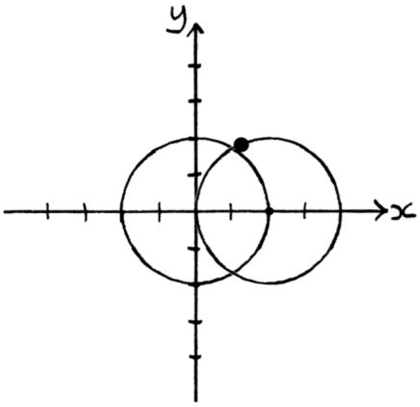
At t = 0.375 seconds:



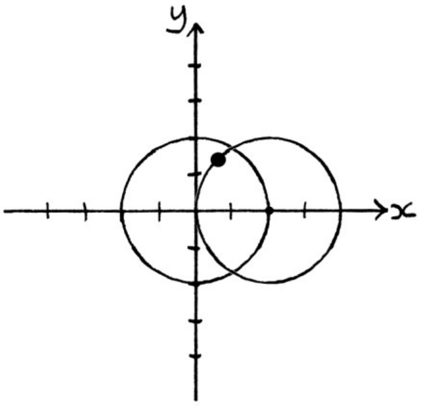
At t = 0.5 seconds:



At t = 0.625 seconds:



At t = 0.75 seconds:



... and so on.

The resulting shape is a perfect circle centred at (2, 0).

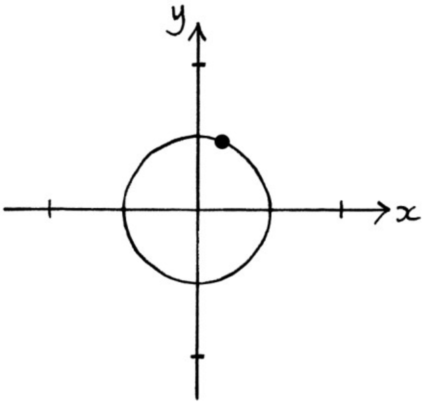
The zero-frequency circle is acting as a mean level, or to be more accurate, the phase point of the zero-frequency circle is acting as a mean level. If we had a phase of 90 degrees, the resulting circle would be centred at (0, 2). If we had a phase of 45 degrees, the resulting circle would be centred at (0.7071, 0.7071).

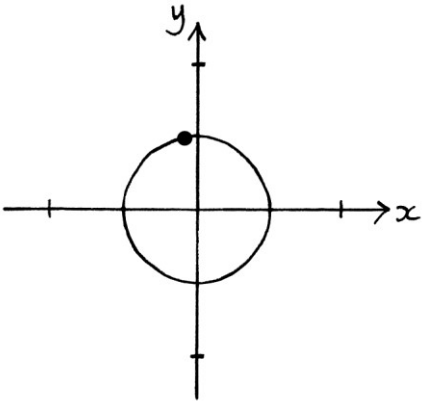
Phase and vectors

Vectors

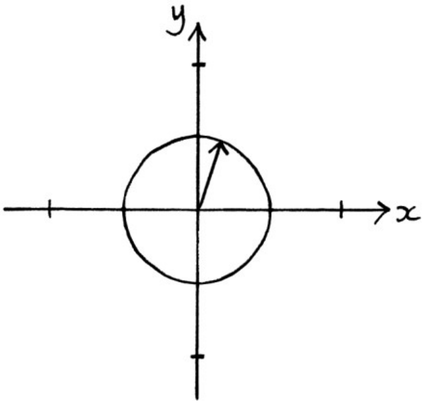
When adding circles with the same frequency and zero mean level, either we can think of the circles, or we can use a second method that amounts to exactly the same thing. This method involves drawing lines from the centre of each circle to its phase point, and then, while taking note of their angles and lengths, redrawing those lines joined together. The end of the joined lines shows the position of the phase point of the outer circle at t = 0, which is also the phase point of the resulting circle. From that, the amplitude and phase of the circle and its underlying waves can be calculated. The basis of this is identical to arranging circles at t = 0, but it is another way to think about it. As the frequency for all the circles is the same, the joined lines will maintain their shape as time progresses. The line will rotate around the origin of the axes with the end of the line marking where the outer object is.

As an example, we will add the following circles together: [They have the same frequency.]

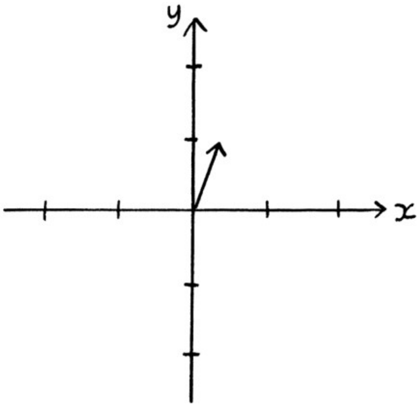




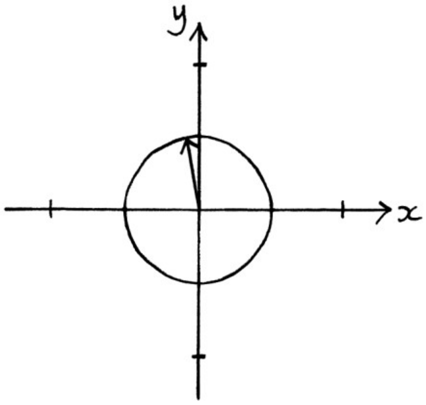
For the first circle, the line drawn from its centre to its phase point looks like this:



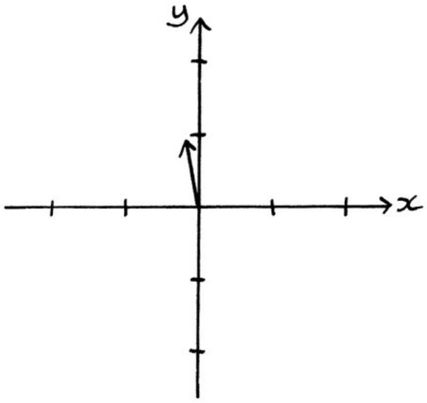
Without the circle, it looks like this:



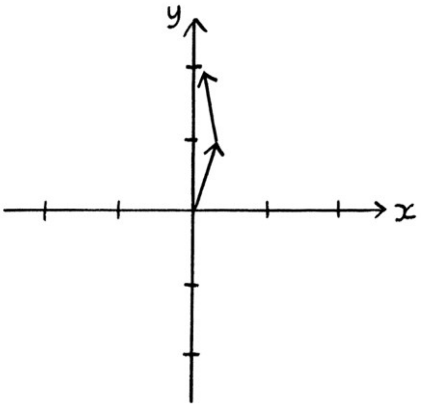
For the second circle, the line drawn from its centre to its phase point looks like this:



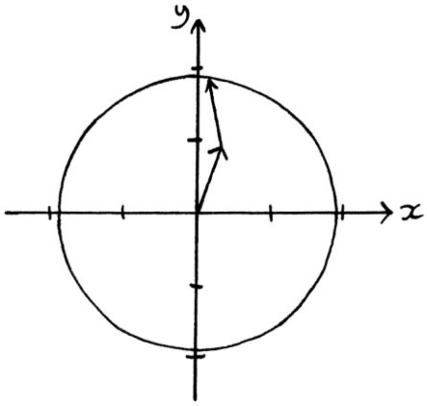
Without the circle, the line looks like this:

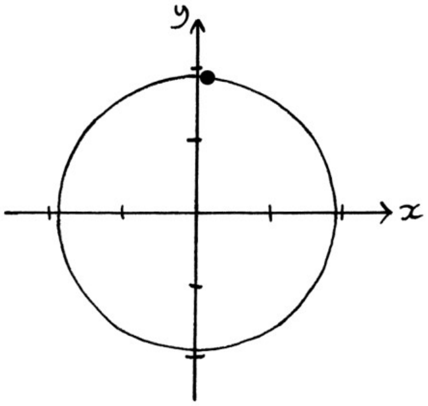


The lines joined together look like this: [They can be placed either away around, and they will still end up at the same place.]

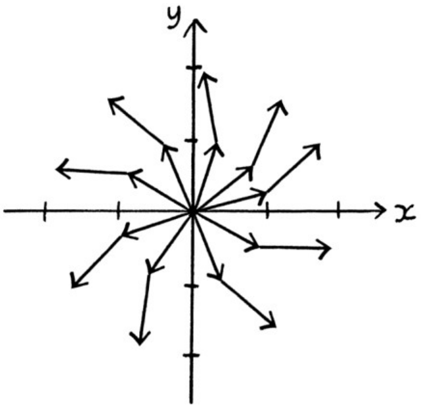


The point at the end of these lines is also the phase point of the resulting circle. That one point is enough to recreate the circle because it indicates the phase and amplitude, and we know that the resulting frequency is the same as that of each of the added circles. [There is no mean level].

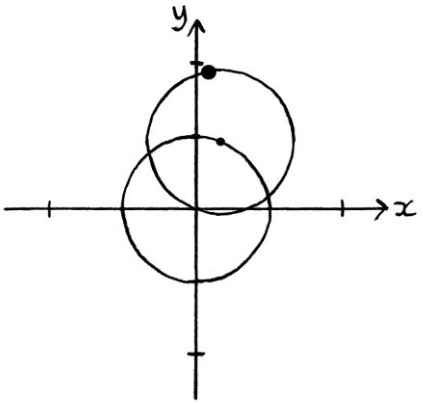


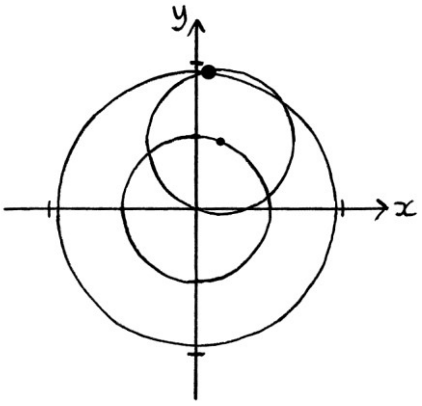


We could also watch the joined-up line rotate around the origin of the axes to draw out the resulting circle too. As the original circles both had the same frequency, the line keeps its shape for all time. It is essentially the same as the "arm" mentioned earlier in this chapter.



The points at the end of the lines are also the positions of the outer object when the circles are arranged for adding:



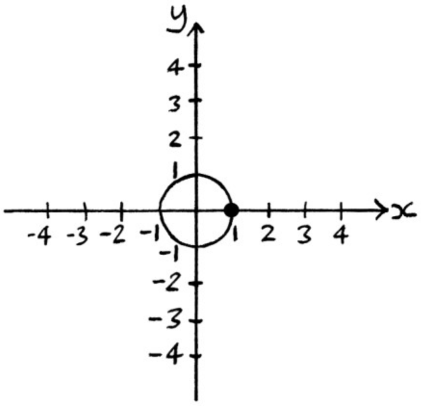


The lines have a direction from the centre of the circle outwards. In this way, these lines can be called "vectors", where a vector is just the name for a line that is treated as having a length and an angle (or a length and a direction, which ultimately means the same thing). Using vectors to work out the details of the resulting circle is another way of doing exactly the same thing we were doing before. Sometimes, thinking about the result of adding circles is made easier when thinking of lines (i.e. vectors), and sometimes, it is made easier when thinking of circles moving around each other.

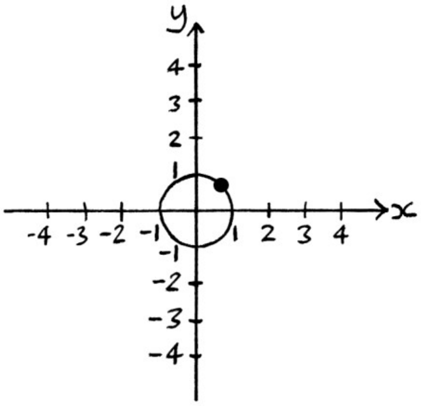
As an example, we will add the following waves:
"y = cos ((360 * 2t) + 0)"
"y = cos ((360 * 2t) + 45)"
"y = cos ((360 * 2t) + 90)"

We will plot these on their own circles, just to show how they look:

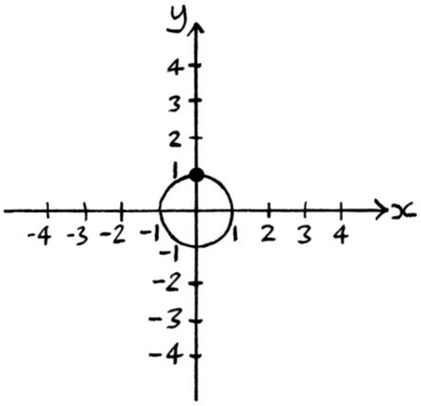
"y = cos ((360 * 2t) + 0)":



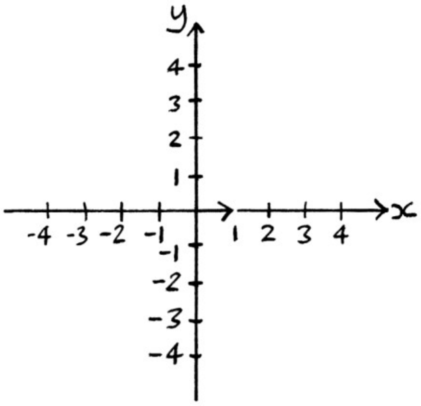
"y = cos ((360 * 2t) + 45)":

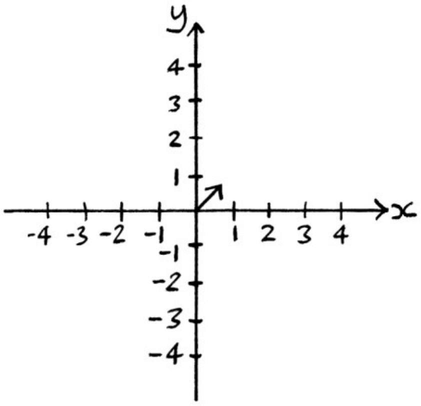


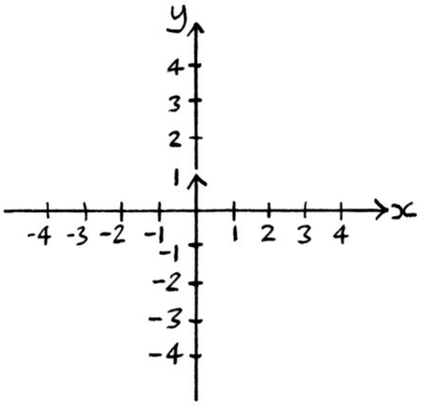
“y = cos ((360 * 2t) + 90)”:



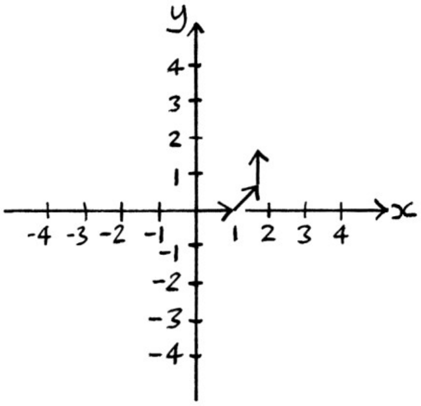
The vectors (in other words, the lines) for the three circles look like this:



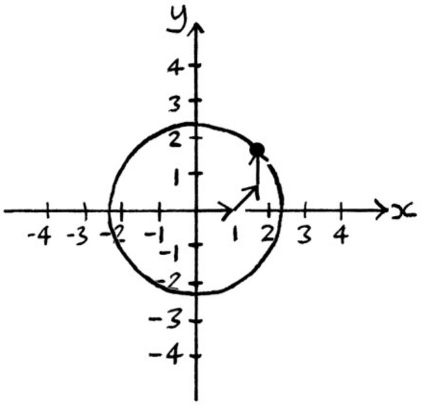


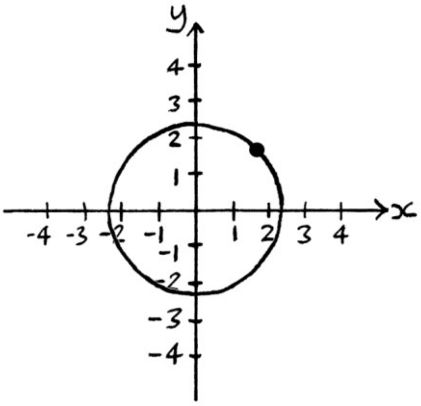


We join them together as so:



The end of the joined-together vectors represents the phase point of the resulting circle.





We can join the vectors in a picture as above, and measure them, or we can use maths to work out where the end is. The first vector is 1 unit long and at an angle of 0 degrees. Therefore, if we follow the line, it moves 1 unit across, and 0 units upwards.

The second vector is 1 unit long and at an angle of 45 degrees. Therefore, this line moves 0.7071 units to the right, and 0.7071 units upwards.

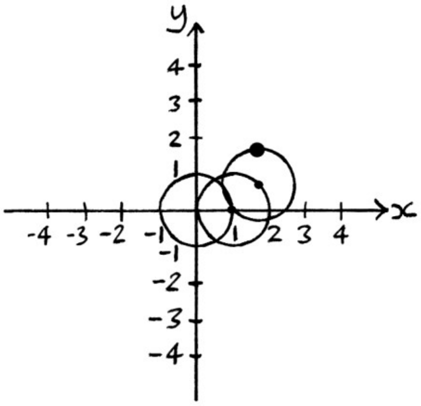
The third vector is 1 unit long and at an angle of 90 degrees. Therefore, this line moves 0 units across, and 1 unit upwards.

We can add up the x-axis movements of all the vectors, and add up the y-axis movements of all the vectors. The x-axis movements are: 1 + 0.7071 + 0 = 1.7071 units. The y-axis movements are: 0 + 0.7071 + 1 = +1.7071 units. This means that the end of the joined line is at (1.7071, 1.7071). The end of the line is the phase point of the resulting circle.

This means that the resulting phase point is:
$\sqrt{1.7071^2 + 1.7071^2}$ = 2.4142 units away from the origin and at an angle of arctan (1.7071 ÷ 1.7071) = 45 degrees [We check this is the answer we want, which it is.]

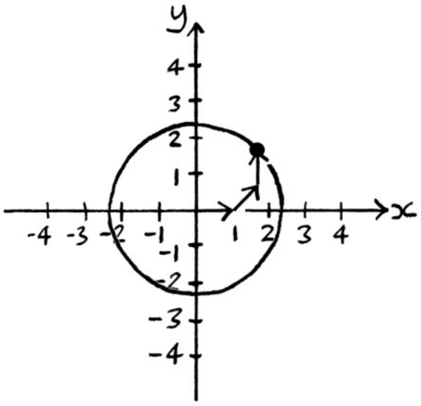
As we were originally adding Cosine waves, we can give the result in the form of a Cosine wave as "y = 2.4142 cos ((360 * 2t) + 45)".

We could have calculated exactly the same result by putting the circles together as they would be at t = 0, and working out the position of the outer object:

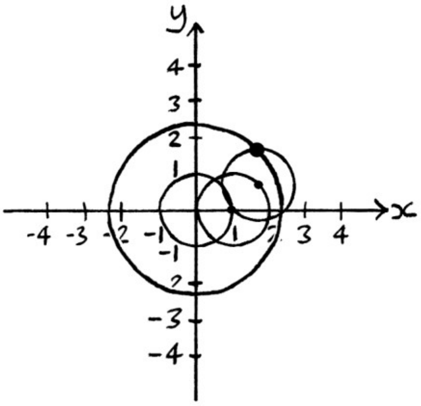


Whether we use vectors or circles, we use the same maths to work out the angle and distance of the resulting phase point from the origin of the axes. The difference is in the way we think about what we are doing.

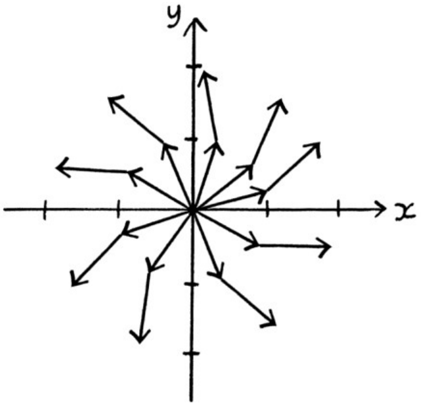
Vectors:



Circles:



A vector that connects the centre of the circle to the object rotating around the circle, *as the object rotates*, is called a “phasor”. A phasor is just a rotating vector. In other words, a phasor is just a rotating line where we pay attention to its angle and length. In this picture from earlier:



... the lines can be thought of as vectors at individual moments in time, or they can all be thought of as showing a single phasor rotating around the centre of the circle. A phasor is just another way of thinking about an object rotating around a circle.

Evenly spaced phases

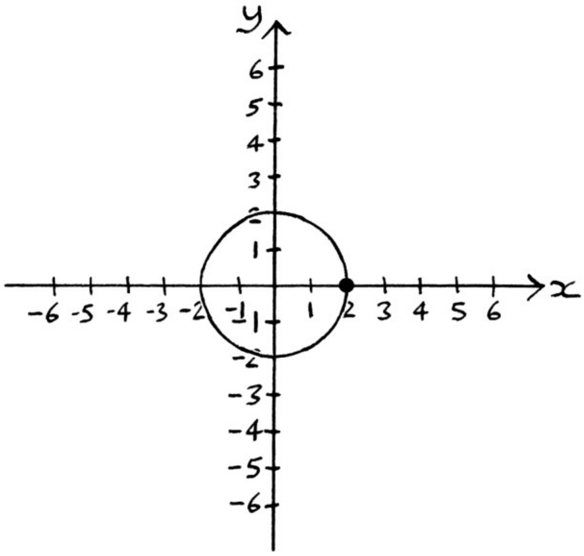
In the chapter on adding waves, I discussed adding two waves of the same frequency and amplitude, but with phases that were 0 and 180 degrees. The result was a flat line with the formula “y = 0” or “y = 0t”, depending on how we want to phrase it. When two waves are added in this way, it is easy to see why they amount to nothing – we can treat the 180-degree phase wave as having zero phase and a negative amplitude. Then if the waves have the same amplitude to start with, making one amplitude negative will make the sum result in zero.

For example, “y = sin 360t” added to “y = sin (360t + 180)” is the same as:
“y = sin 360t” added to “y = −1 sin 360t”
... which is:
“y = sin 360t – sin 360t”
... which results in:
“y = 0 sin 360t”
... or:
“y = 0t”
... or:
“y = 0”.

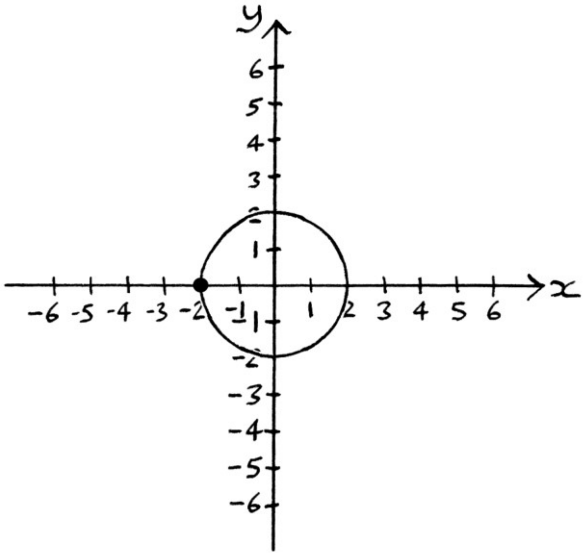
It is also the case that *any* number of waves with the same amplitude and frequency, but with phases that are evenly spaced from 0 to 360 degrees, will also cancel each other out, and result in “y = 0t”. For example, if we have the following three waves, which have phases evenly spaced between 0 and 360 degrees...
“y = sin 360t”
“y = sin (360t + 120)”
“y = sin (360t + 240)”
... then the sum will be “y = 0t”.

The reason why things work in this way becomes clearer when we look at the circle chart and vectors.

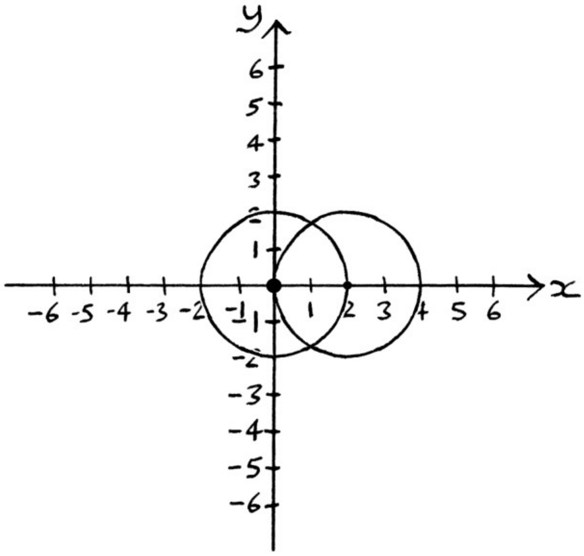
To begin to explain this, we will look at “y = 2 sin 360t” added to “y = 2 sin (360t + 180)”. We will turn these Sine waves into circles. The two circles look like this:



... and:



If we arrange them together for adding, they look like this at t = 0:

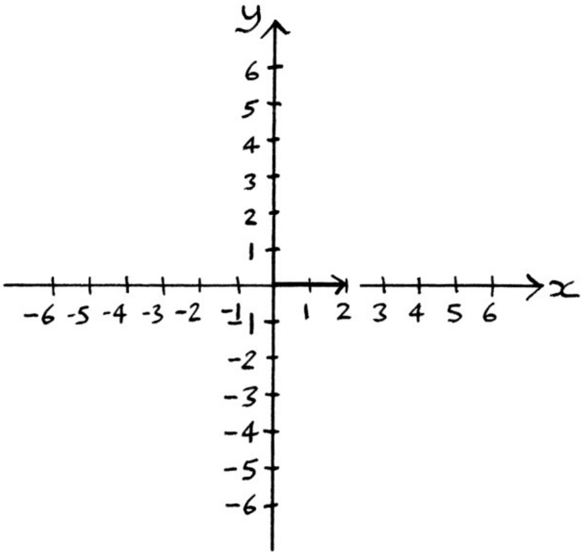


As the circles have the same frequency, it might be clear that as the outer circle rotates around the inner circle, the outer object will remain exactly over the coordinates (0, 0).

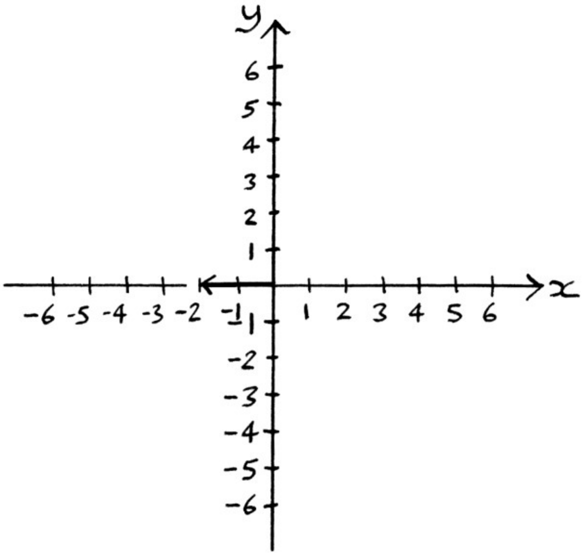
The resulting shape will be an infinitely small point at (0, 0). [One might think that there should be nothing at all on the graph, but really, there should be something at (0, 0) to indicate that all the points of the shape are at zero.] The shape's derived Sine and Cosine "wave" graphs will both just be straight lines at y = 0.

The phase points for the two original circles are at opposite sides. When the circles are added together, we end up with a phase point that is at (0, 0). It is as if adding the circles has averaged the two phase points.

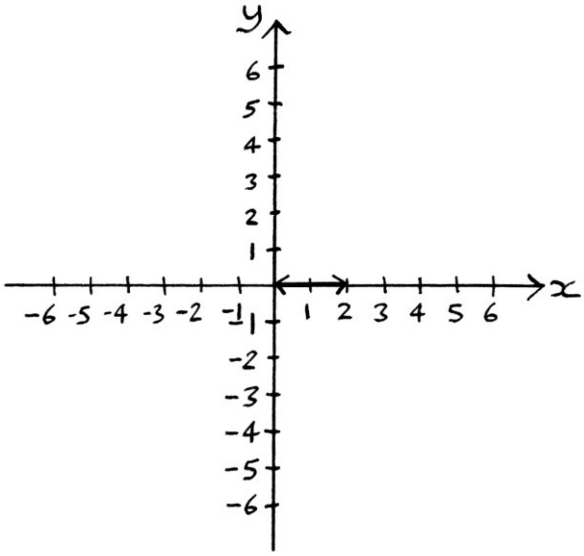
We can also add these waves using vectors. For “y = 2 sin 360t”, the relevant vector moves 2 units to the right and 0 units upwards [Drawn with a gap cut out of the axis to make it clearer]:



For “y = 2 sin (360t + 180)”, the relevant vector moves 2 units to the left and 0 units upwards [Drawn with a gap cut out of the axis to make it clearer]:



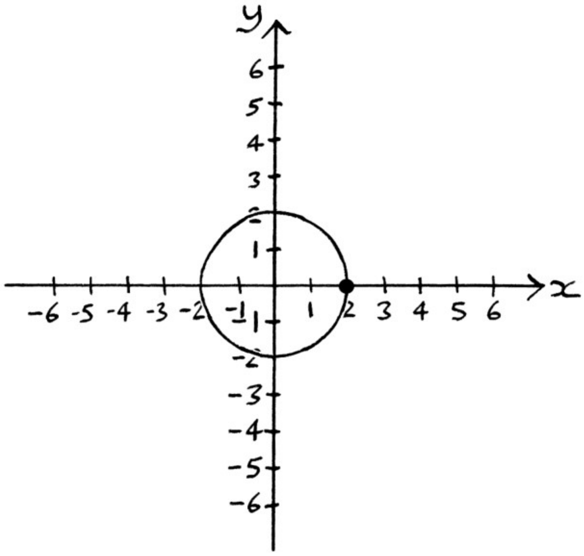
If we join the two vectors together, the first one moves two units to the right and the second one moves two units to the left. The end of the joined lines is at the beginning – they cancel each other out.

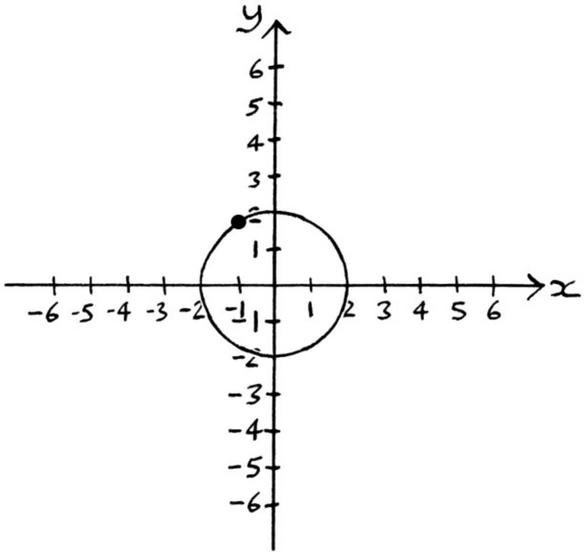


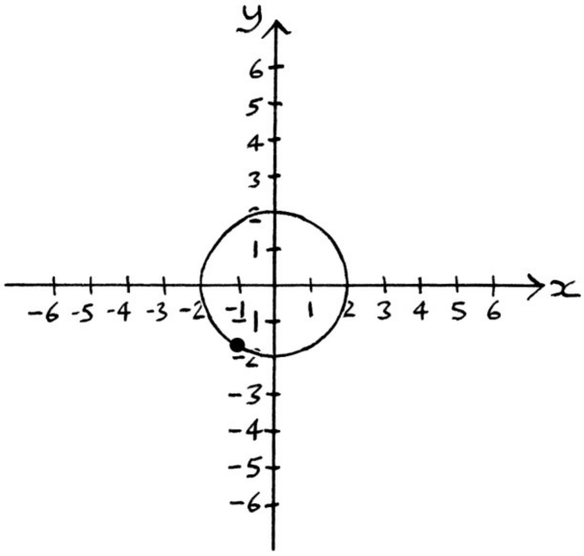
The result of adding the vectors is a vector with no length and no angle. The result represents a circle with no radius, which in turn represents two waves with no amplitude.

Now, we will add three waves with phases of 0, 120 and 240 degrees. We will add:
“y = 2 sin (360t + 0)”
“y = 2 sin (360t + 120)”
“y = 2 sin (360t + 240)”

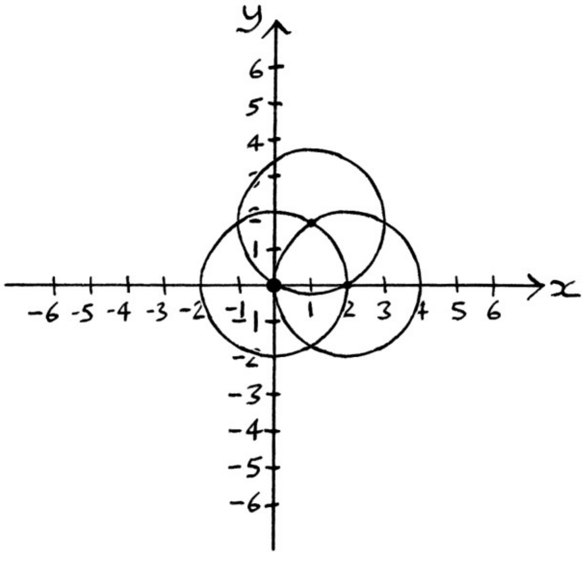
We will turn the Sine waves into circles as so:







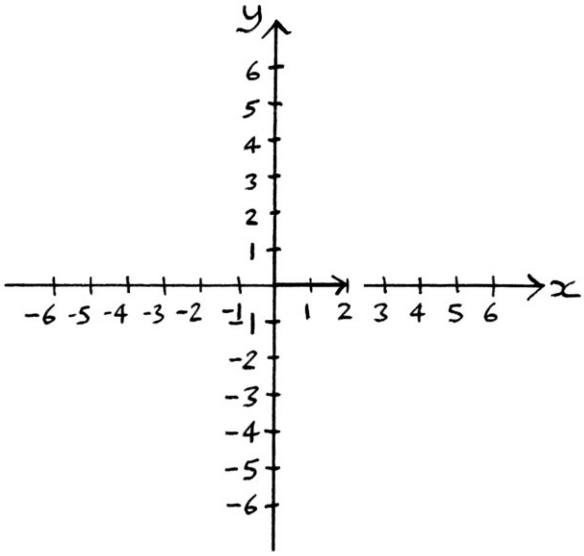
If we arrange the circles together for adding, they will look like this at t = 0:



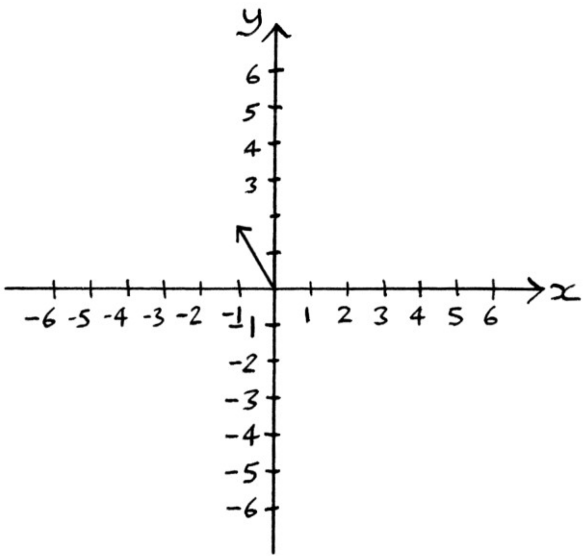
In the above picture, the first circle is centred at (0, 0), and its phase point is at (2, 0). The second circle is centred at (2, 0), and has its phase point at 120 degrees to this. The third circle is centred over the second circle's phase point and has its own phase point, which is the outer object, at 240 degrees to its centre. This happens to be the origin of the axes. As all the circles have the same frequency, the outer object will remain at the origin of the axes for all time. The resulting shape will be a point at (0, 0). The derived Sine and Cosine "waves" will be straight lines at y = 0.

The phase points on the original circles are at even angles around the circle. They are all at the same distance from the origin, but in opposing directions. When the three circles are added together, it is as if we are taking an average of the distances and angles of the phase points from the origin of the axes. Now that we know what is happening, we can guess the results to some other additions by just looking at the phase points. Before we do that, we will calculate the above sum using vectors.

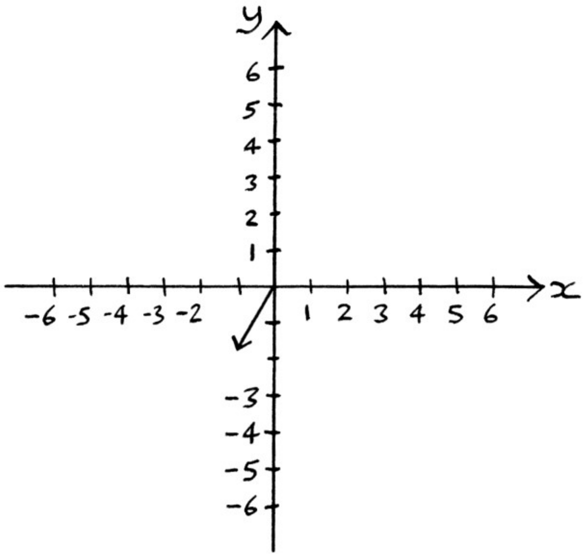
The vector for the first wave is a line that moves 2 units to the right and zero units upwards [drawn with a gap cut out of the x-axis to make it clearer]:



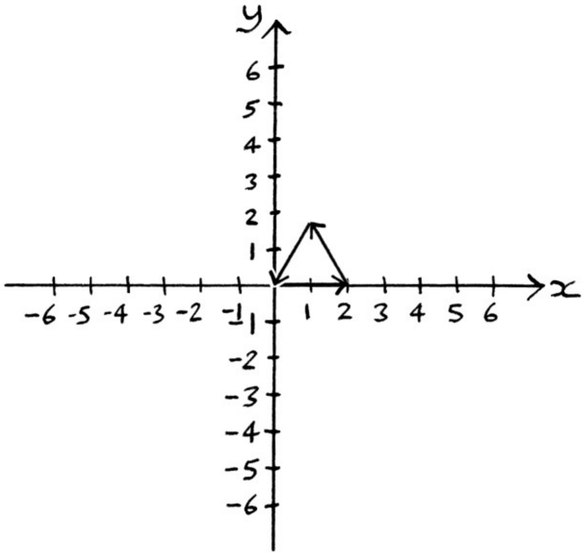
The vector for the second wave is a line that moves 2 units at an angle of 120 degrees:



The vector for the third wave is a line that moves 2 units at an angle of 240 degrees:



If we join the vectors together, we end up with this:



The end of the joined vectors ends up exactly where the start is. Therefore, the result represents a circle with zero radius, and in turn a pair of waves with zero amplitude.

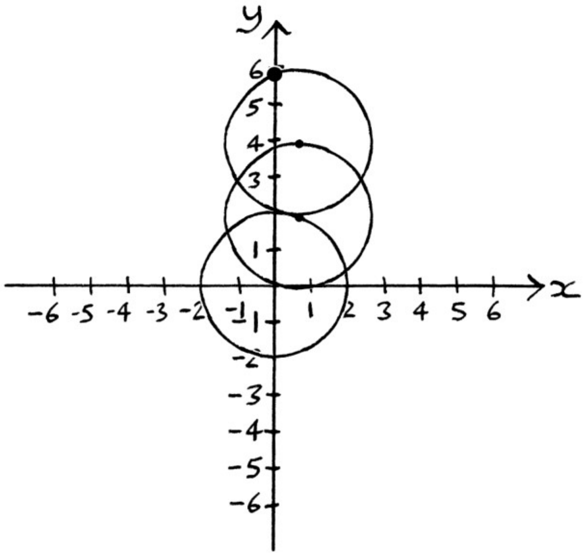
Guessing results

If we draw all the phase points on the same circle, or if we use vectors, it becomes easier to visualise how the result will look. Sometimes, we can intuitively know the result.

As an example of guessing, we will add the following waves:
“y = 2 sin (360t + 70)”
“y = 2 sin (360t + 90)”
“y = 2 sin (360t + 110)”

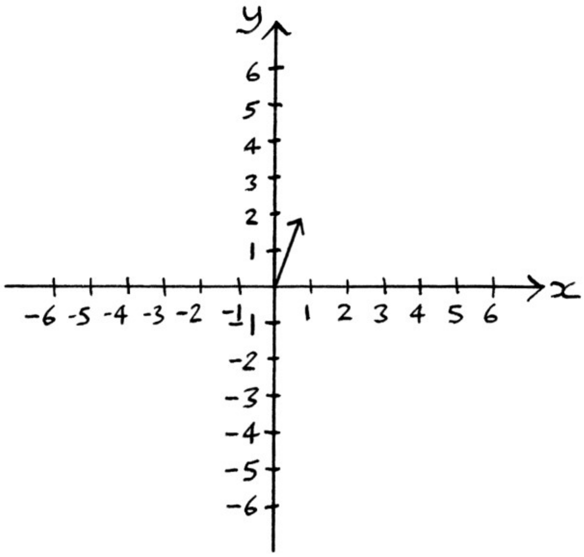
We can guess that the result of the addition of the circles for the three waves will have a phase at the average angle, which is 90 degrees. We can also know that the amplitude will be longer than 2 units because all the phase points are in the same half of the circle. Calculating the actual amplitude would require some maths though.

The circles arranged together for adding look like this:

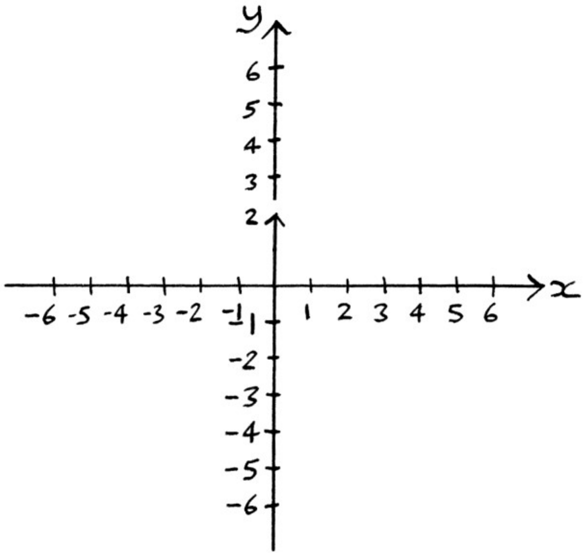


The angle of the outer object at t = 0 is 90 degrees, so the phase point of the resulting circle will be 90 degrees, as we guessed. We could measure the distance from the outer object to the origin of the axes, and it would be 5.7588 units. Therefore, the radius of the resulting circle is 5.7588 units, and therefore, the amplitudes of the resulting derived waves will be 5.7588 units. We probably could not have guessed that.

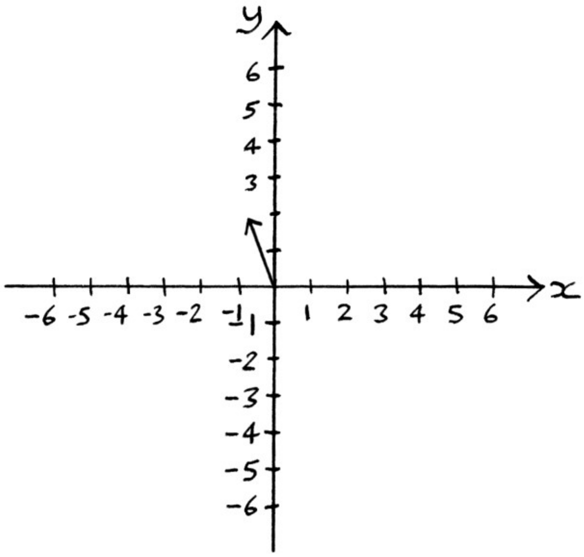
Using vectors, we would have the same result. The first vector is a line that moves 2 units at 70 degrees. This is 2 sin 70 = 1.8794 units upwards and 2 cos 70 = 0.6840 units to the right.



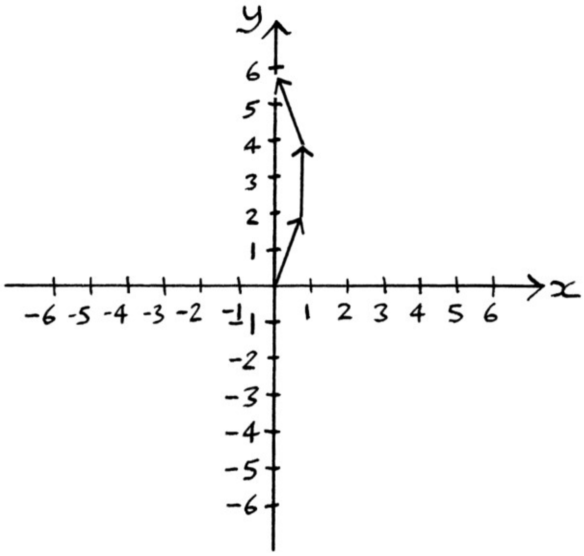
The second vector is a line that moves 2 units at 90 degrees. This is 2 sin 90 = 2 units upwards, and 2 cos 90 = 0 units across.



The third vector is a line that moves 2 units at 110 degrees. This is 2 sin 110 = 1.8794 units upwards, and 2 cos 110 = −0.6840 units across.



Joining the vectors together, we end up with this:



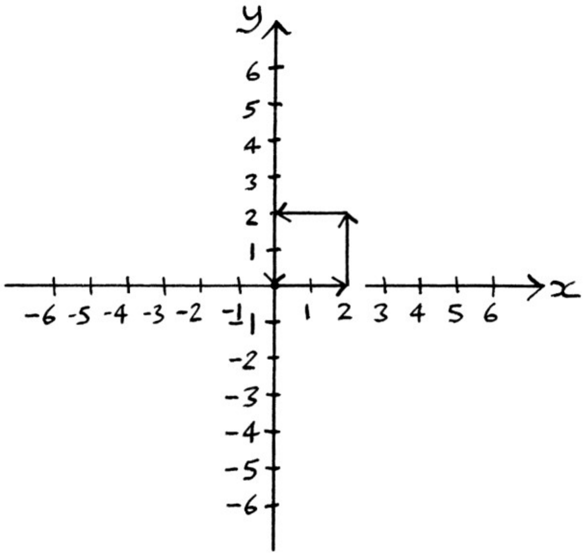
We can calculate the position of the end point of the joined vectors:
The y-axis of the end of the joined up vectors is 1.8794 + 2 + 1.8794 = 5.7588 units.
The x-axis of the end of the joined up vectors is 0.6840 + 0 + −0.6840 = 0 units.

Therefore, the coordinates of the end are (0, 5.7588). The end of the joined up vectors is at the same point as when we did this with circles. Its position represents a circle and its pair of waves with an amplitude of 5.7588 units and a phase point at 90 degrees. The result of adding "y = 2 sin (360t + 80)", "y = 2 sin (360t + 90)" and "y = 2 sin (360t + 100)" is "y = 5.7588 sin (360t + 90)".

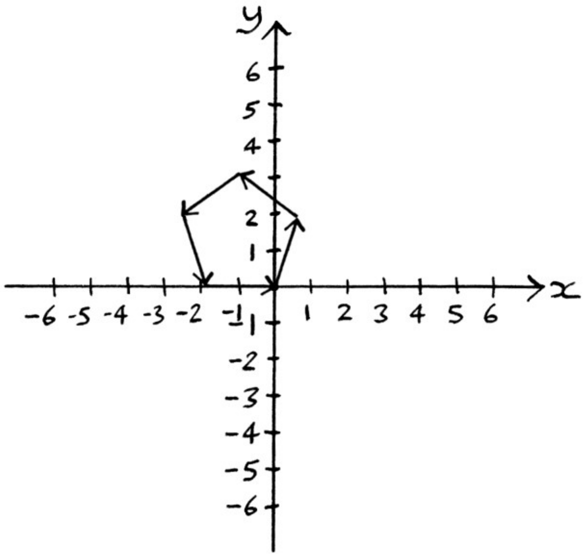
More evenly spaced phases

It is the case that the sum of any number of waves of the same amplitude and frequency, but with phases that are at evenly spaced angles from 0 to 360 degrees, will produce a "wave" with zero amplitude. This might not be immediately clear when illustrated with circles, but it is obvious when we think with vectors. The joined vectors will always end up exactly where they started.

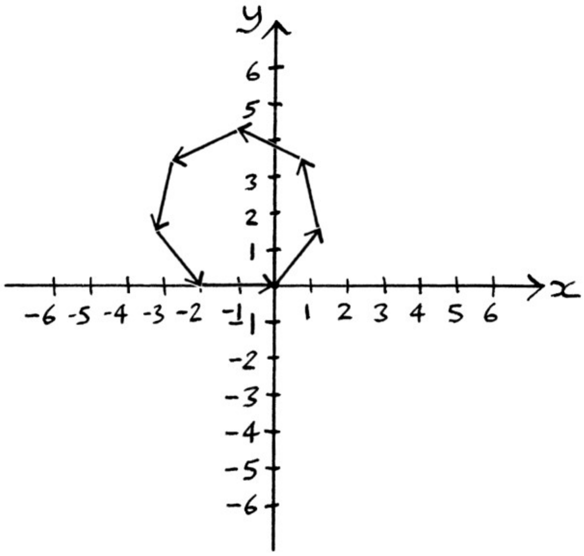
For example, four waves with the phases of 0, 90, 180 and 270 degrees:



Five waves with the phases of 0, 72, 144, 216 and 288 degrees:



Seven waves with phases of 0, 51.4286, 102.8571, 154.2857, 205.7143, 257.1429 and 308.5714 degrees:



Adding up a set of waves

As we just saw, the sum of a series of waves (or circles) with the same frequency and amplitude, zero mean level, and phases that are spaced evenly from 0 to 360 will result in a wave with the formula "y = 0t". We can use this idea to add up particular waves without needing much thought. If we have such a set of waves, but with one missing, we know that the sum of the ones we do have will be the negative of the missing one. It has to be this way because the addition of the last wave of the set must make the sum into zero. We just have to work out what the negative of the missing one is.

As an easy example, supposing we have these waves:
"y = 4 sin (360 * 7t)"
"y = 4 sin ((360 * 7t) + 90)"
"y = 4 sin ((360 * 7t + 270)"
... then we know that these are three of four waves with phases that are 90 degrees apart. The missing one must have a phase of 180 degrees. The sum of all four waves must result in "y = 0t". Therefore, the sum of the three existing waves must be equal to the negative of the missing wave. Therefore, the sum of the three existing waves must be the negative of "y = 4 sin ((360 * 7t) + 180)", which is:
"y = −4 sin ((360 * 7t) + 180)"
... which we can rephrase to be:
"y = 4 sin (360 * 7t)"
Therefore, the sum of the three waves is "y = 4 sin (360 * 7t)".

We have calculated the sum of 3 waves without a calculator. If we had used a calculator, it would have been a more complicated calculation.

As another example, we will add up these waves:
“y = 7.6 sin (360 * 11.5t)”
“y = 7.6 sin ((360 * 11.5t) + 30)”
“y = 7.6 sin ((360 * 11.5t) + 60)”
“y = 7.6 sin ((360 * 11.5t) + 90)”
“y = 7.6 sin ((360 * 11.5t) + 120)”
“y = 7.6 sin ((360 * 11.5t) + 150)”
“y = 7.6 sin ((360 * 11.5t) + 180)”
“y = 7.6 sin ((360 * 11.5t) + 210)”
“y = 7.6 sin ((360 * 11.5t) + 270)”
“y = 7.6 sin ((360 * 11.5t) + 300)”
“y = 7.6 sin ((360 * 11.5t) + 330)”

These are 11 of the 12 waves of a set that has phases that are 30 degrees apart. The missing wave is: “y = 7.6 sin ((360 * 11.5t) + 240)”. The sum of the existing 11 waves must be the negative of the missing wave. It will be:
“y = −7.6 sin ((360 * 11.5t) + 240)”
... which we can rephrase as:
“y = 7.6 sin ((360 * 11.5t) + 240 − 180)” or “y = 7.6 sin ((360 * 11.5t) + 240 + 180)”
... both of which end up as:
“y = 7.6 sin ((360 * 11.5t) + 60)”

Therefore, the sum of the 11 waves is “y = 7.6 sin ((360 * 11.5t) + 60)”. We have added up 11 waves with no effort at all.

This idea is interesting, but it is unlikely that we would ever be in a situation where we are missing one of a set of waves that we need to add up quickly. The idea could be used as part of a slightly strange magic trick.

Converting from waves to vectors

We do not need to convert waves to circles, and then circles to vectors. We can just go from waves to vectors in one step – the amplitude of the wave gives the vector its length; the phase of the wave gives the vector its angle. [Note that generally, the length of the vector is called the "magnitude". The angle can be called the "direction"].

Given all of this, we can also say that we can represent the amplitude and phase of a wave (but not its frequency or mean level) using just a vector. For consistency's sake, it is better to treat all vectors as representing one type of wave – either Sine waves or either Cosine waves – and to be consistent in doing this. If we treat vectors as representing Sine waves, then a vector with an angle of 35 degrees and a magnitude of 102 units could represent a Sine wave such as:
"y = 102 sin (360t + 35)
... or a Sine wave with that amplitude and phase, but a different frequency and mean level. If we treat vectors as representing Cosine waves, then that vector could represent a Cosine wave such as:
"y = 102 cos (360t + 35)
... or a Cosine wave with that amplitude and phase, but a different frequency and mean level.

Given that a vector cannot represent a frequency or mean level, it pays to be careful about the situations when we use them.

Vector summary

A vector is just a line for which we pay attention to its angle and length. We could just call vectors "lines" and it would not change anything. In this section, vectors act as a summary of what the circles are doing. They remove any extraneous information and leave us with the basic situation. Using vectors is essentially the same as using circles. Sometimes, things will be clearer with vectors; sometimes, things will be clearer with circles.

Drawing pictures with circles and waves

As we have seen, the sum of two or more circles is represented in the path taken by the object rotating around the outermost circle. Another way of thinking about this is that if we are given an uninterrupted path that forms a loop, we can treat it as being drawn out by circles added together, and therefore rotating around each other.

This idea is essentially the same as the frequency rule from the last chapter – if a signal is not a pure wave, then it is the sum of two or more pure waves. In this case, if a looped shape is not a circle, we can consider it the sum of two or more circles.

This means that we could draw a reasonably well thought out continuous shape around the origin of the axes, and then treat it as the result of the sum of two or more circles of positive and/or negative frequencies. The shape can be reduced to its vertically and horizontally derived signals, and then those signals can be analysed to find waves that can recreate them. Those waves will make up the circles for the shape.

A possible use for this is treating simple drawings as a sum of circles or waves. A more practical use might be in calculating which gears are required to achieve a particular movement. These two ideas could be used to create a mechanical machine that draws carefully chosen pictures.

A brief mention of adding helices

Given that a circle is really a helix seen end on, the addition of helices is very similar to the addition of circles. One thing to note is that, if a shape on the circle chart is not a circle, then the corresponding helix shape on the helix chart will not be a "pure" helix. The derived waves will not be pure waves.

In Chapter 13 on adding waves, we saw this rule:

> "Any periodic signal that is not a pure wave is the sum, or approximately the sum, of two or more pure waves of different frequencies, and various phases, amplitudes and mean levels."

As I said before, the rule is *mostly* true. A similar rule for adding circles is:

> "Any periodic shape on the circle chart that is not a circle is the sum, or approximately the sum, of two or more circles of different frequencies, and various phases, amplitudes and mean levels."

A similar rule for adding helices would be:

> "Any periodic helix-like shape on the helix chart that is not a pure helix is the sum, or approximately the sum, of two or more pure helices of different frequencies, and various phases, amplitudes and mean levels."

This rule for helices is a significant one to understand. In essence, it is another all-encompassing rule that connects waves, circles and helices. If you can mentally switch between waves, circles and helices, and think of them as different viewpoints of the same thing, then you may gain a better understanding of waves in general.

Conclusion

The results of adding waves are easiest to understand if we think of both the circles and the waves. Understanding why a sum has a particular result is much more straightforward if we can visualise what is happening with circles. Any seemingly strange results or patterns make much more sense when we think of the circles that are involved in the process.

Adding waves or circles that have the same frequency is straightforward – the result will be a pure wave or a perfect circle. Adding waves or circles with different frequencies is more complicated. It is reasonably straightforward to work out the frequency and mean level of the resulting signal, but it is much harder to work out the resulting maximum and minimum points.

If you really want to understand addition of waves and circles, then it pays to become involved. Draw some waves and add them. Draw some circles and add them. Experiment with waves on a graphing calculator.

Chapter 15: The frequency domain

So far in this book, we have shown individual waves as graphs, and we have shown pairs of corresponding waves as circles on the circle chart. In this chapter, we will look at graphs that show the sums of waves or the sums of circles in a way relating to their frequencies. Such a graph is called a “frequency domain graph” because the main purpose of the graph is to illustrate the different frequencies existing in an addition. The frequency domain graph shows a list of all the waves or circles that were added together to make up a particular signal, and orders them by frequency. A frequency domain graph is a graph of addition. Such graphs show the “frequency domain” in the sense that they show the characteristics of added waves or circles as considered by frequency.

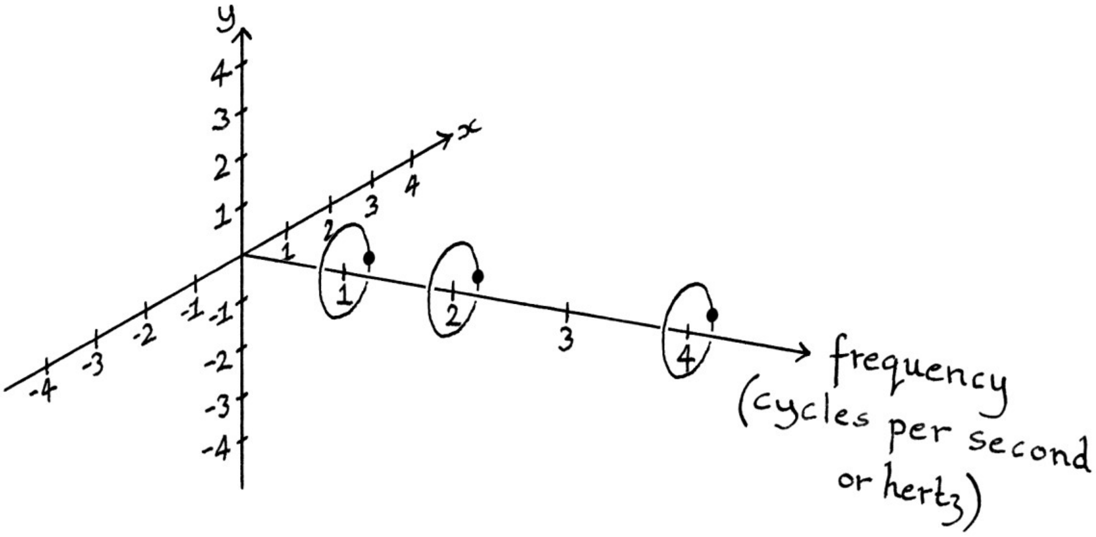
There are many different ways of showing frequency domain graphs, so I will start with what I consider the most straightforward, and move on to the others. What I consider the most straightforward is far from being the most common, so treat it as a step towards the other forms.

Sums of circles in the frequency domain

Basic example:

As a simple example, if we added the circles that represented the following pairs of waves:
“y = 1 sin 360t” and “y = 1 cos 360t”
“y = 1 sin (360 * 2t)” and “y = 1 cos (360 * 2t)”
“y = 1 sin (360 * 4t)” and “y = 1 cos (360 * 4t)”
... we could portray the sum on a frequency graph, or as most people would say, “in the frequency domain”. In this case, the graph will be three-dimensional, so we can show as much information as possible.

The graph looks like this:



The graph shows the added circles set out in a line according to their frequency. The frequency domain is showing the individual parts of the addition. [Note that because the picture is three dimensional, the circles appear as ellipses.]

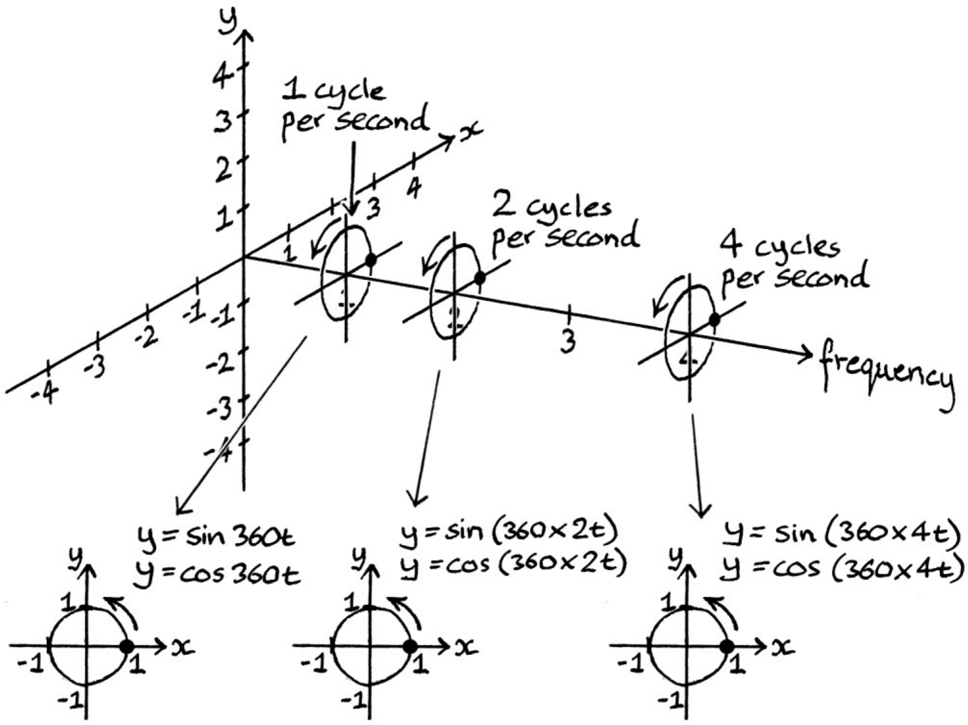
The circle with a 1 cycle-per-second frequency sits with its centre at 1 on the frequency axis. The circle with a 2 cycle-per-second frequency sits with its centre at 2 on the frequency axis. The circle with a 4 cycle-per-second frequency sits with its centre at 4 on the frequency axis.

The units of the frequency axis are in cycles per second or hertz, which are two terms for exactly the same thing. It is more common to see the term "hertz" in this type of graph.

The y-axis and x-axis have the same meaning as they did on the circle chart. They indicate the radiuses of the circles, which are the same as the amplitudes of the derived waves of each of those circles.

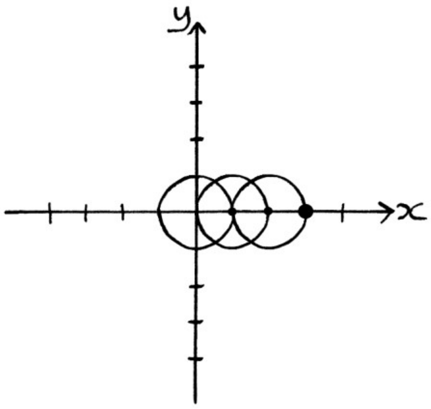
The circles are all the same size because they all have a radius of 1 unit, and they are representing pairs of waves that have amplitudes of 1 unit. The phase points of all the circles are zero. The mean levels of all the circles are zero, as can be deduced by how the circles are centred on the frequency axis. [We could also say that the circles are centred on the coordinates (0, 0) on the x and y-axes.]

Because the graph is three dimensional, it can be slightly difficult to see the radiuses of the circles clearly. Here is the same picture with more information to make it clearer:

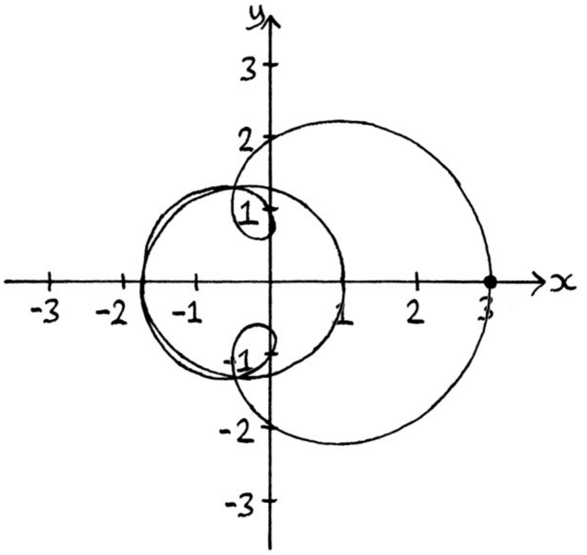


I have shown the actual circles that are represented in the frequency domain.

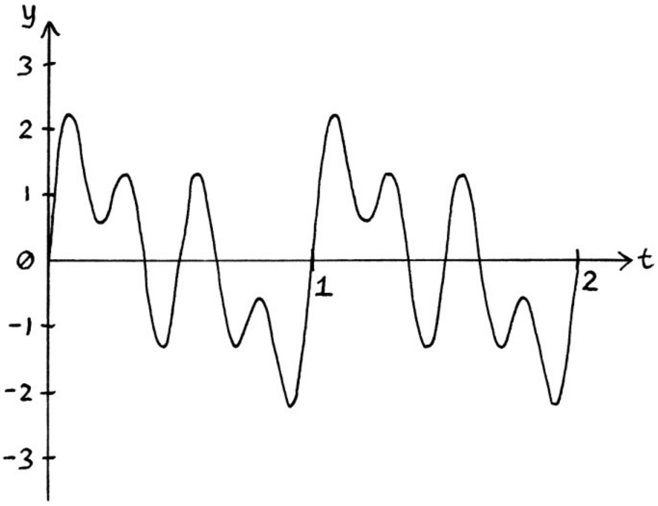
The whole frequency domain graph shows all the circles that are involved in the addition. We can take each circle and join them together, as we did in the last chapter, to produce the resulting sum. At t = 0, the joined circles look like this:



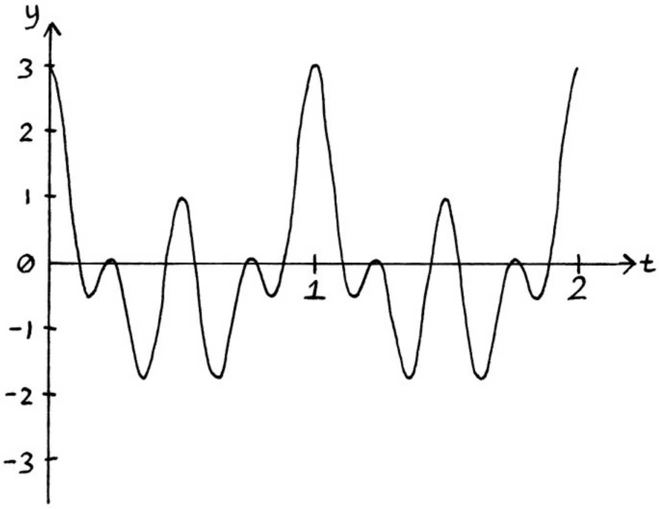
The shape resulting from the addition looks like this:



The vertically derived signal looks like this:



The horizontally derived signal looks like this:



The information in the frequency domain graph contains *all* the information required to create the resulting shape and its two derived signals. The amplitude, frequency, phase and mean levels of each circle are all shown in the graph. Given the graph, we can work out the derived waves from each circle, the circles' helices, the resulting shape, and the vertically and horizontally derived signals from that shape.

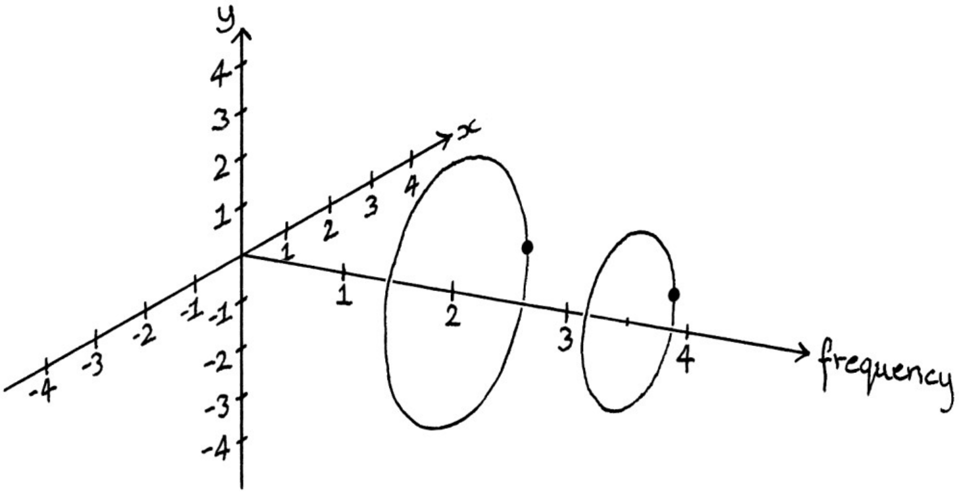
Thinking of all this the other way around, if we did not have any information apart from the resulting shape, we might be able to analyse that shape to discover which circles were added to create it, and then lay out the circles on the frequency domain graph to make the shape easier to comprehend. If we had no other information than one of the derived signals, we might be able to analyse that signal to discover which waves were added to make it, and then portray those waves as circles on the frequency domain graph. This might make the signal easier to understand.

Admittedly, because the frequency graph is a three dimensional graph, it can be slightly difficult to read off the dimensions of the circles, but in my view, this is still the best way to introduce the idea of the frequency domain.

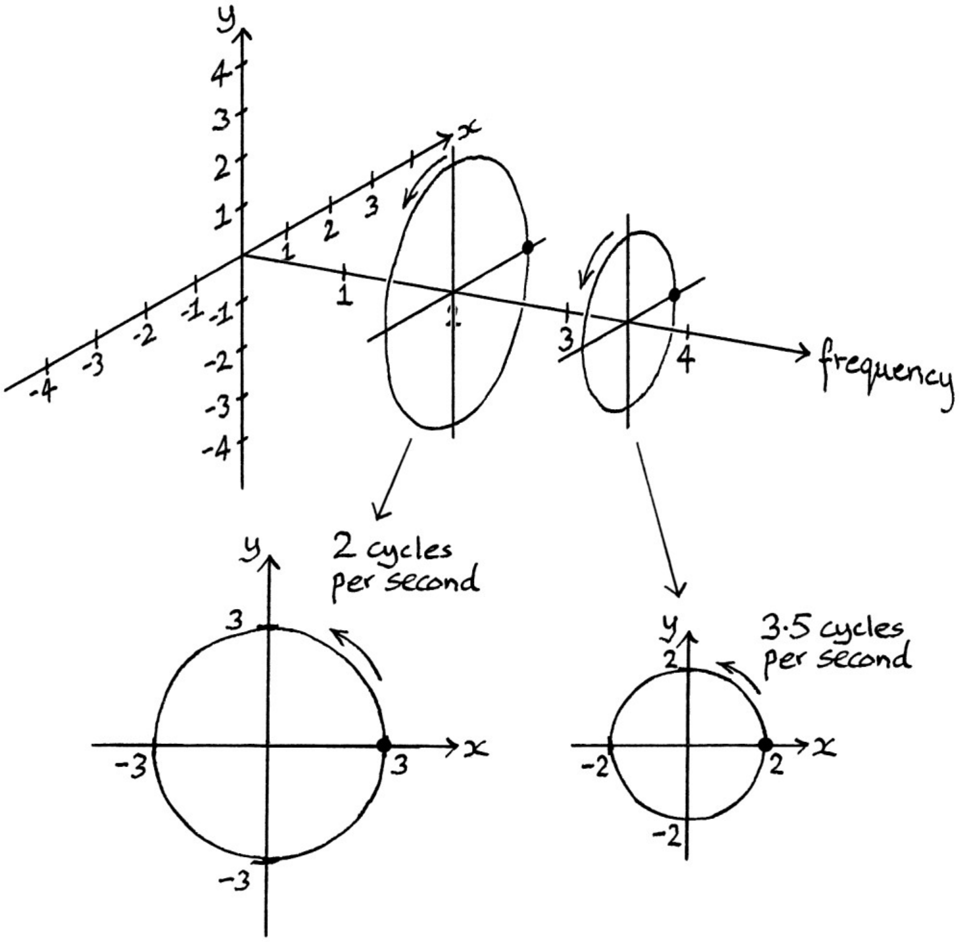
Amplitude

As another example, we will add the circles that represent the following pairs of waves:
"y = 3 sin (360 * 2t)" and "y = 3 cos (360 * 2t)"
"y = 2 sin (360 * 3.5t)" and "y = 2 cos (360 * 3.5t)"

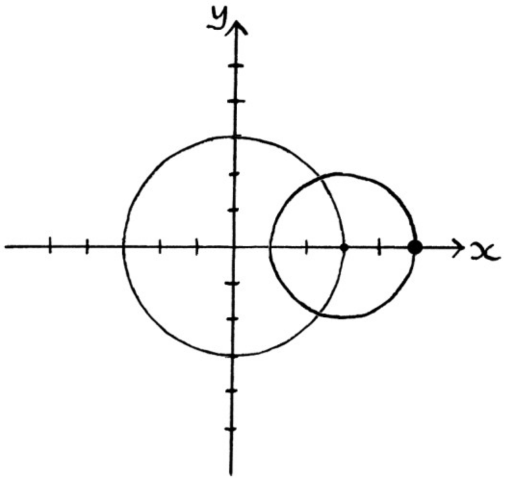
The frequency domain representation of the addition of the circles looks like this:



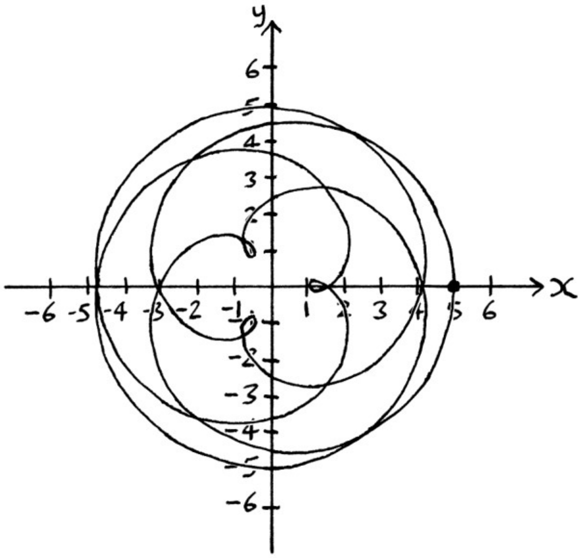
The 2-cycle-per-second circle is centred over 2 on the frequency axis. The 3.5-cycle-per-second circle is centred over 3.5 on the frequency axis. We can write notes on the graph to make its meaning clearer:



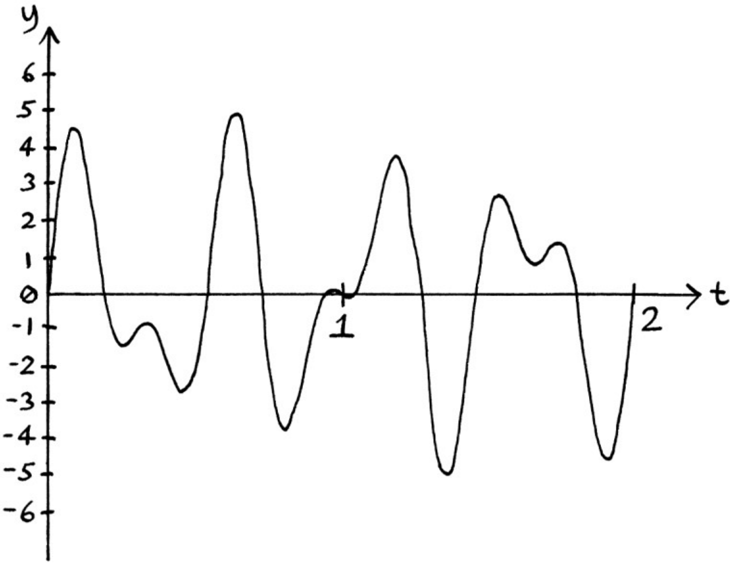
The frequency domain graph lists the circles that are being added together to create a particular shape. We can take each circle and arrange them together for adding:



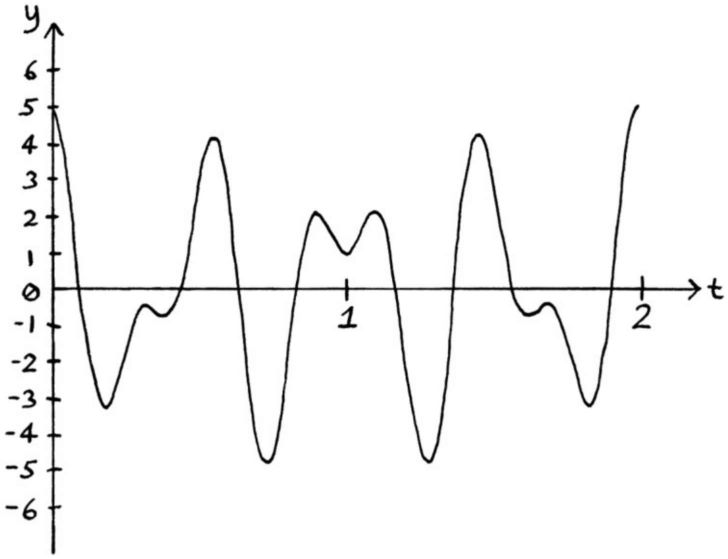
The outer object's movement as it rotates around the outer circle, and as that circle rotates around the inner circle, will indicate the resulting shape, and the two derived signals. The shape that this addition creates looks like this:



The vertically derived signal looks like this:



The horizontally derived signal looks like this:

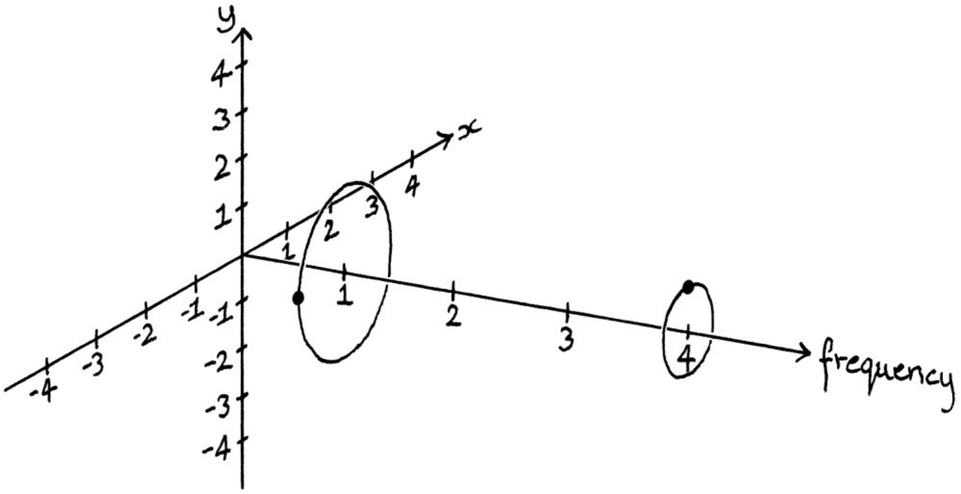


Phase

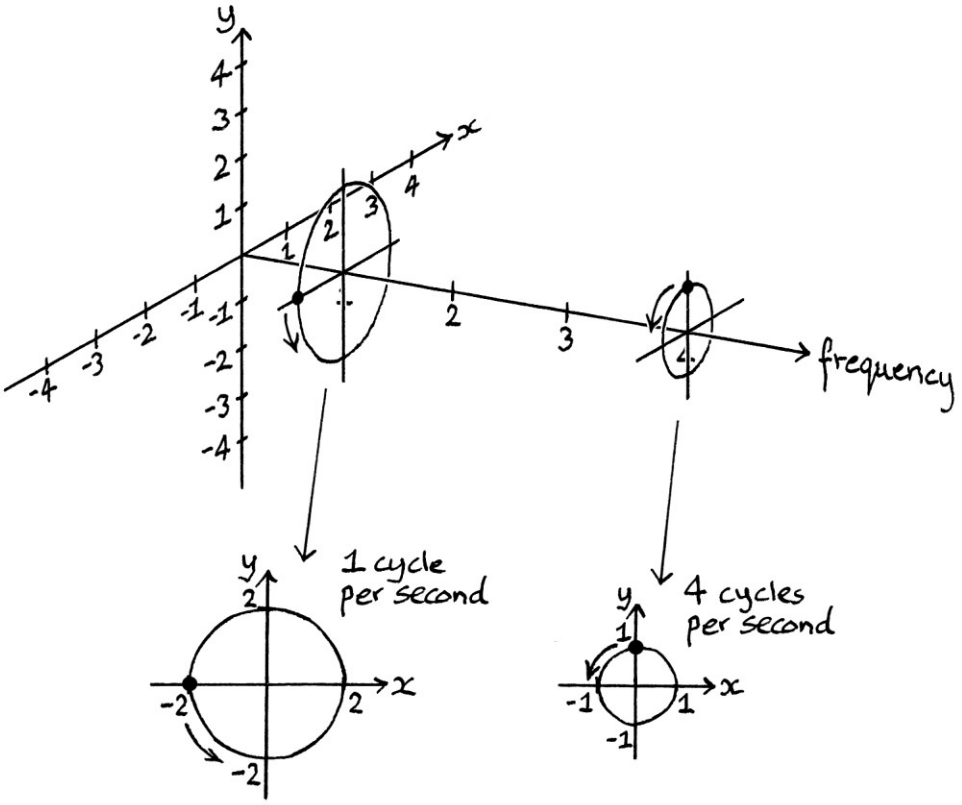
Because we are showing the full circles in these particular frequency domain examples, we can also indicate the phase of each circle.

For example, we will add the circles that represent the following pairs of waves:
“y = 2 sin (360t + 180)” and “y = 2 cos (360t) + 180)”
“y = 1 sin ((360 * 4t) + 90)” and “y = 1 cos ((360 * 4t) + 90)”

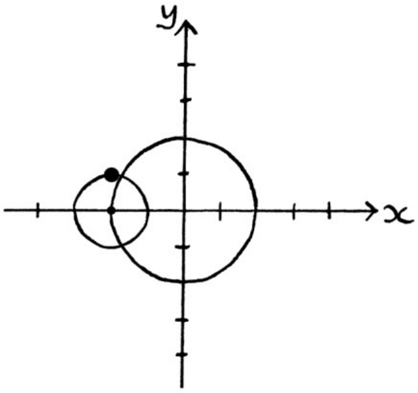
The sum of the two circles is represented in the frequency domain as so:



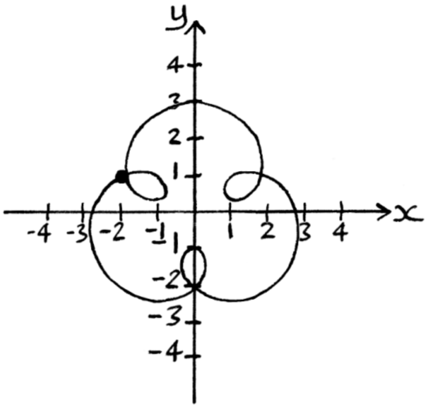
We can draw the graph with extra details to make it clearer:



When adding the circles, their position at t = 0 looks like this:



The resulting shape from adding these two circles looks like this:

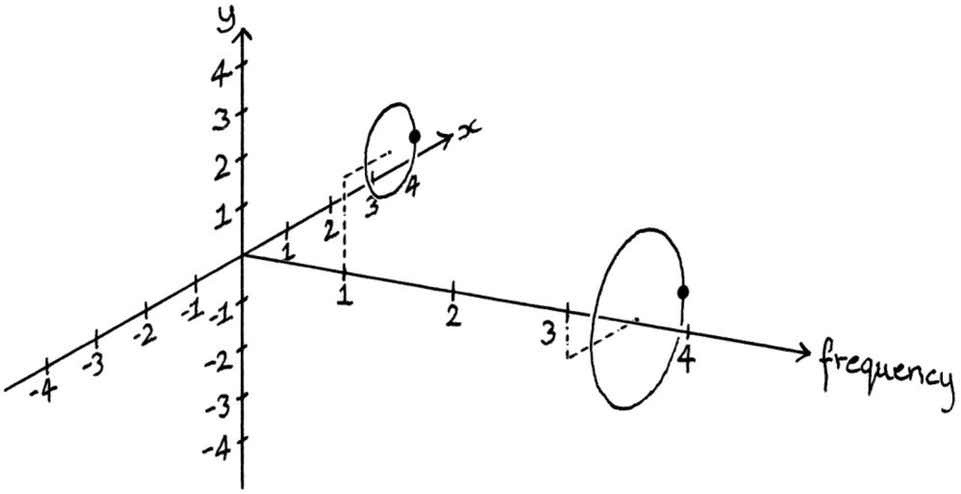


Mean levels

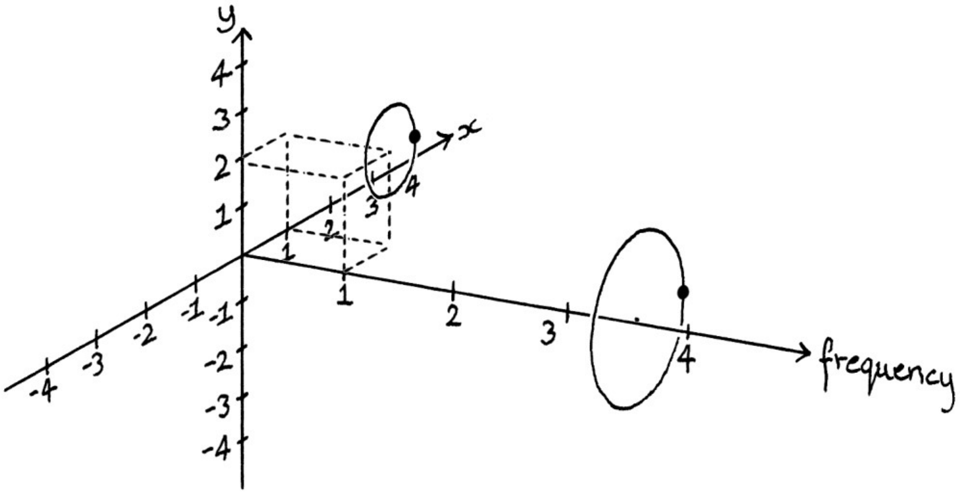
Mean levels can also be represented in this type of frequency domain. There are two ways of showing the mean levels. They can be kept with their circles, in which case they remain truer to the actual sum, but with the drawback that this makes the graph harder to read, or they can be grouped together and placed as a sort of phantom circle with zero amplitude and zero frequency. [The phase of the phantom circle is irrelevant as the amplitude is zero]. This “Mean Level circle” will really be a dot. Despite being placed at zero on the frequency axis, it will not disrupt the addition in the way that a zero-frequency circle would normally as it has zero amplitude. Remember that a zero-frequency circle is made up of a Sine “wave” and a Cosine “wave” that are straight lines. Zero amplitude in this phantom circle means that the straight lines will be at the heights of the mean levels, and nothing else.

As an example, we will add the circles that represent the following pairs of waves:
“y = 2 + 1 sin 360t” and “y = 1 + 1 cos 360t”
“y = −1 + 2 sin (360 * 3t)” and “y = 1.5 + 2 cos (360 * 3t)”

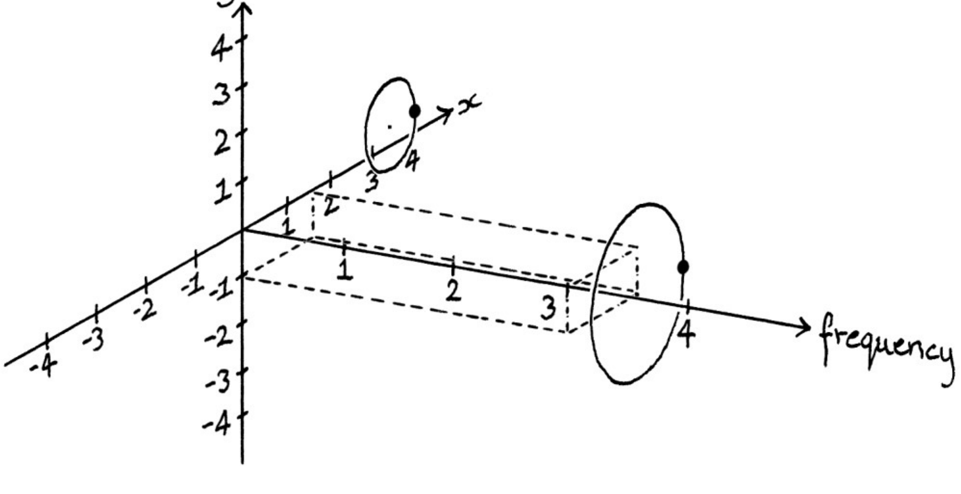
If we keep the mean levels with the circles, we end up with the following frequency domain graph, which is slightly difficult to read:



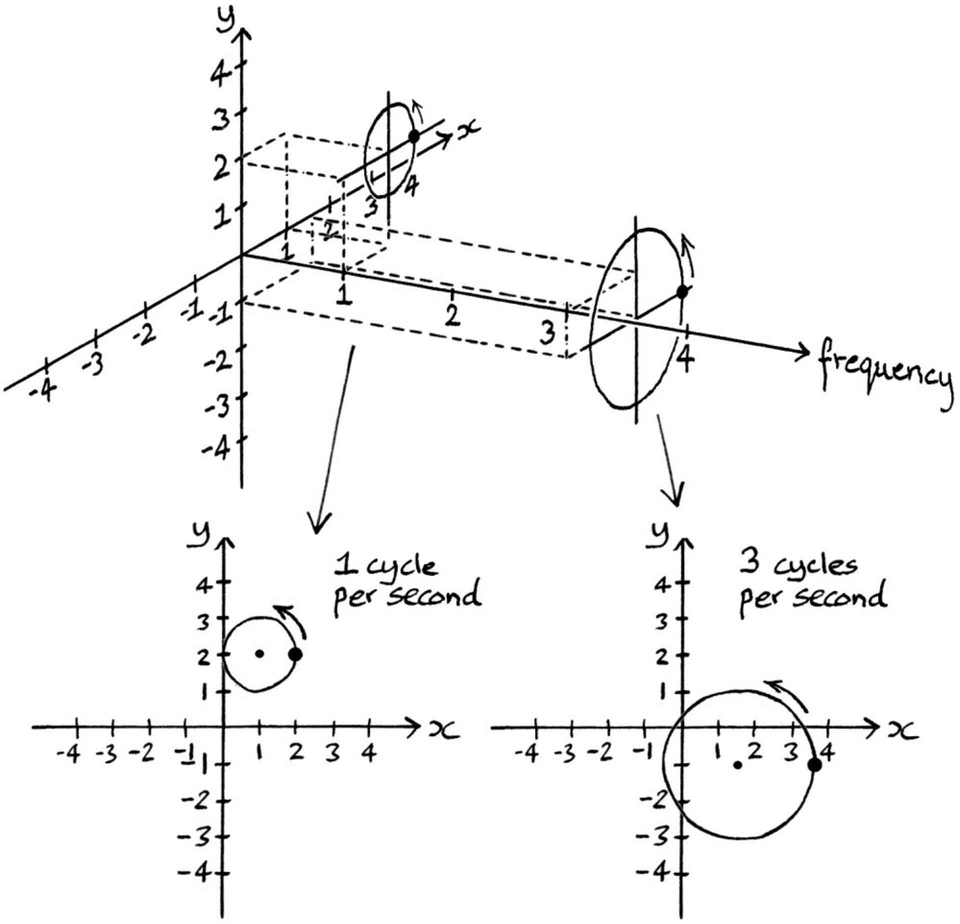
Here is the same picture with more details added to the 1 cycle per second circle to make it clearer as to exactly where it is:



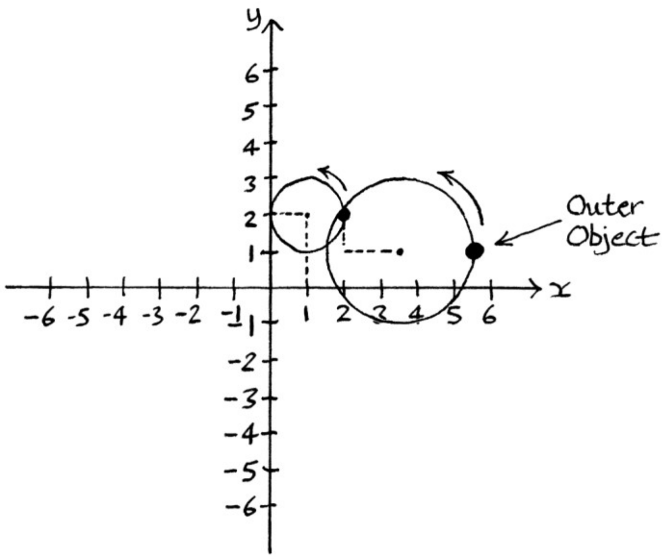
More details added to the 3.5-cycles-per-second circle:



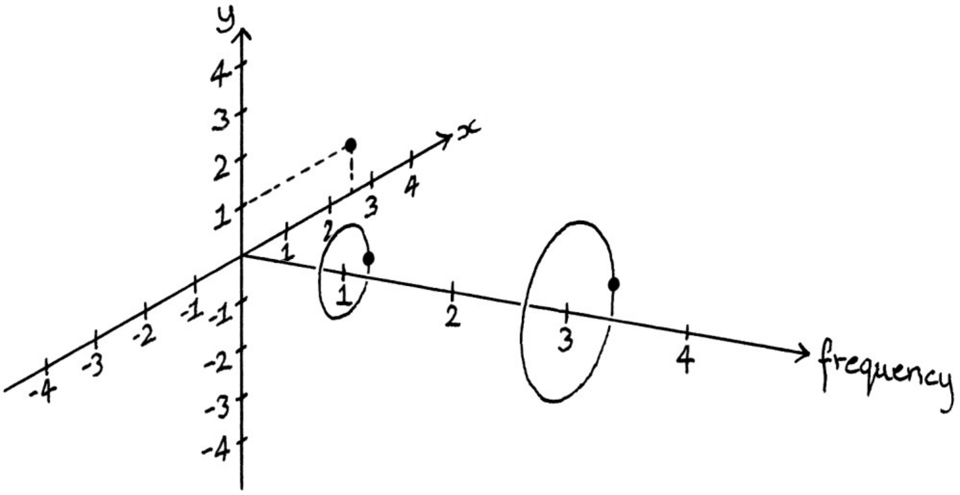
This is the original picture with even more information added to it:



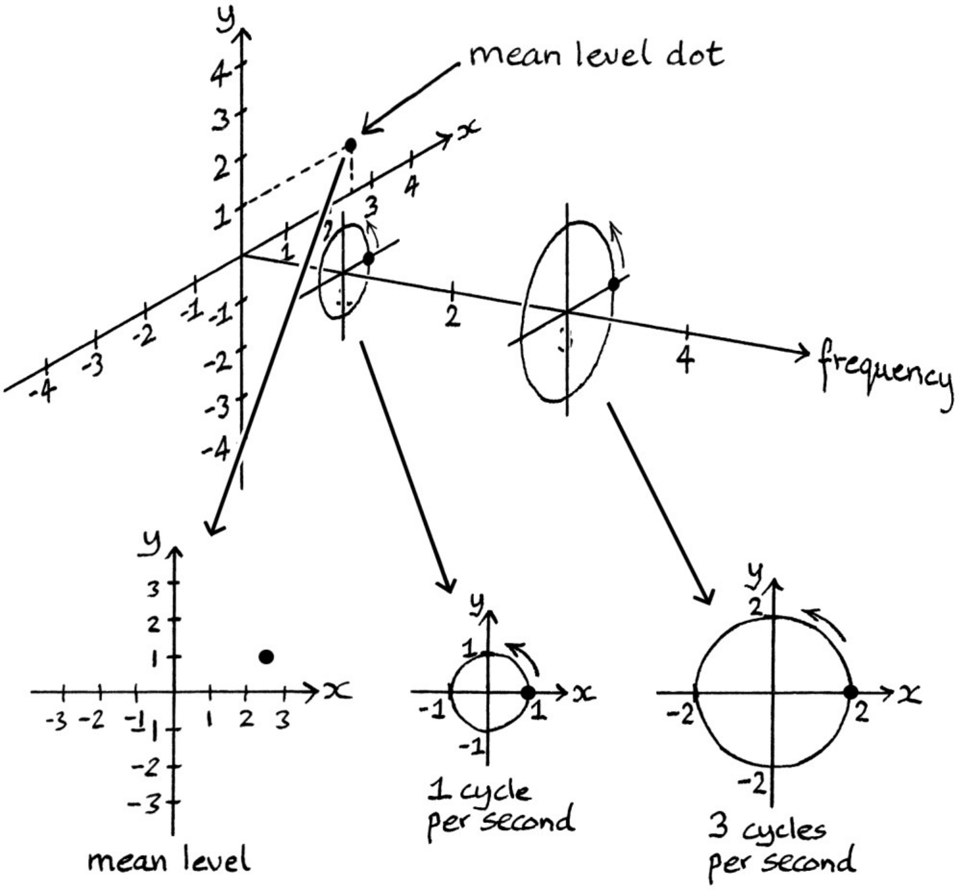
If we were to join the two circles together for the purposes of adding them, while keeping each circle with its own mean level, they would look like this at t = 0:



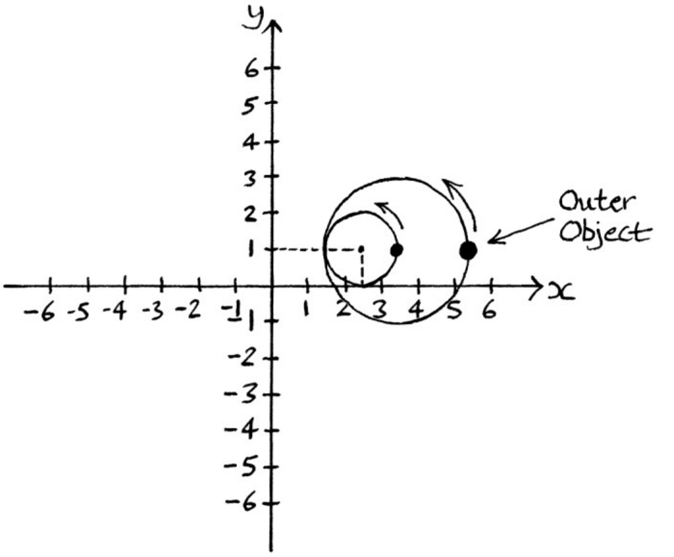
Keeping the mean levels with each circle makes a more accurate portrayal of the situation being explained in both the frequency domain graph and when the circles are joined for adding. However, it also makes everything much harder to read and understand. If we group the mean levels together, and put them at the frequency of zero, as a dot with zero amplitude, while centring each circle directly on the frequency axis, we end up with a much clearer graph:



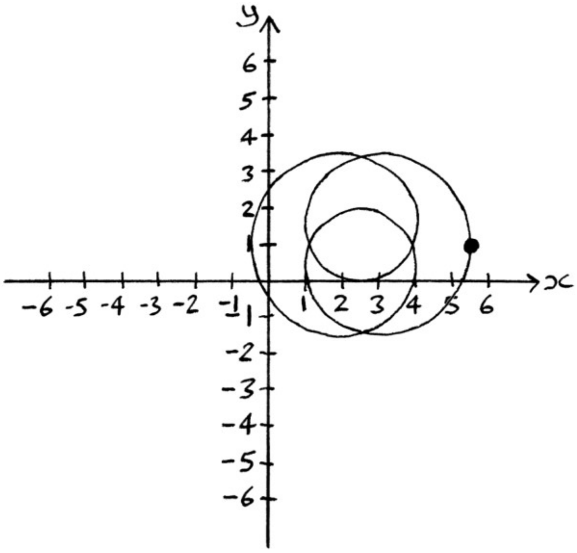
Here is the same graph with more notes:



If we join the two circles for the purposes of adding them, with the mean levels grouped together, they will look like this at t = 0:



No matter which option we choose for the mean levels, the shape created from adding these circles will look like this:



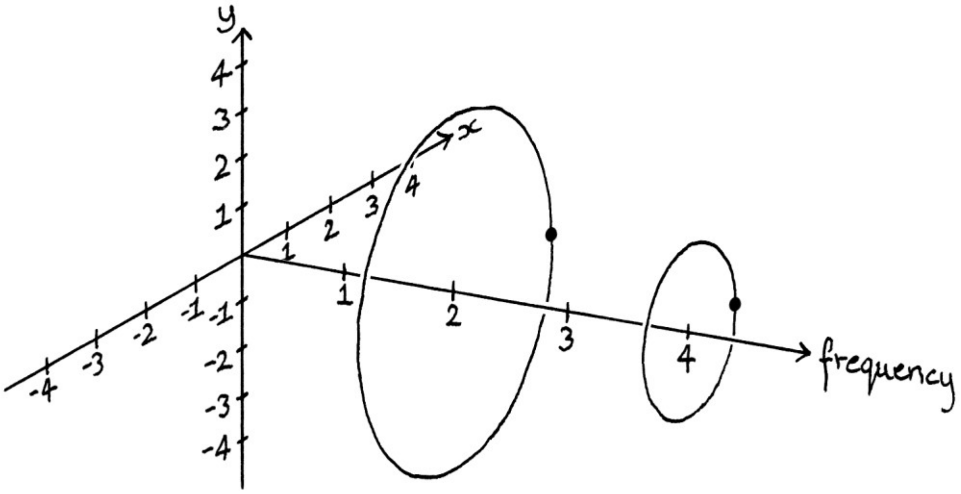
Duplicate frequencies

One unavoidable problem with the frequency domain graph is that it cannot distinguish between two circles that have the same frequency. If we have two circles with the same frequency, they have to be added together and the resulting circle put at that point on the frequency axis. Because of this, the frequency domain really shows the addition of the different frequencies after all identical frequency circles have been added together. It does not necessarily show *all* the original circles.

As an example, we will add the circles that represent the following pairs of waves:
"y = 3 sin (360 * 2t)" and "y = 3 cos (360 * 2t)"
"y = 1 sin (360 * 2t)" and "y = 1 cos (360 * 2t)"
"y = 2 sin (360 * 4t)" and "y = 2 cos (360 * 4t)"

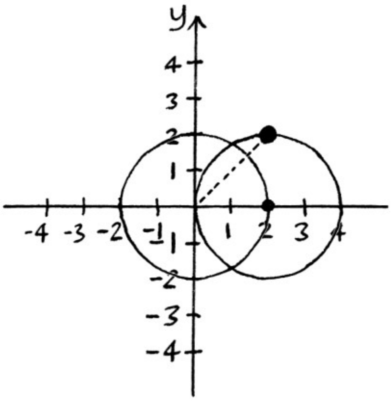
The first two circles have the same frequency, so must be added together before they can be placed on the frequency axis – the frequency axis can only have one circle at any particular frequency. We can tell from the formulas of the first two circles that their sum will be an individual circle that represents the single pair of waves: "y = 4 sin (360 * 2t)" and "y = 4 cos (360 * 2t)"

Therefore, we put the circle that represents the waves "y = 4 sin (360 * 2t)" and "y = 4 cos (360 * 2t)" at the 2 cycles per second point on the frequency axis, and we put the other circle, with a frequency of 4 cycles per second, at the 4 cycles per second point.

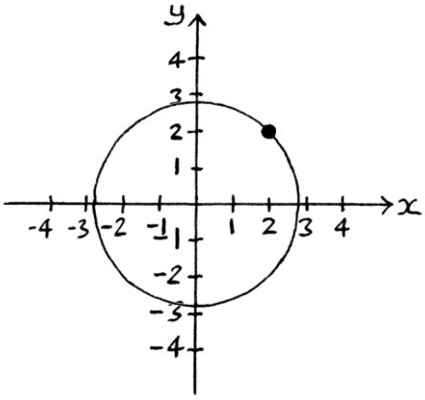


If we have two circles with the same frequency but different phases, again, they must be added together before they can be plotted on to the frequency domain. As an example, we will add the circles that represent the following pairs of waves:
“y = 2 sin (360 * 2t)” and “y = 2 cos (360 * 2t)”
“y = 2 sin ((360 * 2t) + 90)” and “y = 2 cos ((360 * 2t) + 90)”

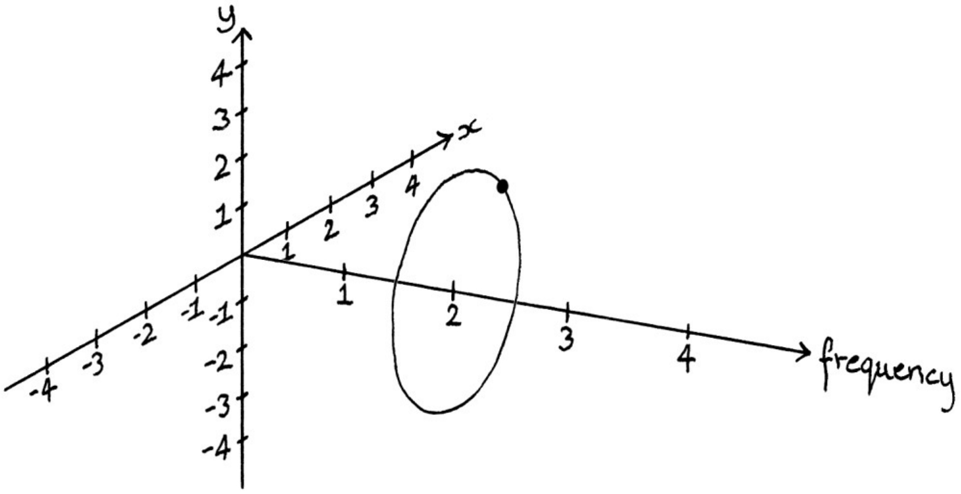
We can calculate what the sum of these two circles will look like by drawing them joined together at t = 0, and paying attention to the angle of the outer object and its distance from the centre of the inner circle:



The outer object is at an angle of 45 degrees from the centre of the inner circle, and is 2.8284 units from the centre of the inner circle. [2.8284 is the square root of the sum of the squares of the y-axis and x-axis position of the outer object, so, in other words, it is the square root of 8]. Therefore, the sum of the two circles will represent the pair of waves:
“y = 2.8284 sin ((360 * 2t) + 45)” and “y = 2.8284 cos ((360 * 2t) + 45)”
... and will look like this:



The resulting circle looks like this on the frequency domain graph:

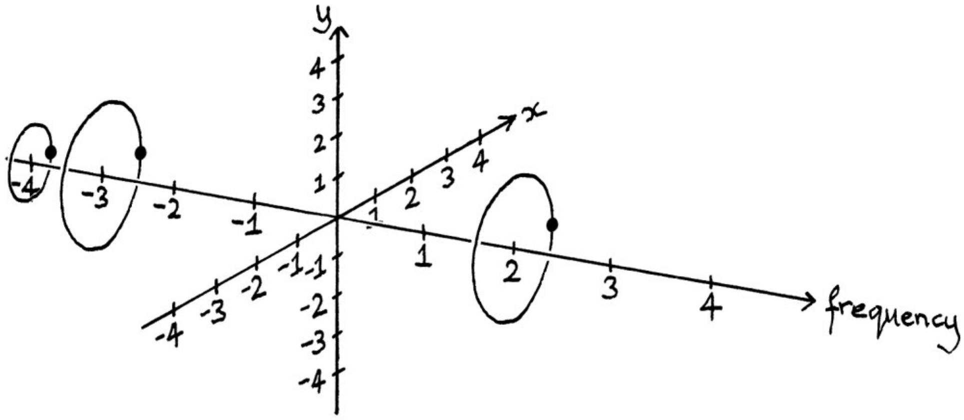


Having duplicate frequencies blended into each other in the frequency domain is not so much of a problem because of how the frequency domain is usually used. The frequency domain is generally used for showing the analysis of received sound waves or radio waves. In those cases, waves of the same frequency would have been blended together before they were received, and any analysis of the signals would be unable to distinguish between separate waves of the same frequency. This idea is related to how I described adding waves of the same frequency as being similar to pouring containers of water into a bucket – by looking into the bucket, we can tell how much water there is, but we cannot tell how many containers were used to fill it. Pouring different containers of unmixable liquids into a bucket will mean that looking into the bucket will allow us to tell how many different liquids there are, and the quantity of each, but it will not tell us how many containers of each unmixable liquid were used to fill it. In the frequency domain, we can tell how much (as in the amplitude) of each frequency there is, but we cannot tell how many circles or waves of that frequency were added together to make up each amplitude.

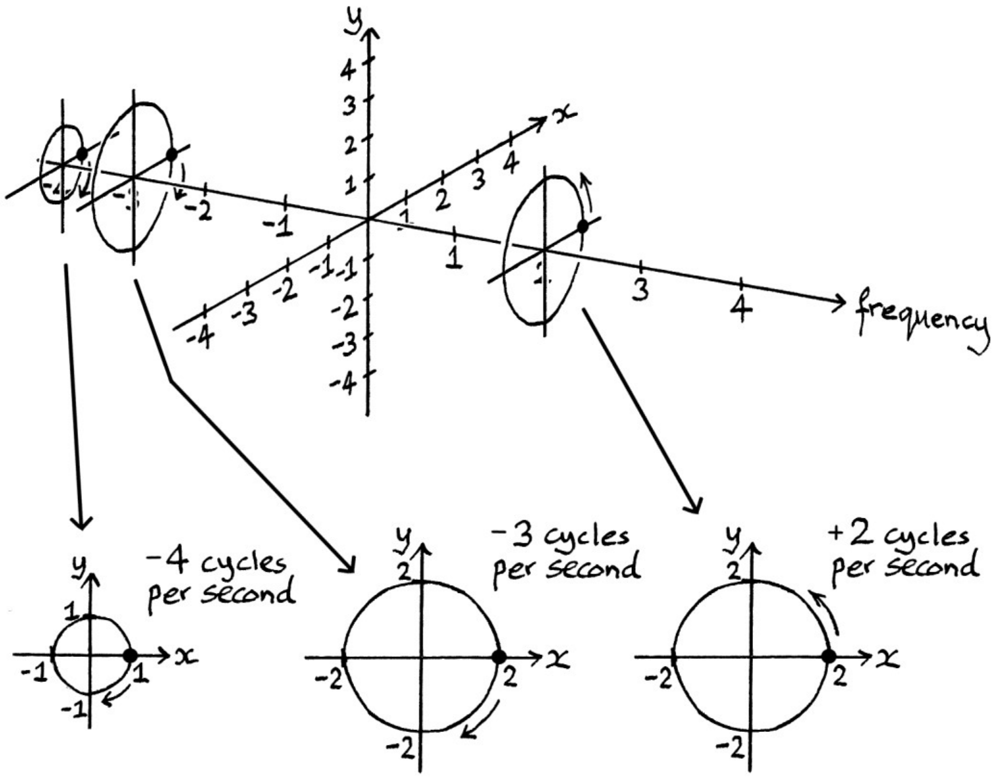
Negative frequencies

We can extend the frequency axis of the frequency domain graph to show negative frequencies too. As an example, we will add the circles that represent the following pairs of waves:
"y = 2 sin (360 * 2t)" and "y = 2 cos (360 * 2t)"
"y = 2 sin (360 * −3t)" and "y = 2 cos (360 * −3t)"
"y = 1 sin (360 * −4t)" and "y = 1 cos (360 * −4t)"

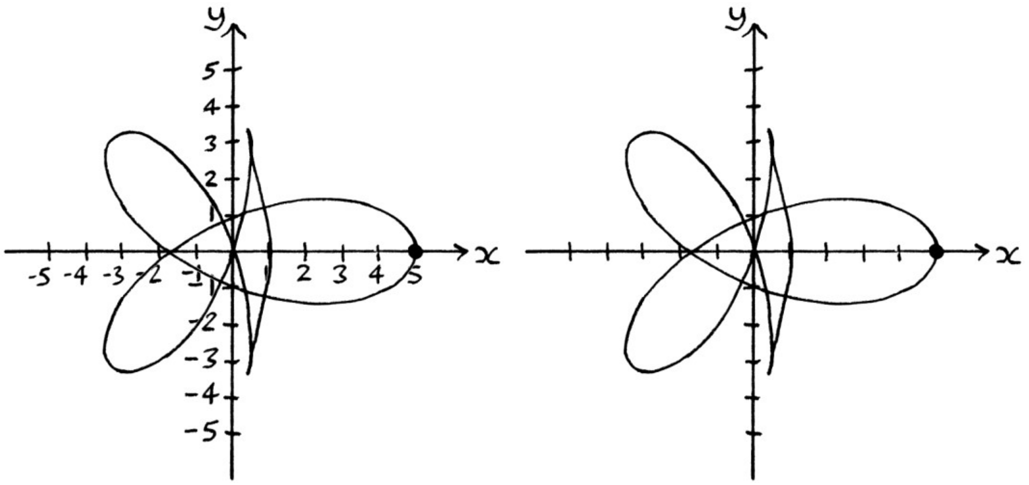
The frequency domain graph looks like this:



The objects rotating around the circles in the left hand half of the graph are rotating clockwise, and so have a negative frequency; the objects rotating around the circles in the right hand half are rotating anticlockwise, and so have a positive frequency. A more descriptive version of this picture is as follows:



The shape resulting from the addition represented in this frequency domain graph looks like this, shown with axis numbering and without:



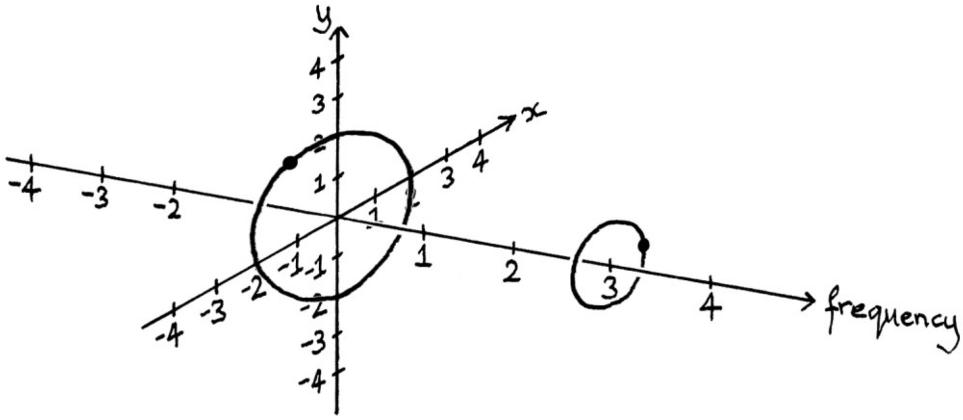
Zero frequencies

The frequency domain graph can show zero frequencies, in which case the circle at zero cycles per second will represent a stationary object on a circle with that radius and phase. In the sum represented on the frequency domain, the zero-frequency circle has the same effect as the mean level "dot" from earlier.

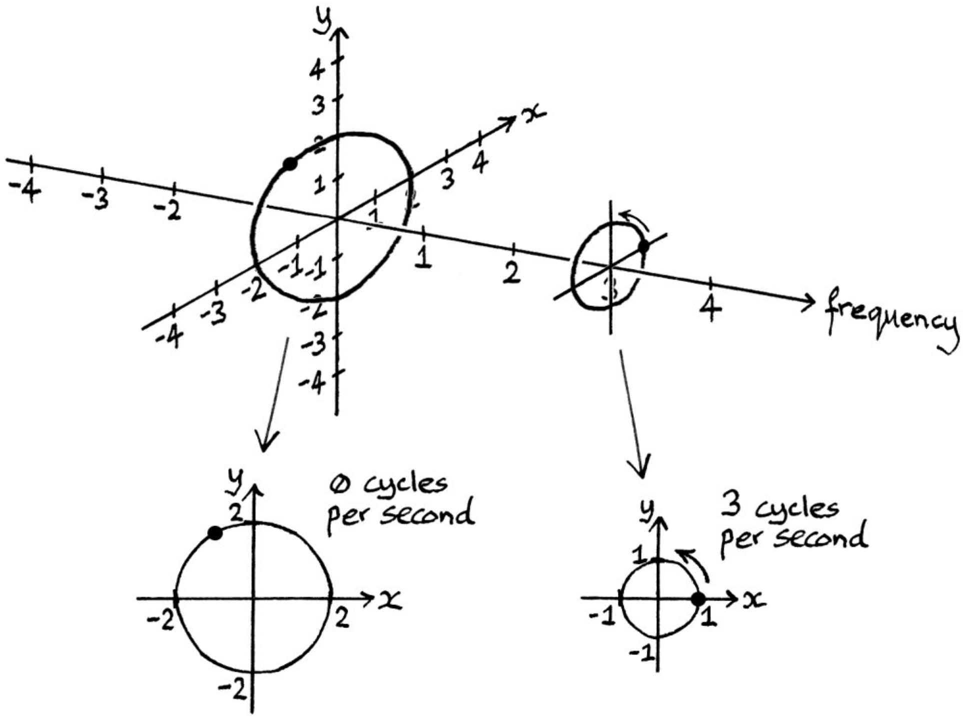
As an example, we will show on the frequency domain graph the circles that represent the following pairs of waves:
"y = 2 sin ((360 * 0t) + 120)" and "y = 2 cos ((360 * 0t) + 120)"
"y = 1 sin (360 * 3t)" and "y = 1 cos (360 * 3t)"

The first circle represents an object that is stationary on a 2-unit radius circle with a phase of 120 degrees. In other words, the object is sat at 120 degrees on this circle's edge for all moments in time. The Sine "wave" derived from the object's position will be a straight line at y = 2 * sin 120 = 1.7321 units. The Cosine "wave" derived from this circle will be a straight line at y = 2 * cos 120 = −1 unit. The effect of the zero-frequency circle in this case would be identical to that of a Mean Level circle with zero amplitude (i.e. a dot) at the coordinates (−1, 1.7321).

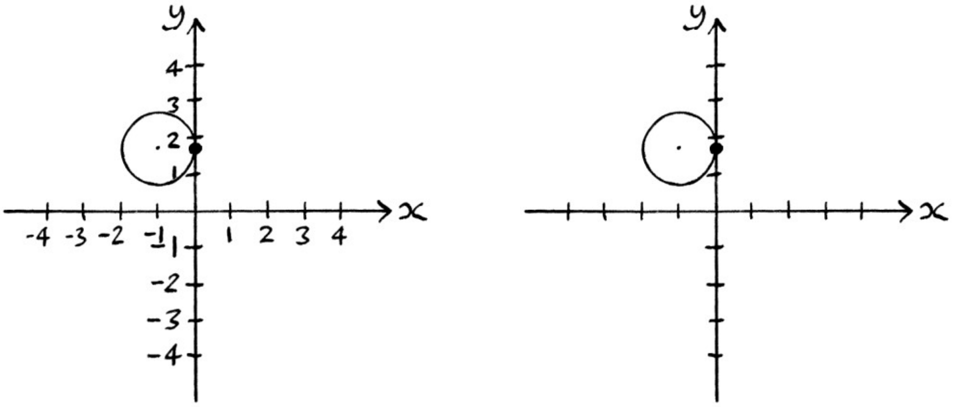
If we ignore the faulty perspective in the picture, the frequency domain graph for both the circles looks like this:



The same graph with more information shown on it looks like this:



The resulting shape looks like this, shown with axis numbering and without:

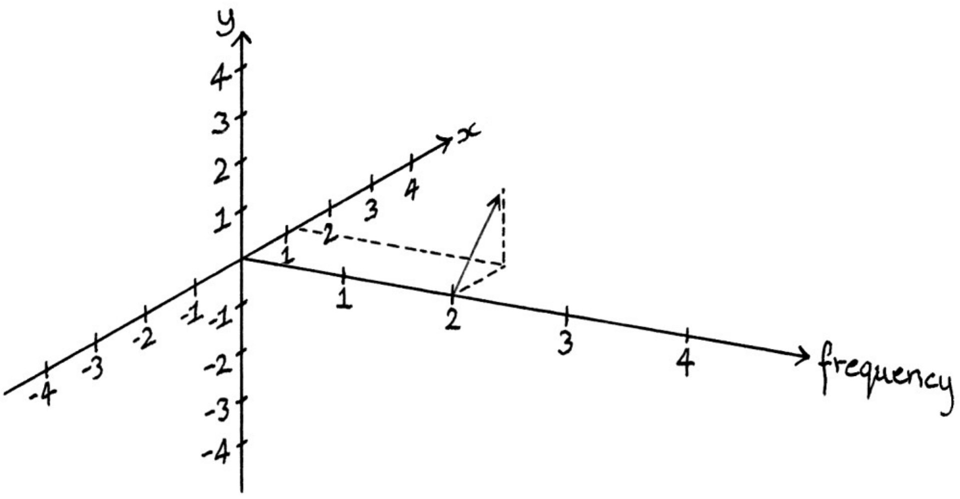


The shape consists of the 3-cycles-per-second circle centred over the phase point of the zero-frequency circle.

Arrows

Instead of drawing circles, we could draw arrows pointing to where the phase points of the circles would be. Such a graph has exactly the same meaning, but it is slightly less intuitive. The arrows are essentially vectors that indicate the phase point and amplitude of the circle.

In this picture, I have added dotted lines to make it slightly easier to see where the arrow is pointing:



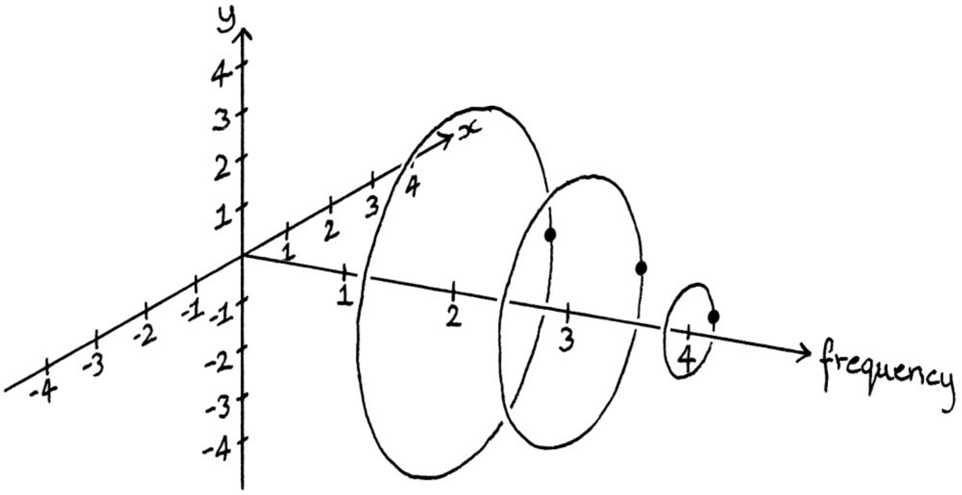
Sums of waves in the frequency domain

In nearly all cases where there is a frequency domain graph, it will not be referring to the addition of circles, but to the addition of either Sine waves or Cosine waves. The graph will not be three-dimensional, but instead two-dimensional.

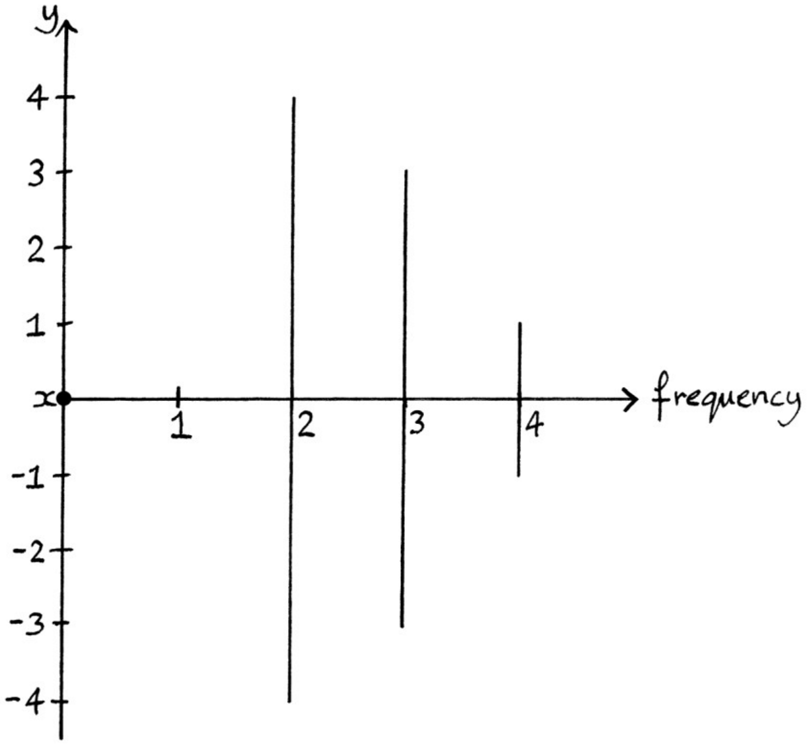
Sums of *Sine* waves in the frequency domain

A Sine wave frequency graph is a graph that shows which Sine waves are being added together to make a signal, ordered by frequency.

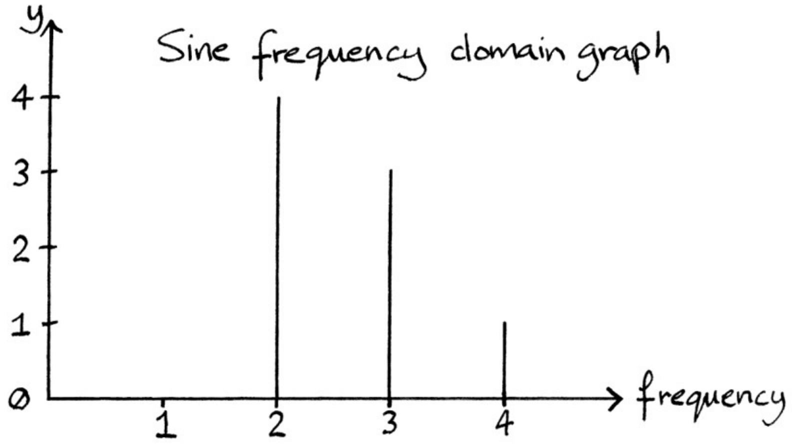
It is not difficult to turn a three-dimensional circle frequency graph into a two-dimensional Sine wave graph. We start with the circle frequency domain graph:



Then we rotate it so that the x-axis is pointing directly away from us. The frequency axis points to the right, and the y-axis still points directly upwards:

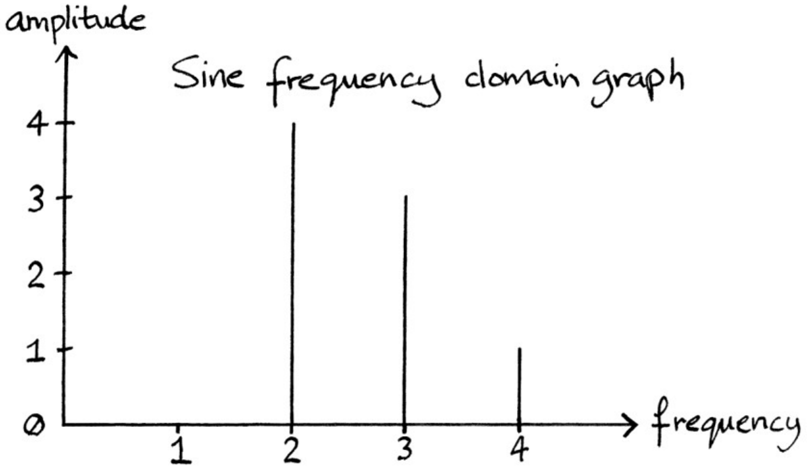


Then, we remove the parts of the circles (which are now side on) that are below y = 0. We end up with this:



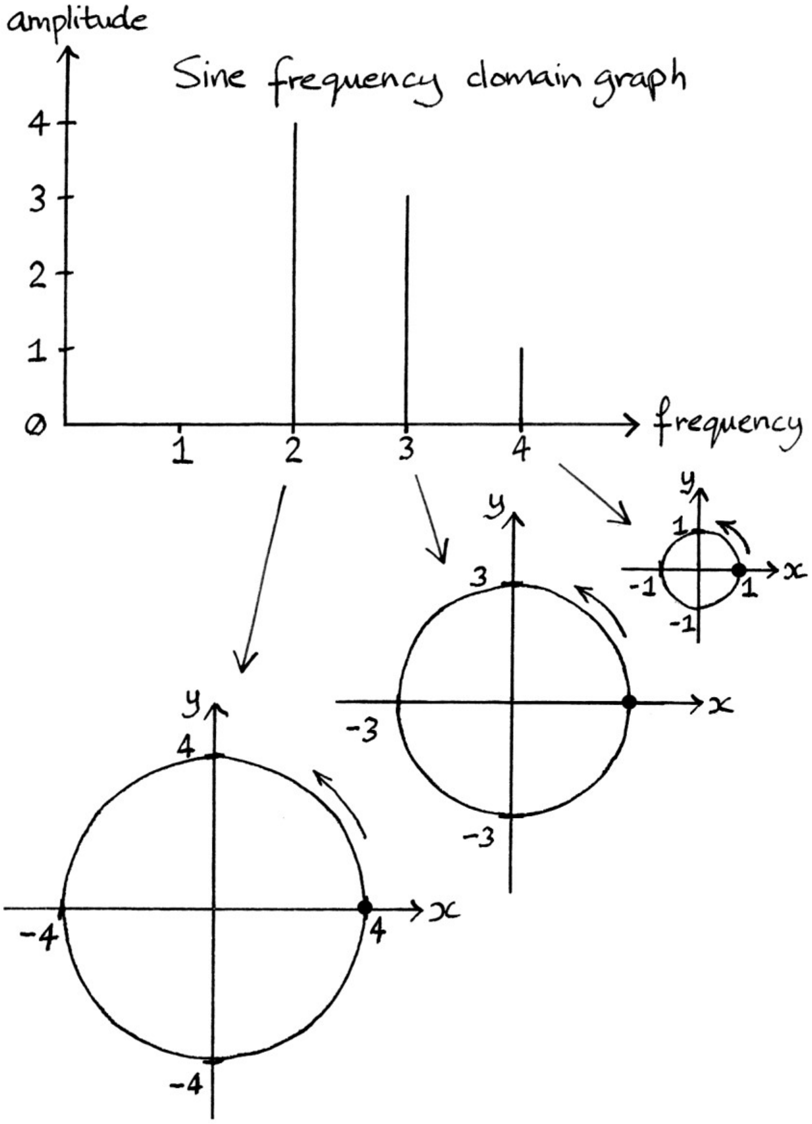
We are now looking at the top half of each circle viewed side on.

Each vertical line represents the *amplitude* of the wave for that frequency. Note that it does not represent the height of the phase point, unless the phase happened to be 90 degrees. As we are no longer dealing with circles, for which the y-axis represented the radius (which was equal to the amplitude), the y-axis now represents just the amplitude of the waves. We can therefore call the y-axis, the "amplitude axis":

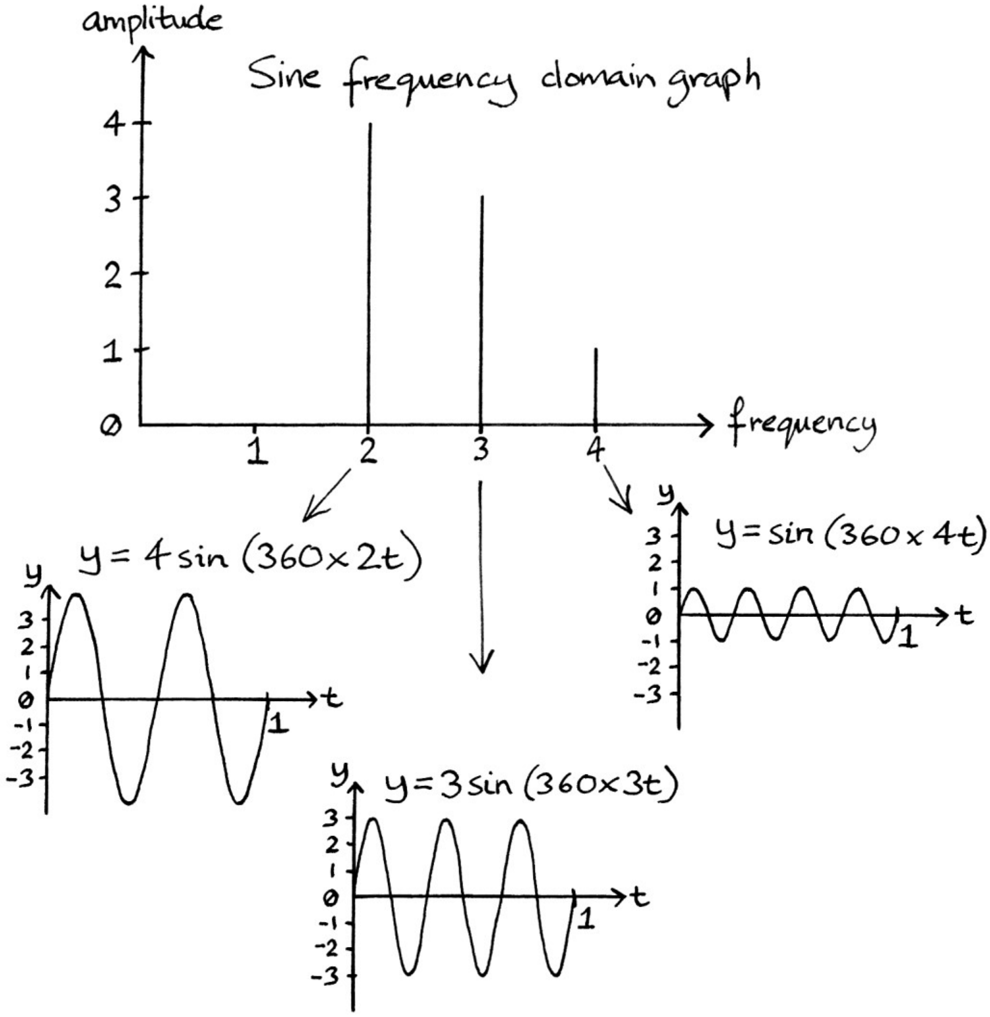


This graph is easier to read more accurately than the three-dimensional circle frequency graph, but it contains less information. For one thing, we can only see the heights of the circles and not the widths, so we cannot tell if the corresponding Cosine waves have a mean level or not. Another thing is that we cannot see what the phase is. It is a big compromise, but in everyday signal processing with sound or radio waves, this is not considered a problem because sound and radio waves are received, not as circles, but as summed waves. Of course, there are types of waves other than sound and radio waves, but for most of the time, such a graph will be sufficient as a rough guide. It is important to realise that because we cannot see the phase of each wave, there is a (probably wrong) assumption that all phases are zero. Therefore, the addition of all the waves shown on the graph might not result in exactly the correct signal.

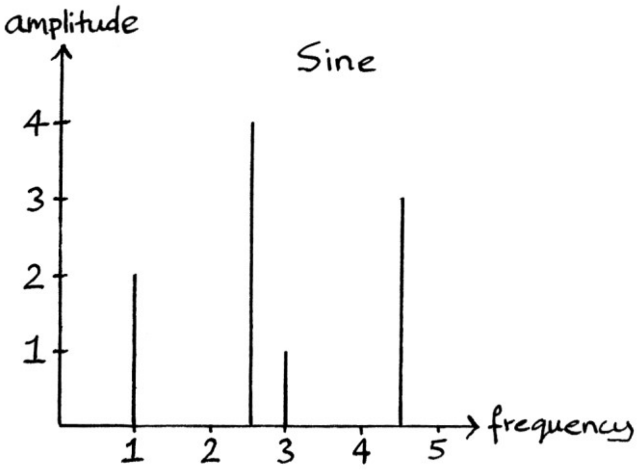
Here is the same graph with notes comparing the lines with circles:



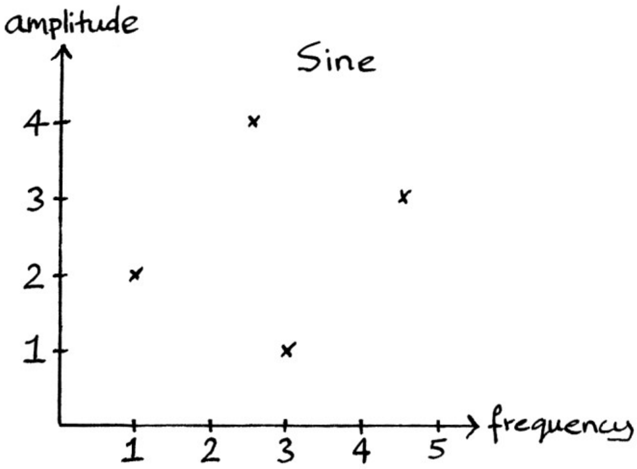
Here is the same graph with notes comparing the lines with the waves represented by those lines:



If we were adding the following *waves*:
“y = 2 sin (360t)”
“y = 4 sin (360 * 2.5t)”
“y = 1 sin (360 * 3t)”
“y = 3 sin (360 * 4.5t)”
... then we could portray the addition with this Sine wave frequency domain graph:



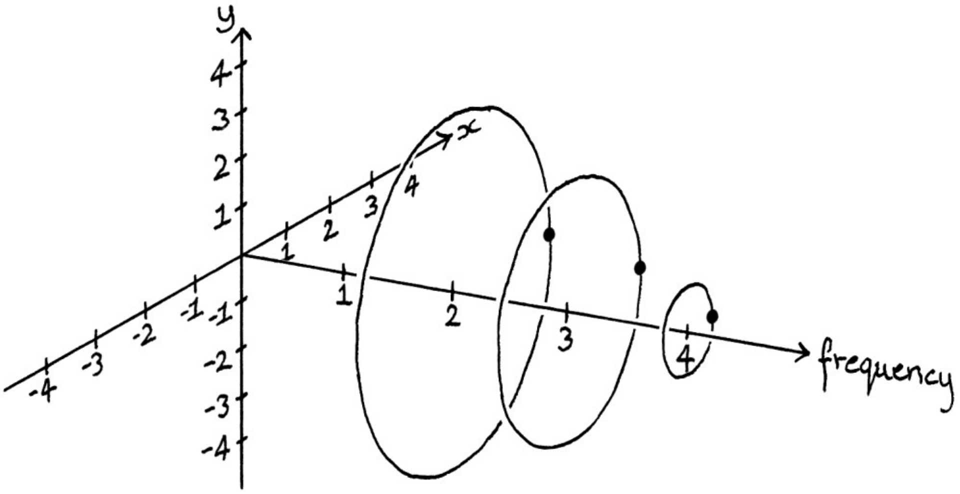
There are two ways of drawing a wave-based frequency domain graph. Either we can use vertical lines to indicate the amplitudes, as used in the picture above, or we can just use points or crosses to show where the top of that line will be. Both methods give the same amount of information. Here is the same graph with crosses:



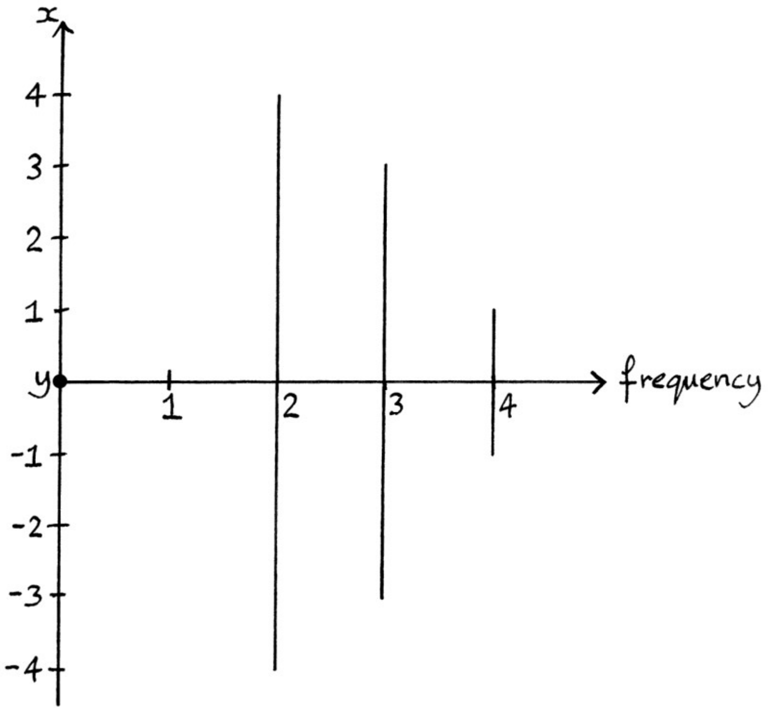
Remember that the lines and crosses represent the *amplitude* of each wave, which is also the radius of the circle from which that wave was, or could have been, derived. They do not represent the height of the phase point on the circle. This type of graph cannot indicate phase, so there is the assumption that all phases are zero or irrelevant.

Sums of *Cosine* waves in the frequency domain

As well as having a frequency graph showing the addition of Sine waves, we can also have one that shows the addition of Cosine waves. Starting with the circle frequency graph:

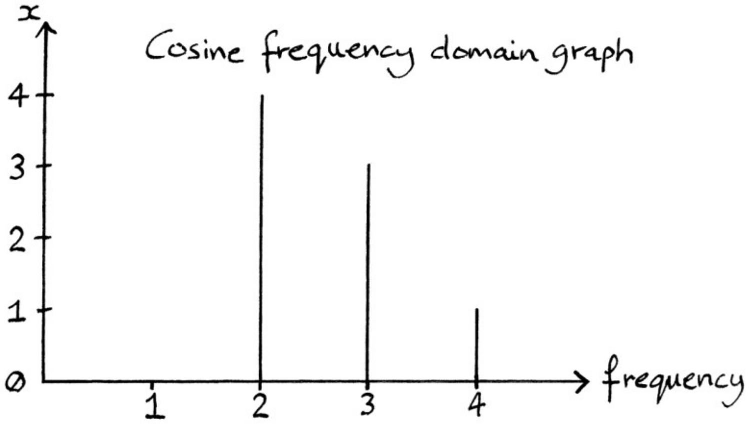


... we rotate and turn it so that the y-axis is pointing directly at us. The x-axis points upwards, and the frequency axis points to the right.

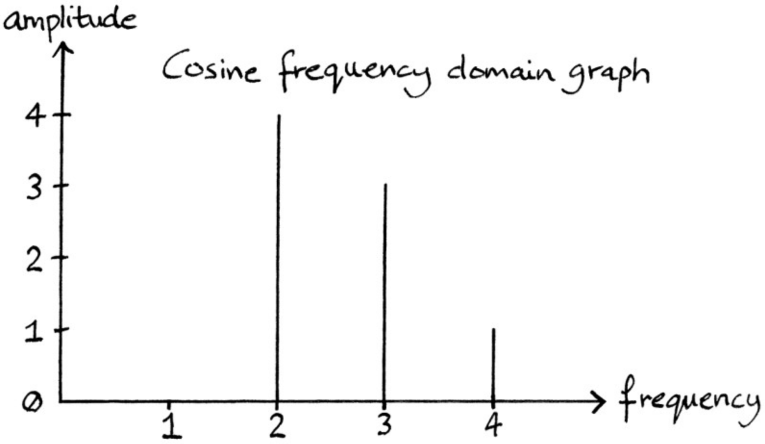


The view is the same as if we were looking at the circle frequency domain graph from the top.

We then remove the parts of the side-on circles that are below the frequency axis (which, because the graph is a rotated version of the circle chart, is the area underneath x = 0). We end up with this:

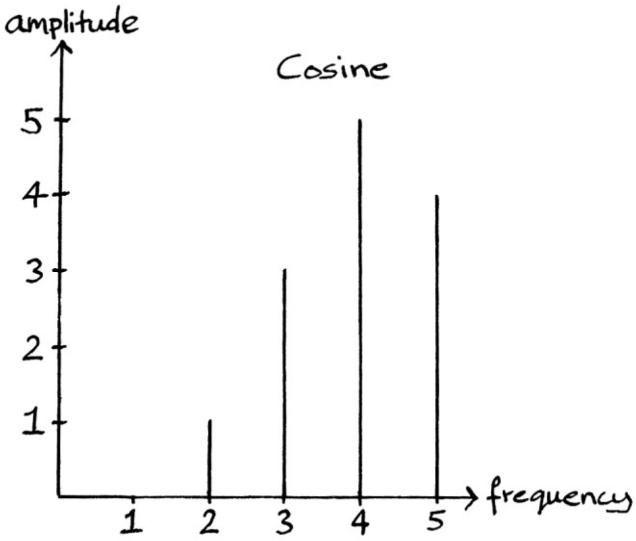


Each vertical line now represents the *amplitude* of the Cosine wave for that frequency. As we are no longer dealing with circles, for which the x-axis represented the radius (which was equal to the amplitude), the x-axis (as in what was the x-axis before we rotated everything) now represents just the amplitude of the waves, so we can rename it to be the amplitude axis:



As with Sine, the Cosine frequency graph only shows amplitude – it does not show phase. Therefore, the sum of the Cosine waves from the graph may or may not be exactly correct. Another point is that it is not possible to know if the corresponding Sine waves have a mean level or not.

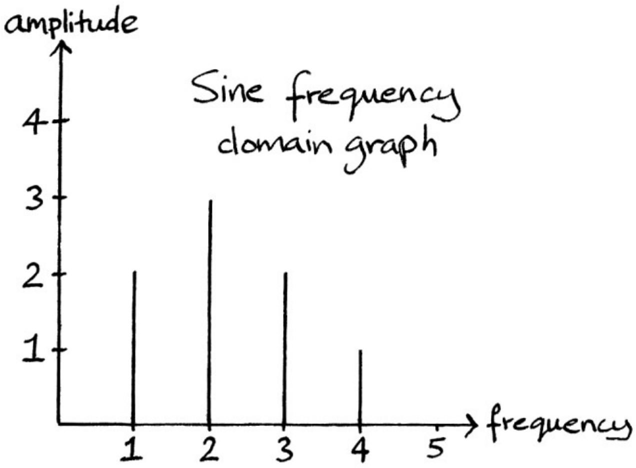
If we were adding the following waves:
“y = 1 cos (360 * 2t)”
“y = 3 cos (360 * 3t)”
“y = 5 cos (360 * 4t)”
“y = 4 cos (360 * 5t)”
... we could portray the addition with a Cosine wave frequency graph as so:



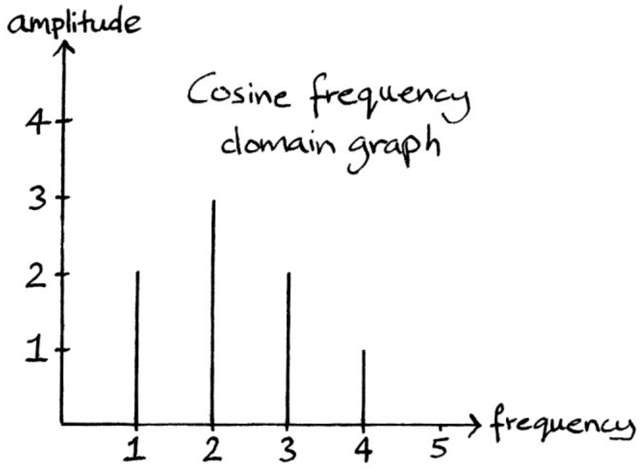
Sine and Cosine

If we have a frequency domain graph for the sum of Sine waves, then the graph for the corresponding Cosine waves will look identical. Similarly, if we have a frequency domain graph for the sum of Cosine waves, then the graph for the corresponding Sine waves will look identical.

As an example, this Sine wave frequency domain graph:



... implies the layout of the corresponding Cosine wave frequency domain graph:

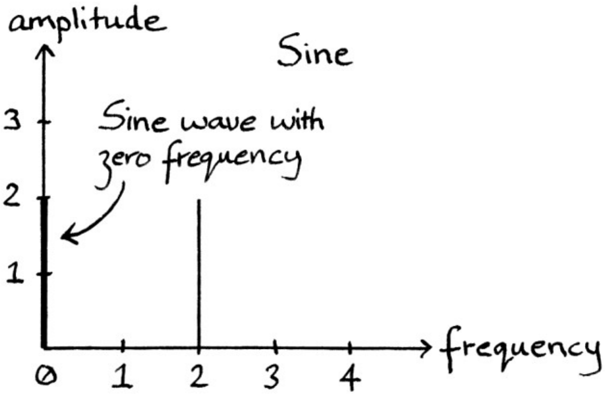


The way the two graphs imply each other might be obvious from how the graphs are showing the amplitudes and frequencies of the circles, so the amplitudes and frequencies of the derived pairs of waves will be the same.

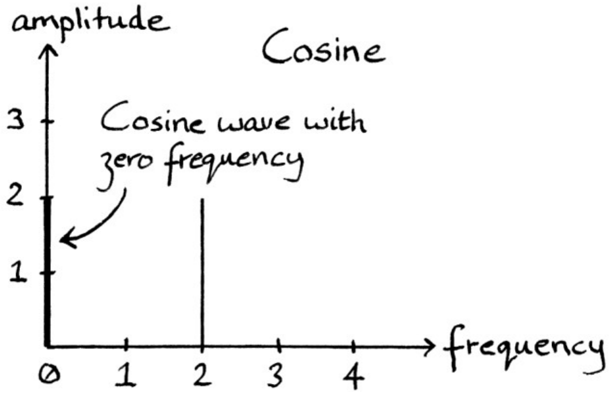
Zero frequencies for Sine and Cosine

Because the standard wave frequency domain graph cannot show phase, one has to assume that all waves have zero phase. On a Sine wave frequency graph, this means that if there is a wave with a frequency of zero, then that wave can only represent a “wave” where y = 0 for all time. On a Cosine wave frequency graph, a wave with a frequency of zero can only represent a “wave” where “y” is the amplitude of the wave for all time. This also means that if we have a Sine wave frequency domain graph with a wave at zero cycles per second, it may as well not be there, as it will have no effect whatsoever on the sum.

In the following Sine wave frequency graph, there is a Sine wave with a frequency of zero cycles per second and an amplitude of 2 units. As we cannot show phase on the graph, the Sine wave must be presumed to have a phase of zero degrees. An object moving around the circle from which such a wave is, or could be said to be, derived would be stationary at 0 degrees for all time. There is also a Sine wave with a frequency of 2 cycles per second with an amplitude of 2 units (and again, a presumed phase of zero degrees). If we were to add the two waves together, the zero-frequency wave would have no effect on the sum.



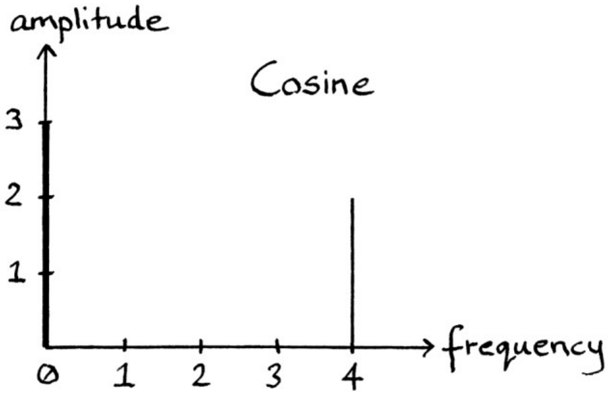
In the following Cosine wave frequency graph, we have the corresponding Cosine waves to the Sine waves from the last example. As in the last example, these must be presumed to have zero phase, as such a graph cannot indicate phase. Therefore, the zero-frequency Cosine wave is, or could be said to be, derived from a circle where the object is stationary at 0 degrees for all time. Because it is a Cosine wave, this means that the object will be at y = 2 for all time. Therefore, when we add the two waves in the graph together, we end up with the second wave added to the number 2. The zero-frequency wave has the effect of a 2-unit mean level.



Mean levels for Sine and Cosine

The following repeats what I have just said, but from the point of view of portraying mean levels. Mean level on a *Cosine* wave frequency graph can be shown by having a zero-frequency wave entry with an amplitude of the required mean level (and an implied zero phase). If the other actual waves had had mean levels, the mean levels can all be grouped together to end up as the amplitude of this zero-frequency Cosine wave.

For example, for the wave “y = 3 + 2 cos (360 * 4t)”, we would put the wave part at 4 cycles per second, and we would put the mean level part as a line with an amplitude of 3 units, as so:

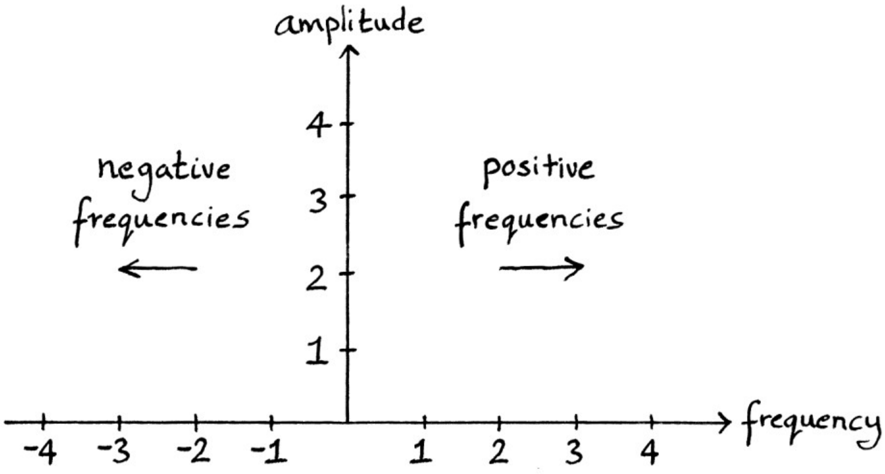


A Cosine “wave” with zero frequency and zero phase results in a “wave” graph that is just a straight line at that amplitude. Remember that the lines on the wave frequency domain graphs represent *amplitude* and not the *phase point* of the circle. However, for a Cosine wave with zero phase, the amplitude will actually be equal to the x-axis position of the phase point on the circle, so the amplitude will be the same as the phase point. Therefore, for Cosine waves, a zero-frequency wave in the sum of waves with a particular amplitude will act as a mean level of that amplitude.

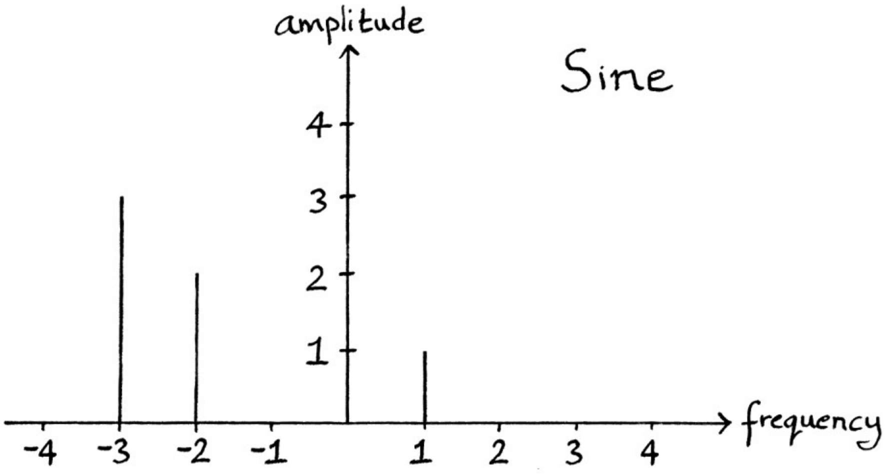
As we saw before, there is a problem in doing the same thing for a *Sine* wave because a Sine wave with zero frequency and zero phase results in a “wave” graph that is a straight line at y = 0. Its value is zero, no matter what the amplitude. Therefore, in the sum of waves, it would have no effect whatsoever. Therefore, if we want to portray a mean level on the wave frequency domain graph, we cannot do it with Sine waves, and we would have to come up with a more contrived and less useful method. [Somehow including a phase of 90 degrees for the wave at 0 cycles per second would work]. This is one of several reasons that most of the time, whenever you see a wave frequency graph, it will be a Cosine wave frequency graph. It is also one of several reasons that, generally, a received radio or sound signal is treated as the sum of Cosine waves (with zero phases), and not the sum of Sine waves (with zero phases).

Negative frequencies for Sine and Cosine

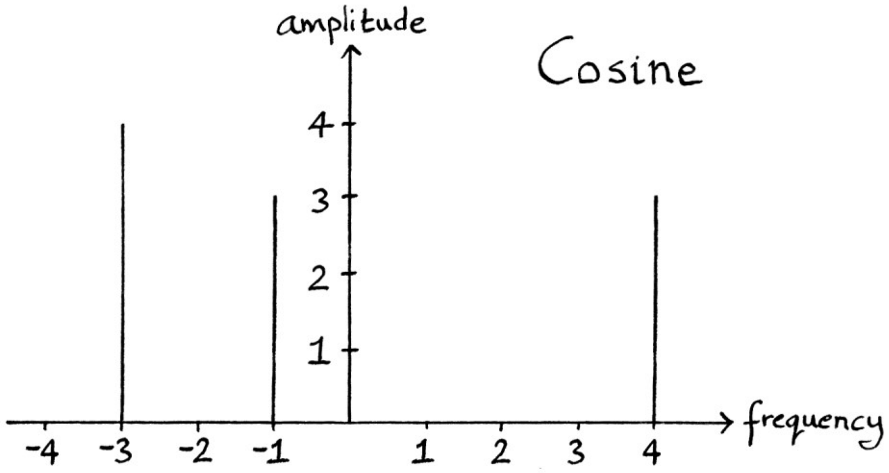
It is possible to extend the wave frequency domain graphs to show negative-frequency waves.



For example, if we were adding the following waves:
“y = 1 sin (360t)”
“y = 2 sin (360 * −2t)”
“y = 3 sin (360 * −3t)”
... then the addition could be summarised by this frequency domain graph:



If we were adding the following waves:
“y = 3 cos (360 * −1t)”
“y = 4 cos (360 * −3t)”
“y = 3 cos (360 * 4t)”
... then the addition could be summarised by this frequency domain graph:



These extended wave graphs raise important ideas about the difference between frequency domain graphs showing waves and those showing circles.

First, a Sine wave with a negative-frequency formula can also be portrayed with a positive-frequency formula with a different phase [As explained in Chapter 11]. Therefore, if phase is irrelevant to whoever constructed the graph, which it might be because the wave frequency domain graphs cannot show phase, then a negative-frequency Sine wave could just as easily be placed in the positive half of the frequency axis. For example, “y = 2 sin (360 * −2t)” is the same as “y = 2 sin ((360 * 2t) + 180), so it can be placed at +2 cycles per second on the frequency axis, in which case the +180 degree phase is forgotten about. Therefore, if you really do not care about phase, which is implied by the use of such a frequency domain graph, then it is debatable as to whether you really need a negative-frequency half of the frequency axis for Sine waves.

Second, a negative-frequency Cosine wave *with zero phase* is identical to a positive-frequency Cosine wave *with zero phase*. They are the same thing. Therefore, a negative-frequency Cosine wave *with zero phase* can be placed in the positive half of the frequency axis, and in that case, we definitely do not need the negative half of the frequency axis for Cosine waves *with zero phases*. [If there is already a Cosine wave with the positive frequency of the negative-frequency Cosine wave, we add the amplitudes together, and put the result in the positive half]. For Cosine waves with non-zero phases, we end up in the same situation as that for Sine waves.

The above example of adding:
“y = 3 cos (360 * −1t)”
“y = 4 cos (360 * −3t)”
“y = 3 cos (360 * 4t)”
... is identical to adding:
“y = 3 cos (360 * 1t)”
“y = 4 cos (360 * 3t)”
“y = 3 cos (360 * 4t)”

When the wave frequency domain graphs are used to show the addition of known waves, and phase is being ignored, it does not really make sense to have the negative frequency half of the graph. You will, however, still see the negative frequency half used in practice, as will be seen when I explain different types of graphs later in this chapter.

The negative half does have a use for showing phase if we are dealing with Sine waves with phases of either 0 or 180 degrees, and we want to show that. We put Sine waves with a phase of 0 degrees in the right half as usual, and, because a phase of 180 degrees is the same as a negative-frequency Sine wave with a phase of zero degrees, we put the Sine waves with 180-degree phases in the left half.

Graph variations: y-axis

When you see wave frequency domain graphs in practice, you will encounter different types of values on the y-axis, and even different types of y-axis.

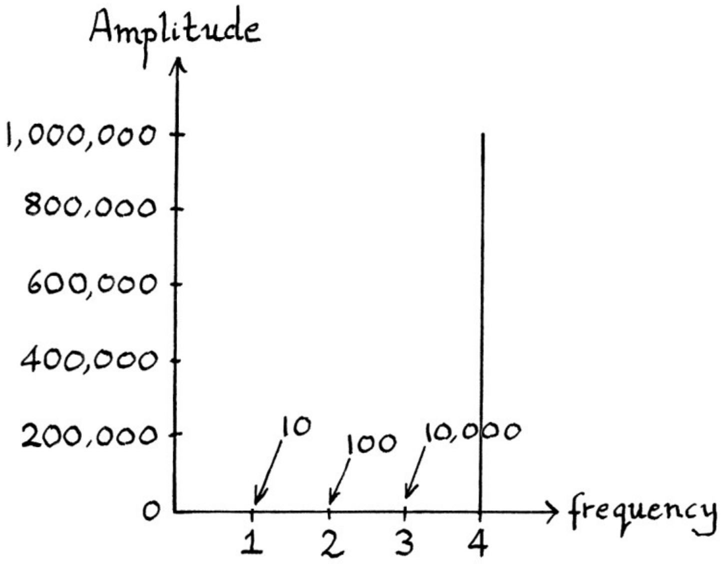
Alternatives to amplitude

The amplitude of a wave on a frequency domain graph will most likely be measured in a form dependent on the way it is created or received, and not as generic “units”. For example, a sound starts as changes in air pressure, and it might be observed via a microphone as changes in electrical current. Generally, amplitude might be measured in many units, including amps, volts, pascals and so on.

An alternative to amplitude on the y-axis is “power”. Power is the amount of energy transferred over a set time, and is measured in watts. When it comes to waves, power is the square of the amplitude. To be more specific, the overall power of a wave is the square of the overall amplitude; the instantaneous power of a wave is the square of the instantaneous amplitude at that moment in time. The overall power is also called the “peak power”. The differences in power of various waves will be proportional to the difference in the areas of the circles from which those waves are, or could be said to be, derived. Power is the square of the amplitude; the area of a circle is the square of the radius (which is the same as the amplitude) multiplied by π.

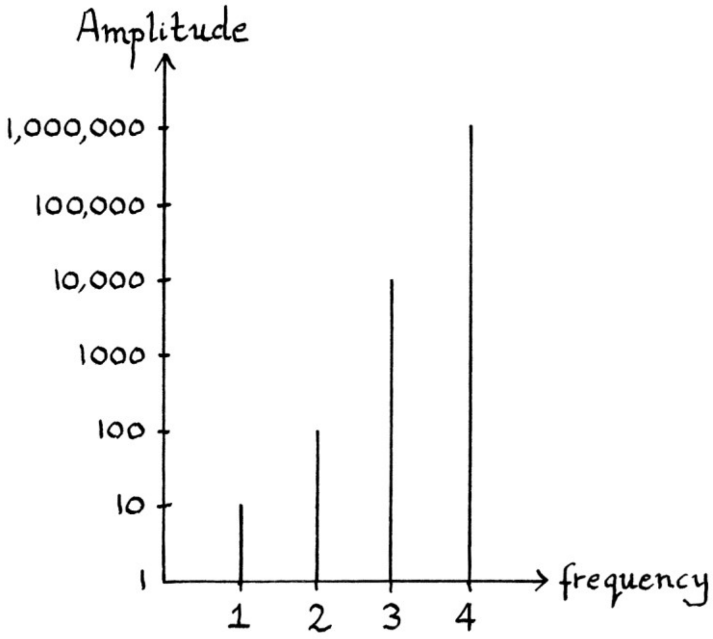
Variations in scale

Another way of showing the y-axis on the wave frequency domain graph is to use a “logarithmic scale”. [The name “logarithmic” makes what is essentially a simple concept sound much more complicated than it really is]. Imagine that the amplitudes of some waves shown on a frequency domain graph range from 10 up to 1,000,000 units, and that we are equally interested in all the values. The y-axis of that graph will have to be extremely long for us to be able to see all the values in any details. It is unlikely we would be able to draw such a graph on a normally sized piece of paper. Such a graph drawn to the normal size of graphs in this book would look like this:



If it were not for the arrows indicating the values of the amplitudes, it would be impossible to read three of the values on this graph because they are so small in comparison to the 1,000,000-unit amplitude.

One solution to this is to squash the y-axis in a way that the numbering does not increase at a constant rate, but instead at an ever-increasing rate. Such a y-axis is called a "logarithmic scale". The information from the above graph, plotted with a logarithmic scale for the y-axis, looks like this:



It is now easy to see the value of the small amplitudes and the largest amplitude all at the same time. The downside is that is harder to see how the actual values of the amplitudes compare with each other without some thought.

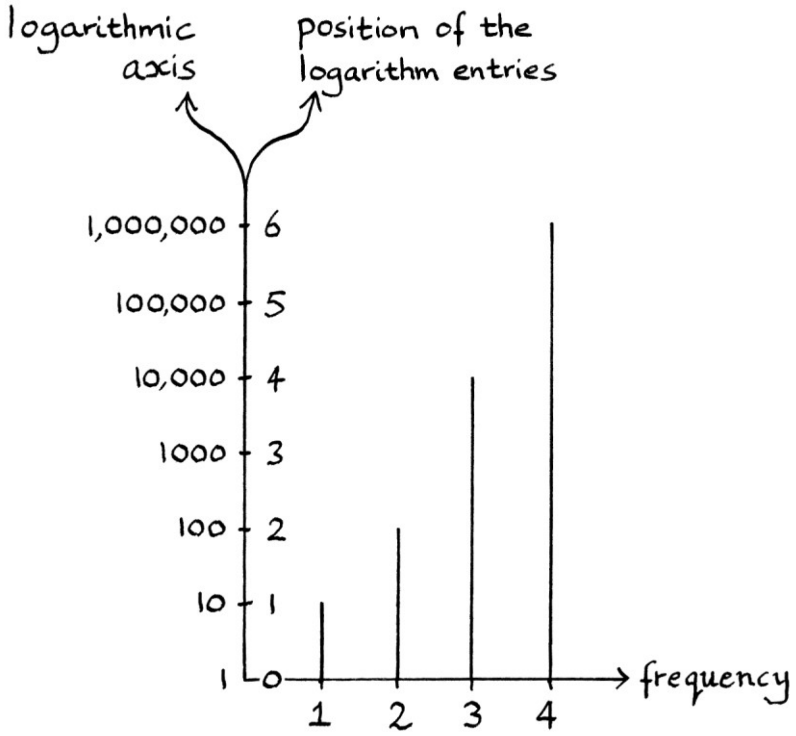
Each labelled point on the y-axis scale is ten times the size of the one before it. The rules for how the scale progresses are as follows:

- The first entry on the scale is 1, which is 10 to the power of 0.
- The second entry on the scale is 10, which is 10 to the power of 1.
- The third entry on the scale is 100, which is 10^2.
- The fourth entry on the scale is 1000, which is 10^3.
- The fifth entry on the scale is 10,000, which is 10^4.
- The sixth entry on the scale is 100,000, which is 10^5.
- The seventh entry on the scale is 1,000,000, which is 10^6.

Note how the scale does not start at 0. This is because all the entries on the y-axis scale are based on 10 raised to a power, and there is no number to which 10 can be raised that results in 0. If we need to, however, we can get closer to the number zero by having entries such as 10^{-1}, which is 0.1, 10^{-2}, which is 0.01, and so on.

Although the scale goes from 1 to 1,000,000, and increases at ever-increasing steps, it helps to notice that the labelled scale *entries* are evenly spaced.

To make the logarithmic axis more self-explanatory, we could have a parallel axis that indicates the position of the labelled entries. We will call this the "position-of-the-logarithm-entries" axis, and we will call the amplitude axis, the "logarithmic axis". Note that I am only naming these axes in this way to explain the idea, and normally, there would not be two parallel axes, and they would not be so-named.



Doing this, or more realistically, thinking in this way, lets us easily calculate where amplitudes should be plotted on the y-axis. To calculate where a particular amplitude should be placed on the axis, we need to find the value to which the number 10 must be raised to produce that value. In other words, given an amplitude, we have to find out what value of "x" in the calculation "10^x" results in that amplitude.

To put this slightly more mathematically:
10^x = amplitude
... where "x" will be the place on the "position-of-the-logarithm-entries" axis that the amplitude should be marked.

The way to find the result is to use logarithms. [I will explain logarithms in Chapter 26, and it does not matter if this does not make sense yet.] Logarithms find the value to which a number must be raised to be equal to a given value. As we are dealing with powers of the number 10, we want to find the value to which 10 must be raised to be equal to our given amplitude. When it comes to logarithms in this case, we will be use the base 10 logarithm because we are using powers of 10. On a calculator, this will be the logarithm button, or $\log_{10}$ button.

The formula "10^x = amplitude" can be rephrased as:
x = $\log_{10}$ (amplitude)
In fact, "$\log_{10}$ (amplitude)" is the inverse of "$10^{\text{amplitude}}$".

"$\log_{10}$ (amplitude)" can be solved on a calculator by typing in the amplitude and pressing the logarithm or $\log_{10}$ button.

If we had an amplitude of 10 units, we would want to find "x" in the calculation:
$10^x = 10$. We can calculate "x" in this case without a calculator, but using one anyway, we would use $\log_{10}$ (10), which would produce the result 1. The value of 10 units is, therefore, placed at 1 on the "position-of-the-logarithm-entries" axis, which is at 10 on the logarithmic axis.

If we have an amplitude of 5, we need to find "x" in this calculation: $10^x = 5$, which is the same as finding $\log_{10}$ (5). A calculator will give the result 0.6990. This means that $10^{0.6990} = 5$. Therefore, the value of 5 units is placed at 0.6990 on the "position-of-the-logarithm-entries" axis, which is 0.6990 of the way between 0 and the first logarithmic axis label. Note how this is not halfway between 0 and 10 on the logarithmic axis – the logarithmic axis does not increase at a constant rate.

If we have an amplitude of 2000 units, we would need to find "x" in this calculation: $10^x = 2000$, which is the same as finding $\log_{10}$ (2000). A calculator will give the result as 3.3010. Therefore, $10^{3.3010} = 2000$. The value of 2000 units is placed at the position of 3.3010 on the position axis, which is about a third of the way between 1,000 and 10,000 on the logarithmic axis.

There are other ways to lay out a logarithmic scale instead of using powers of ten. For example, we could use powers of two, in which case the axis numbering would proceed as: 1, 2, 4, 8, 16, 32, 64, 128 and so on. [If we did that, we would use $\log_2$ on a calculator]. We could use powers of any number, but powers of ten are the easiest to understand, and in practice, these are the most common.

[If you do not understand logarithms at this point, it really does not matter, as long as you can see that the y-axis scale is squashed.]

Graph variations: decibels

As well as seeing the y-axis as a logarithmic scale, you will often see the y-axis showing "decibels", which are based on a similar idea.

The concept of decibels is often badly explained, badly understood, and badly used. I only want to introduce the idea of decibels to indicate how they work on a graph, which means all that might be true here, too. Decibels are less difficult to understand than they appear at first glance, but on the other hand, there is much more to know about them than one might think. In this explanation, I will introduce all the characteristics of decibels gradually to make them easier to digest.

Bels

A decibel is a tenth of a bel. This is similar to how a decimetre is a tenth of a metre. However, generally the unit bel is not used, as it is more practical to use a decibel. This section could be called "Graph variations: bels", but the term "decibel" has become so common that calling it "Graph variations: decibels" is more apt. I will still make this section generally about bels, as it makes things a lot easier to understand.

Bels are a "comparing unit" – they do not refer to any measurable quantity, but instead, they indicate the size of a value in terms of its ratio to another value. To make things more complicated, they do this with a logarithmic scale. A bel is the base ten logarithm of the ratio between a given value and the value with which it is being compared. To put this a different way, a bel is the value to which 10 must be raised to produce a value equal to a particular ratio.

Strictly speaking, when we looked at the logarithmic scale before, it was indicating the size of values with reference to how they compared with the value 1. Bels are basically doing the same thing, but can compare against other reference values.

When using bels, there are generally accepted reference values against which given values are compared, and they depend on the subject being discussed. If we are looking at voltages, then we might compare against the value of 1 volt. If we are looking at power in watts, we might compare against 1 milliwatt. If we are looking at sound *pressure*, we might compare against the quietest sound that people can generally hear (20 micropascals). If we are looking at sound *intensity*, we might compare against 10^{-12} watts per square metre. A bel does not have to indicate a comparison with one of the generally accepted reference points – it can indicate a comparison with any desired value. However, it is always important to mention what that value is, or else a bel becomes meaningless. Not saying to what comparison a bel is based on is like saying, "I am 20% more...," and not finishing the sentence. Bels and decibels are frequently seen without mentioning their context, which is one of the countless reasons they can be confusing or ambiguous. This mistake has led many people to think, wrongly, that decibels are a unit that solely measures the "loudness" of sound.

When thinking of tenths of a bel, or in other words, decibels, the word "decibel" is abbreviated to "dB", and often that is suffixed with one or more letters to indicate the type of values being discussed. For example, "dBv" refers to volts (or more specifically, decibels comparing against 1 volt), and "dBm" refers to milliwatts (or more specifically, decibels comparing against 1 milliwatt). The abbreviation "dB SPL" refers to sound pressure (or more specifically, decibels comparing against a sound pressure of 20 micropascals). The abbreviation "dB SIL" refers to sound intensity (or more specifically, decibels comparing against a sound intensity of 10^{-12} watts per square metre).

Another important thing to know about bels is that they deal solely with comparisons of *power*. If a value is not one that refers to power, then it must be squared before it can become described using bels. Given that, it is necessary to categorise the type of value being considered before using bels. There are two types of value:

Type 1: These types are called "Root Power Quantities" or "Field Quantities". We can think of these types as being related to *amplitude*. Voltage, current and sound *pressure* belong in this group.

Type 2: These types are called "Power Quantities". We can think of these types as being related to *power*, or to put it differently, *amplitude squared*. Light intensity, sound *intensity*, and values measured in watts belong in this group.

[Note that I am deciding to call them *Type 1* "Root Power" and *Type 2* "Power" quantities to make them easier to remember and distinguish. Usually, they would just be called "Root Power" (or "Field") and "Power" quantities, without the prefixes "Type 1" and "Type 2".]

As I said earlier, the “overall power” or “peak power” of a wave is the square of the amplitude of a wave. The two different types of value here are consistent with that idea. We can think of Type 1 as the “amplitude type”, and Type 2 as the “amplitude-squared type”.

To use a Type 1 “Root power” value with bels, it and the value to which it is being compared must first be squared to become Type 2 “Power” values.

Type 2 “Power” formula

To convert a value’s relationship with a reference value into bels for a Type 2 “Power” value, we do the following calculation:
$\log_{10}$ (given value ÷ reference value)

From this, we can see that we are just taking the base 10 logarithm of our value after it has been divided by the reference value.

As an example, if we were comparing 12 watts with the generally accepted reference point of 1 milliwatt, we would calculate this:
$\log_{10} (12 \div 0.001) = 4.07918$ bels.

Type 1 “Root Power” Formula

To convert a value’s relationship with a reference value into bels for a Type 1 “Root Power” value, we have to square the given value and the reference value before we take the logarithm. The formula is as so:
$\log_{10}$ ((given value)^2 ÷ (reference value)^2)

Due to the way that logarithms work, we can rephrase this to be:
2 $\log_{10}$ (given value ÷ reference value)

[Squaring each of the values in the division is the same as not squaring them and multiplying the result of the logarithm by 2. Why this works requires slightly more knowledge of logarithms than I am going to give here.]

As an example, if we were comparing 5 volts with the generally accepted reference of 1 volt, we would proceed as follows:

$\log^{10} (5^2 \div 1^2) = \log (25 \div 1) = 1.3979$ bels.
... or we could do:
$2 \log_{10} (5 \div 1) = 2 \log (5) = 1.3979$ bels.

Decibels

In each of the above examples, to find the number of *decibels*, we need to multiply the answer by 10 (because there are 10 decibels in a bel: 1 bel = 10 decibels).

Therefore, the result of the first example (4.07918 bels) given as decibels is 40.7918 decibels. We can phrase this as 40.7918 dB, or to emphasise how we are dealing in watts in that example and comparing against 1 milliwatt, 40.7918 dBm (where dBm is the standard way of saying decibels comparing against 1 milliwatt).

The result of the second example was 1.3979 bels. We multiply this by 10 and we have 13.9794 decibels. We can write this as 13.9794 dB, or because we are dealing in volts and comparing against 1 volt, we can write it as 13.9794 dBv.

Given that people generally use decibels instead of bels, we can give the formula for calculating decibels for Type 2 “Power” values as:
$10 * \log_{10}$ (given value ÷ reference value)

This is the same formula as for bels, but multiplied by 10.

We can give the formula for Type 1 “Root Power” values for decibels, which is the Type 1 bels formula multiplied by 10, as:

$10 * \log_{10} ((\text{given value})^2 \div (\text{reference value})^2)$
... or if we use the other formula:
$10 * 2 * \log_{10}$ (given value ÷ reference value)
... which is:
$20 * \log_{10}$ (given value ÷ reference value)

Graphs

Now that I have briefly explained decibels, we can go back to our earlier example of amplitudes in a logarithmic frequency domain graph, and plot the amplitudes as decibels on a graph. The amplitudes on the graph were:

10 units at 1 cycle per second.
100 units at 2 cycles per second.
1000 units at 3 cycles per second.
1,000,000 units at 4 cycles per second.

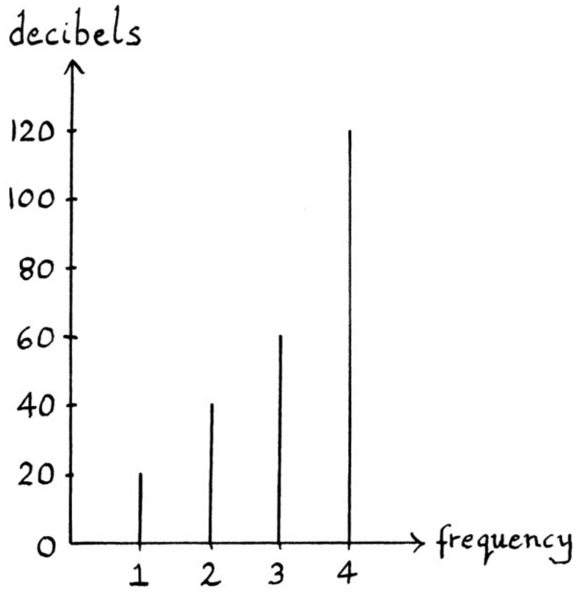
We will set a reference value of 1 unit and base the results around that. As these are amplitudes, they are Type 1 "Root Power" values. Therefore, we will use the formula: 10 * $\log_{10}$ ((given value)^2 ÷ (reference value)^2)

10 units becomes: 10 * $\log_{10}$ ($10^2 \div 1^2$) = 20 dB
100 units becomes: 10 * $\log_{10}$ ($100^2 \div 1^2$) = 40 dB
1000 units becomes: 10 * $\log_{10}$ ($1000^2 \div 1^2$) = 60 dB
1,000,000 units becomes: 10 * $\log_{10}$ ($1000000^2 \div 1^2$) = 120 dB

[As the context of the decibels is obvious within this example, I have not written any suffix or explanation as to what they are being compared.]

The results would have been the same if we had used the other formula:
20 * $\log_{10}$ (given value ÷ reference value)

Our graph is as so:



The most significant thing to realise from this graph is that decibels represent logarithmic values, but decibels *themselves* increase at an even rate. Therefore, when drawn on a graph, decibels increase at an even rate up the y-axis (and therefore, so do bels). When we drew a logarithmic scale on the graph, the y-axis scale increased logarithmically in the form of 1, 10, 100, 1000 and so on. Now, we are using decibels, the y-axis scale increases at an even rate: 0, 1, 2, 3, 4, 5 and so on.

Another thing to note is that the decibel scale shown here starts at zero, while the logarithmic scale from before had to start at a power of 10, and therefore could not start at zero.

Looking at the graph, it is harder to see the underlying relationship between the amplitudes. This is partly because the values were squared before they became decibels, and partly because of the logarithmic nature of bels and decibels. As you become more used to bels and decibels, such a graph will still allow you to understand the relationship between amplitudes.

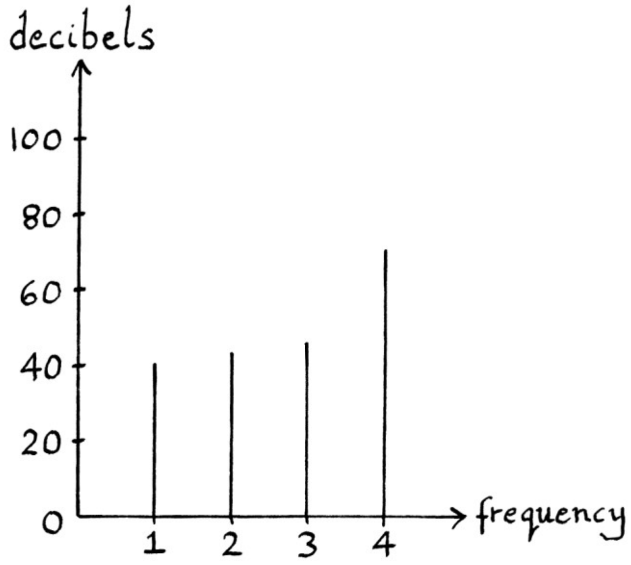
Supposing we had a graph showing power in watts of the waves at different frequencies (so in other words, Type 2 “Power” values), the underlying relationship would be more obvious. We will say we have the following values:

10 watts at 1 cycle per second
20 watts at 2 cycles per second
40 watts at 3 cycles per second
10,000 watts at 4 cycles per second.

We will calculate the decibels compared with the general standard reference of 1 milliwatt. As these are Type 2 “Power” units, we will use this formula:
10 * $\log_{10}$ (given value ÷ reference value)

10 watts becomes: 10 * log (10 ÷ 0.001) = 40 dBm
20 watts becomes: 10 * log (20 ÷ 0.001) = 43.0103 dBm
40 watts becomes: 10 * log (40 ÷ 0.001) = 46.0206 dBm
10,000 watts becomes: 10 * log (10,000 ÷ 0.001) = 70 dBm

Plotting these on a graph, we end up with this:



Converting decibels back to the original values

If we are given a decibel value, and we know to what it is being compared, then we can work out the actual value it represents. To do this, we work backwards using the formulas.

If we know that the value was a Type 2 "Power" value, we can use the formula:
$10 * \log_{10}$ (original value ÷ reference value) = given decibel value

As an example, we will find the actual value represented by 3 dBm. As this is 3 dBm, we know that it is comparing against 0.001 milliwatts (because dBm is the abbreviation for this generally accepted standard). We also know that this is a Type 2 "Power" value because watts relate to power, and power is a Type 2 unit.

We can therefore fill in the formula:
$10 * \log_{10}$ (original value ÷ 0.001) = 3

We will go through the steps for solving this:
$\log_{10}$ (original value ÷ 0.001) = 3 ÷ 10
$\log_{10}$ (original value ÷ 0.001) = 0.3
(original value ÷ 0.001) = $10^{0.3}$
[The above step is possible because $\log_{10}$ (x) is the inverse of 10^{x}.]
original value = 0.001 * $10^{0.3}$
original value = 0.001995 watts, or 1.995 milliwatts.

If we know that the value was a Type 1 “Root Power” value, we could use the formula:
10 * $\log_{10}$ ((original value)^2 ÷ (reference value)^2) = given decibel value
... however, it is simpler to use the other version of this formula:
20 * $\log_{10}$ (original value ÷ reference value) = given decibel value

As an example, we will find the actual value represented by 20 dBv. As this is 20 dBv, we know that it is comparing against 1 volt (because dBv is the abbreviation for this generally accepted standard). We also know that this is a Type 1 “Root Power” value because volts relate to amplitude.

We fill the values into the formula and solve it:
20 * $\log_{10}$ (original value ÷ 1) = 20
20 * $\log_{10}$ (original value) = 20
$\log_{10}$ (original value) = 1
original value = 10^1
original value = 10 volts.

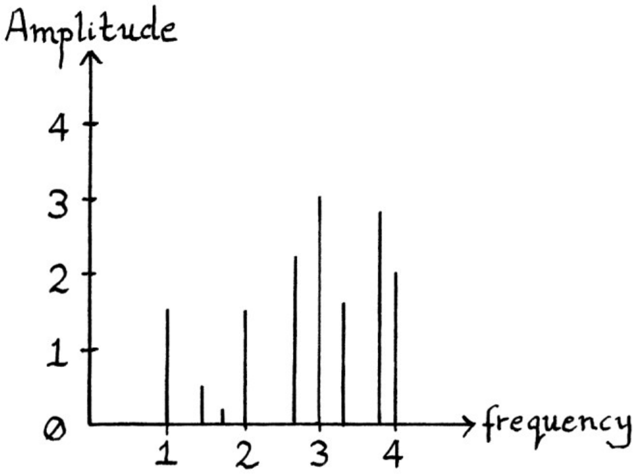
Types of wave frequency domain graphs

Most of the time when we see a wave frequency domain graph, it will not be showing a list of known waves that are being added together as a sum. Instead, it will be showing the result of mathematical processes to determine which frequencies of waves were added together to create a received radio or sound signal. In other words, it will not be demonstrating what we already know, but instead it will be demonstrating what has been calculated. It will still show a list of frequencies that were added together, but everything will have been calculated by analysing a signal. These two ideas are *nearly* the same, but the method used to calculate the underlying waves of a signal may or may not have been accurate. This might be because there was a compromise between speed and accuracy; it might be because accuracy is not needed for the task in hand; or it might be because the signal being analysed is one where total accuracy is impossible to achieve.

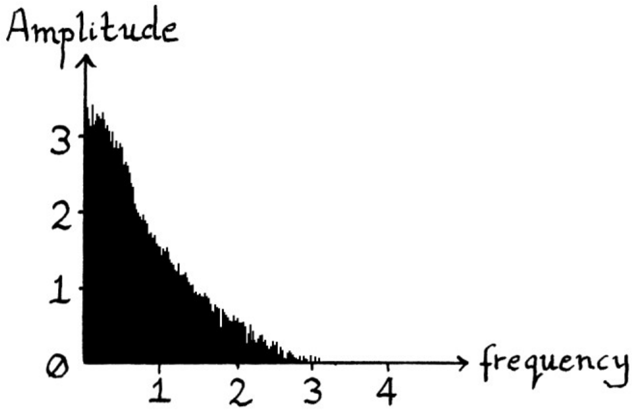
Besides that difference in the contents of the graph, there are several variations of wave frequency domain graphs, including the following.

1: Basic

This is the version I have shown already where the y-axis is amplitude, and the waves have the same amplitude for all time. Here is a simple version:

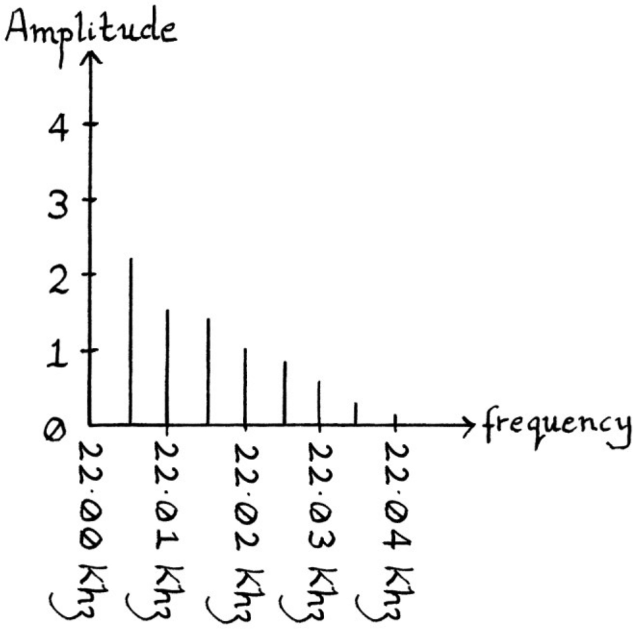


Here is a more complicated version showing many waves at the same time. It has much more information in it because of the number of waves, and it would be impossible to recreate the signal from looking at the graph. However, such a graph can still be useful for understanding aspects of the represented signal.

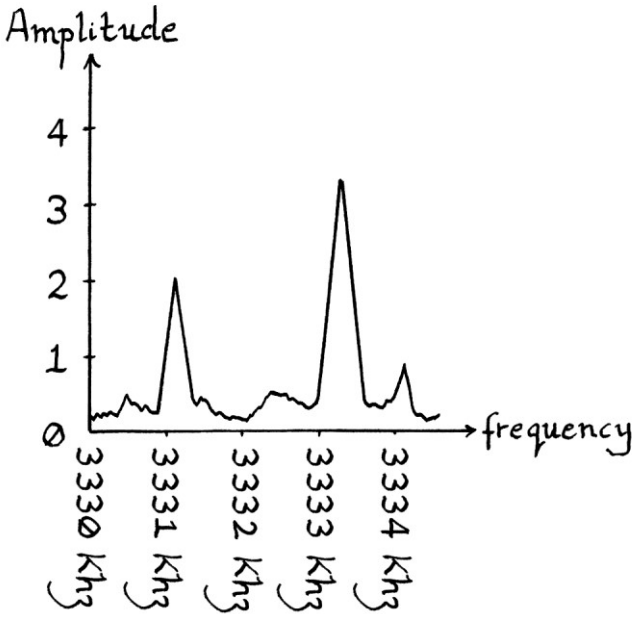


The frequency axis does not have to start at zero. In many cases, frequency domain graphs just show a section of the possible frequencies, which makes it easier to examine a particular section. This means that the graph could possibly be used to recreate the section shown, but could not be used to recreate the entire signal (unless the entire signal exists in just that section).

An example is as so:

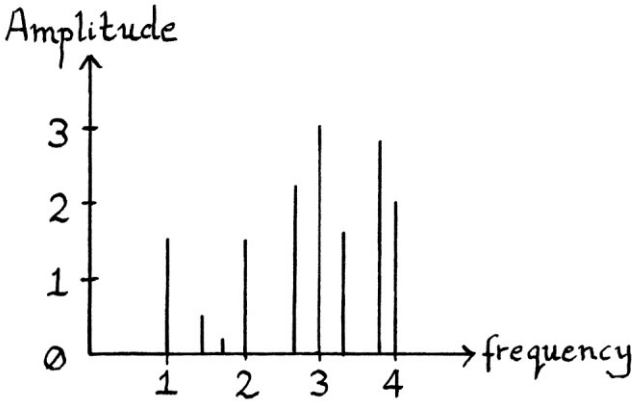


In practice, with *received* radio or sound signals, there will always be some unwanted background noise, so none of the frequency axis will ever be completely blank. There will always be something there, even if it is relatively quiet compared with the wanted parts of the signal. It is also the case that the limitations of a receiver and the method used to analyse the signal might blur any distinctive boundaries between the frequency of a wave and its neighbours. All this means that, in the real world, received radio waves in a frequency domain graph will not appear as straight lines, but instead as wider sections with peaks:

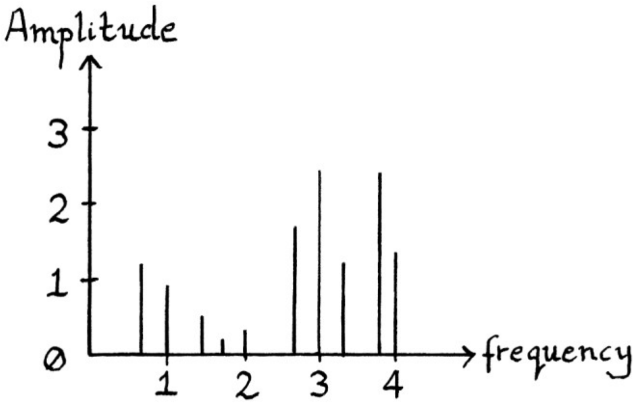


2: Time

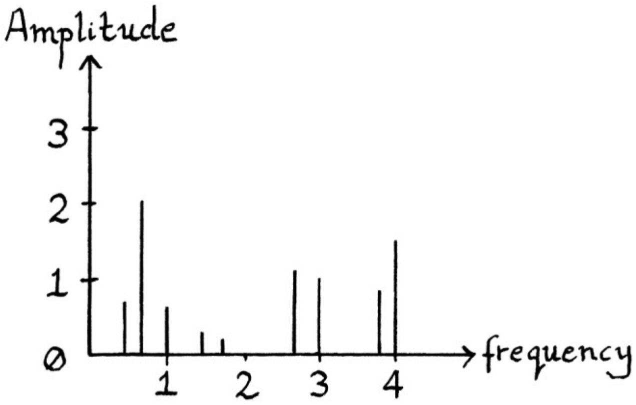
If the observed signal is one that changes over time, and it is shown on a computer screen, then we might see a basic amplitude graph that is constantly updated. For example, at one moment in time it might look like this:



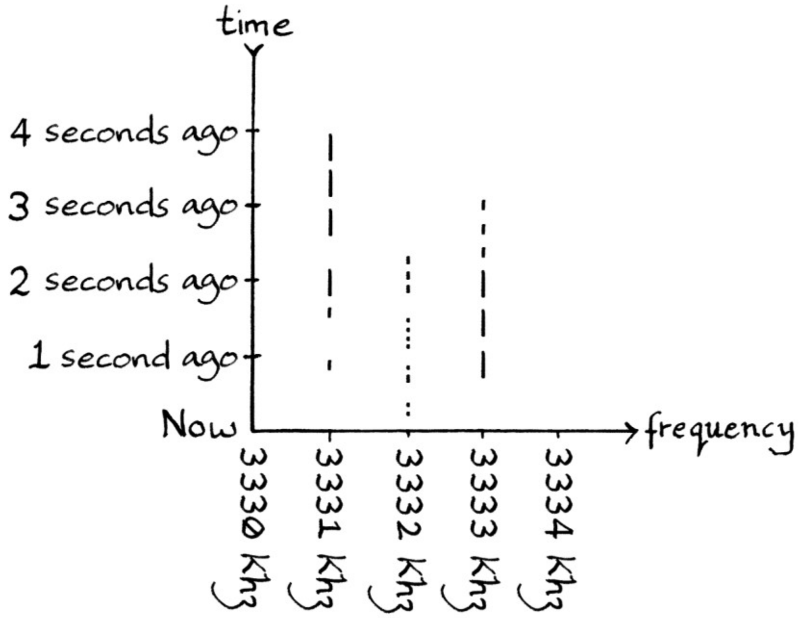
A tenth of a second later, it might look like this:



A tenth of a second later still, it might look like this:



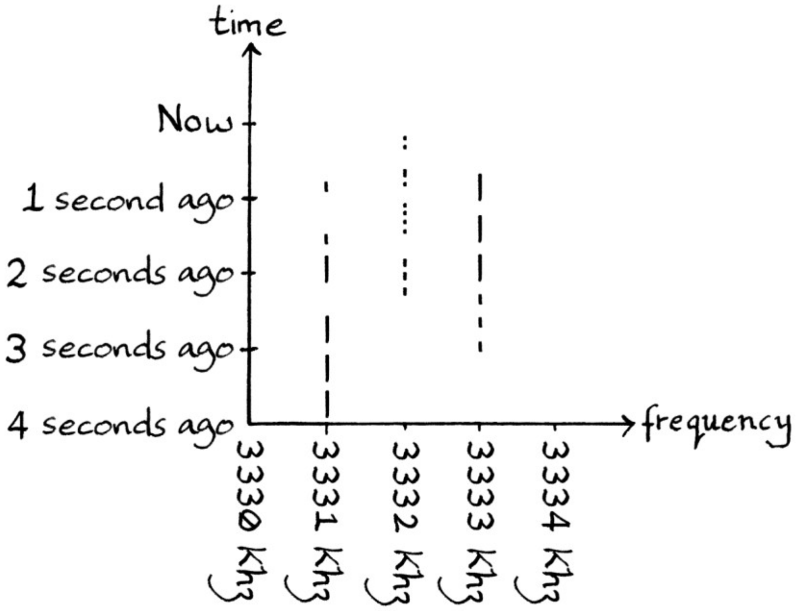
Another way of showing changes over time is by having the y-axis as "time" instead of "amplitude". This type of graph is commonly used in radio receiving software where signals change from one moment to the next. The graph at any one moment in time represents the sum of waves used to create the signal at that particular moment in time. [Or more usually, the sum for that portion of the shown frequency range]. The graph enables us to keep track of changes. Generally, when we see such a graph in software, the time axis will scroll as time passes, so we see the current wave state at the bottom of the graph, and the picture moves upwards off the graph. The graph, as I am drawing it here, merely shows the presence of waves of a particular frequency – we cannot tell what their amplitude is. In this particular drawing, the overall signal contains Morse code messages, so it does not matter that we cannot see the amplitude of the waves – the meaning of the signal is in the presence or absence of the underlying waves at moments in time, not in their amplitude. This is a good example of how we do not need to know every attribute of a wave for it still to be useful.



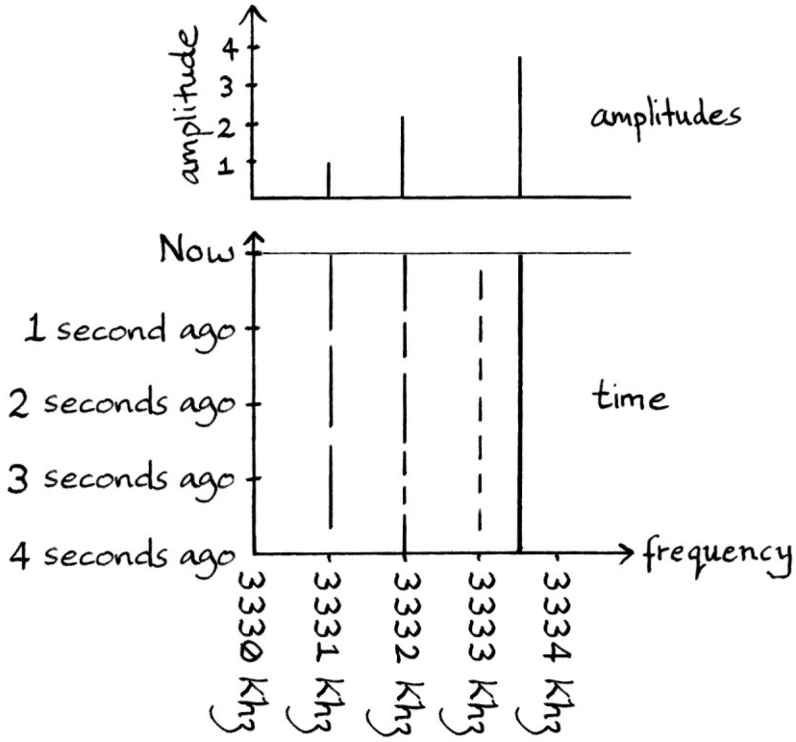
[Side note: in the real world, such Morse code messages might be being transmitted from places hundreds of miles away from each other, but from the point of view of a radio receiver at one particular place, they are still added together to form part of the gigantic received signal.]

Such a graph is often called a "waterfall" due to its similarity to a real-world waterfall, even if it is moving in the opposite way to a real waterfall.

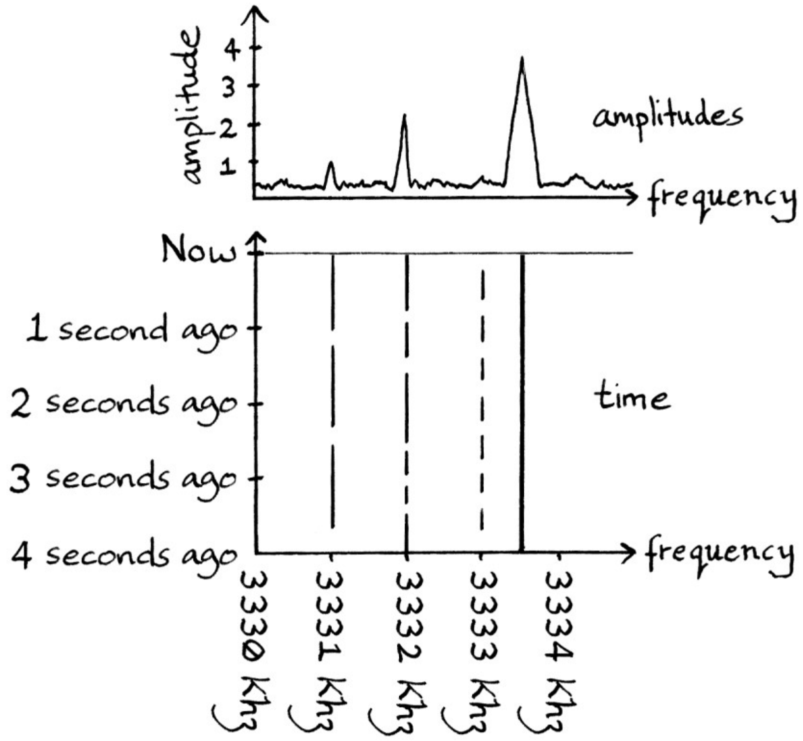
You may see the graph move downwards off the graph instead, in which case, it does vaguely resemble a real waterfall:



You may see graphs drawn or plotted side-on, so that the frequency is along the y-axis, and the time is along the x-axis. Often, if you see the graph on a computer screen, it will have different colours to give a guide to the different amplitudes or powers of the waves. You may see a graph with a second graph drawn at the top, showing the current amplitudes (or powers) of all the frequencies:



In practice, such a graph would also show noise, and would reflect the limitations of the processes used to analyse the signal to create it, so it would look more like this:



[One thing to notice from thinking about these graphs is that there are really two meanings of the word “signal”. The word “signal” as I have been using it so far, has referred to a wave or sum of waves. When a radio receiver is working, it is really receiving waves from every radio frequency in “hearing” distance, all added together to make up one very large signal. Part of the radio receiver’s job is to filter out the unwanted frequencies – some of this is done by the design of the antenna; some is done by electronics. In my use of the word “signal” as the sum of waves, we can say that the actual signal that a radio receiver is receiving is huge.

Another way of using the word “signal” is to mean a particular message that someone is sending. In the Morse code graphs above, each Morse code transmission could be called a “signal”. In this sense, we can say, “a radio transmitter transmits the Morse code signal,” or, “a radio operator decodes the Morse code signal.”

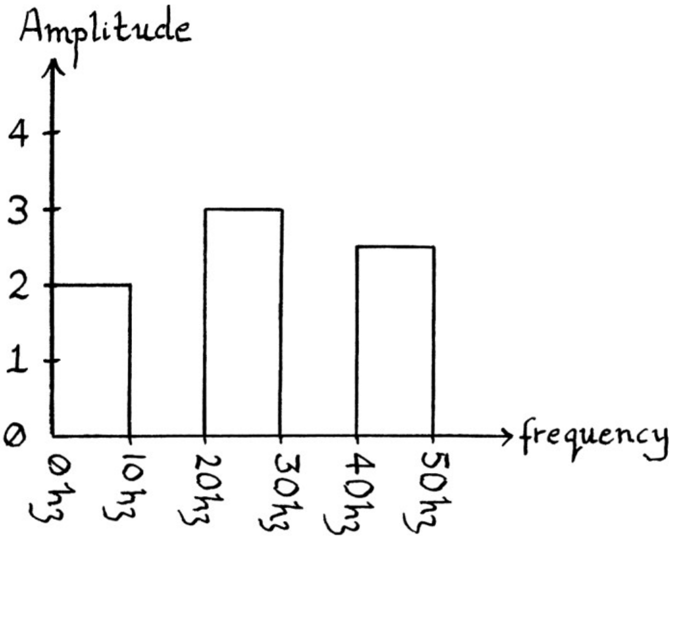
In the more specific sense, one could say that a radio station broadcasts a signal, and a radio receiver receives that signal. In the broader sense, one could say that the sum of all radio waves is a signal, and that a radio receiver has to filter out the parts of that signal to obtain a radio station’s broadcast.

The two meanings allow us to say something such as, “the signal (being the sum of countless added waves of countless frequencies) contains a signal (being a message sent using waves over a limited range of frequencies).

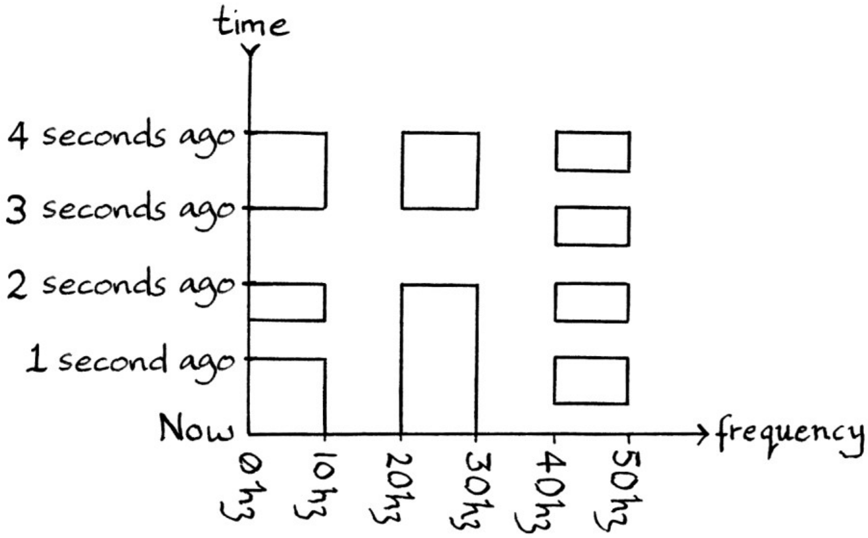
The two meanings of “signal” are different enough that they are unlikely to be confused, but there are probably situations when there is an ambiguity.]

3: Block graph

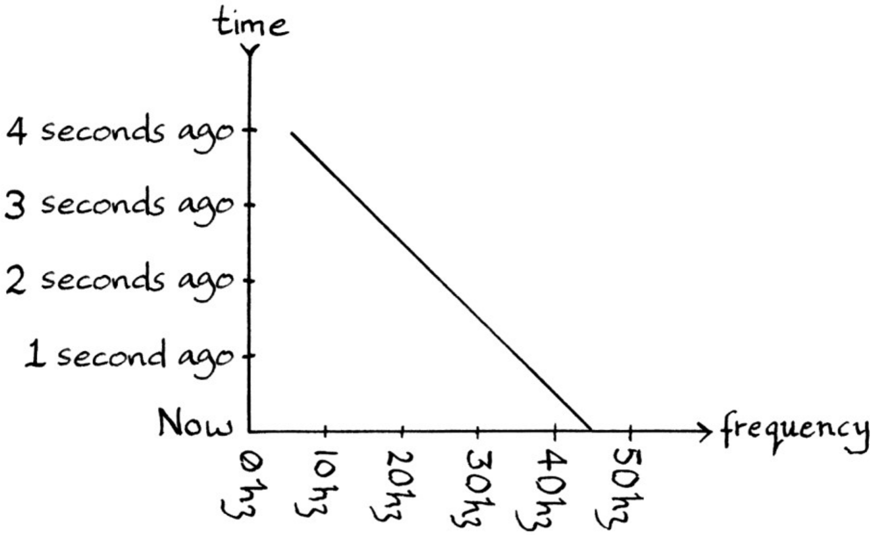
The mathematical processes to reduce a received signal into its constituent parts can be time consuming. Therefore, computer programs that are trying to show the changes in frequency in real time often use a method that groups the waves into frequency bands that appear as a block graph. This saves time and processing power, and is useful if a high level of detail is not needed. For example, if there are added waves with the frequencies:
1, 2, 4, 22, 25, 29, 45, 46, 47
... and various amplitudes, then the graph might be put into blocks, and we will see a graph similar to that shown here (if the y-axis is amplitude): [Note that this exaggerates the idea, and in practice, the graph would not be so blocky].



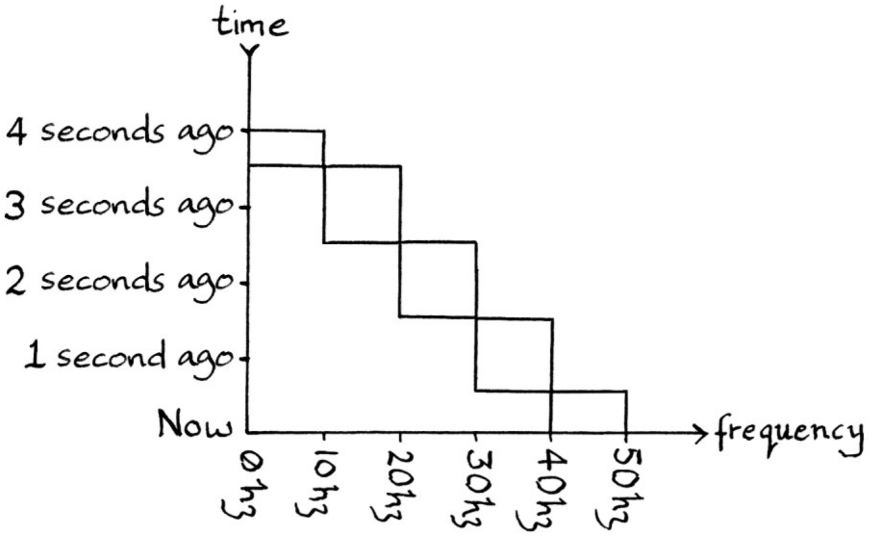
If the y-axis is time, and the waves appear and disappear as time progresses, we might see a graph as shown here:



Supposing we had a wave with a gradually increasing frequency, and the y-axis is time, then in a basic, non-block graph, it might appear as so:



... but in a block-type graph, it might appear as so:



All the points of the diagonal line become hidden within the blocks. We can still tell that the frequency of the wave has increased over the last 4 seconds, but there is less detail.

The blocks of frequencies for these graphs are called “bins” [Where the word “bin” relates to the meaning of “bin” as in “storage container”, as is used in the term “bread bin”.] The more bins that are used, the higher the detail of the graph, but on the other hand, the more work that the computer has to do. There is always a compromise between the detail required and the effort needed to achieve that detail.

A graph such as this could not be used to recreate the original signal. Its main use is as a guide as to what is happening. It can still be useful for analysing a signal.

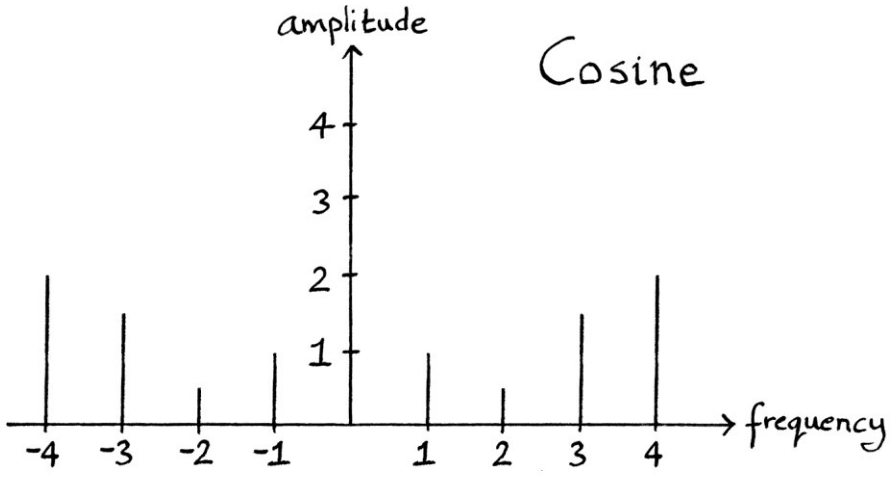
4: Mirrored negative frequencies

In slightly more advanced maths, waves are treated as if they exist in the three dimensional helix chart. In this way, the curve of a Cosine wave is created by adding two helices of identical but opposing frequencies. It will still be a two dimensional Cosine wave, but it will be lying flat in a three dimensional chart. The curve of a Sine wave is created by subtracting one helix from another where the helices have identical but opposing frequencies, and then rotating it so that it lies flat. This will be a two dimensional Sine wave lying flat in a three dimensional chart. [The waves are generally created using formulas in terms of Imaginary powers of “e”.] For each pair of helices used to create a wave, there will be a positive frequency and a negative frequency – the frequencies are the same but opposites of each other. We saw the basis of this idea in Chapter 14 when we added two identical circles with opposing frequencies. [We will explore the idea more in later chapters.] In this way of thinking, if we have a signal made from adding waves, it will be thought of as being made from adding pairs of helices.

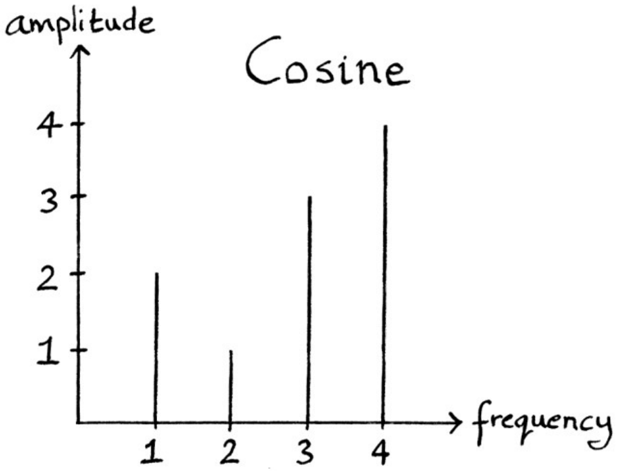
Creating waves using helices in the helix chart is really a way of creating a portrayal of the waves to make them easier to deal with. A consequence of this way of thinking is that the portrayal of each of the added waves in a signal will have both positive and negative frequencies. In reality, a received radio signal, for example, will consist only of positive frequencies, but when the signal is portrayed as being in the helix chart, there will be positive *and negative* frequencies. These positive and negative frequencies only exist due to how we are thinking about the waves. For most cases, these frequencies will be mirrored across the y-axis. The positive and negative frequencies do not reflect reality, but the way we are portraying reality. One might say that they are illusionary frequencies.

It is important to remember that there are not literally any negative frequencies in a real-world signal thought about in this way. There are no objects that are rotating the "wrong way" around circles. Both the positive and negative frequencies only apply to the *portrayal* of the waves, and not to the actual waves. Typically, each amplitude for the illusionary positive and negative frequencies will be half the actual amplitude of the underlying wave. When the corresponding waves from each side of the y-axis are added together, the result will just be positive-frequency waves with the correct amplitude for the actual waves.

If you see such a thing on a computer, some computer programs show you both the negative and positive halves, while some programs helpfully add the two halves together for you. An example graph is as so:



The following is the result of adding both halves together, which gives the actual frequencies for the actual signal:



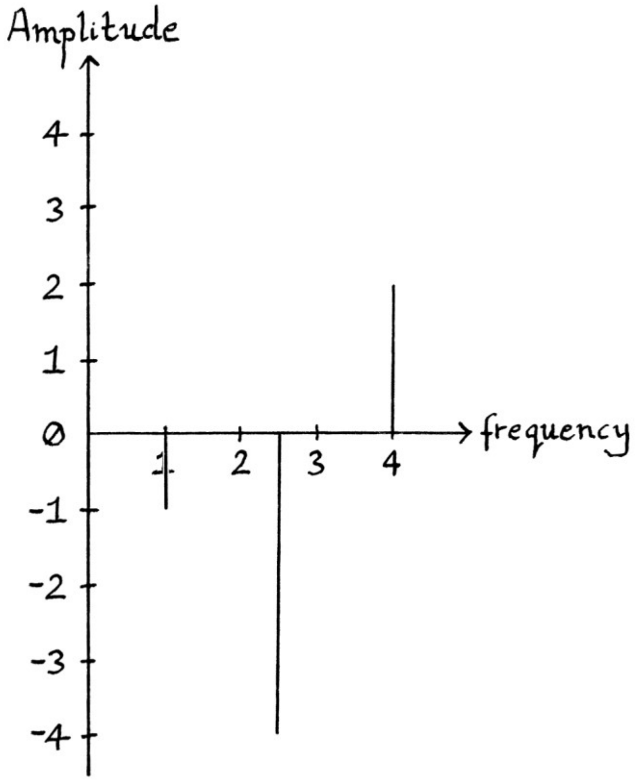
Note that if the waves were all Sine waves with zero phases, adding the corresponding halves together would result in nothing at all on the frequency domain graph. [This is because a negative-frequency Sine wave with zero phase is an upside down version of that wave with the frequency made positive. Adding them together results in “y = 0” for all time.]

Sometimes, you might see frequency domain graphs that show actual negative frequencies, but these are rare in comparison with the graphs mentioned in this section.

5: Negative amplitudes

In some frequency domain graphs, you might see negative amplitudes. As I have said before, while you are learning, it is helpful to minimise the use of negative amplitudes because they conflict with the idea of the amplitude being equal to the radius of the circle from which the wave is, or could be said to be, derived. However, it is often convenient to use negative amplitudes in certain situations.

On a frequency domain graph, negative amplitudes are portrayed using lines that go downwards from the frequency axis. In the following graph, there are two negative amplitudes and one positive amplitude.



A negative-amplitude Sine wave formula refers to the same curve as that formula with the amplitude made positive and 180 degrees added to the phase.

For example:
“y = −2 sin ((360 * 5t) + 25)”
... refers to the same curve as:
“y = +2 sin ((360 * 5t) + 205)”

6: Negative frequencies and negative amplitudes

You might see frequency domain graphs with both negative frequencies and negative amplitudes. It is not difficult to work out how the graphs could be redrawn to have positive frequencies and positive amplitudes.

For the same basic formula:

- a positive-amplitude, positive-frequency Sine wave *with zero phase*, e.g:
  “y = 2 sin (360 * 5t)”
- a negative-amplitude, negative-frequency Sine wave *with zero phase*, e.g:
  “y = −2 sin (360 * −5t)”
- a negative-amplitude, positive-frequency Sine wave with a phase of +180 degrees, e.g:
  “y = −2 sin ((360 * 5t) + 180)”
- a positive-amplitude, negative-frequency Sine wave with a phase of +180 degrees, e.g
  “y = 2 sin ((360 * −5t) + 180)”

... are all identical, and look like a correct-way-up Sine wave.

It is also the case that:

- a positive-amplitude, negative-frequency Sine wave *with zero phase*, e.g:
  “y = 2 sin (360 * −5t)”
- a negative-amplitude, positive-frequency Sine wave *with zero phase*, e.g:
  “y = −2 sin (360 * 5t)”
- a positive-amplitude, positive-frequency Sine wave with a phase of +180 degrees, e.g:
  “y = 2 sin ((360 * 5t) + 180)”
- a negative-amplitude, negative-frequency Sine wave with a phase of +180 degrees, e.g:
  “y = −2 sin ((360 * −5t) + 180)”

... are all identical, and look like an upside down Sine wave.

When it comes to Cosine waves, the equivalences are slightly different:

- a positive-amplitude, positive-frequency Cosine wave *with no phase*, e.g:
  "y = 2 cos (360 * 5t)"
- a positive-amplitude, negative-frequency Cosine wave *with no phase*, e.g:
  "y = 2 cos (360 * −5t)"
- a negative-amplitude, positive-frequency Cosine wave with a phase of +180 degrees, e.g:
  "y = −2 cos ((360 * 5t) + 180)"
- a negative-amplitude, negative-frequency Cosine wave with a phase of +180 degrees, e.g:
  "y = 2 cos ((360 * −5t) + 180)"

... are all identical, and look like a correct-way-up Cosine wave.

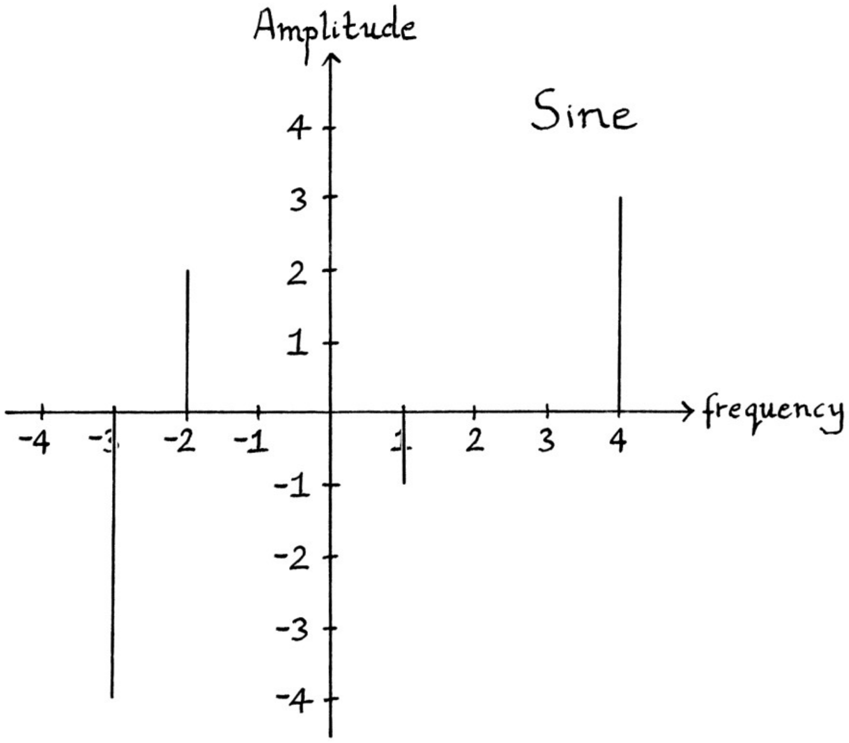
It is also the case that:

- a negative-amplitude, positive-frequency Cosine wave *with no phase*, e.g:
  "y = −2 cos (360 * 5t)"
- a negative-amplitude, negative-frequency Cosine wave *with no phase*, e.g:
  "y = −2 cos (360 * −5t)"
- a positive-amplitude, positive-frequency Cosine wave with a phase of +180 degrees, e.g:
  "y = 2 cos ((360 * 5t) + 180)"
- a positive-amplitude, negative-frequency Cosine wave with a phase of +180 degrees, e.g:
  "y = 2 cos ((360 * −5t) + 180)"

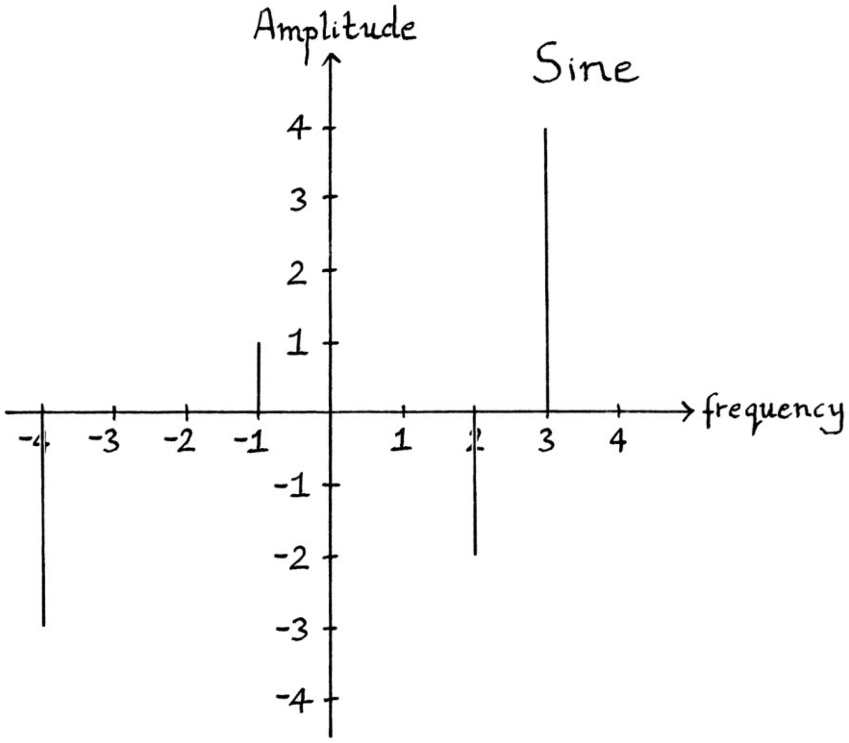
... are all identical, and look like an upside-down Cosine wave.

In the subject of waves as a whole, sometimes it can be easier to describe a wave in a way that does not involve just positive frequencies and positive amplitudes. Choosing one description over the others in particular explanations can make the point you are trying to make more simple and intuitive.

If you see a frequency domain graph for Sine waves, where all of them have zero phases, such as this:

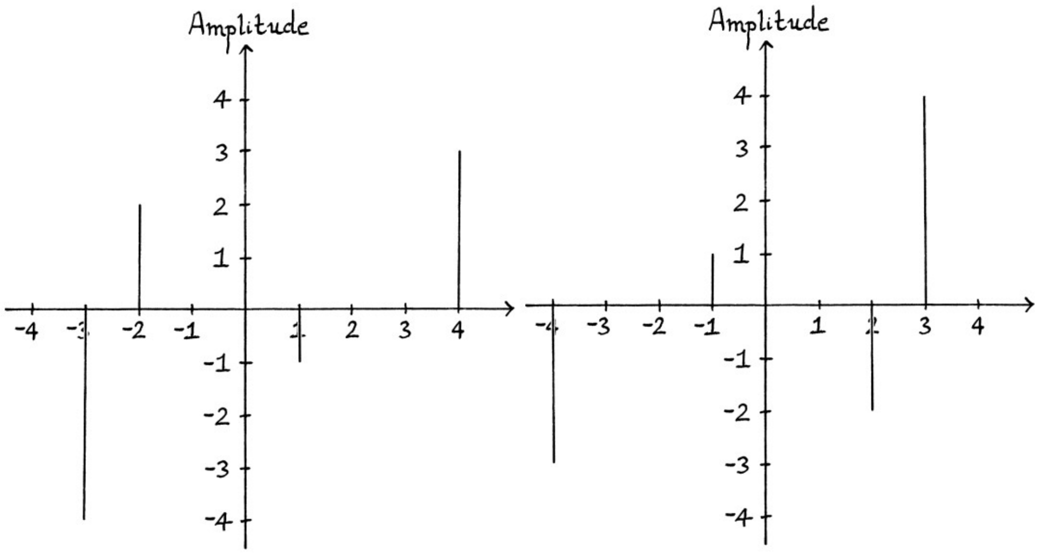


... then you should be able to tell that it is identical in meaning to this (where all the Sine waves also have zero phases):



The second graph is really a flipped and mirrored version of the first graph.

The two graphs next to each other look like this:



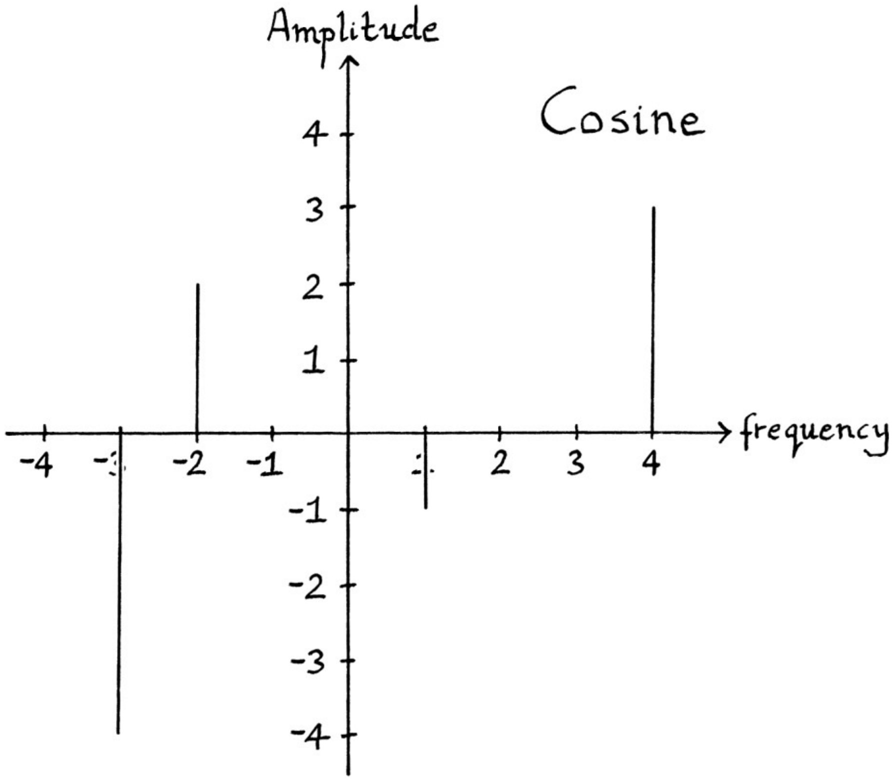
In the first graph, there is a zero-phase Sine wave with an amplitude of −4 units and a frequency of −3 cycles per second. This is the same as a zero-phase Sine wave with an amplitude of +4 units and a frequency of +3 cycles per second. In other words, “y = −4 sin (360 * −3t)” is the same as “y = 4 sin (360 * 3t)”.

In the first graph, there is a zero-phase Sine wave with an amplitude of +2 units and a frequency of −2 cycles per second. This is the same as a zero-phase Sine wave with an amplitude of −2 units and a frequency of +2 cycles per second. In other words, “y = 2 sin (360 * −2t)” is the same as “y = −2 sin (360 * 2t)”.

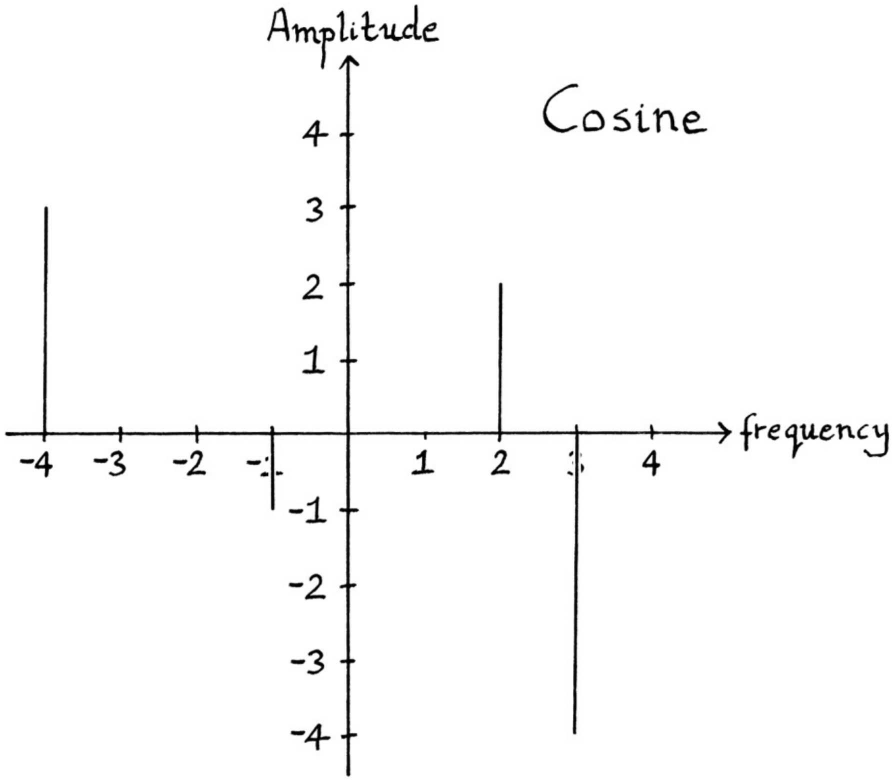
In the first graph, there is a zero-phase Sine wave with an amplitude of −1 units, and a frequency of +1 cycles per second. This is the same as a zero-phase Sine wave with an amplitude of +1 units and a frequency of −1 cycles per second. In other words, “y = −1 sin (360 * 1t)” is the same as “y = 1 sin (360 * −1t)”.

In the first graph, there is a zero-phase Sine wave with an amplitude of +3 units and a frequency of +4 cycles per second. This is the same as a zero-phase Sine wave with an amplitude of −3 units and a frequency of −4 cycles per second. In other words, “y = 3 sin (360 * 4t)” is the same as “y = −3 sin (360 * −4t)”.

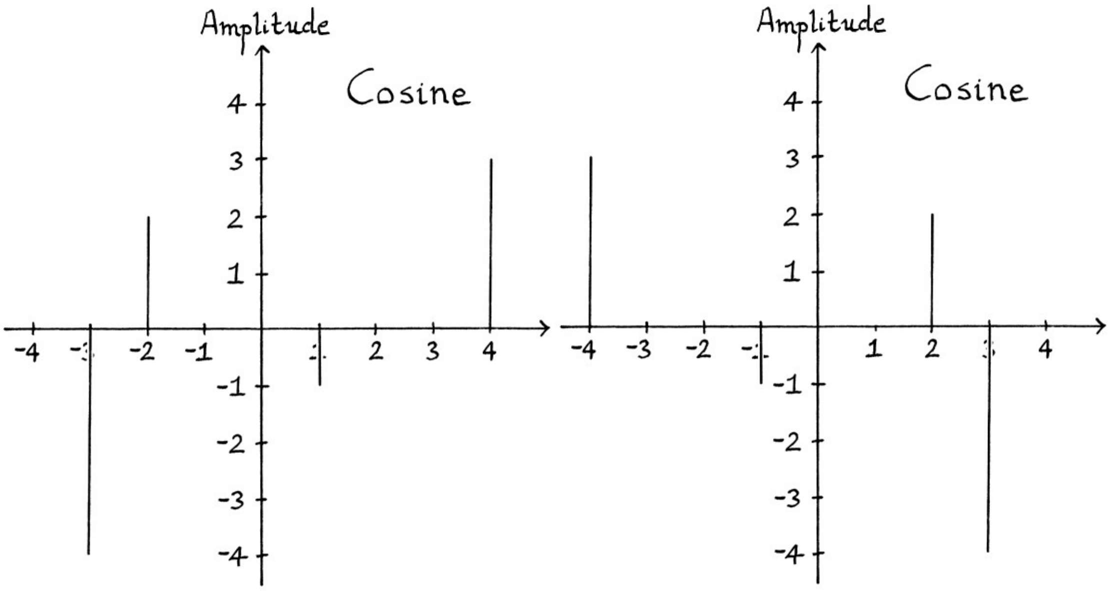
If we are thinking of Cosine waves *with zero phases*, then we could have this graph:



... and it would be identical in meaning to this graph:



The two graphs side by side look like this:



The graphs are mirror images of each other reflected over the y-axis.

On the first graph, there is a zero-phase Cosine wave with an amplitude of −4 units and a frequency of −3 cycles per second. This is the same as a zero-phase Cosine wave with an amplitude of −4 units and a frequency of +3 cycles per second.

On the first graph, there is a zero-phase Cosine wave with an amplitude of +2 units and a frequency of −2 cycles per second. This is the same as a zero-phase Cosine wave with an amplitude of +2 units and a frequency of +2 cycles per second.

On the first graph, there is a zero-phase Cosine wave with an amplitude of −1 units and a frequency of +1 cycles per second. This is the same as a zero-phase Cosine wave with an amplitude of −1 units and a frequency of −1 cycles per second.

On the first graph, there is a zero-phase Cosine wave with an amplitude of +3 units and a frequency of +4 cycles per second. This is the same as a zero-phase Cosine wave with an amplitude of +3 units and a frequency of −4 cycles per second.

Possible sources of confusion

Negative frequencies and negative amplitudes

As I said earlier, a formula with a negative *amplitude* and a particular phase refers to exactly the same wave curve as that formula with a positive amplitude and 180 degrees added to that phase.

As we saw in Chapter 11, a Sine wave formula with a negative *frequency* and a phase so many degrees above or below 90 degrees refers to exactly the same wave curve as that formula with a positive frequency and a phase that number of degrees *below* or *above* 90 degrees. Similarly, a Cosine wave formula with a negative *frequency* and a phase so many degrees above or below 0 degrees refers to exactly the same wave curve as that formula with a positive frequency and a phase that number of degrees *below or above* 0 degrees.

Where the confusion lies for Sine waves is that a negative-frequency Sine wave formula with a phase of 0 degrees happens to refer to the same curve as that formula with a positive frequency and a phase of 180 degrees, and vice versa. This is because 0 and 180 are equidistant from 90 degrees. It is also the case that a Cosine wave formula with a negative frequency and a phase of 90 degrees refers to the same curve as that formula with a positive frequency and a phase of 270 degrees. This is because 90 and 270 are equidistant from 0 degrees.

Given the two specific frequency examples, it is easy to think, wrongly, that to convert a negative-frequency wave formula to a positive-frequency formula, one just has to add 180 degrees, in the same way as happens with positive and negative amplitudes. However, doing this only works if a Sine wave has a phase of specifically 0 degrees or 180 degrees, or if a Cosine wave has a phase of specifically 90 degrees or 270 degrees. It will not work for any other phase.

If we are using Sine waves *with zero* phase, then we can convert from positive to negative frequency or negative to positive frequency by adding 180 degrees to the phase, which can lead us to misunderstand the underlying rule.

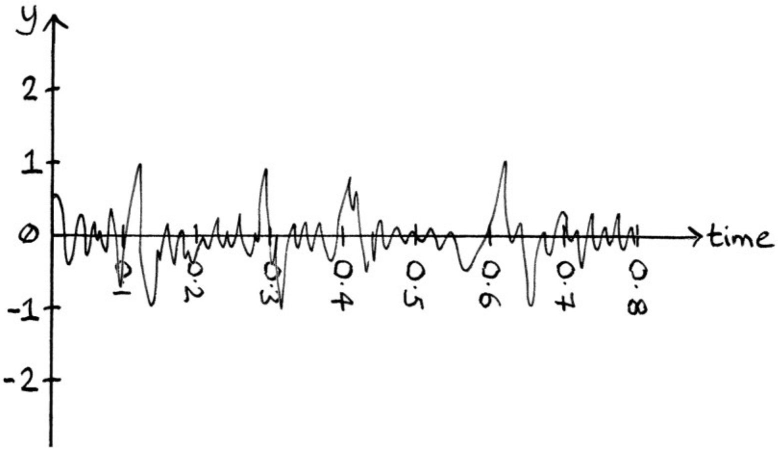
When converting between positive and negative frequencies, it is much better, even for waves with zero phase, to imagine flipping the circle from which those waves are derived, either left to right, or top to bottom, depending on whether the waves are Sine waves or Cosine waves.

The name of the frequency domain graph

The frequency domain graph as a whole is often called “the FFT”. The letters “FFT” stand for “Fast Fourier Transform”, which is one of the mathematical methods used to determine which waves were added to create an aperiodic discrete signal. (A discrete signal is one that is portrayed as a sequence of y-axis values at evenly spaced moments in time for the purposes of making it easier for computers to analyse it). People will sometimes refer to a frequency domain graph as the “FFT” even when that process was not used to work out the underlying waves, or when the graph is just showing known waves added together.

The signal is not the frequency domain

One easy way to be confused by frequency domain graphs is to forget that a signal does not appear how it is portrayed in such a graph. A frequency domain graph shows the individual waves that were added to make a signal, sorted by frequency. The signal does not look like the frequency domain graph. The signal will be a sum of waves and might have an appearance similar to this:



It is the signal that is analysed to find the underlying waves or circles (if those are not known). It is the signal that travels through the air (if it is a radio or sound signal). It is not the frequency domain representation of the signal that travels through the air. In, for example, a WAV file, the data in the file is in the form of the signal. [It is actually in a *discrete* form of the signal, as explained in a later chapter]. Mathematical procedures have to be performed on the data to reduce it to the frequency domain view. It is easy to be confused and think that a WAV file, for example, is a sequence of frequencies, which it is not, or that a radio signal travels in the form of a sequence of frequencies, which it does not.

Not all types of waves add together

It is easy to forget that not all waves are radio or sound waves. Not all types of wave add together in the way of superposition. For example, if we have a wave describing the position of a point on a spinning disc, it will not increase in amplitude if we put a second spinning disc next to it. Waves showing the characteristics of several spinning discs could be placed on the frequency domain graph, but the graph would not indicate a sum – it would just be a list of independently occurring waves.

Line thickness

Theoretically, the lines representing individual waves or circles on a basic frequency domain graph are infinitely thin. They represent a value at that exact frequency. If a line is at 4.5 cycles per second, then it represents solely 4.5 cycles per second, it does not represent, for example, the range of frequencies from 4.499999999 to 4.500000001 cycles per second. In practice, a line cannot be drawn so that it is infinitely thin, so there is some artistic interpretation in drawing or reading such a graph. [However, if the graph is a block graph as mentioned earlier in this chapter, then a thick line or block will represent a range of frequencies].

The type of waves that make up a signal

When receiving, say, a radio or sound signal, and reducing it to its constituent waves so that it can be portrayed on a wave frequency domain graph, it is necessary to decide whether to treat that signal as being the sum of Cosine waves or the sum of Sine waves. We cannot know which type of wave made up the original signal, and the perceived phase of the underlying waves will vary depending on how far away the receiver is from the source of the signal. Whichever we choose, we must be consistent. If we decide that one of the constituent waves is a Sine wave, then we must treat all of the constituent waves as Sine waves. If we decide that one of the constituent waves is a Cosine wave, then we must treat all the constituent waves as Cosine waves.

Generally, it is more common for people to treat a signal as being made up of Cosine waves. Mathematically, this choice makes absolutely no difference to anything because a Cosine wave is just a Sine wave with a 90-degree phase. The reason for this choice is mostly convention, and partly related to how some people are slightly fearful of non-zero phases. If we are dealing with waves with zero phases, then some advantages of a Cosine wave are:

- Negative-frequency Cosine waves *with zero phase* are the same as positive-frequency Cosine waves *with zero phase*.
- Negative-frequency Cosine waves *with zero phase* can be added to positive-frequency Cosine waves *with zero phase* of the same amplitude and frequency without resulting in zero (unlike Sine waves with zero phase).
- A zero-frequency Cosine wave *zero phase* can be used as a mean level, while a zero-frequency Sine wave *with zero phase* cannot.

These advantages are not a particularly good reason for choosing Cosine waves over Sine waves though because they all apply to Sine waves with 90-degree phases too.

People often use the term “Cosine wave” in these situations without emphasising the “with zero phase” part, which gives Cosine waves an air of, first, being different from Sine waves, and, second, being better. There is no difference except that it is quicker to say the word “Cosine” than it is to say “Sine with a 90-degree phase”. If people did use the term “Sine waves with a 90-degree phase”, then it would emphasise exactly what was going on, and give people a better understanding of waves. Generally, people might find things easier to learn if the term “Cosine” did not exist at all, and everything was done using Sine.

Amplitude and phase

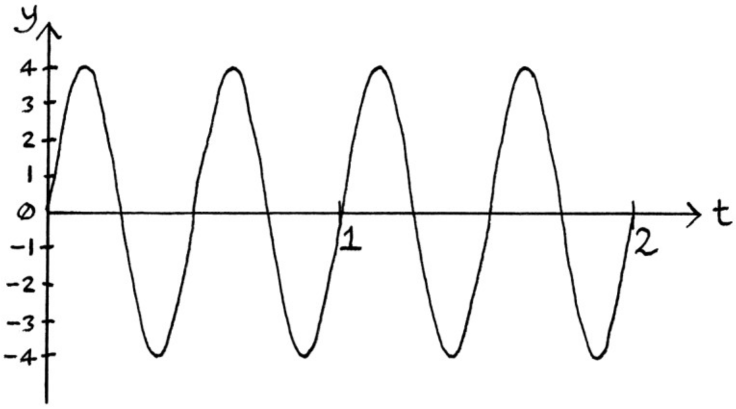
Most of the times when you see a frequency domain graph it will be referring to received sound or received radio waves. In such cases, the graph will not be showing the attributes of waves at the exact time and exact place that those waves were *created*. Instead, it will be showing the attributes of waves at the time and place that they were *received*. These may or may not be the same.

For waves such as radio or sound waves, the measured amplitude of a specific wave will be lower at the time and place that it is received, than at the time and place from which it came. Such waves fade as they travel. For example, if someone speaks to you from a distance, the sound you hear will be quieter the further away you are. A similar issue is that a sound or radio signal will be received in a form dependent on the receiving equipment – for example, sound’s amplitude exists in nature as variations in air pressure, but might be received by a recording device as variations in electrical current.

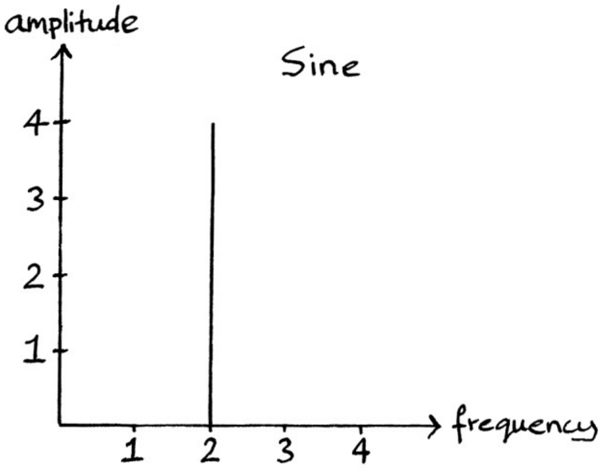
Impure signals

As I have said before, any periodic signal that is not a pure wave is the sum of, or approximately the sum of, two or more pure waves. This means that if we alter a pure wave even slightly in a way that stops it being a pure wave, it will immediately become the sum of two or more pure waves. This idea becomes more obvious when we look at wave frequency domain graphs.

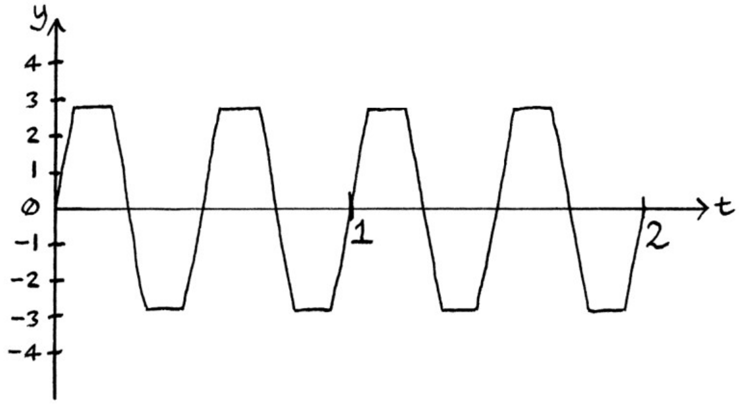
As an example, we will start with the wave “y = 4 sin (360 * 2t)”. The wave graph looks like this:



On the wave frequency domain graph, the wave appears as so:



We will squash the wave so that its peaks and dips are flattened, and so end up with this signal:

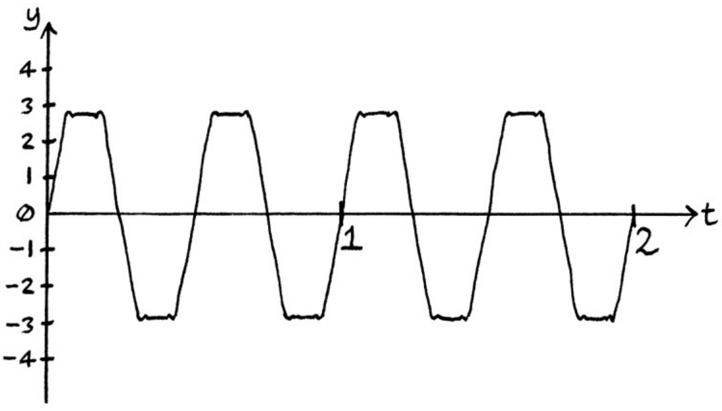


The above signal is not a pure wave, and therefore, it must instead be the sum of two or more pure waves (or approximately the sum). Therefore, it will not appear as just one vertical line in the frequency domain, but as several. From looking at the signal, we can see that it repeats twice a second, so it has a frequency of 2 cycles per second. This means that the waves that would need to be added to create it must have frequencies that are integer multiples of 2 cycles per second. Therefore, their frequencies will all be in the subset of: 2, 4, 6, 8, 10, 12, 14 etc cycles per second. I will explain how to calculate the exact details of the underlying waves in Chapter 18, but for now, I can say that they are:

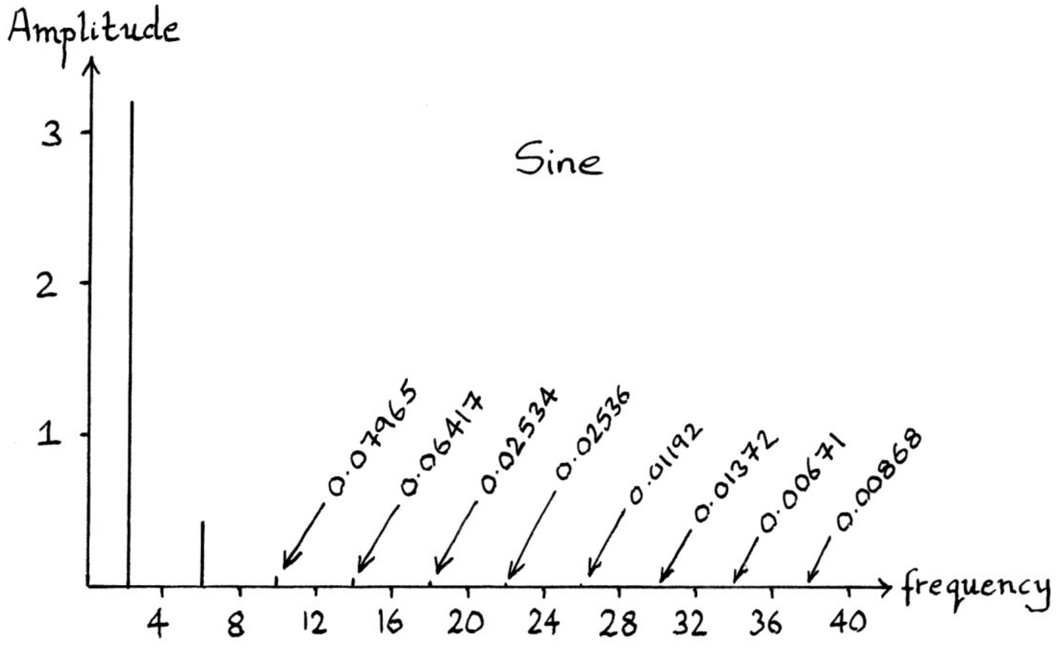
“y = 3.24754 sin (360 * 2t)”
“y = 0.43283 sin (360 * 6t)”
“y = 0.07965 sin ((360 * 10t) + 180)”
“y = 0.06417 sin ((360 * 14t) + 180)”
“y = 0.02534 sin (360 * 18t)”
“y = 0.02536 sin (360 * 22t)”
“y = 0.01192 sin ((360 * 26t) + 180)”
“y = 0.01372 sin ((360 * 30t) + 180)”
“y = 0.00671 sin (360 * 34t)”
“y = 0.00868 sin (360 * 38t)”
... and so on.

The amplitudes in the list become smaller, and the frequencies become higher, but there might never be an end to the list of waves that make up this particular signal.

The signal created by adding the waves in the list is as so:



The signal is close but not an exact match of the original "squashed" wave. The frequency domain portrayal of the waves that make up the signal is as so:



[Remember that this type of graph cannot show the phase of the waves.]

A slight alteration of a single pure wave has created a signal that is made up of many pure waves.

A consequence of how any impure periodic signal is the sum of two or more pure waves of various frequencies is that a signal can take up more bandwidth in the frequency spectrum than might be intended. In the above example, the "squashed" signal could have been created by accident by limiting the amplitude of the original wave. Instead of the signal being one wave, it instantly becomes a great number of waves (although admittedly, most would have very small amplitudes). If a radio transmitter had transmitted the “squashed” signal by mistake, it would actually have been transmitting countless waves of other frequencies at the same time, and it would risk interfering with other broadcasts on nearby frequencies.

Phase on wave graphs

The three-dimensional graphs in the first section showed circles with amplitude, phase and mean level, but because the graphs were three-dimensional, they were difficult to read accurately. The two-dimensional graphs in the last section were easier to read, but could not show non-zero phases (apart from phases of 180 degrees if we use negative amplitudes or negative frequencies). It is still possible to indicate the phase using two-dimensional graphs, although no method is perfect.

Phase ticks

The first method that I will explain uses little ticks on the sides of the vertical lines to indicate the phase. First, we extend the vertical lines below the amplitude axis so they match all of the circles viewed side on or from the top view.

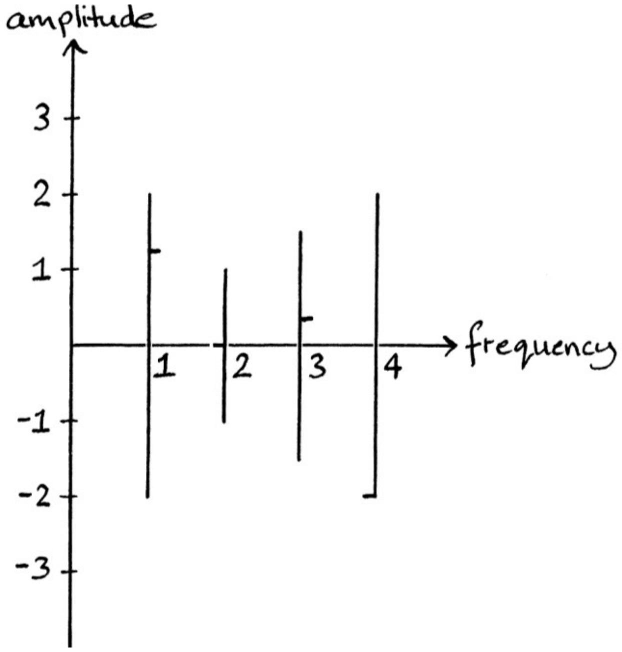
Then, if we are thinking of the graph for Sine (as in the circle viewed side on):

- We put a tick on the right-hand side of each line to indicate the phase point’s original y-axis value on the circle if it is between 270 degrees and 90 degrees (in other words in the right-hand half or far side of the circle).

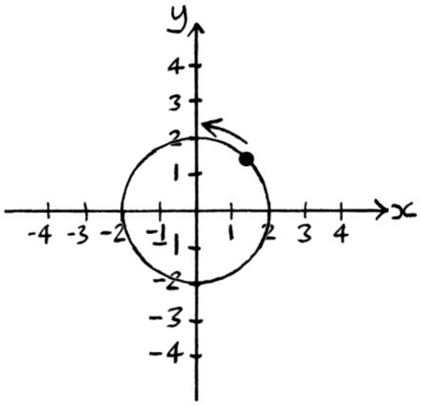
- We put the tick on the left-hand side if the phase is between 90 degrees and 270 degrees (in other words in the left-hand half or near side of the circle).

- If the phase is at 0 or 360 degrees, we have to cut a section out of the frequency axis line to make it visible.

An example graph is as so:

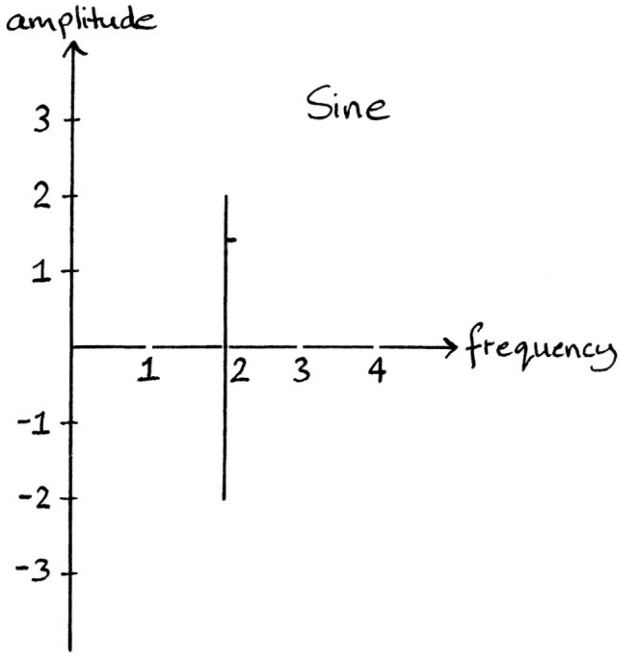


We will imagine that we have a circle with a radius of 2 units, a frequency of 2 cycles per second, and a phase of 45 degrees. The phase point will be at 45 degrees. This is on the right-hand side of the circle at a y-axis height of 1.4142 units:

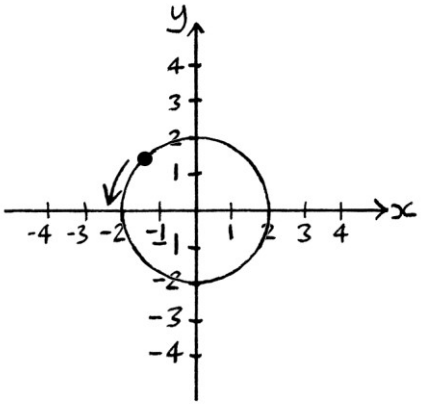


Therefore, for the circle viewed side on (in other words for the Sine perspective), the graph would have a line at 2 cycles per second with a tick on the right-hand side at 1.4142 units on the amplitude axis.

It would appear as in the following picture, with the graph labelled as “Sine” to indicate that it is the side view of the circle:

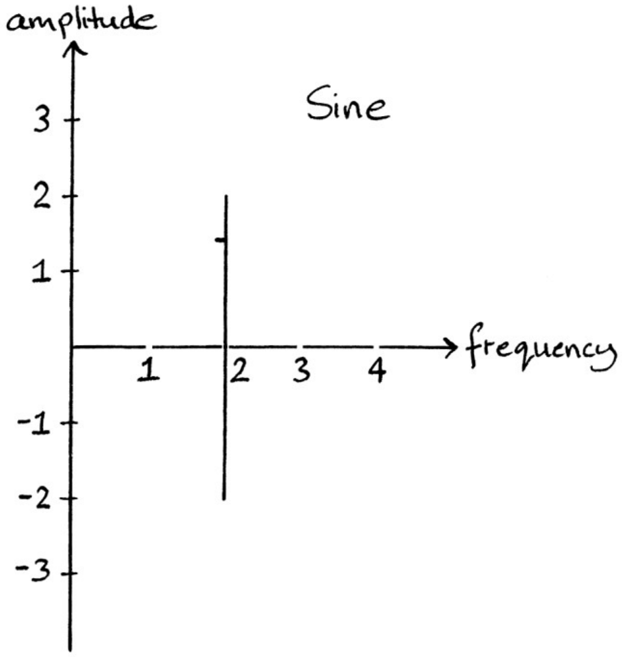


For the same circle with a phase of 135 degrees, the phase point would be at 135 degrees, which is on the *left-hand* side of the circle at a y-axis height also of 1.4142 units:

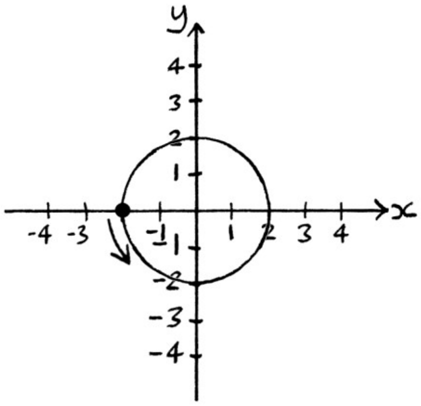


Therefore, for the circle viewed side on, the graph would have a line at 2 cycles per second with a tick on the *left-hand* side at 1.4142 units on the amplitude axis.

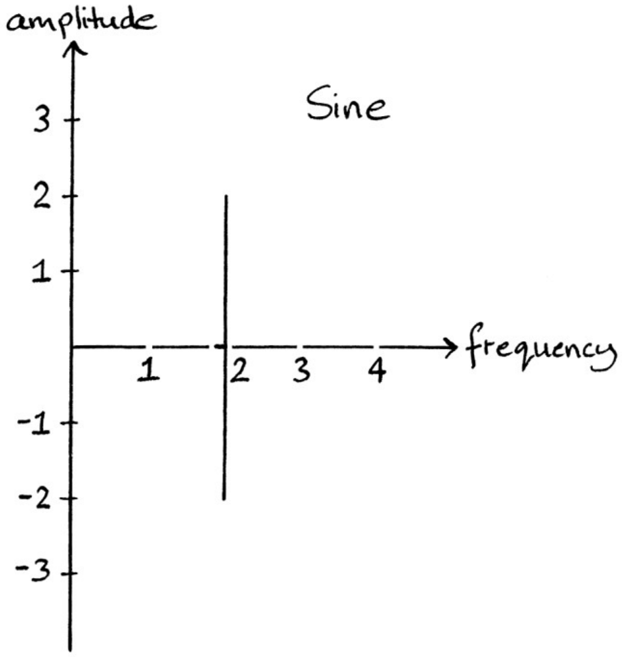
The graph would be as so:



For the same circle with a phase of 180 degrees, the circle would look like this, with the phase point on the left in the middle:



The relevant graph looks like this with the phase tick on the left in the middle too:

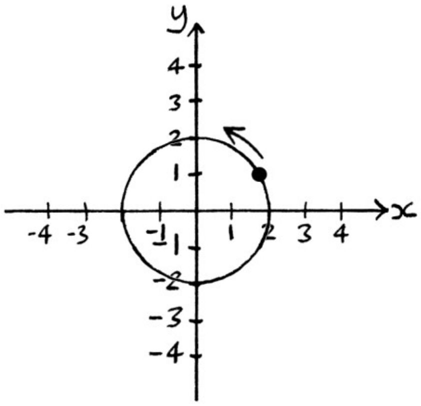


If we are doing graphs for the circle viewed from the top (in other words, from the Cosine point of view):

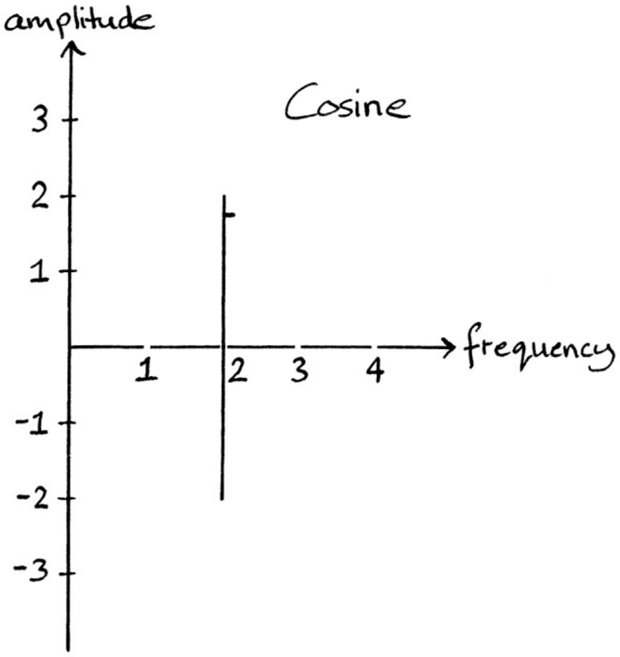
- We put a tick on the right-hand side of each line to indicate the phase point's original *x-axis* position on the circle if it is between 0 degrees and 180 degrees (in other words in the top half of the circle).

- We put the tick on the left hand side if the phase is between 180 degrees and 0 degrees (in other words in the bottom half of the circle).

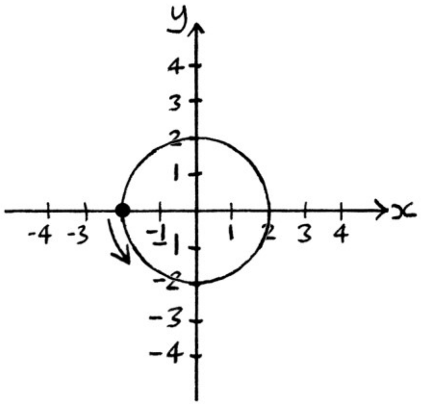
As an example, we will use the circle from the previous example, and give it a phase of 30 degrees:



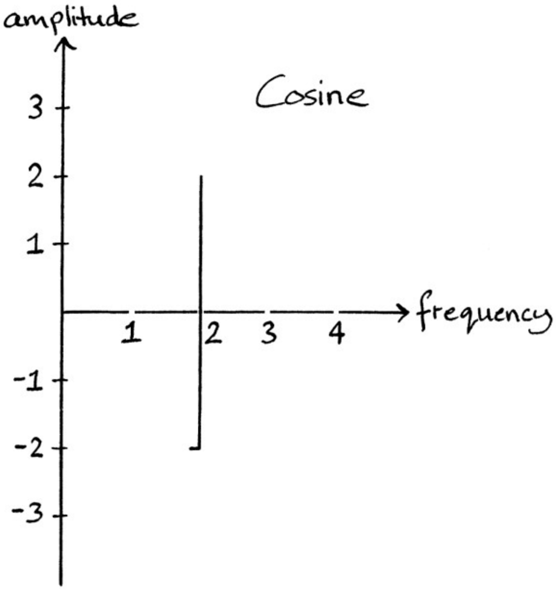
The phase point is in the top half of the circle at an *x-axis* value of 1.7321 units, so on the “top view” or “Cosine perspective” frequency-phase graph, the tick is on the *right-hand* side at a height of 1.7321 units:

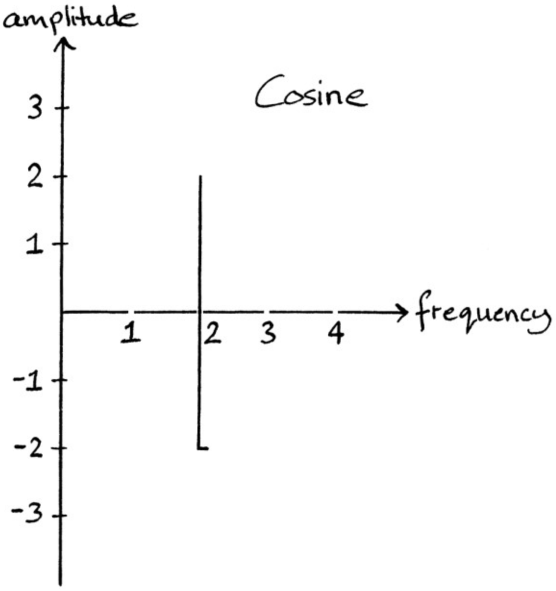


If the circle has a phase of 180 degrees, the circle looks like this:

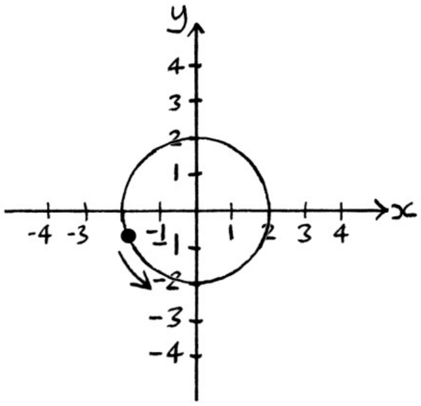


The phase point could be said to be in the top or bottom half of the circle. It has an *x-axis* value of −2 units. The graph could have the tick on the left hand side or the right hand side and have the same meaning:

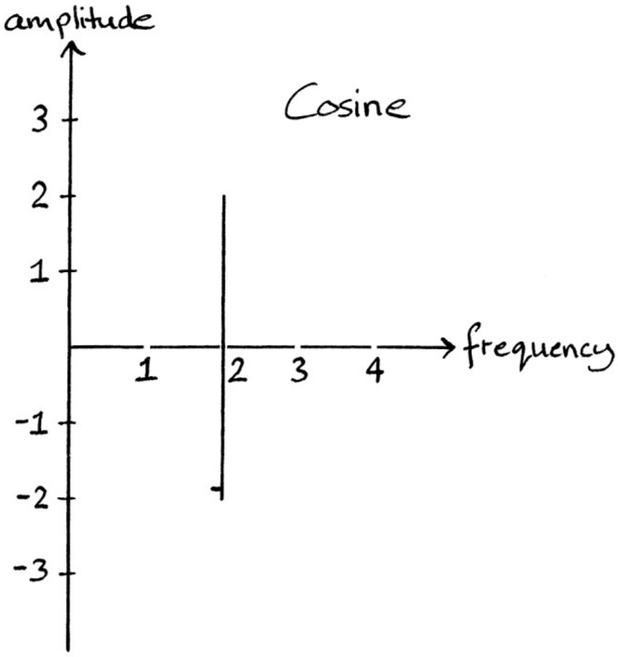




If the circle has a phase of 200 degrees, it will look like this:



The phase point is in the bottom half of the circle, so the tick will be on the left-hand side. The *x-axis* value of the phase point is −1.8794 units, so the tick will be at that height on the graph:



To read the angles that the ticks represent, it would be necessary to measure their position on the line, and then use inverse Sine or Cosine to find the angle.

In practice, you will seldom see graphs drawn this way, but it is a reasonably good way of demonstrating the phase of a wave in the frequency domain without confusing the graph too much.

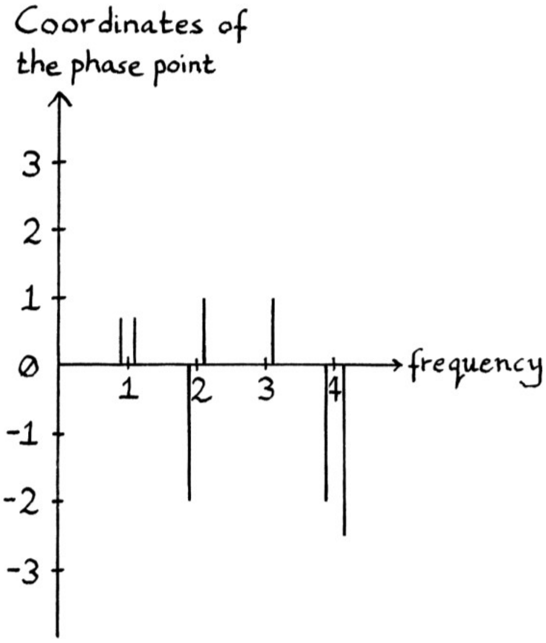
Phase point coordinate pairs

One type of graph you will sometimes see for showing phase is more complicated to understand. This graph has two lines for each existing frequency entry. This type of graph can be useful for explanations, but because there are lines on the graph that are not positioned exactly at the correct frequency, but instead either side of it, the graph is inherently inaccurate. However, it is still useful as long as it is recognised for what it is.

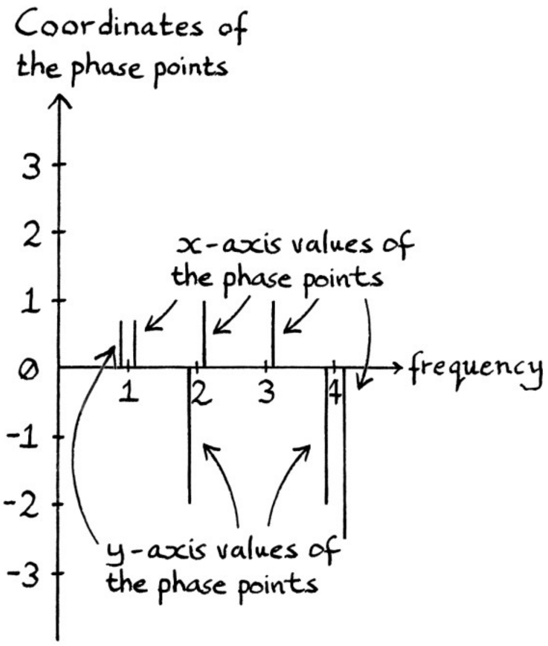
There are two ways to think about this graph. In this section, I will treat it as showing the coordinates of the phase points of the circles, which is the most intuitive way to think about it, given everything we have seen in this book so far. In the next section, I will treat it as showing other information, but both ways amount to the same thing.

The easiest way to understand the graph is as follows: the first vertical line of each pair indicates the y-axis value of the phase point of the circle, and the second line indicates the x-axis value of the phase point. [Which way around the lines are placed is arbitrary, but I will do it this way]. This type of graph allows us to calculate the radius of the circle and angle of the phase point.

The y-axis on this type of graph, *as I am showing it here*, should really be called "the axis that shows the y-axis value of the circle's phase point if we are looking at the first line in each pair, and the x-axis value of the circle's phase point if we are looking at the second line in each pair". I will give it the more concise name of the "Coordinates of the phase point" axis. An example graph looks like this:



The line on the *left-hand* side of each pair shows the y-axis position of the phase point of the circle for the frequency in the middle of the lines. The line on the *right-hand* side of each pair shows the x-axis position of the phase point on the circle for the frequency in the middle of each pair. Each pair of lines is centred around the actual frequency value assigned to them. Therefore, the first pair of lines in the above graph is applied to the frequency of one cycle per second, and not to the frequencies just above and below that, despite how the graph looks. The same graph with notes added to it looks like this:



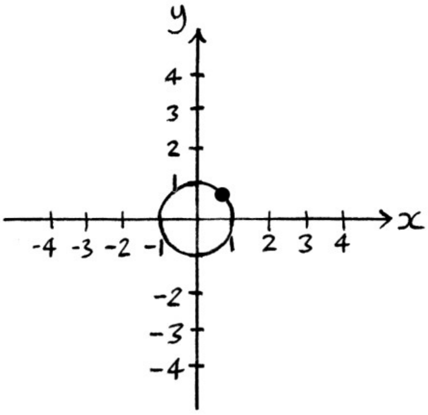
[Note that the graph still represents a sum of circles – the circles represented by each *pair* of lines are added together to create a signal.]

We will look at the two lines at the frequency of 1 cycle per second. They are both 0.7071 units. This means that the y-axis value of the phase point on the circle is 0.7071 units, and the x-axis value of the phase point is also 0.7071 units. Knowing where the phase point is situated allows us to calculate the radius of the circle, as the circle's perimeter must pass through that point. We can also calculate the angle of the phase point from the origin of the axes. [There is the assumption that all the waves have zero mean level, so the circle is centred on the origin of the axes].

We can use Pythagoras's theorem to calculate the radius of the circle. It is the square root of $0.7071^2 + 0.7071^2 = 1$ unit. Therefore, the circle has a radius of 1 unit, and the waves derived from that circle both have an amplitude of 1 unit.

We can use arctan to calculate the angle of the phase point. The angle is arctan (0.7071 ÷ 0.7071) = arctan 1 = 45 degrees. [As we are using arctan, we need to check that this is the angle we want and not the other angle that would produce the same gradient (45 + 180 = 225 degrees). As the phase point is in the top right quarter of the circle, we can tell that 45 degrees is the correct answer.]

Therefore, the circle has a radius of 1 unit, and the phase point is at 45 degrees. Such a circle looks like this:



This circle represents the two waves:
“y = 1 sin ((360 * 1t) + 45)”
... and:
“y = 1 cos ((360 * 1t) + 45)”.

Now, we will look at the lines at the frequency of 2 cycles per second. The first line, which represents the y-axis of the phase point, has a length of −2 units; the second line, which represents the x-axis of the phase point, has a length of +1 unit. This means that the phase point of the circle that these lines represent is at the coordinates (1, −2). Pythagoras’s theorem will show that this point is 2.2361 units from the origin of the axes. Therefore, the circle has a radius of 2.2361 units, and its derived waves have amplitudes of 2.2361 units. The angle of this point is arctan (−2 ÷ 1) = arctan −2 = −63.4349 degrees. We will convert this to a positive angle, and we have −63.4349 + 360 = 296.5651 degrees. From the coordinates of the phase point, we know it is in the bottom right-hand quarter of the circle. Therefore, this is the angle that we want, and not the other angle that would produce the same gradient (296.5651 – 180 = 116.5651 degrees). Therefore, the angle of the phase point is 296.5651 degrees, and the phases of the derived waves in their formulas are also 296.5651 degrees.

The circle represents the two waves:
“y = 2.2361 sin ((360 * 2t) + 296.5651)”
... and:
“y = 2.2361 cos ((360 * 2t) + 296.5651)”.

Now we will look at the lines at the frequency of 3 cycles per second. The y-axis line has a length of zero units; the x-axis line has a length of 1 unit. The coordinates of the phase point on the circle will be (1, 0). Thinking of the circle, we know that it will have a radius of 1 unit, and a phase of zero degrees. However, we will use Pythagoras’s theorem to work out the radius. It will end up as the square root of 1, which is 1. To calculate the angle, we need to calculate arctan (0 ÷ 1). Given that this is an arctan calculation with a zero in it, we do not need to use arctan, and we can work out the result by thinking of the circle. However, the arctan of (0 ÷ 1) = arctan 0 = 0 degrees. [This is the result we want, and not 0 + 180 = 180 degrees because the phase point is in the right hand half of the circle]. As we could tell before, the circle has a radius of 1 unit and a phase of 0 degrees.

The circle represents the two waves:
“y = 1 sin ((360 * 3t) + 0)”
... and:
“y = 1 cos ((360 * 3t) + 0)”

Now we will look at the two lines at the frequency of 4 cycles per second. The y-axis line is at −2 units; the x-axis line is at −2.5 units. This means the phase point on the circle is at the coordinates of (−2.5, −2). Pythagoras’s theorem gives us the radius as 3.2016 units. We use arctan to calculate the angle of the phase point: arctan (−2 ÷ −2.5) = 38.6598 degrees. Thinking about the position of the phase point on the circle, which is in the bottom left-hand quarter of the circle, we know that the result we want is 38.6598 + 180 = 218.6598 degrees (which has the same gradient as 38.6598 degrees). Therefore, our circle has a radius of 3.2016 units, and the derived waves have amplitudes of 3.2016 units. The phase point is at 218.6598 degrees on the circle, so the derived waves both have phases of +218.6598 degrees in their formulas.

The circle represents the two waves:
“y = 3.2016 sin ((360 * 4t) + 218.6598)”
... and:
“y = 3.2016 cos ((360 * 4t) + 218.6598)”

Two zero-phase wave pairs

We will now look at another type of graph, which is *identical* in layout to the "phase point coordinates" graph, but the interpretation of what it means is different. It is, in essence, exactly the same graph. This interpretation of what it means is the way that most people will think of the graph, and the concept behind it raises important ideas about circles and waves with phase.

The type of graph relies on how any Sine wave with a non-zero phase can also be represented as a zero-phase Sine wave added to a zero-phase Cosine wave. Similarly, any Cosine wave with a non-zero phase can be represented as a zero-phase Sine wave added to a zero-phase Cosine wave. The zero-phase waves might have negative amplitudes. We can express the idea mathematically:
"y = A sin ((360 * ft) + ϕ)"
... is equal to:
"y = B sin (360 * ft) + C cos (360 * ft)
... where "B" and "C" are the amplitudes of the zero-phase waves, and we would need to calculate their values. [We will see how to do this later in this section.]

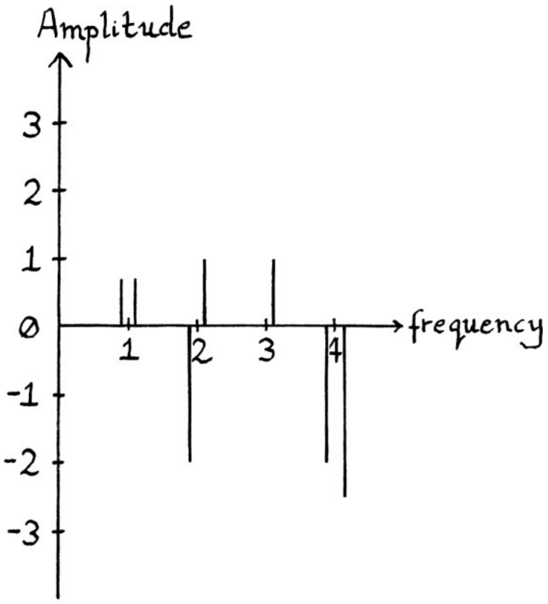
In this type of graph, the pairs of vertical lines represent the *amplitudes* of first a Cosine wave, and then a Sine wave. The amplitudes can be positive or negative on the graph. Both waves have zero phases. The frequencies of the waves are indicated by the point on the frequency axis between the two lines.

The y-axis of the graph is labelled "amplitude". In the last section, we were considering the first line of each pair as showing the vertical axis position of the phase point, but now we are treating it as showing the *amplitude* of the horizontal wave (the Cosine wave). In the last interpretation, we were treating the second line of each pair as being the horizontal axis position of the phase point, but now we are treating it as showing the *amplitude* of the vertical wave (the Sine wave). In a way, the meaning of the two lines of each pair has swapped around.

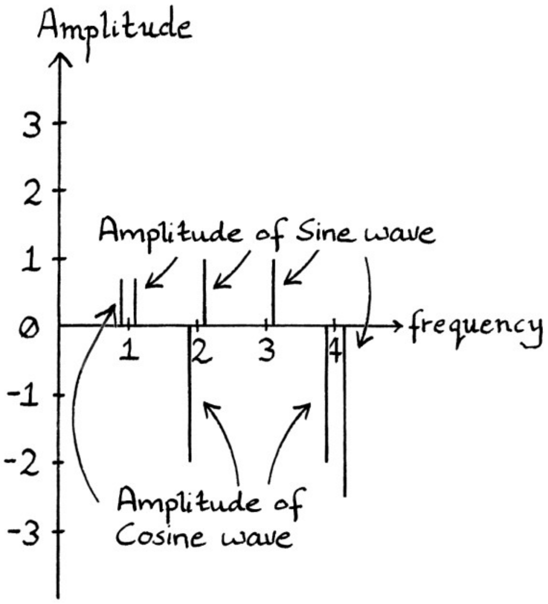
Note that I said that the vertical lines represent the amplitudes of the waves and that both waves have zero phases. In practice, sometimes the lines go downwards, which means they have negative amplitudes. Either we can think of them as having negative amplitudes, or we can think of them as having positive amplitudes and phases of 180 degrees. This is a situation where using negative amplitudes makes things simpler. Given all of this, we could describe the graph in two different ways (which mean exactly the same thing):

- If we want to use the idea of negative amplitudes, the lines represent waves with no phase, and positive and negative amplitudes. A negative amplitude is represented by a line pointing downwards.
- If we do not want to use the idea of negative amplitudes, then the lines represent waves with phases of 0 or 180 degrees, and positive amplitudes. A phase of 180 degrees is represented by a line pointing downwards.

A graph showing exactly the same details as in the previous section is as so:



The graph is identical to the one from the previous section, but the y-axis is now labelled "amplitude". The graph still represents a sum of waves, but now the waves within each pair are added together. With more notes, the graph is as so:



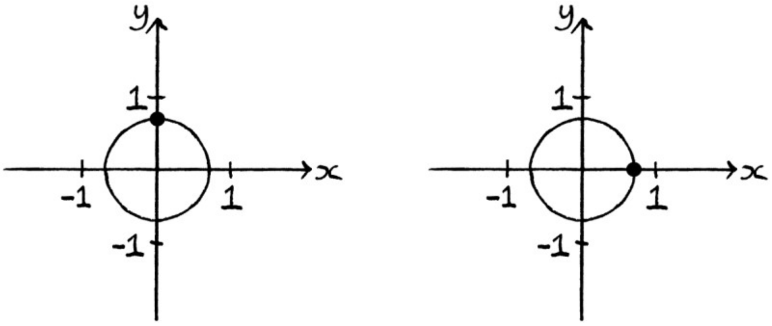
Now that the lines represent the amplitudes of Cosine and Sine waves (*with zero phases*) for a particular frequency, we can find out the amplitude and phase of the wave or circle that they represent by adding them together.

First, we will look at the two lines at the frequency of 1 cycle per second. Both lines are 0.7071 units long. This means that the Cosine wave represented by the first line has an amplitude of 0.7071 units (and a phase of zero degrees), and the Sine wave that the second line represents also has an amplitude of 0.7071 units (and a phase of zero degrees).

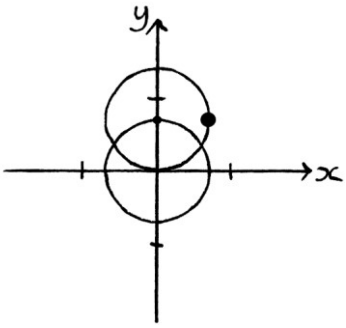
We could add them as waves, but it is more intuitive to add them as circles.

We first need to convert both waves into the same type – the Cosine wave therefore becomes a Sine wave with a +90 degree phase. We can then draw the circles. The first circle has a radius of 0.7071 units and a phase of 90 degrees (because it represents the Cosine wave). The second circle has a radius of 0.7071 units and a phase of zero degrees.

The circles look like this:



Arranged for adding together, at t = 0, they look like this:



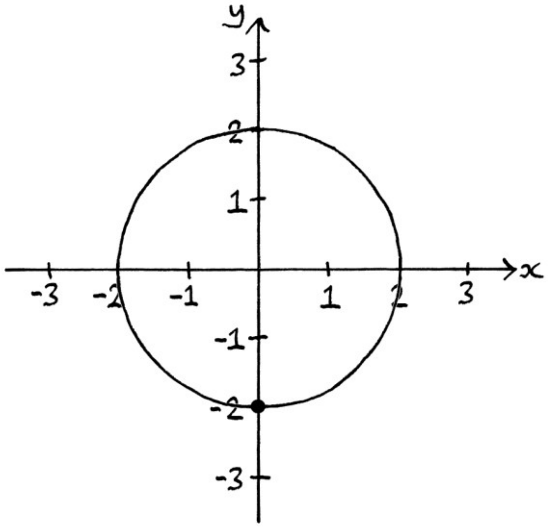
To calculate the radius of the resulting circle, we calculate the distance of the outer circle's phase point from the origin of the axes. This will be the square root of $0.7071^2 + 0.7071^2 = 1$ unit. The angle of the outer circle's phase point from the origin of the axes will be arctan (0.7071 ÷ 0.7071) = arctan 1 = 45 degrees. [We check to see if arctan has given us the angle we want, which it has]. Therefore, the resulting circle has a radius of 1 unit and a phase point at 45 degrees. The derived waves will have amplitudes of 1 unit and phases of +45 degrees in their formulas. Their frequencies will be the same as that of the resulting circle, which is the same as that of the two added circles, which is the same as the point on the frequency axis between the two lines on the original graph – in other words, it is 1 cycle per second. The formulas for the two waves derived from the resulting circle will be:
"y = 1 sin ((360 * 1t) + 45)"
... and:
"y = 1 cos ((360 * 1t) + 45)"

[Generally, for this type of graph, only one of these waves would be treated as the required answer, instead of giving the result as a circle or two corresponding waves.]

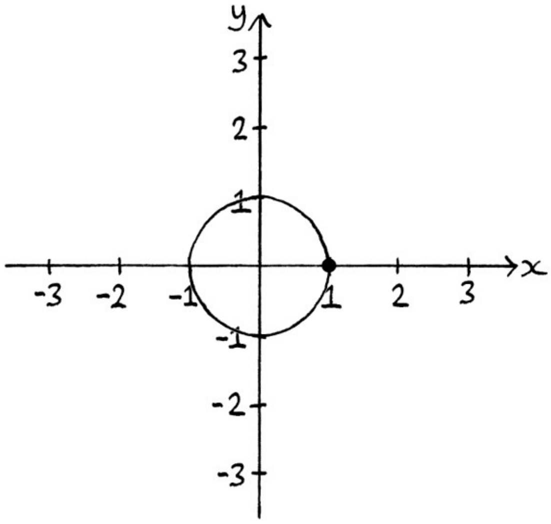
We have arrived at exactly the same answer as when we treated the pairs of vertical lines as representing the coordinates of phase points of circles. We have also used the same basic method to achieve the answer (although via a slightly more convoluted path). The graph is the same graph, but with different labels, and a different interpretation, but the same methods are used to calculate the circles or waves that it represents.

Now we will look at the pair of lines at 2 cycles per second. The first line, which represents the amplitude of a Cosine wave with zero phase, has a length of −2 units; the second line, which represents the amplitude of a Sine wave with zero phase, has a length of +1 unit.

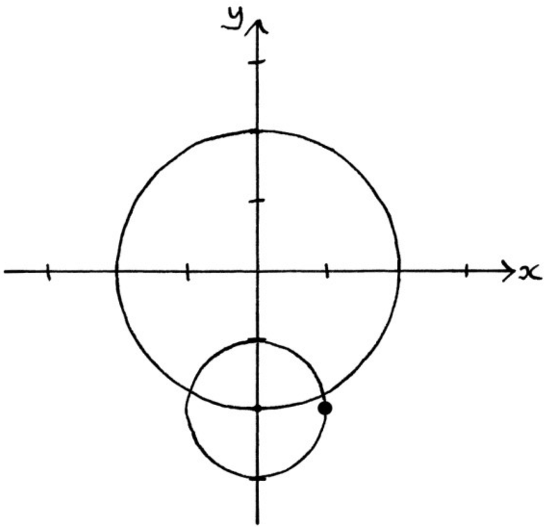
To turn these into circles of the same type, we need to treat the Cosine wave as a Sine wave with a phase of +90 degrees. The Cosine wave here has a negative amplitude, which, in our way of thinking, has to be adjusted to be a positive amplitude with 180 degrees added on to the phase. Therefore, the Cosine wave ends up as a circle with a radius of +2 units, and a phase of 90 + 180 = 270 degrees:



The circle for the Sine wave has a radius of 1 unit and a phase of zero degrees:



We join the circles together for adding, and they look like this at t = 0:



We calculate the amplitude of the resulting circle by seeing how far away the outer phase point is from the origin of the axes at t = 0 using Pythagoras's theorem. It is the square root of $1^2 + -2^2 = 2.2361$ units. The angle of the outer object from the origin of the axes is arctan $(-2 \div 1) = -63.4349$ degrees, which is the same as $-63.4349 + 360 = 296.5651$ degrees. [As the outer circle's phase point is in the bottom right hand quarter of the circle chart, this is the result we want.]

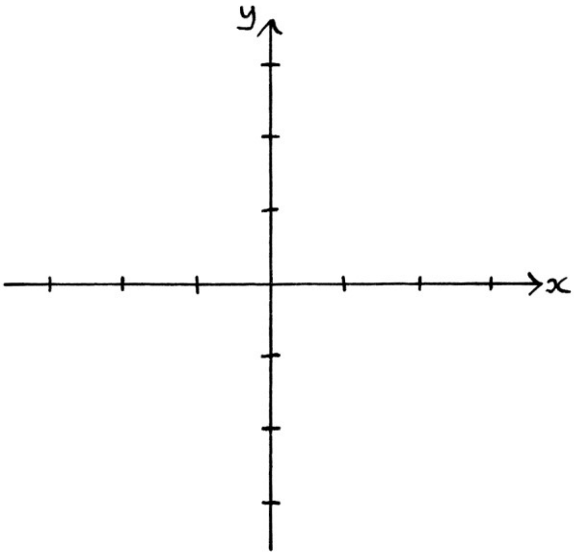
The circle has a radius of 2.2361 units and a phase point at 296.5651 degrees. The derived waves have amplitudes of 2.2361 units and phases of +296.5651 degrees in their formulas. This is the same result as when we treated the graph as showing the coordinates of the phase point in the last section.

The resulting circle represents the two waves:
"y = 2.2361 sin ((360 * 2t) + 296.5651)"
... and:
"y = 2.2361 cos ((360 * 2t) + 296.5651)"

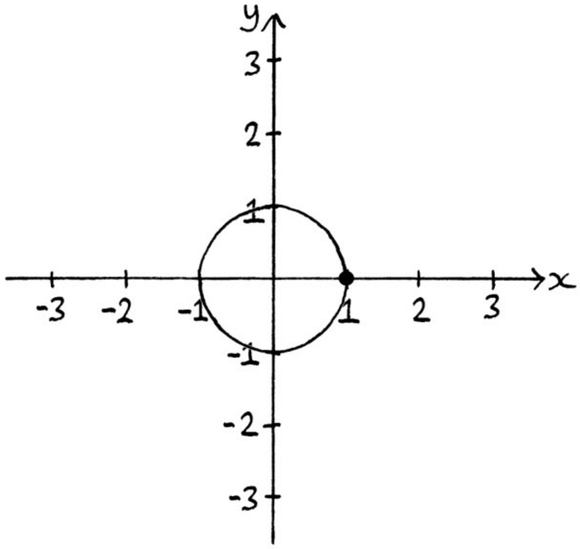
These are the same as when we calculated them with the first interpretation.

Now, we will look at the pair of lines at 3 cycles per second. The first line (the Cosine line) has a length of 0 units; the second line (the Sine line) has a length of +1 unit. We will turn these into circles.

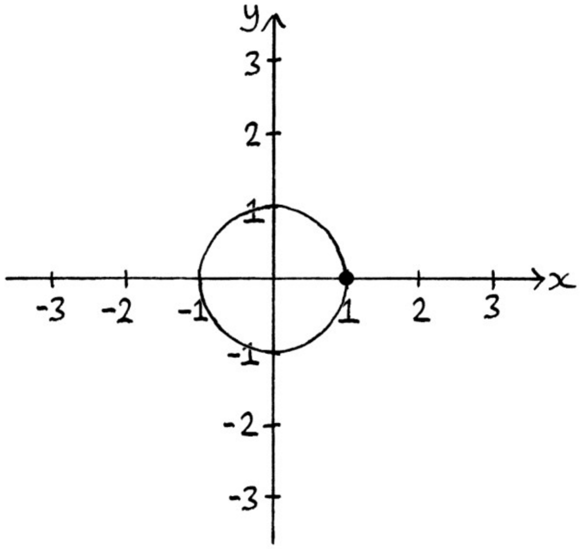
The circle for Cosine will have a phase of 90 degrees and an amplitude of zero units. Such a circle is really no circle at all:



The circle for Sine has a phase of 0 degrees and an amplitude of 1 unit:



If we add them together, we end up with a circle with an amplitude of 1 unit and a phase of zero degrees:

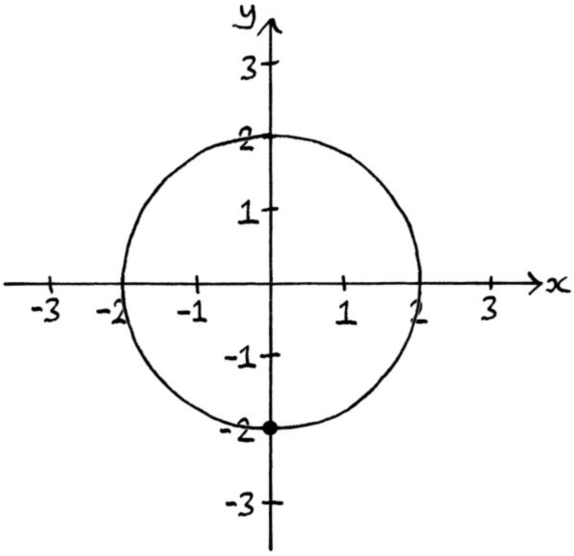


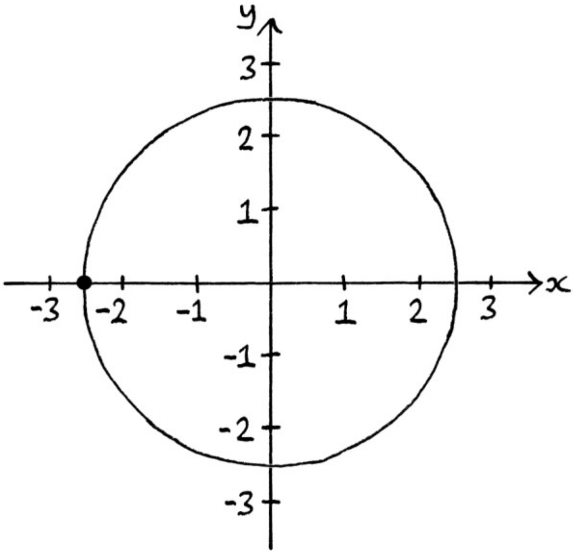
We do not need to do any maths to calculate the amplitude and phase for this circle.

The resulting circle represents the waves:
“y = 1 sin ((360 * 3t) + 0)”
... and:
“y = 1 cos ((360 * 3t) + 0)”

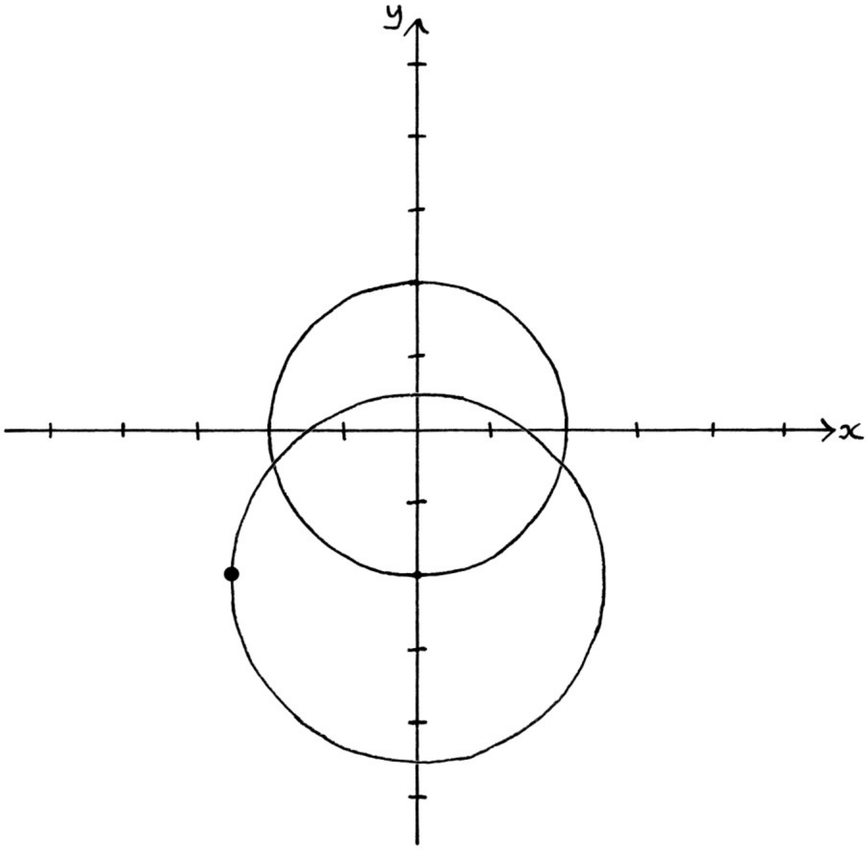
This is the same result as before.

Next, we will look at the pair of lines at 4 cycles per second. The first line (the Cosine line) has a length of −2 units; the second line (the Sine line) has a length of −2.5 units. To convert these into the same type of circle, we will give them both in terms of Sine, so the first line is a Sine wave with a 90-degree phase, and the second line is a Sine wave with a zero degree phase. We turn them into circles, which means we have to turn both the negative amplitudes into positive amplitudes, so we add 180 degrees to the phase points of each circle. We end up with the first circle having a radius of 2 units and a phase of 270 degrees, and the second circle having a radius of 2.5 units and a phase of 180 degrees:





At t = 0, the two joined circles look like this:



We can use Pythagoras's theorem to find the distance of the outer circle's phase point from the origin of the axes. We need the square root of $(-2)^2 + (-2.5)^2$, which is 3.2016 units. We use arctan to find its angle: arctan (−2 ÷ −2.5) = 38.6598 degrees. We check the circle to see if this is the result we want from arctan. As the outer circle's phase point is in the bottom left hand quarter of the circle, it is not the correct result, so we want the other angle that would produce that gradient. We add 180 degrees to end up with 218.6598 degrees. Therefore, the circle has an amplitude of 3.2016 units and a phase of 218.6598 degrees. This is the same result as we had in the last section.

The resulting circle represents the two waves:
"y = 3.2016 sin ((360 * 4t) + 218.6598)"
... and:
"y = 3.2016 cos ((360 * 4t) + 218.6598)"

Again, these are the same results as before.

Thoughts so far

As we can see, it does not matter whether we treat the pairs of lines on the graph as showing the “y” and “x” coordinates of the phase points, or as showing the positive or negative amplitudes of Cosine and Sine waves with zero phase. We end up with the same resulting circle for each pair of lines, and the method to achieve the results is identical.

The reason that the graph works in the same way if we consider the lines to show the coordinates of the phase points, or if we consider the lines to show the amplitudes of a Sine and Cosine wave, is reasonably straightforward.

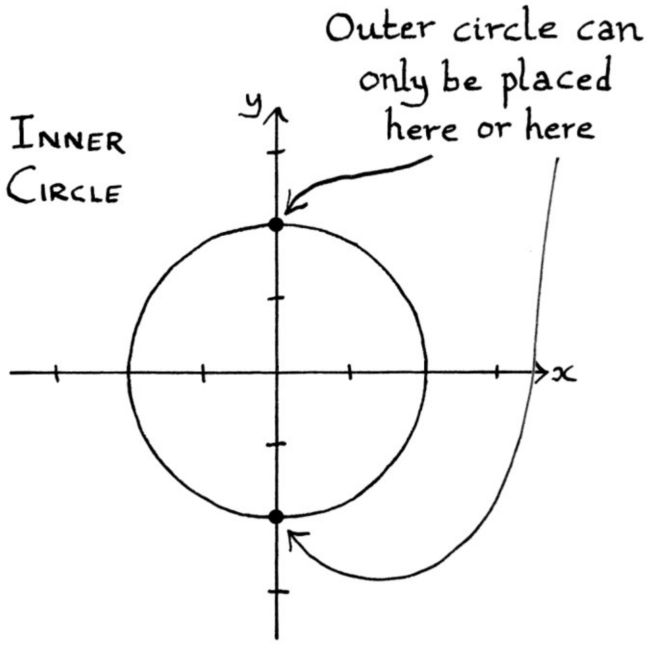
If we let the lines represent Sine and Cosine waves, then either they have no phase (if the amplitude is positive) or they have 180 degrees added on to their phase (if the amplitude is negative).

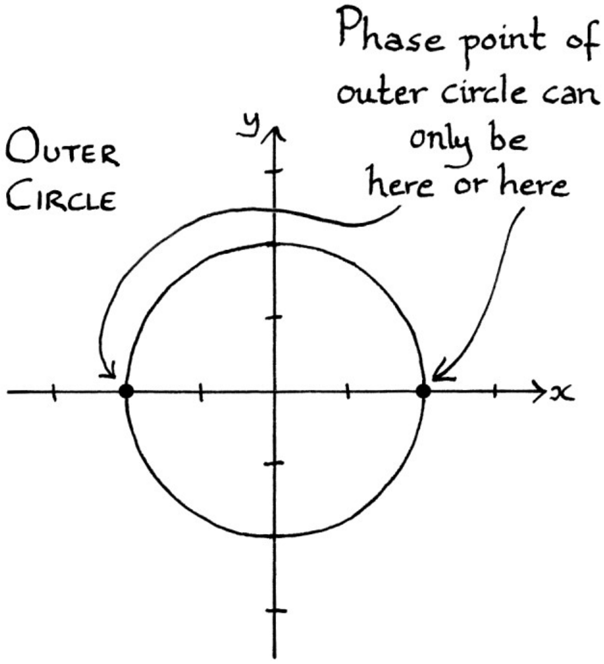
If we turn the Cosine waves into Sine waves to make them compatible, then we add 90 degrees to the phase. Therefore, those waves will end up having a phase of either 90 degrees or 270 degrees. The phases cannot be anything else. The circles that represent those former Cosine waves will have a phase point that is at a y-axis position either equal to the amplitude of the Cosine wave, or equal to the negative of that amplitude.

The original Sine waves can only have a phase of 0 degrees or 180 degrees. Therefore, the phase point of the circle that represents them can only ever be at an x-axis position equal to the amplitude of the Sine wave, or its negative.

When adding circles together, with the circle that represents the Cosine wave innermost, and the circle that represents the Sine wave outermost, the centre of the outer circle is placed over the phase point of the inner circle. Therefore, either it will be placed at exactly 90 degrees or it will be placed exactly at 270 degrees. There is nowhere else it can be placed. The centre of the outer circle will be along the y-axis, and on the circumference of the inner circle. Its x-axis coordinate will be 0; its y-axis coordinate will either be equal to the amplitude of the inner circle, or the negative of that amplitude.

As the outer circle has a phase of 0 or 180 degrees, the phase point of the outer circle, when it is placed on the inner circle, must have the y-axis coordinate of that of the inner circle’s phase point. The x-axis position of the outer circle’s phase point can only ever be at 0 or 180 degrees to the inner circle’s phase point, so it has to be equal to either the amplitude of the outer circle's amplitude, or the negative of its amplitude.





The result of all this is that the y-axis value of the outer circle's phase point will be equal to the first line on the frequency graph, and the x-axis value of the outer circle's phase point will be equal to the second line on the frequency graph. In this way, the first line on the frequency graph shows the y-axis value of the phase point of the outer circle, which is identical to the amplitude of the Cosine wave. The second line on the frequency graph shows the x-axis value of the phase point of the outer circle, which is identical to the amplitude of the Sine wave.

The amplitude of the Sine wave is the x-axis of the phase point; the amplitude of the Cosine wave is the y-axis of the phase point.

It is worth drawing such a frequency domain graph and some circles for yourself to become used to how this all works.

Significance

What is happening in the "two zero-phase circle pairs" interpretation of this type of graph is fairly significant. It demonstrates that we can represent a wave with any phase as being the sum of a Cosine wave with zero phase and a Sine wave with zero phase (or at least, with zero phases if we use negative amplitudes).

The idea could be rephrased to say that we can represent a wave with any phase as being the sum of a Sine wave with zero phase, and a Sine wave with a 90 degree phase (if we use negative amplitudes).

This, again, could be rephrased to say that we can represent a wave with any phase as being the sum of a Sine wave with a positive amplitude and a phase of 0 or 180 degrees, and a Sine wave with a positive amplitude and a phase of 90 or 270 degrees.

We could also say that we can represent a *circle* with any phase as being the sum of two circles – one with a phase of 0 or 180 degrees, and the other with a phase of 90 or 270 degrees.

The idea of turning a wave with a non-zero phase into a Sine wave with zero phase and a Cosine wave with zero phase is often used in the analysis of waves. Frequently, the process of reducing a signal to its constituent parts has a step where a constituent wave is shown to be the sum of a Sine wave with zero phase and a Cosine wave with zero phase (and possibly negative amplitudes). Often, the actual finished result will be given in terms of the sum of a Sine wave with zero phase and a Cosine wave with zero phase, instead of taking it one step further to produce the actual constituent wave. This is where you will see this type of graph the most.

Previously in this book, I have said that any periodic signal that is not a pure wave is, or approximately is, the sum of two or more pure waves of different frequencies, and possibly different phases and amplitudes. The graph in this section shows that we can rephrase that statement to be:

> "Any periodic signal that is not a pure wave is, or approximately is, the sum of two or more *pairs* of waves, where the different pairs will be of different frequencies, but the waves *within* each pair will be of the same frequency, maybe different amplitudes (which might be positive or negative), and will be a Sine wave with zero phase and a Cosine wave with zero phase."

The new statement is not as succinct as the previous statement, and in practice, the use of the knowledge contained in the new statement might not save any time. However, in everyday signal processing, the idea of a wave with a phase being the sum of a Cosine wave and a Sine wave with zero phases (but possibly negative amplitudes) is one that you will often see.

Thoughts on the two graph types

It is significant that the "phase point coordinate" form of the graph and the "two zero-phase wave" form of the graph show the same information but with different interpretations.

A deduction that we can make from the idea is that if we know the coordinates of the phase point of a circle with any phase, we instantly know the radiuses of the two "zero-phase" circles that when added would result in that circle. [Technically, they are not zero phase – one of the circles will have a phase of 90 or 270 degrees; the other will have a phase of 0 or 180 degrees]. It also means that we instantly know the amplitudes of the two pairs of zero-phase waves that would make up the single pair of any-phase waves derived from the original circle. [Technically, these waves are only zero-phase if we are using negative amplitudes].

To find the radiuses of the two "zero-phase" circles, we just have to swap the coordinates around. For example, if the phase point is at the coordinates (2, 4) on the original circle, then we know that the radius of the first "zero-phase" circle will be 4 units, and the radius of the second "zero-phase" circle will be 2 units. The first circle will have its phase point at 90 degrees, and the second circle will have its phase point at 0 degrees. The frequencies of the circles will be the same. Adding those two circles together will result in the original circle.

If we have a wave derived from a circle with a phase point at the coordinates:
(6, 11)
... then we instantly know that the wave is the sum of:

- a Cosine wave with zero phase and an amplitude of 11 units

... and:

- a Sine wave with zero phase and an amplitude of 6 units.

We can also say that it is the sum of:

- a Sine wave with an amplitude of 11 units and a phase of 90 degrees

... and:

- a Sine wave with an amplitude of 6 units and zero phase.

If we have a wave derived from a circle with a phase point at the coordinates:
(−5, −4)
... then we instantly know that that wave is the sum of:

- a Cosine wave with zero phase and an amplitude of −4 units

... and:

- a Sine wave with zero phase and an amplitude of −5 units.

We can also say that the wave is the sum of:

- a Cosine wave with an amplitude of 4 units and a phase of 180 degrees

... and:

- a Sine wave with an amplitude of 5 units and a phase of 180 degrees.

We can also say that the wave is the sum of:

- a Sine wave with an amplitude of 4 units and a phase of 270 degrees

... and:

- a Sine wave with an amplitude of 5 units and a phase of 180 degrees.

Calculating the zero-phase waves

As we have seen, any wave with any phase is equal to a zero-phase Sine wave added to a zero-phase Cosine wave. If we ignore circles, it is simple to find the zero-phase Sine and Cosine waves for any wave. The rule for Sine waves is as follows:

If we have the Sine wave:
“y = A sin ((360 * ft) + ϕ)”
... then it will be equal to:
“y = A cos (ϕ) * sin (360 * ft)”
... added to:
“y = A sin (ϕ) * cos (360 * ft)”

In other words, the amplitude of the zero-phase Sine wave will be our original amplitude multiplied by the Cosine of the original phase. The amplitude of the zero-phase Cosine wave will be our original amplitude multiplied by the Sine of the original phase. The frequencies will be the same as that of our original wave. The phases will be zero. If our original wave had a mean level, then the mean level can be placed with either wave or split between them – it will not make any difference. It is worth noting that the amplitudes of the zero-phase waves might be negative, and one of them might be zero.

As an example, we will calculate the zero-phase Sine wave and Cosine wave that are equal to:
“y = 3.3 sin ((360 * 11t) + 22)”

The zero-phase Sine wave will have the formula:
“y = 3.3 cos (22) * sin (360 * 11t)”
... which is:
“y = 3.0597 sin (360 * 11t)”

The zero-phase Cosine wave will have the formula:
“y = 3.3 sin (22) * cos (360 * 11t)”
... which is:
“y = 1.2362 cos (360 * 11t)”

Therefore, we can say that:
“y = 3.3 sin ((360 * 11t) + 22)”
... is equal to:
“y = 3.0597 sin (360 * 11t)”
... added to:
“y = 1.2362 cos (360 * 11t)”

We could also phrase that as:
3.3 sin ((360 * 11t) + 22) = 3.0597 sin (360 * 11t) + 1.2362 cos (360 * 11t)

To convert a Cosine wave to be the sum of a zero-phase Sine wave and Cosine wave, the rule is slightly different:

If we have the Cosine wave:
“y = A cos ((360 * ft) + ϕ)”
... then it will be equal to:
“y = A sin (−ϕ) * sin (360 * ft)”
... added to:
“y = A cos (−ϕ) * cos (360 * ft)”

In other words, the amplitude of the zero-phase Sine wave will be our original amplitude multiplied by the Sine of the negative of the original phase. The amplitude of the zero-phase Cosine wave will be our original amplitude multiplied by the Cosine of the negative of the original phase. The frequencies will be the same as that of our original wave. The phases will be zero. As before, if our original wave had a mean level, it can be added to either wave or split between them. The amplitudes of the zero-phase waves might be negative or zero. As the Cosine of a negative value is the same as the Cosine of that value made positive, we could give the Cosine wave part as “y = A cos (ϕ) * cos (360 * ft)”, but it is easier to remember the rule if both phases are negative.

As an example, we will convert:
“y = 17 cos ((360 * 4.123t) + 275)”
... into the sum of a zero-phase Sine wave and Cosine wave.

The Sine wave will be:
“y = 17 sin (−275) * sin (360 * 4.123t)”
... which is:
“y = 16.9353 sin (360 * 4.123t)”

The Cosine wave will be:
"y = 17 cos (−275) * cos (360 * 4.123t)"
... which is:
"y = 1.4816 cos (360 * 4.123t)"

Therefore, we can say that:
"y = 17 cos ((360 * 4.123t) + 275)"
... is equal to:
"y = 16.9353 sin (360 * 4.123t)"
... added to:
"y = 1.4816 cos (360 * 4.123t)"

We can also phrase that as:
17 cos ((360 * 4.123t) + 275) =
16.9353 sin (360 * 4.123t) + 1.4816 cos (360 * 4.123t)

As another example, for the wave:
"y = 2 cos ((360 * 1t) + 150)"
... the Sine wave will be:
"y = 2 sin (−150) * sin (360 * 1t)"
... which is:
"y = −1 sin (360 * 1t)"

The Cosine wave will be:
"y = 2 cos (−150) * cos (360 * 1t)"
... which is:
"y = −1.7321 cos (360 * 1t)"

Therefore, we can say that:
"y = 2 cos ((360 * 1t) + 150)"
... is equal to:
"y = −1 sin (360 * 1t)"
... added to:
"y = −1.7321 cos (360 * 1t)"

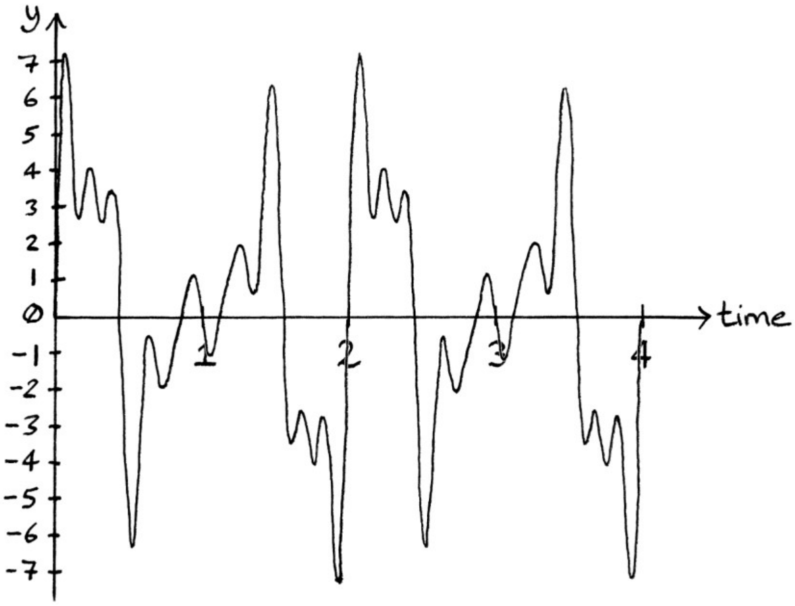
We can also phrase that as:
2 cos ((360 * 1t) + 150) = −1 sin (360 * 1t) + −1.7321 cos (360 * 1t)

[Note how the amplitudes of both zero-phase waves are negative. This is a situation where it is easiest to leave the amplitudes as negative values. We could rephrase each wave to have a positive amplitude, but then they would have non-zero phases, and the point of this particular exercise is to find the zero-phase waves that, when added, produce our original wave.]

Filtering out unwanted frequencies

If a signal is reduced to the individual waves that were added to create it, then it is possible to perform processes on the signal that would be more difficult or impossible otherwise. The process of filtering out unwanted frequencies can be done without reducing a signal to its constituent waves, but reducing it to its constituent waves makes it much easier to visualise what is happening, and allows us to filter frequencies by hand. The frequency domain graph is the best way for understanding the basic concept of filters. In this section, I will explain *what* filters do, but not the maths behind how they do it.

As an example, we will say that we have been given a signal made up of Sine waves of a constant frequency and amplitude, with zero phase and zero mean level. The signal looks like this:

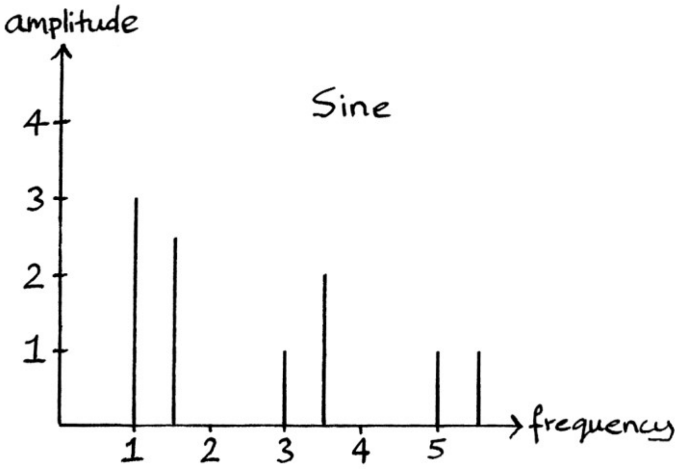


The signal repeats once every 2 seconds, so it has a frequency of 0.5 cycles per second. Given what we know about adding waves together, this means that the waves that were added together to make it must all have had frequencies that were multiples of 0.5 cycles per second.

After analysis, which I will explain in Chapter 18, we would discover that the waves that were added to create the signal are:

"y = 3 sin (360 * 1t)"
"y = 2.5 sin (360 * 1.5t)"
"y = 1 sin (360 * 3t)"
"y = 2 sin (360 * 3.5t)"
"y = 1 sin (360 * 5t)"
"y = 1 sin (360 * 5.5t)"

The frequency domain graph of the signal looks like this:



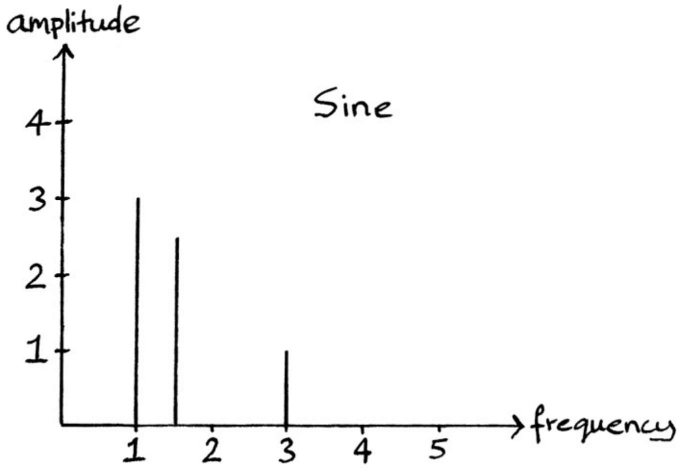
We will use this signal in the following examples.

Filters allow us to remove ranges of frequencies that we do not want. There are three main types of filters: low pass, band pass and high pass.

Low pass filter

A low pass filter allows waves of a frequency below a chosen value to *pass*, while blocking higher frequencies. The word "pass" is used in the name because a filter will usually be a part of a chain of processes that deal with a signal. The filter receives the signal, deals with it, and then *passes* the result on to the next stage. A low pass filter passes on the part of the signal made up of frequencies lower than a chosen amount on to the next stage.

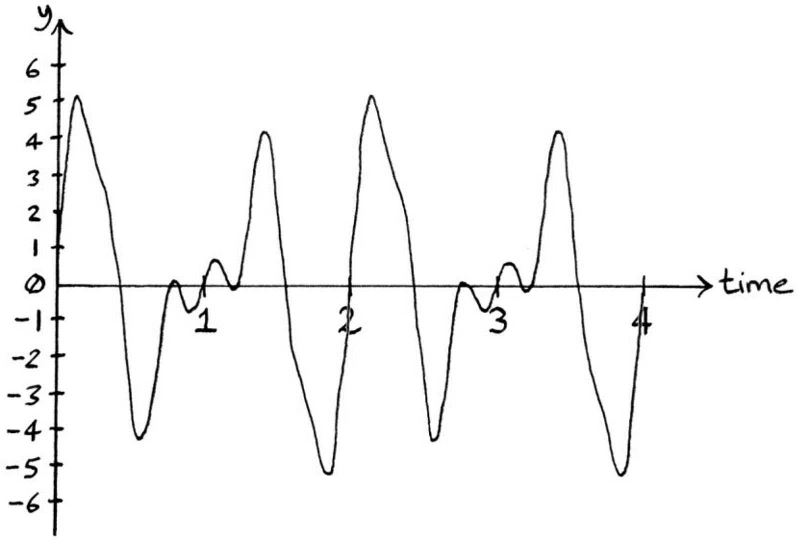
Supposing we had a low pass filter set at 3.25 cycles per second, then anything below that frequency would be allowed to pass on to the next stage, but anything above that frequency would be blocked. In this example, the "next stage" is really just the showing of the frequency domain graph after the filter has done its work. The frequency domain graph for our signal after a low pass filter at 3.25 cycles per second looks like this:



Every wave with a frequency higher than 3.25 cycles per second has been removed from the sum of waves. This means the signal after the filter now just contains the following waves:
"y = 3 sin (360 * 1t)"
"y = 2.5 sin (360 * 1.5t)"
"y = 1 sin (360 * 3t)"

All these waves are passed to the next stage of processing the signal (which in this case is just the drawing of the signal). The new signal that the frequency domain represents is a simpler signal because there are fewer waves making it up.

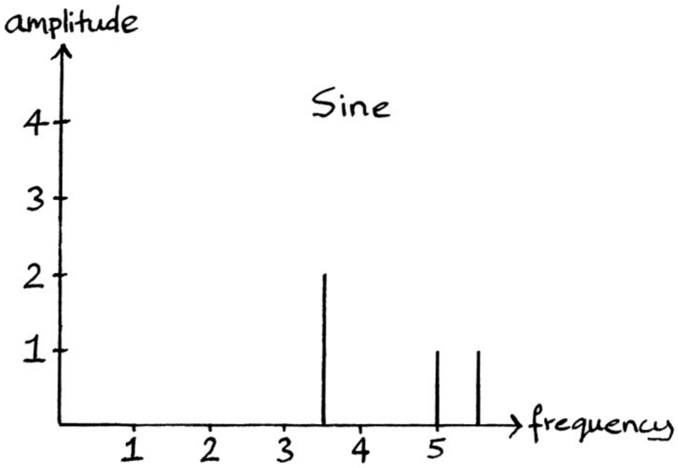
The new signal looks like this:



High pass filter

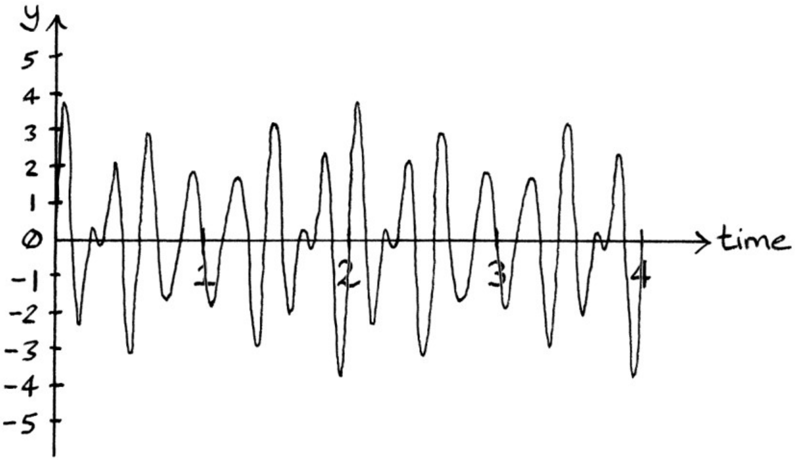
A high pass filter allows waves of a frequency *above* a chosen value to pass, while blocking *lower* frequencies.

Supposing we had a high pass filter set at 3.25 cycles per second, then anything above that frequency would be allowed to pass on to the next stage, but anything below that frequency would be blocked. The frequency domain graph after a high pass filter at 3.25 cycles per second will look like this:



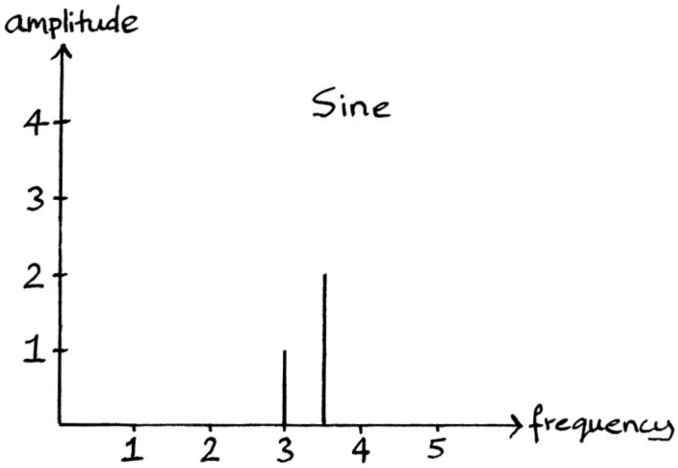
Everything *lower* than 3.25 cycles per second has been removed from the sum of waves. This means the signal after the filter will just contain the following waves:
"y = 2 sin (360 * 3.5t)"
"y = 1 sin (360 * 5t)"
"y = 1 sin (360 * 5.5t)"

After the filter has affected the signal, the signal will look like this:



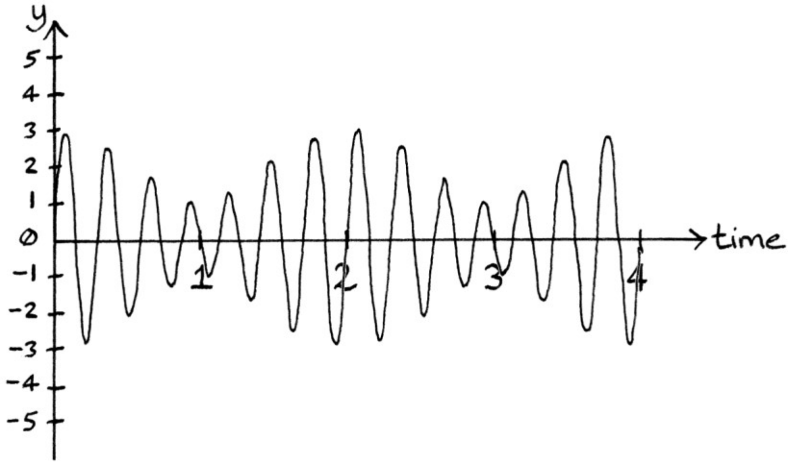
Band pass filter

A band pass filter allows waves of frequencies between two chosen values to pass, while blocking frequencies outside that range. We can think of it as a low pass filter and a high pass filter combined. Supposing we had a band pass filter to include the range of frequencies between 2 and 4 cycles per second, then anything between those frequencies would be allowed to pass though, but anything outside that range would be blocked. The frequency domain graph after such a band pass filter would look like this:



The resulting signal would consist of just these two waves:
“y = 1 sin (360 * 3t)”
“y = 2 sin (360 * 3.5t)”

... and would look like this:



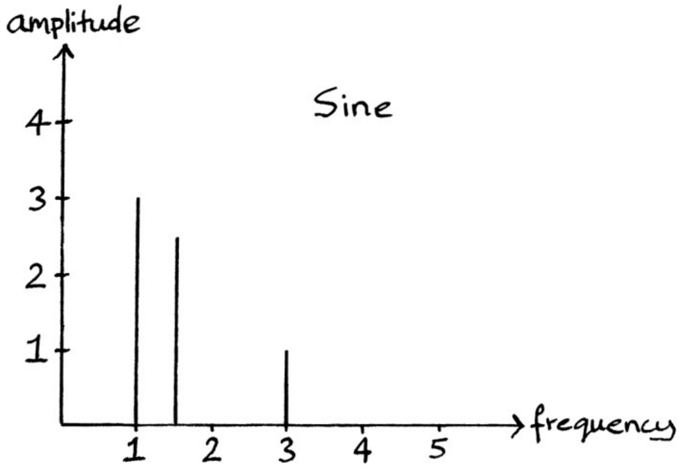
The opposite of a Band Pass filter is a “Band Reject filter”. A Band Reject filter blocks a chosen range of frequencies.

Filtering in practice

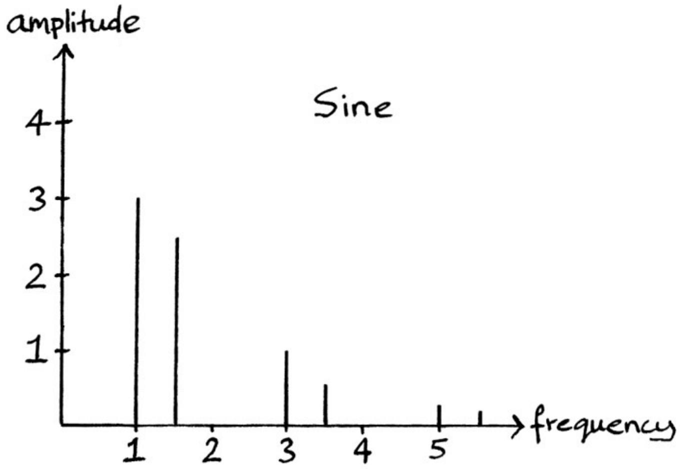
In practice, filtering can be a lot more complicated.

In the real world, it is rare that a signal would consist of just a handful of waves added together. In a sound or radio signal, there would be countless added waves, and it would take a lot of effort to work out exactly which ones they were. Therefore, filtering a signal by reducing it to its constituent waves and then removing the unwanted waves would take too long to be useful for most situations. If the signal is in the form of a discrete signal (in other words, one where the signal is treated as a series of values at evenly spaced moments in time, as discussed in a later chapter), then there are quicker, but less exact, methods of filtering. These methods do not reduce the signal to its constituent parts, but instead perform mathematical processes against the signal, which have the result of filtering it in particular ways. These kinds of filters have less abrupt cut-off points. Instead of a low pass filter blocking all waves above, say, 3.25 cycles per second, it might *reduce* the amplitude of waves just above 3.25 cycles per second, in a way that the much higher frequencies become reduced more.

For example, the graph of the low pass filter with a cut-off at 3.25 cycles per second from the previous example looked like this:



... but in the real world, a low pass filter might not be able to eliminate the frequencies immediately next to the cut-off point completely, and so the graph would end up as so:



Examples of filters

A radio antenna acts as a band pass filter – the design and length of the antenna let a particular range of frequencies reach the radio's circuitry, while reducing or eliminating those from outside that range.

The electronics inside a radio acts as a band pass filter, so that we can tune to a particular radio broadcast of interest.

A microphone acts as a band pass filter, in that the abilities of the microphone mean it is only receptive to a certain range of frequencies.

Human ears act as a band pass filter, in that the internal workings of the ear can only work with frequencies between certain frequencies. A sound signal enters the ears, but human ears can only pass on to the brain that part of the signal between about 20 hertz and 20,000 hertz. As people grow older, the maximum frequency reduces.

A wall can act as a low pass filter for sound. Lower frequency sounds (deeper sounds) can travel through more easily than higher frequency sounds.

We could say that distance is a low pass filter for sound. Being far away from the source of a radio transmitter or source of sound means that only the lower frequencies will reach us.

Away from the world of waves, we could say that a kitchen sieve is analogous to a low pass filter if we are sifting flour – the sieve blocks particles of a higher size. Those that pass through the sieve are passed on to the next stage of cooking.

Conclusion

The purpose of a frequency domain graph is to show the frequencies of all the waves that were added together to make up a signal.

Usually, the frequencies shown on the frequency domain graph would have been calculated by analysing a given signal. In practice, frequency domain graphs are used more as guides to the waves in one frequency band of the signal, than as completely accurate portrayals, and it is rare that they would be detailed enough to reconstruct the original signal exactly.

Portraying a signal in the frequency domain allows us to visualise and analyse aspects of it in ways that would be much harder to do in the time domain. As with so much to do with waves, different approaches to viewing them gives us more ways to deal with them.

Chapter 16: Multiplication with waves

Multiplying a wave by a value or by another wave is slightly more complicated than addition, but still straightforward. As with addition, it helps if we remember that we are multiplying every y-axis value from one wave by the corresponding y-axis value from the other wave for the same moment in time.

In this chapter, we will look at various multiplications, and we will find the underlying rules as the chapter progresses.

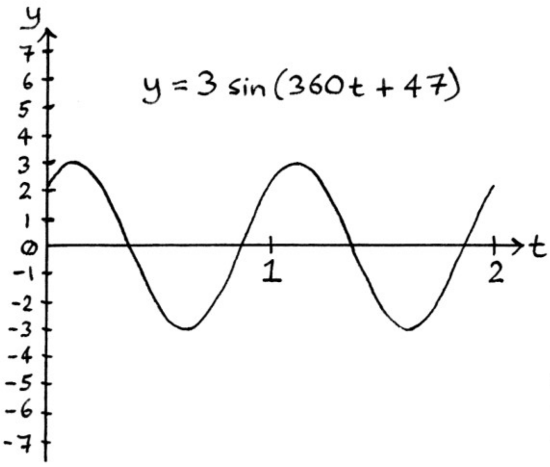
Multiplication by numbers

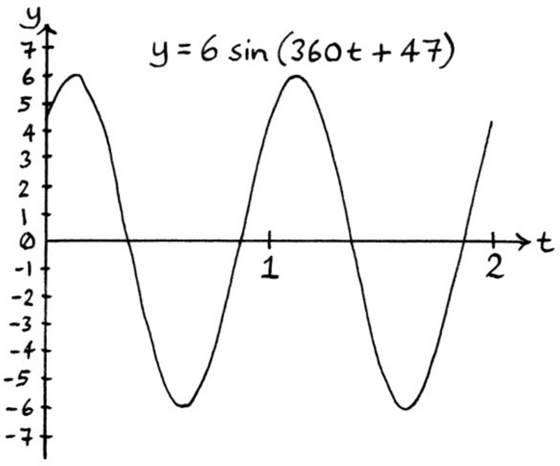
Multiplication of a wave by a number results in every y-axis value in that wave being multiplied by that number. Every value is scaled accordingly.

Zero mean level

For waves with zero mean level, the effect on the overall wave is that its amplitude is multiplied by the number, and therefore scaled. The frequency and phase remain the same.

For example, “y = 3 sin (360t + 47)” multiplied by 2 is “y = 6 sin (360t + 47)

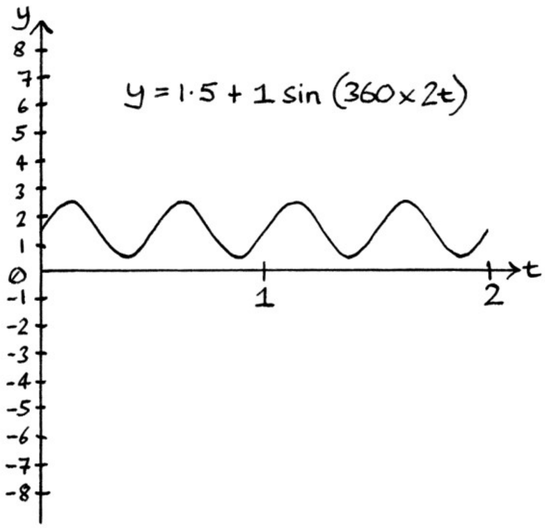


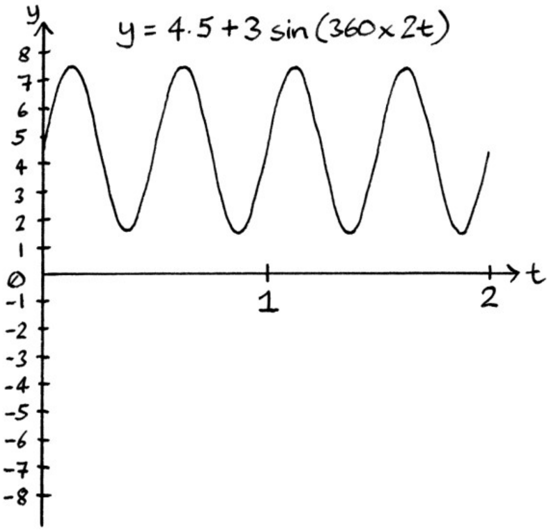


Non-zero mean level

If a wave has a non-zero mean level, then a multiplication by a value scales both the amplitude of the wave *and* its mean level.

For example:
“y = 1.5 + 1 sin (360 * 2t)”
... multiplied by 3 results in:
“3 * (1.5 + 1 sin (360 * 2t)”
... which is:
“4.5 + 3 sin (360 * 2t)”.



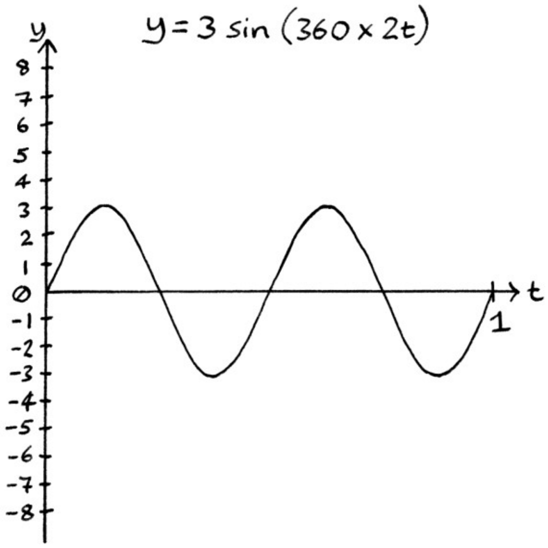


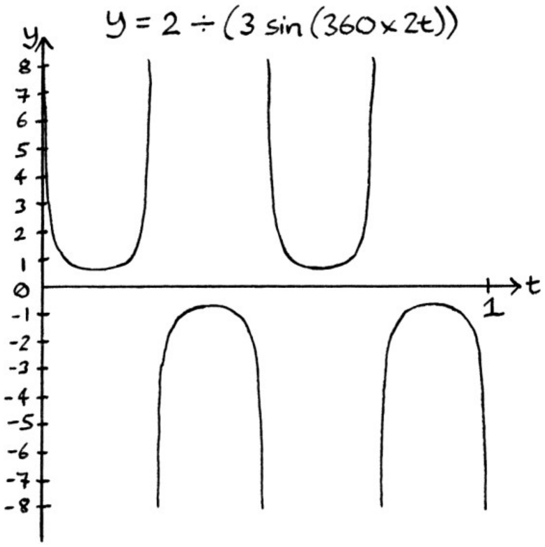
Division

Division is essentially the same as multiplication. If we divide a wave by 2, it is the same as multiplying it by 0.5.

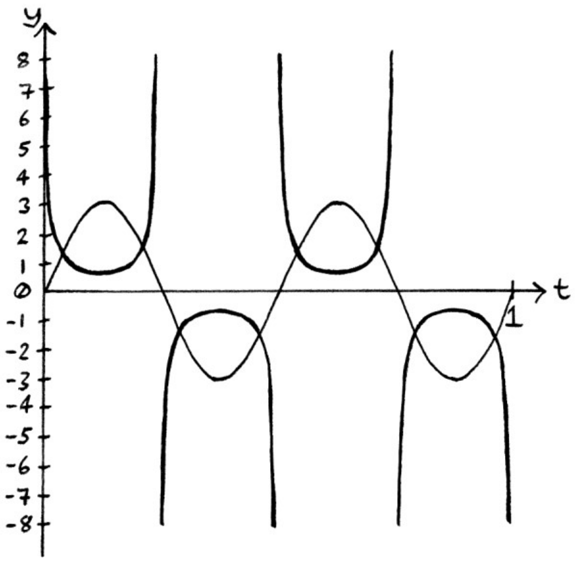
Things become more complicated when we divide a value by a wave. For example, if we divide 2 by “y = 3 sin (360 * 2t)”, then it is the same as taking the reciprocal of every y-axis value along the wave, and then multiplying each result by 2. [The reciprocal of a number is 1 divided by that number].

The result of doing this is not a wave at all:





The result has this shape because if we take the reciprocal of small numbers, the result is very high. For example 1 ÷ 0.001 is 1000, and 1 ÷ 0.0001 is 10,000. The reciprocal of 0 is infinitely high. [To emphasise how "infinity" is more of a concept than a number, some people would consider the result undefined]. Therefore, whenever the original wave's curve is near zero, the result of the division at that point becomes extremely high. This becomes clearer if we draw the graphs over each other as so:



Multiplication by waves

Things become more complicated when we multiply two waves together. It is easiest to visualise such a multiplication if you realise that the process just involves taking the corresponding y-axis values from each wave for every moment in time and multiplying them together. Given that, it pays to remember a few of the basic rules of everyday multiplication:

- If two values being multiplied together are positive, the result will be positive.
- If two values being multiplied together are both negative, the result will be positive.
- If one or both values are zero, the result will be zero.

Given that multiplication of a wave (with no mean level) by a *number* changes the *overall* amplitude, we can think of multiplication of one wave by a second wave as changing the instantaneous amplitude at every moment in time in a way dependent on the second wave.

When considering the multiplication of waves, it pays to remember that multiplication is really just repeated addition. When we were adding waves, it turned out that if we added waves with the same frequency, the result would be a pure wave, and that if we added waves with different frequencies, the result would *not* be a pure wave. When it comes to multiplication, the same is true because multiplication is just repeated addition. Therefore, if we multiply two waves of the same frequency (and zero mean levels), the result will be a pure wave, and if we multiply two waves of different frequencies (with any mean levels), the result will *not* be a pure wave.

It is easiest to understand the multiplication of two waves with examples.

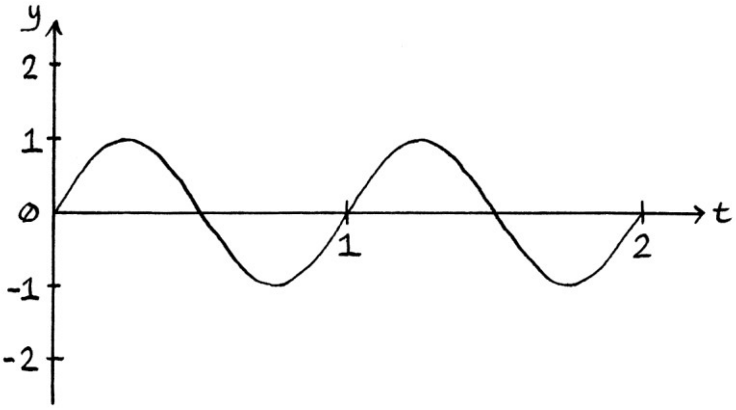
Identical waves with zero mean level

Multiplying two *identical* waves with zero mean levels is the easiest form of multiplication to do. It is also one of the most significant forms of multiplication when it comes to waves. The process is the same as multiplying a wave by itself, which is the same as squaring every single y-axis value of a wave.

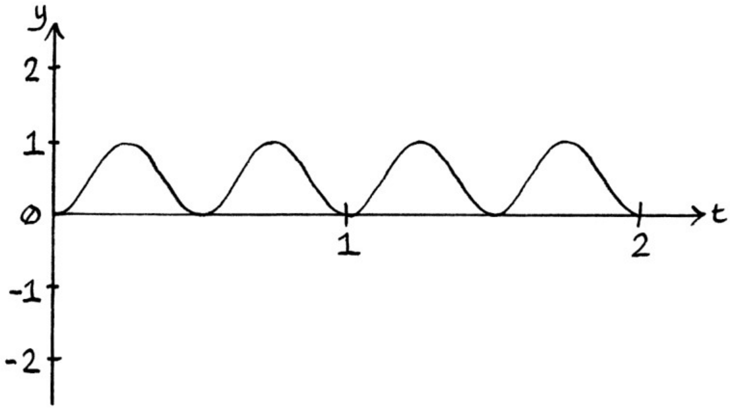
As we are essentially squaring every y-axis value, it means that the resulting signal will consist entirely of positive values (because a negative value multiplied by a negative value results in a positive value).

Example 1

We will start with the wave “y = sin 360t”. Note that this wave has a zero mean level and zero phase.



If we multiply it by itself, we end up with this:

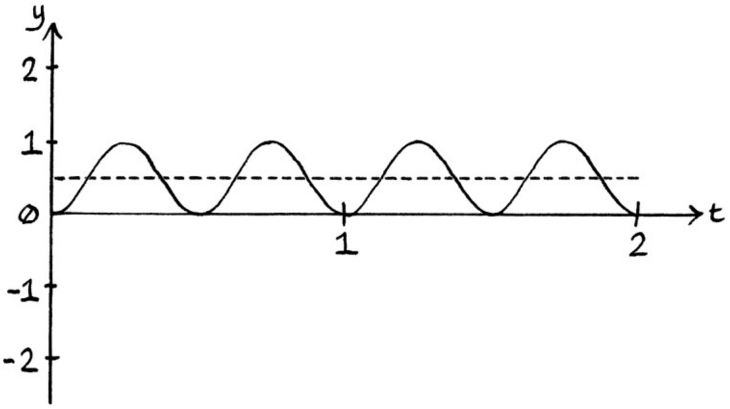


There are several things to notice about this resulting signal:

- The resulting signal is still a pure wave. It can be described with the formula for a Sine wave or a Cosine wave. The result can still be said to be derived from an object's rotation around a circle.

- Every y-axis value is 0 or positive. This is because squaring any number will always produce a result that is 0 or positive. All the points of the wave are either on or above the x-axis, which means that the graph has a non-zero mean level. It is the case that squaring any wave will result in a signal that is all above the x-axis, and that has a mean level that is above the x-axis.

- The resulting wave's mean level is 0.5 units:



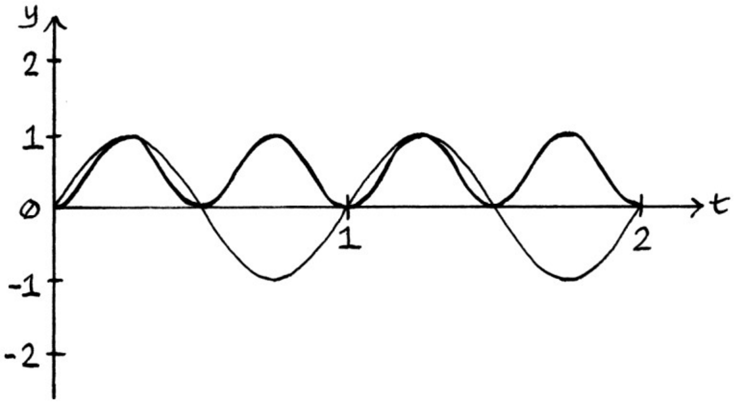
- The resulting wave repeats its shape twice as quickly as the original wave – its frequency has doubled. This has happened because the negative parts of the original wave are now positive *peaks* instead of dips. There are also equally high peaks made from squaring the positive peaks from the original wave. There are twice as many peaks as before. The points that were halfway between the original peaks and dips (in other words, the points at y = 0), are now the dips because everything else has become higher than they were.

- The amplitude of the resulting wave is half that of the original wave. This is because the previous maximum values were 1, and the square of 1 is 1. No point can be higher than 1. The new lowest point is now 0. Therefore, the midpoint is now 0.5, and so the amplitude is half what it was before. [Actually, the resulting amplitude is half the square of the original amplitude, but it is impossible to know this from this example.]

- The resulting wave's formula has a phase of +270 degrees. The resulting wave has a y-axis value of 0 at t = 0 (because it is the square of the original wave at t = 0, which was 0), yet y = 0 is not the middle of the wave – in fact, it is the lowest point of the wave. A square of a Sine wave with no phase and no mean level will always have a y-axis value of 0 at t = 0. If we visualise the circle from which the resulting wave could be said to be derived, then we can tell that the object rotating around the circle will be at its lowest point in this case, which is at 270 degrees. The object will be rising from its lowest point (270 degrees) at t = 0, and the wave's curve will also be rising from its lowest point too at that time. The phase cannot be anything but 270 degrees for this resulting wave.

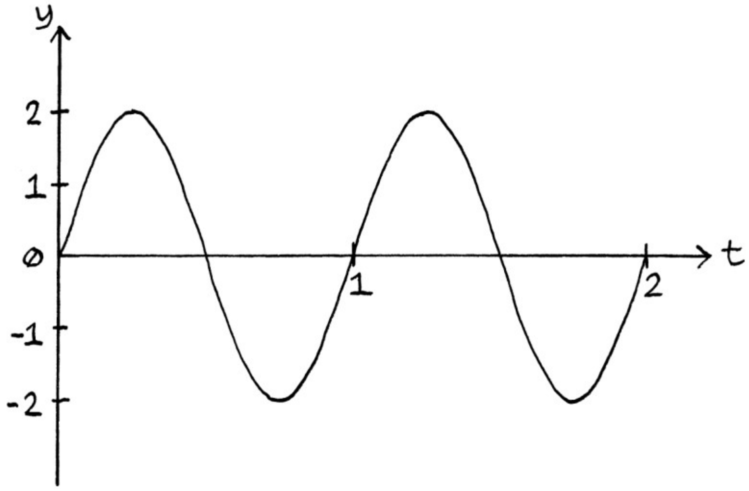
- The resulting wave's formula is "y = 0.5 + 0.5 sin ((360 * 2t) + 270)"

If we draw both the original wave and the resulting wave on the same graph, we can have a better idea of what has happened:

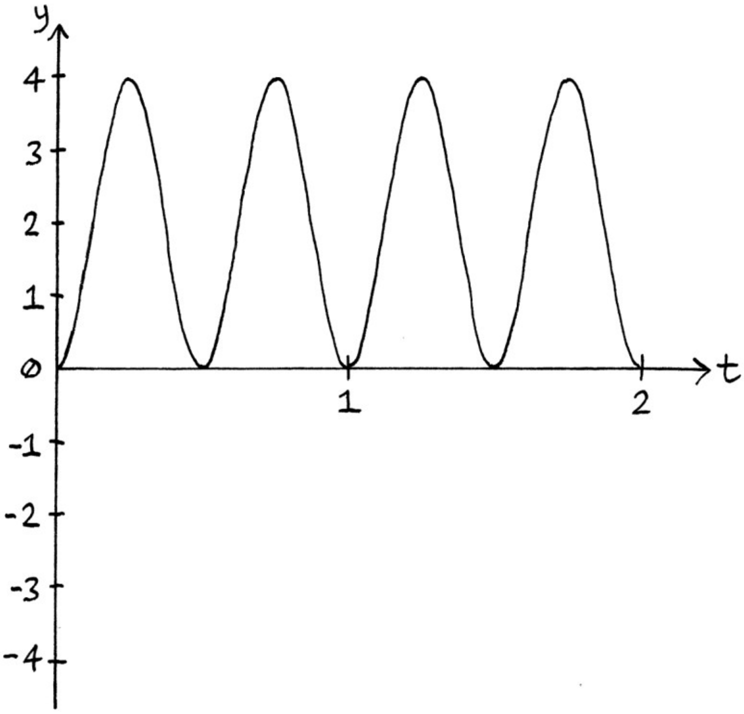


Example 2

Now, we will multiply “y = 2 sin 360t” by itself. This is the same as the wave from the previous example, except it has an amplitude of 2. The wave looks like this:



The result of multiplying this wave by itself looks like this:



Things to note about this resulting signal are:

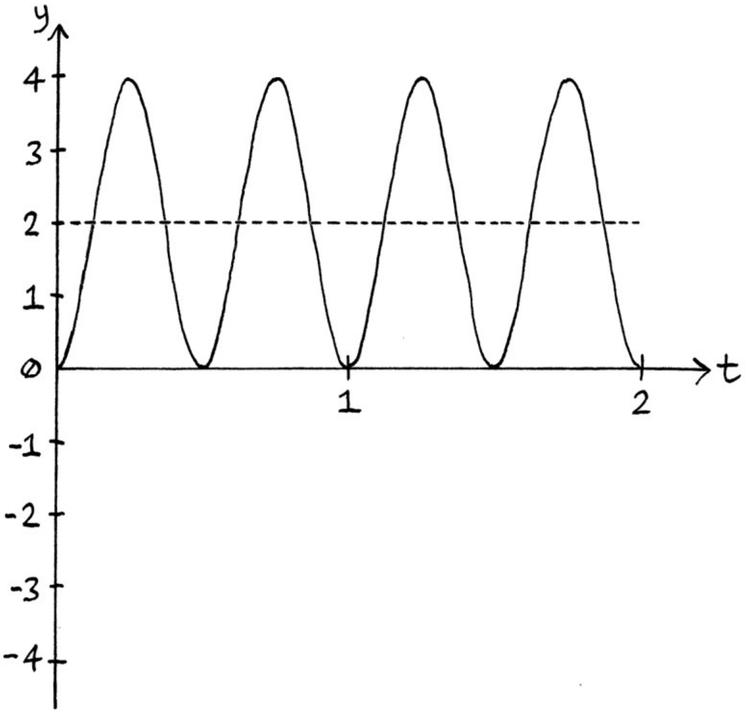
- It is a pure wave again.

- Its frequency is twice the frequency of the original wave – it is 2 cycles per second.

- Its highest y-axis value is 4. This is because the maximum y-axis value of the original wave was 2, so the maximum y-axis value of the result will be 2 squared. The resulting signal's lowest y-axis value is 0. This is because any value below 0 in the original wave becomes greater than zero after it is squared.

- Its amplitude is 2. This is actually half the square of the original wave's amplitude, although it is impossible to know this from this example.

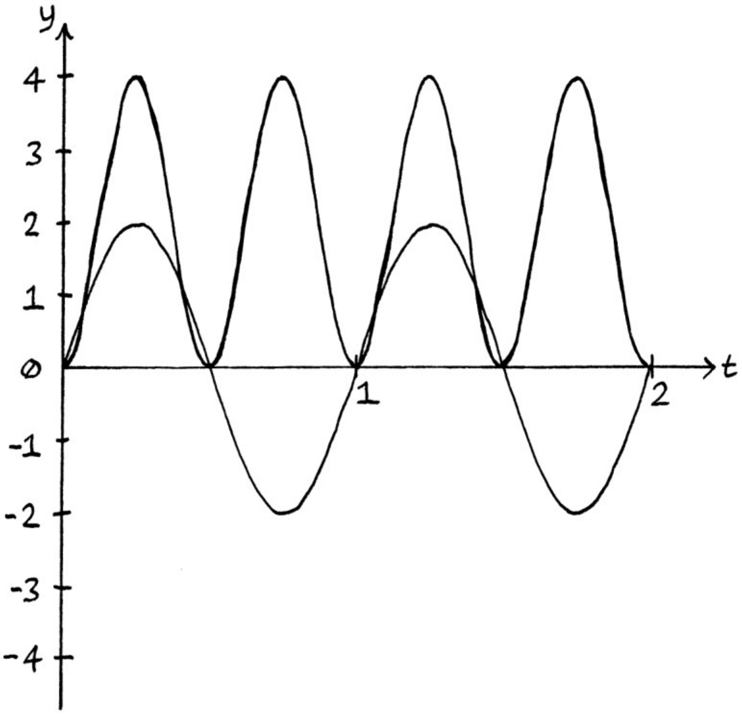
- Its midpoint is at y = 2, so its mean level is 2:



- Its phase is 270 degrees again. This is because the y-axis value of the result at t = 0 is zero, and y = 0 is now the lowest point of the wave. On the circle from which this wave could be said to be derived, the object would be at the lowest point, which is at 270 degrees. Therefore, the object has a starting point of 270 degrees, and the derived wave has a phase of 270 degrees. For any Sine wave with no phase, and zero mean level, that is multiplied by another Sine wave with no phase and zero mean level, the y-axis value at t = 0 will be zero, and also the minimum point.

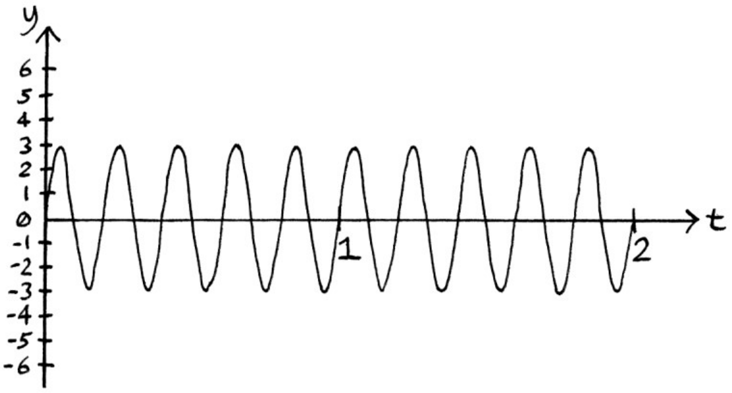
- The resulting wave's formula is "y = 2 + 2 sin ((360 * 2t) + 270)"

Overlaying the original wave and the resulting wave on the same graph, we can examine how the two relate to each other:

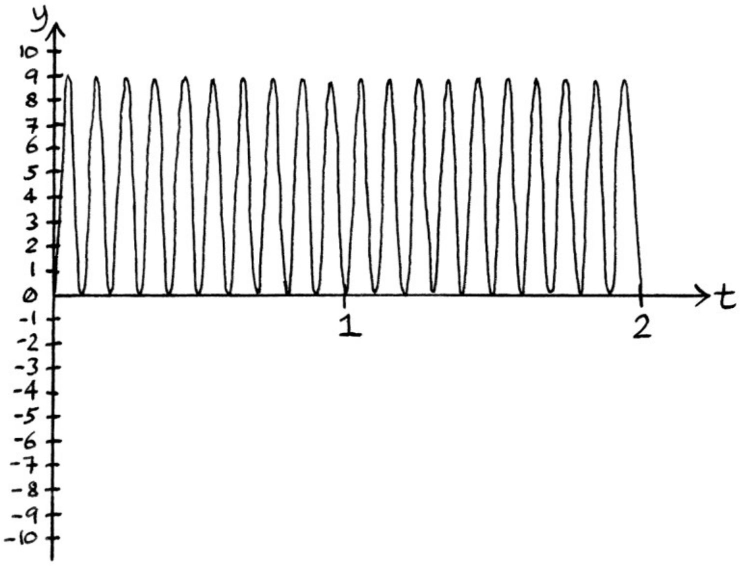


Example 3

We will multiply "y = 3 sin (360 * 5t)" by itself. The wave looks like this:



The result looks like this:



Things to notice about this resulting signal are:

- It is still a pure wave.

- The wave's amplitude is 4.5. This is half of 3 * 3. In the previous examples, the amplitude was also half the square of the original amplitude, but it was harder to be sure that this was the case. You might be able to guess that the amplitude would end up like this – the maximum y-axis value of the result will be the maximum y-axis value of the original wave squared, regardless of whether it was positive or negative. Therefore, it will be 9 units. The minimum value will be zero, because all values will end up at either zero or higher. The amplitude will be half the distance from the minimum to the maximum, so will be 9 ÷ 2 = 4.5.

- This wave's mean level is also 4.5, which again is half the square of the original amplitude. If you can see why the amplitude of the result is half the square of the original amplitude, you should be able to see that the mean level will be the same – it is halfway between the maximum and minimum values.

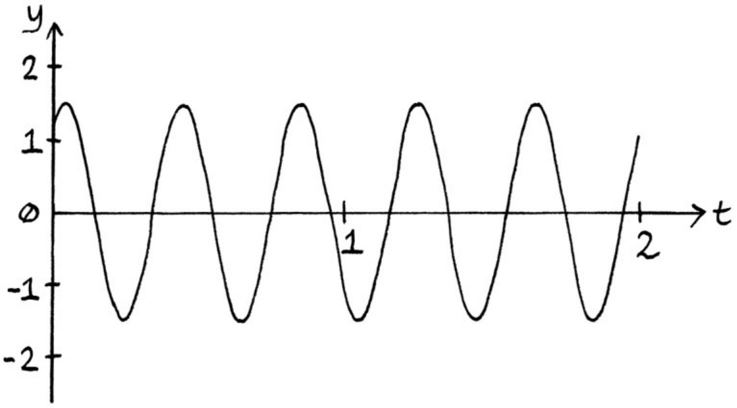
- The wave's frequency is twice the original frequency: 10 cycles per second.

- The wave's phase is +270 degrees again.

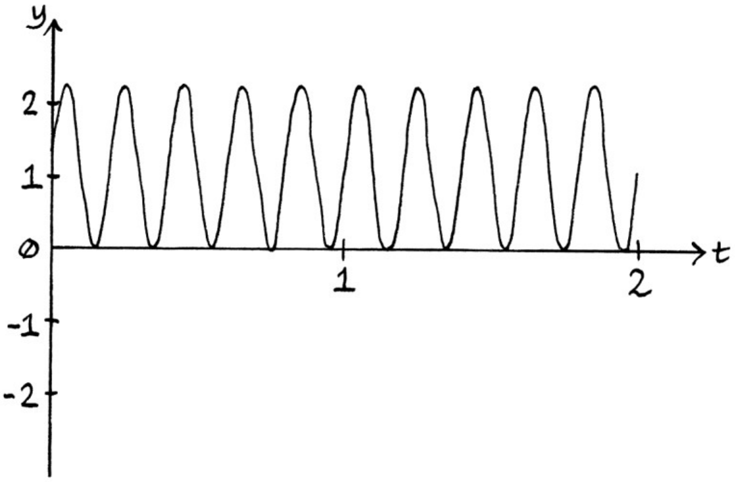
- This resulting wave's formula is: "y = 4.5 + 4.5 sin ((360 * 10t) + 270)".

Example 4

We will multiply “y = 1.5 sin ((360 * 2.5t) + 45)” by itself. This wave has a phase of +45 degrees. The original wave looks like this:



The resulting wave looks like this:



Significant facts about this wave are:

- It is a pure wave.

- It has an amplitude of 1.125, which is half the square of the original wave’s amplitude.

- It has a mean level of 1.125, which, again, is half the square of the original wave’s amplitude.

- It has a frequency of 5 cycles per second, which is twice the frequency of the original wave.

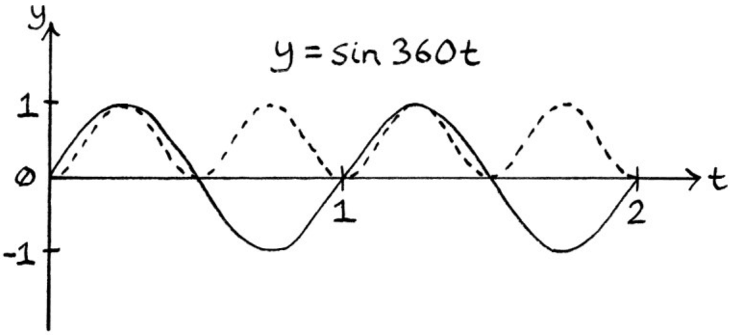
- It has a phase of 0 degrees. At t = 0, the resulting wave's y-axis value is the square of: "y = 1.5 sin ((360 * 0) + 45)", which is the square of "y = 1.5 sin (0 + 45)", which is the square of "y = 1.5 sin 45", which is 1.125. The value 1.125 happens to be the midpoint of the wave (which is the mean level), and the wave is rising at that point. On the circle from which this wave could be said to be derived, the object rotating around it would be at 0 degrees for it to have a height equal to the mean level. Therefore, the object starts at 0 degrees, the circle has a phase point of 0 degrees, and the resulting wave has a phase of 0 degrees.

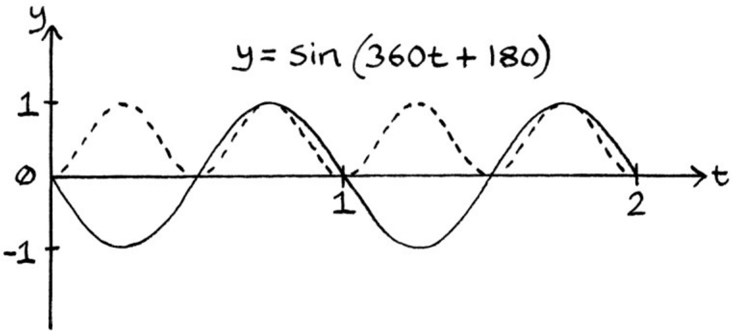
- Its formula is "y = 1.125 + 1.125 sin (360 * 5t)"

Example 5

The wave "y = sin (360t + 180)" multiplied by itself results in:
"y = 0.5 + 0.5 sin ((360 * 2t) + 270)"

This is noteworthy because the result is identical to the square of "y = sin 360t". They produce the same result because "y = sin (360t + 180)" is an upside down version of "y = sin 360t". The peaks of one wave align with the dips of the other wave, so when they are squared, the peaks and dips both create peaks in the same place. This means the results are the same.

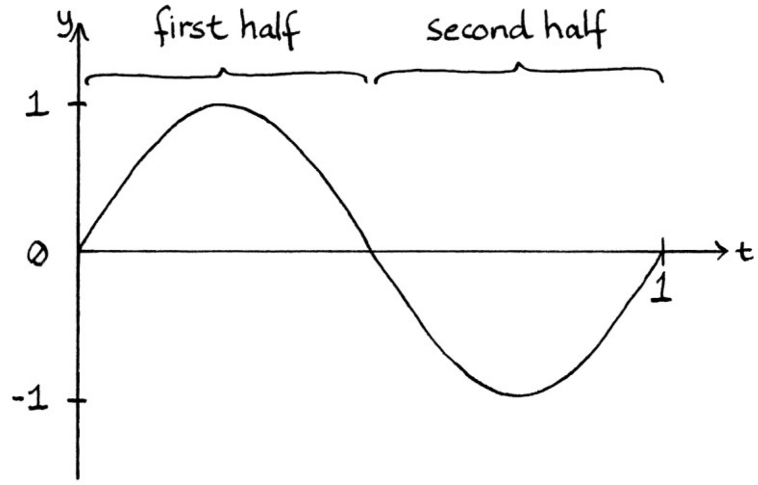


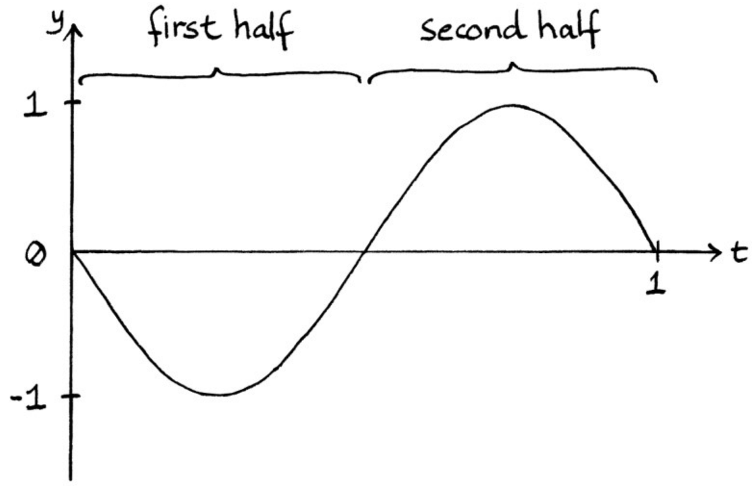


Resulting phase

The first thing to notice about the resulting phases is that if the original wave has a phase of 0 degrees, the result has a phase of +270 degrees (or −90 degrees). The phase of the result is like this because the y-axis value of the result at t = 0 is the minimum possible value, and the curve is increasing at this point.

For a wave with zero phase, the second half (the negative half) of a cycle of the original wave becomes positive when it is squared, and so ends up matching the square of the first half. Both the following waves will produce the same result when squared:



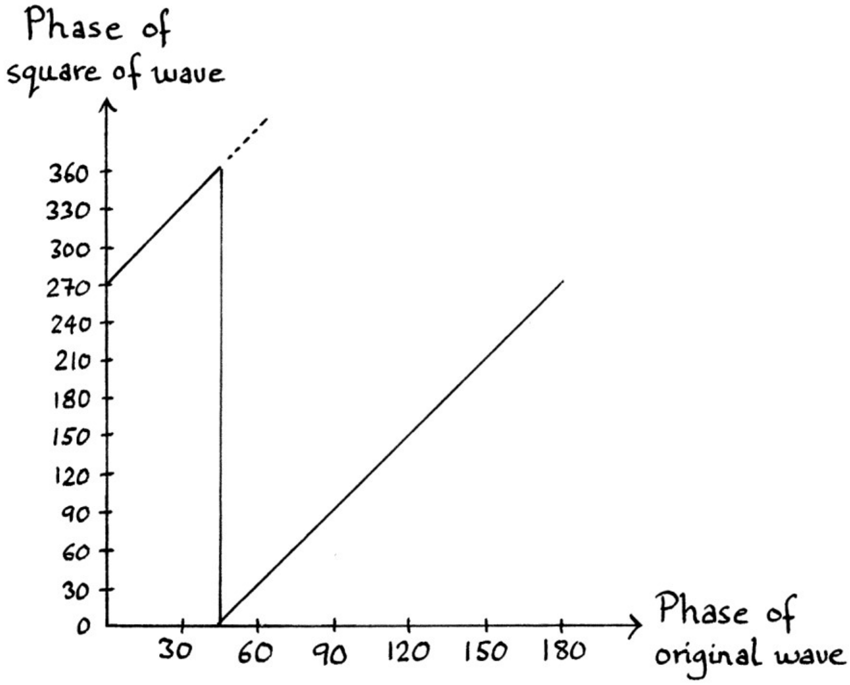


For original waves with phases other than zero, the part underneath the x-axis will also end up matching that of the part above the x-axis, once they both have been squared. This means that the square of any wave will match the square of that wave with a phase 180 degrees higher or lower.

Here is a table of calculated phases as a guide to what is happening. The first column has the phase of the original wave from 0 to 180 degrees; the second column has the phase of the wave resulting from squaring that original wave. [The results for phases over 180 degrees are the same as those for that phase minus 180 degrees.] The results in this table are valid, no matter what the amplitude or frequency of the original wave.

Original wave's phase	**Squared wave's phase**
0	270
9	288
18	306
27	324
36	342
45	0
54	18
63	36
72	54
81	72
90	90
99	108
108	126
117	144
126	162
135	180
144	198
153	216
162	234
171	252
180	270

These values plotted on a graph look like this:



Note how the graph drops when it reaches 360 degrees. This is just because I am keeping the values below 360 degrees. The angle of 360 degrees is the same as 0 degrees. I could just as easily have drawn the graph going upwards above 360 degrees, but doing that makes take up more space on the page.

From this graph, we can see the following things:

- The resulting phases in the graph increase at an even rate – in fact they increase at twice the rate of the original waves' phases. If the original wave has a phase 1 degree higher, the resulting wave will have a phase 2 degrees higher. In the graph, the resulting phases increase at twice the rate because the results have twice the frequency. An object rotating around the circle from which a result derives travels twice as fast. It moves through the angles twice as quickly.
- If the original wave has a phase of 0 degrees, the resulting wave has a phase of 270 degrees (which is also the same as a phase of −90 degrees).
- If the original wave has a phase of 45 degrees, the resulting wave has a phase of 0 degrees.
- A phase of 90 degrees ends up as a phase also of 90 degrees.
- A phase of 180 degrees ends up as a phase of 270 degrees (which should be expected as the second half of the original wave produces the same results as the first half).

We can make a formula based on the graph:
“resulting phase = (2 * original phase) + 270”
... which we could also phrase as:
“resulting phase = (2 * original phase) − 90”
... because the angles of +270 degrees and −90 degrees are the same thing.

We will test the formula with an original phase of 162 degrees. In other words, we will calculate what the phase of a resulting wave will be if we square a wave that has a phase of 162 degrees:
resulting phase = (2 * 162) – 90 = 234 degrees, which the table and graph confirm as being correct.

We will try an original phase of 9 degrees:
resulting phase = (2 * 9) – 90 = −72 degrees. This is the same as 288 degrees (−72 + 360 = 288 degrees), which the first table confirms to be correct too.

We will try 90 degrees:
resulting phase = (2 * 90) – 90 = 90 degrees.

General rules for calculating the result

Some simple rules for calculating the formula for a wave multiplied by itself *when the wave has a zero mean level* are:

- The resulting amplitude will be half the square of the original amplitude.

- The resulting mean level will also be half the square of the original amplitude.

- The resulting frequency will be double the original frequency.

- The resulting phase will be twice the original phase minus 90 degrees. [This is the same as the original phase plus 270 degrees.]

The rule tested

We will test the rule using the following wave:
"y = 12.45 sin ((360 * 1.35t) + 210.48)"

The resulting amplitude should be: 12.45^2 ÷ 2 = 77.50125 units.

The resulting frequency should be: 1.35 * 2 = 2.7 cycles per second.

The resulting mean level should be the same as the amplitude: 77.50125 units.

The resulting phase should be: (2 * 210.48) – 90 = 330.96 degrees.

The formula of the resulting wave should, therefore, be:
"y = 77.50125 + 77.50125 sin ((360 * 2.7t) + 330.96)"

We can check this on a graphing calculator, which will confirm it is correct.

Given how the second half of one cycle of a wave will produce the same result as the first half, another wave that when squared would achieve the same result is:
"y = 12.45 sin ((360 * 1.35t) + 30.48)"
[This is because 210.48 – 180 = 30.48.]

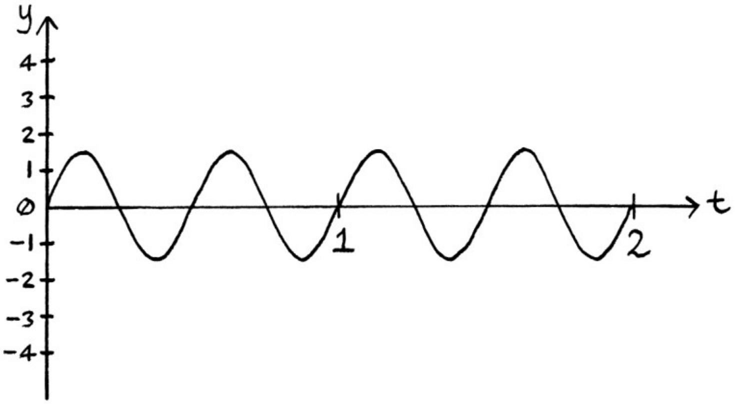
Different amplitudes

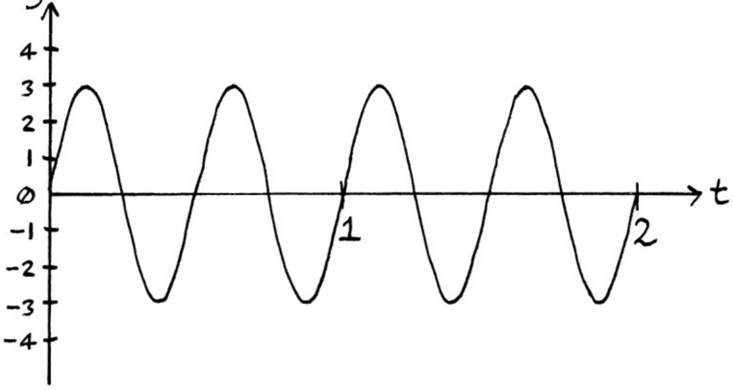
Now we will look at multiplying two waves that are the same except for their amplitudes, and which both have zero mean levels. This works in a similar way to when the amplitudes are the same.

Example 1

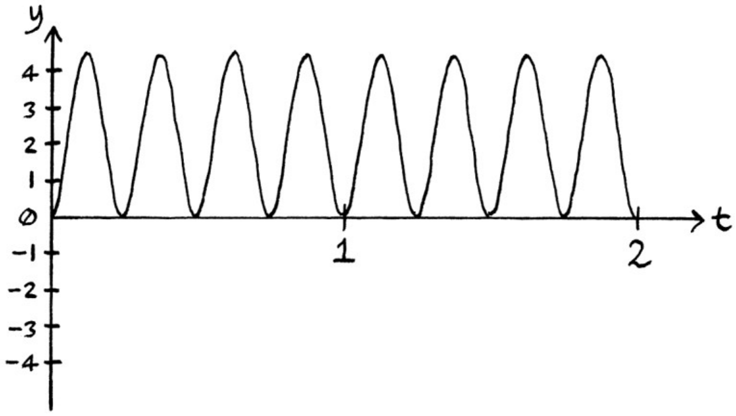
We will multiply:
“y = 1.5 sin (360 * 2t)”
... by:
“y = 3 sin (360 * 2t)”

The waves look like this:





The resulting signal looks like this:



Things to note about the resulting signal are:

- It is a pure wave.
- The highest y-axis value is 4.5, which is 1.5 * 3. This should be expected, as the maximum peaks of each original wave are multiplied by each other. As the frequencies are the same, the cycles align with each other, and the peaks and dips all align with each other.
- Its lowest y-axis value is 0.
- The amplitude is half of the maximum y-axis value of the result: 4.5 ÷ 2 = 2.25.
- The mean level is also 2.25.
- The frequency is twice the original frequencies: 4 cycles per second.
- The phase is 270 degrees.
- The resulting formula is “y = 2.25 + 2.25 sin ((360 * 4t) + 270)”

Example 2

Now we will multiply two waves that are identical except for their amplitudes, and that have matching non-zero phases. We will multiply:
“y = 2.5 sin ((360 * 1.5t) + 22)”
... by:
“y = 4 sin ((360 * 1.5t) + 22)”

The result has the formula:
“y = 5 + 5 sin ((360 * 3t) + 314)”

Things to notice about this result are:

- The amplitude is (2.5 * 4) ÷ 2 = 5 units.
- The mean level is also (2.5 * 4) ÷ 2 = 5 units.
- The frequency is 1.5 * 2 = 3 cycles per second.
- The phase is 314 degrees. This shows that our previous rule for calculating the phase if a wave is squared remains true for multiplication *if the phases are the same in each wave*. The calculation that we could have used to work out the resulting phase is: (2 * 22) – 90 = –46 = 314 degrees.

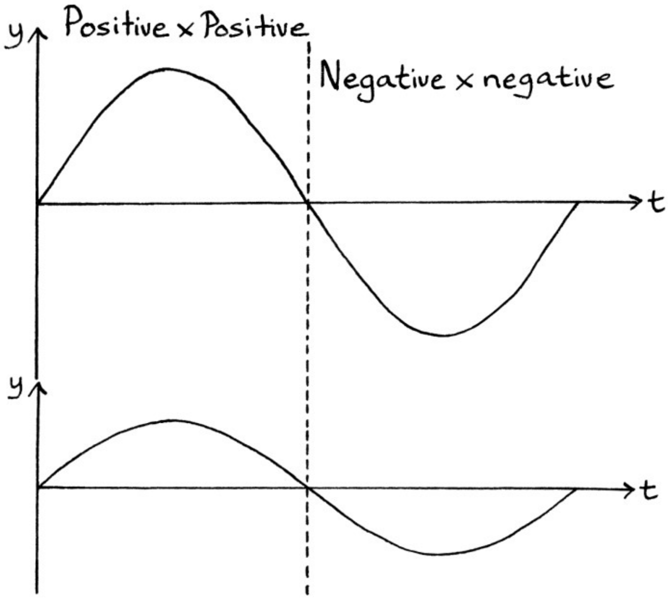
General Rules

The rules for multiplying two waves where the frequencies and phases are the same, the amplitudes are different, and both mean levels are zero are as follows:

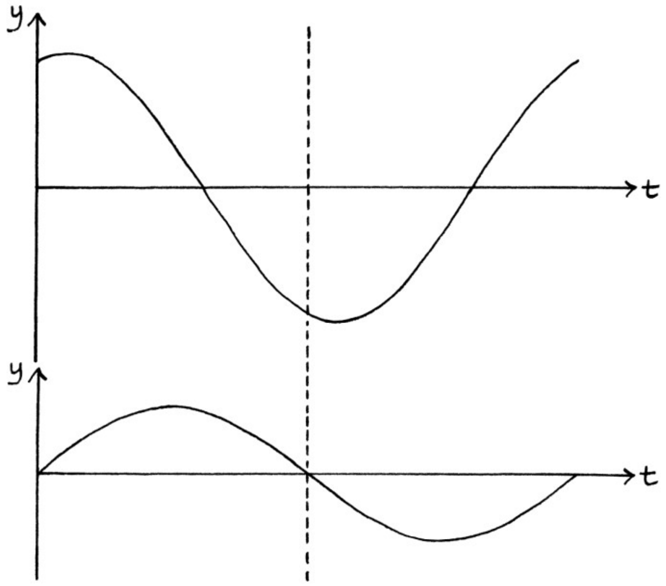
- The resulting amplitude will be two amplitudes multiplied by each other and then halved.
- The resulting mean level will be the same as the resulting amplitude.
- The resulting frequency will be the twice the original frequencies.
- The resulting phase will be (2 * original phase) – 90.

Different phases

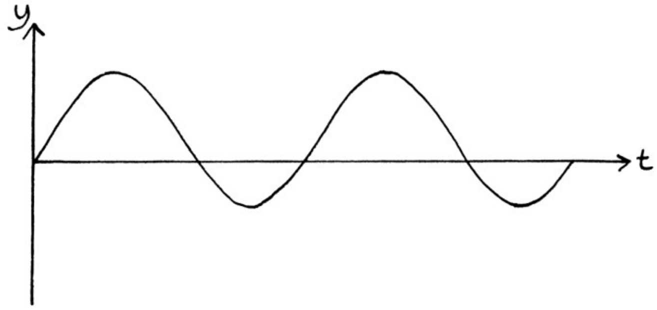
In this section, we will look at multiplying waves that are identical, except for their phase. In previous examples, when the phase was the same, all the resulting y-axis values were positive. This is because having the same frequency and phase meant that positive values from one wave were always multiplied against positive values from the other wave, and negative values from one wave were always multiplied against negative values from the other wave.



When the phases are different, whether a positive value will be multiplied against a positive value or not depends on how the two cycles align with each other. This means that the resulting wave will not necessarily be entirely above y = 0.



The resulting wave might be all above, all below, or somewhere in between.



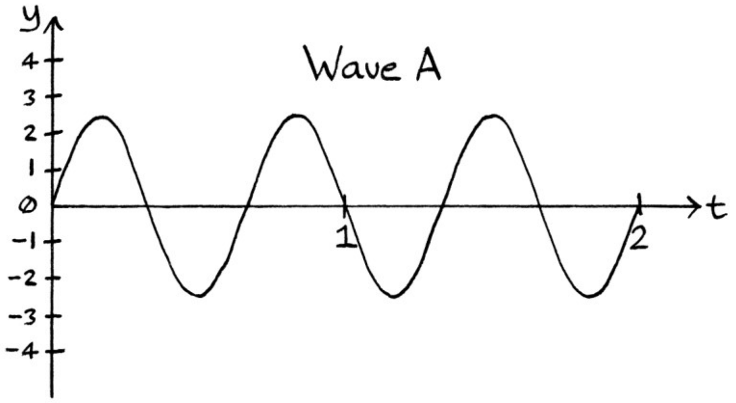
To investigate the result of multiplying waves of different phases, we will look at a range of phases while keeping all other attributes the same.

Example – same amplitude, a range of phases

We will repeatedly multiply a wave by a second wave that is identical in every way except for its phase. We will see what happens as the phase of the second wave changes.

We will multiply the wave:
"y = 2.5 sin (360 * 1.5t)"
... by:
"y = 2.5 sin ((360 * 1.5t) + ϕ)"
... for values of ϕ from 0 to 360 degrees, with steps of 22.5 degrees. In other words, we will multiply the first wave by "y = 2.5 sin (360 * 1.5t)", and then we will multiply it by, "y = 2.5 sin ((360 * 1.5t) + 22.5)", and then we will multiply it by "y = 2.5 sin ((360 * 1.5t) + 45)", and so on, and look at each result in turn.

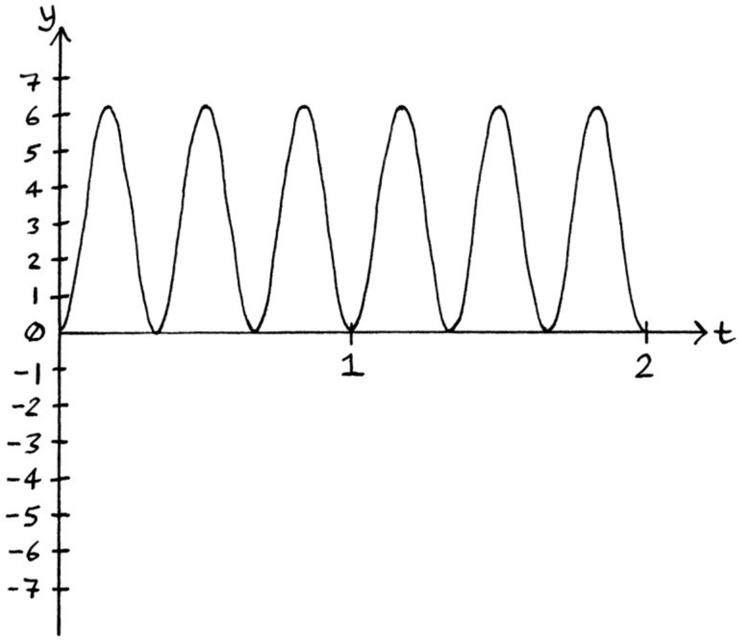
We will call the first wave, "Wave A", and the second wave with the changing phase, "Wave B". Wave A will look like this for the duration of this example:



Wave B with 0 degree phase

If Wave B has a zero phase, then it is identical to Wave A. We are calculating:
“y = 2.5 sin (360 * 1.5t)”
... multiplied by:
“y = 2.5 sin (360 * 1.5t)”
... which results in:
“y = 3.125 + 3.125 sin ((360 * 3t) + 270)”.

The result looks like this:



Characteristics of the resulting wave are:

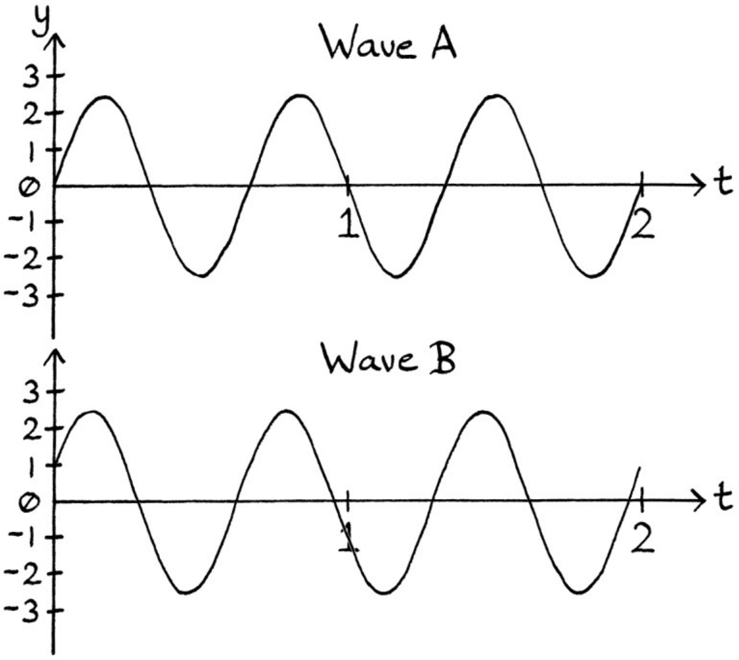
- It has an amplitude of 3.125 units.
- It has a frequency of 3 cycles per second.
- It has a phase of 270 degrees.
- It has a mean level of 3.125 units.
- Its maximum point is at y = 6.25.
- Its minimum point is at y = 0.

Given what we have learnt so far in this chapter, this is all to be expected.

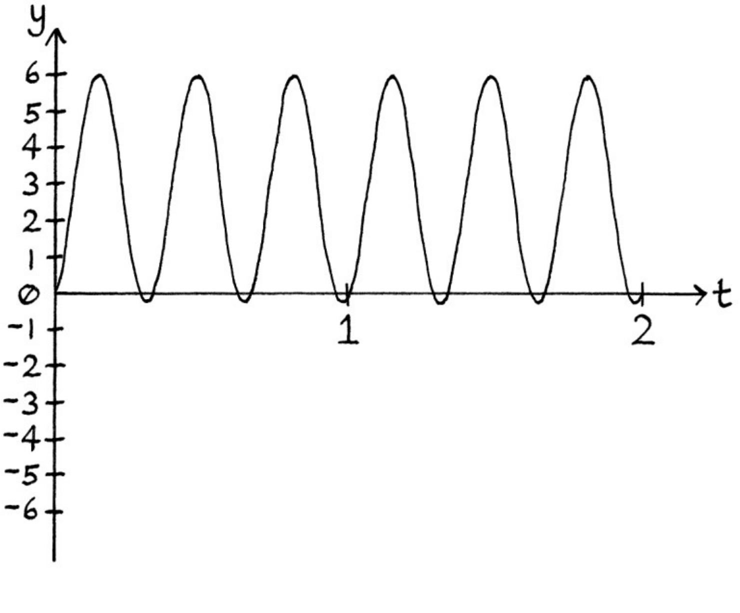
Wave B with 22.5 degree phase

Next, we will give Wave B a phase of 22.5 degrees. We will be multiplying the following two waves:
Wave A: “y = 2.5 sin (360 * 1.5t)”
... and:
Wave B: “y = 2.5 sin ((360 * 1.5t) + 22.5)”

The two waves look like this:



The result looks like this:



Notice how the resulting wave is not entirely above y = 0.

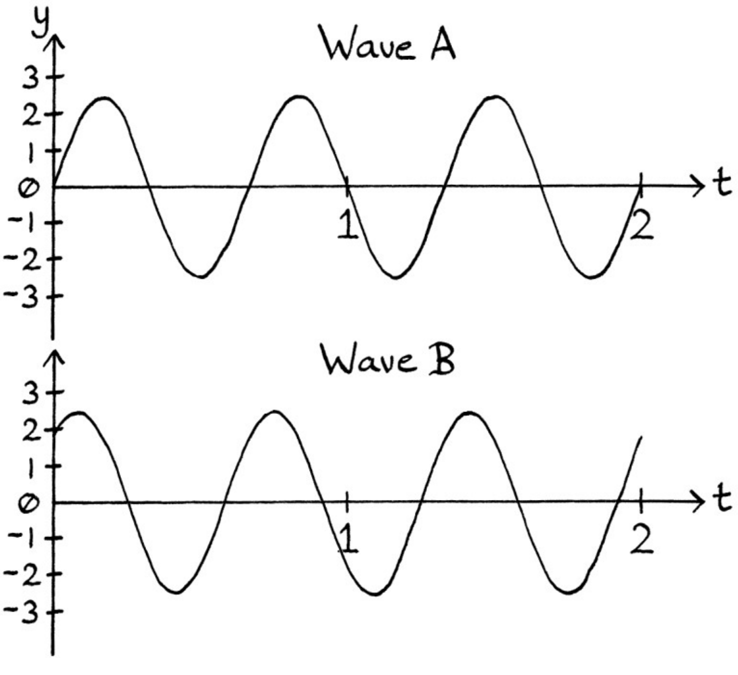
Interesting aspects of this result are:

- At t = 0, the y-axis value is 0. This is because Wave A at t = 0 is 0. Zero multiplied by any number is zero. This means that for each Wave B in this example, the resulting y-axis value at t = 0 will always be zero.
- The resulting wave has the same shape and size as when both phases were zero, but it is now shifted further down the y-axis slightly.
- The amplitude is still 3.125 units. The frequency is still 3 cycles per second.
- The maximum point of the wave is +6.01212 units, and the minimum is −0.2379 units. This means that the mean level is halfway between these – it is 2.8871 units. From this, we can see that giving Wave B a phase of 22.5 degrees has lowered the mean level by 0.2379 units.
- The phase of the resulting wave is 292.5 degrees.

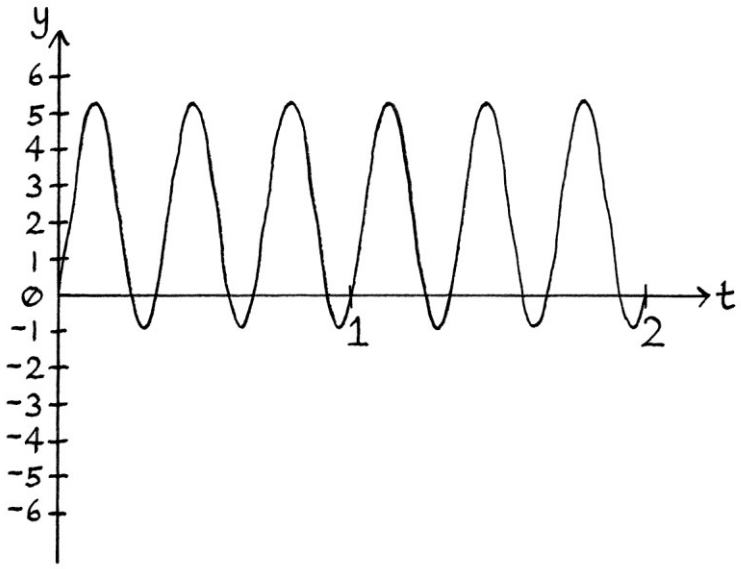
The formula of the resulting wave is:
"y = 2.8871 + 3.125 sin ((360 * 3t) + 292.5)"

Wave B with 45 degree phase

Next, we will give Wave B a phase of 45 degrees. We will be multiplying the following two waves:
Wave A: "y = 2.5 sin (360 * 1.5t)"
Wave B: "y = 2.5 sin ((360 * 1.5t) + 45)"



The result looks like this:



Interesting aspects of this result are:

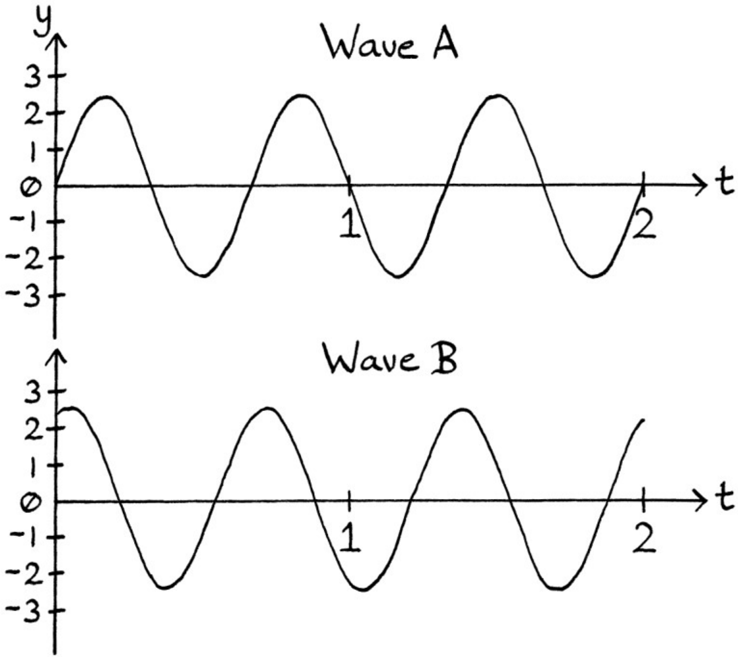
- The resulting wave has the same shape and size as when both phases were zero, but it is now shifted further down the y-axis.
- The y-axis value at t = 0 is still zero, as it will be for all these examples, because Wave A is zero at t = 0, and zero multiplied by any number results in zero.
- The amplitude is still 3.125 units and the frequency is still 3 cycles per second.
- The maximum point of the wave is +5.3347 units, and the minimum is −0.9153 units. This means that the mean level is halfway between these – it is 2.2097 units.
- The phase of the resulting wave is 315 degrees.

The formula of the resulting wave is:
"y = 2.2097 + 3.125 sin ((360 * 3t) + 315)"

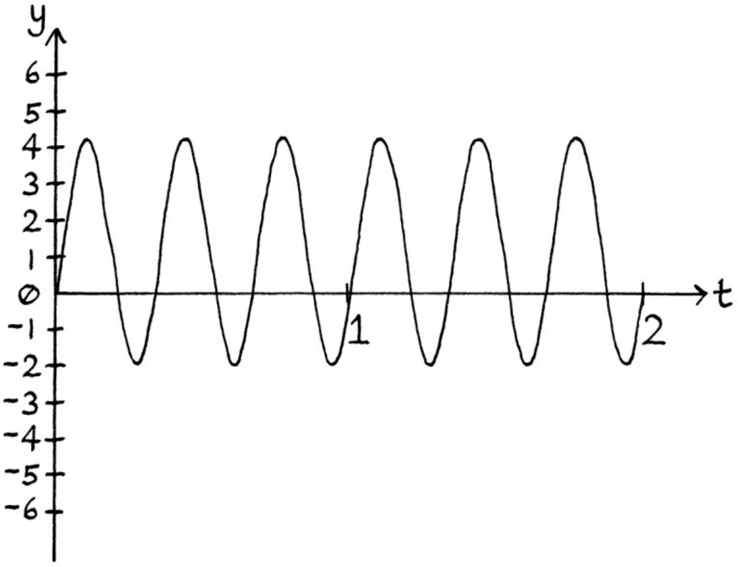
Wave B with 67.5 degree phase

Next, we will give Wave B a phase of 67.5 degrees. We will be multiplying the following two waves:
Wave A: “y = 2.5 sin (360 * 1.5t)”
... and:
Wave B: “y = 2.5 sin ((360 * 1.5t) + 67.5)”

The two waves look like this:



The result of the multiplication looks like this:



Interesting aspects of this result are:

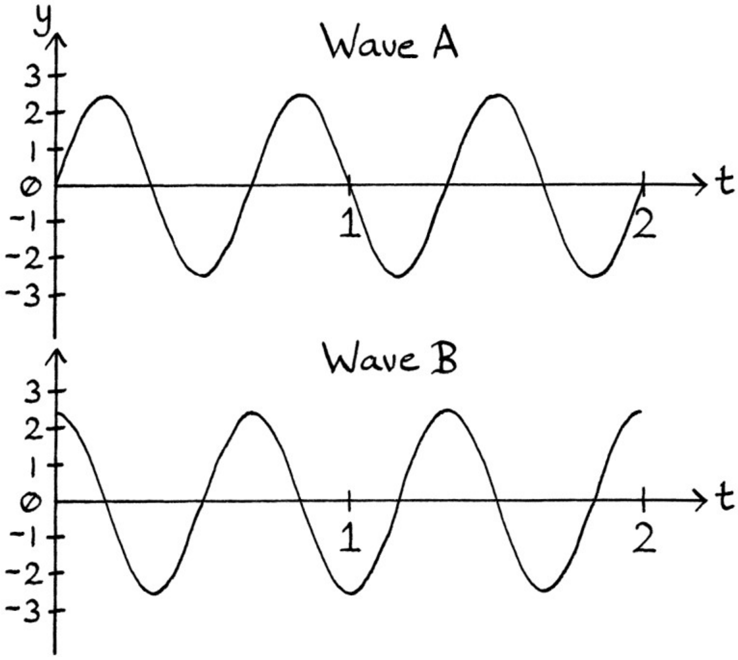
- The resulting wave has the same shape and size as when both phases were zero, but it is shifted further down the y-axis.
- The amplitude is still 3.125 units and the frequency is still 3 cycles per second.
- The maximum point of the wave is +4.3209 units, and the minimum is −1.9291 units.
- The mean level is halfway between these, and so is 1.1959 units.
- The phase of the resulting wave is 337.5 degrees.

The formula of the resulting wave is:
"y = 1.1959 + 3.125 sin ((360 * 3t) + 337.5)"

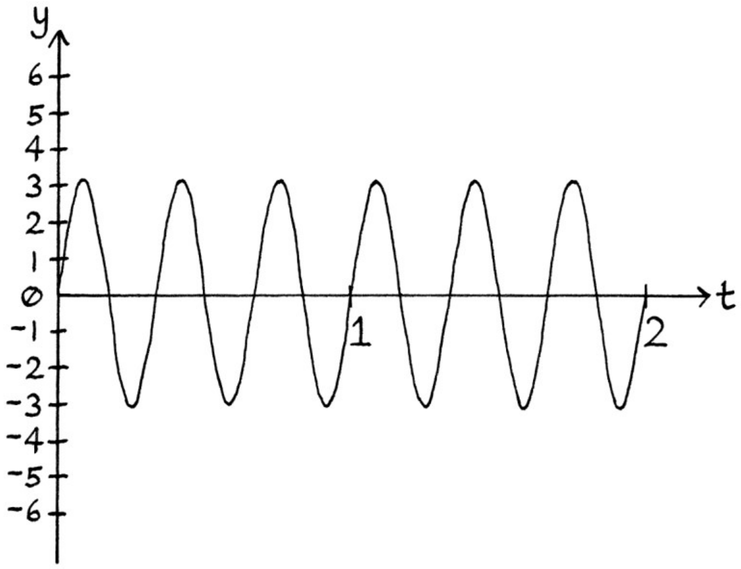
Wave B with 90 degree phase

Next, we will give Wave B a phase of 90 degrees. We will be multiplying the following two waves:
Wave A: "y = 2.5 sin (360 * 1.5t)"
... and:
Wave B: "y = 2.5 sin ((360 * 1.5t) + 90)"

The two waves look like this:



The result looks like this:



Interesting aspects of this result are:

- The resulting wave still has the same shape and size as when both phases were zero.
- The resulting wave is centred exactly over y = 0, meaning its mean level is 0 units.
- The amplitude is still 3.125 units and the frequency is still 3 cycles per second.
- The maximum point of the wave is +3.125 units, and the minimum is −3.125 units.
- The phase of the resulting wave is 0 degrees.

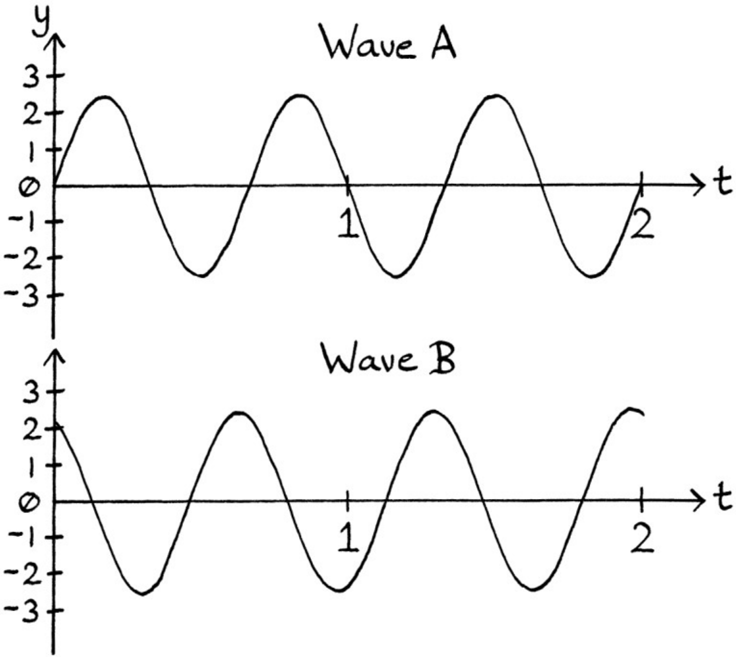
An important thing to notice here is that a wave with zero mean level multiplied by a wave that is identical except for having a phase +90 degrees higher, will result in a wave with zero mean level. This means that if we multiply a Sine wave (with zero phase) by a Cosine wave (with zero phase), we will end up with a wave with zero mean level.

The formula of the resulting wave is:
"y = 3.125 sin (360 * 3t)"

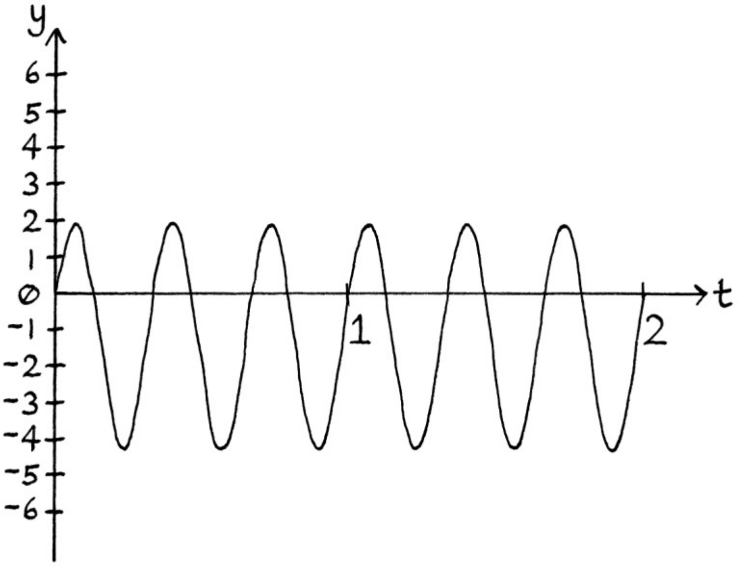
Wave B with 112.5 degree phase

Next, we will give Wave B a phase of 112.5 degrees. We will be multiplying the following two waves:
Wave A: “y = 2.5 sin (360 * 1.5t)”
... and:
Wave B: “y = 2.5 sin ((360 * 1.5t) + 112.5)”

The two waves look like this:



The result looks like this:



Interesting aspects of this result are:

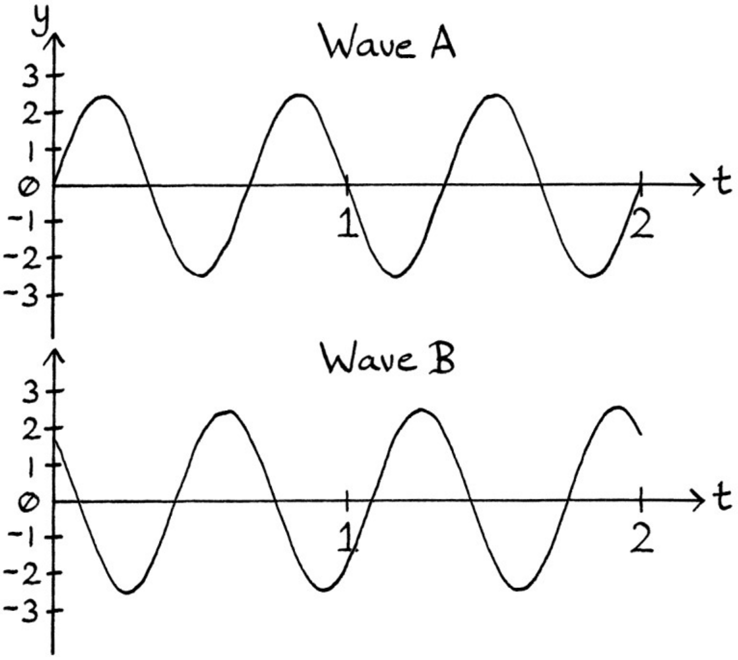
- The resulting wave has the standard shape and size.
- The amplitude is still 3.125 units and the frequency is still 3 cycles per second.
- The maximum point of the wave is +1.9291 units, and the minimum is −4.3209 units.
- The mean level is −1.1959 units.
- The phase of the resulting wave is 22.5 degrees.

The formula of the resulting wave is:
"y = −1.1959 + 3.125 sin ((360 * 3t) + 22.5)"

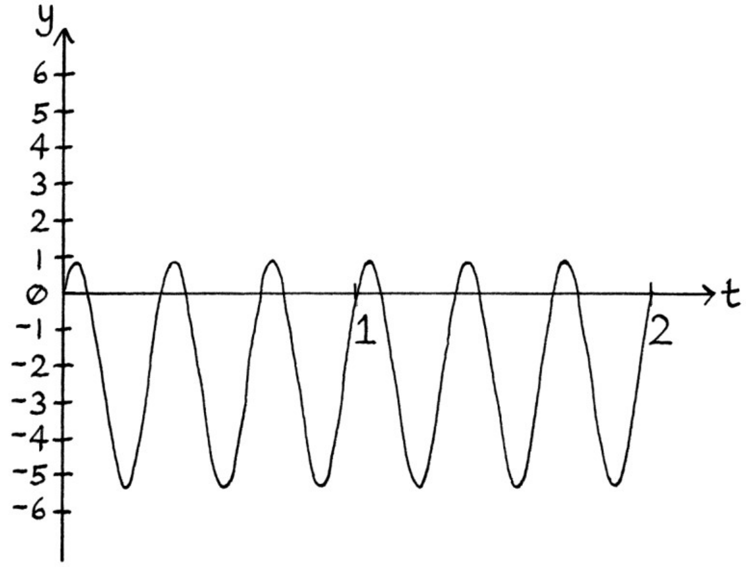
Wave B with 135 degree phase

Next, we will give Wave B a phase of 135 degrees. We will be multiplying the following two waves:
Wave A: "y = 2.5 sin (360 * 1.5t)"
... and:
Wave B: "y = 2.5 sin ((360 * 1.5t) + 135)"

The two original waves look like this:



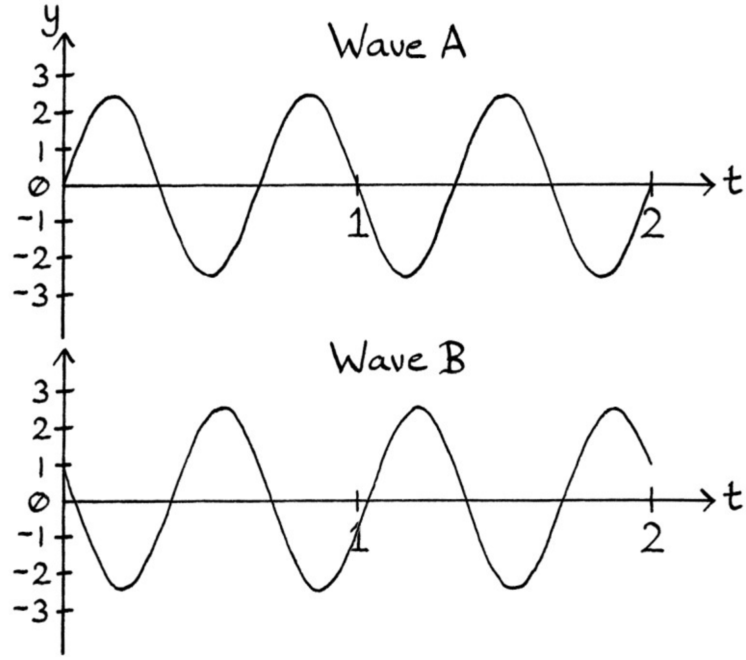
The result of the multiplication looks like this:



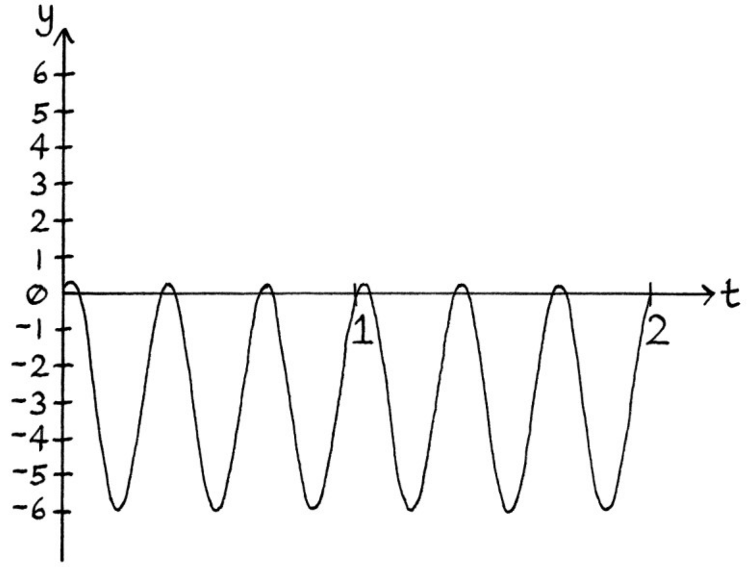
The formula of the resulting wave is:
"y = −2.2097 + 3.125 sin ((360 * 3t) + 45)"

Wave B with 157.5 degree phase

Next, we will give Wave B a phase of 157.5 degrees. We will be multiplying these waves:
Wave A: "y = 2.5 sin (360 * 1.5t)"
Wave B: "y = 2.5 sin ((360 * 1.5t) + 157.5)"



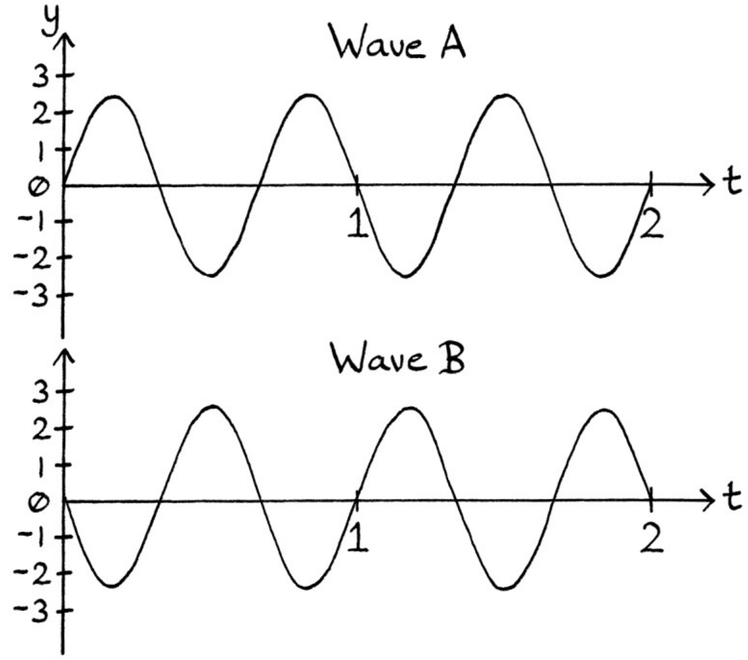
The result looks like this:



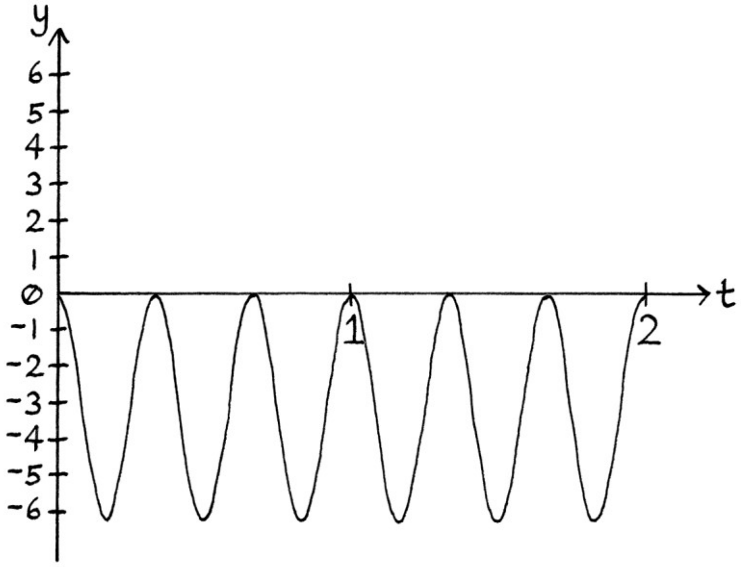
The formula of the resulting wave is:
“y = −2.8871 + 3.125 sin ((360 * 3t) + 67.5)”

Wave B with 180 degree phase

Next, we will give Wave B a phase of 180 degrees. We will be multiplying these two waves:
Wave A: “y = 2.5 sin (360 * 1.5t)”
Wave B: “y = 2.5 sin ((360 * 1.5t) + 180)”



The result looks like this:



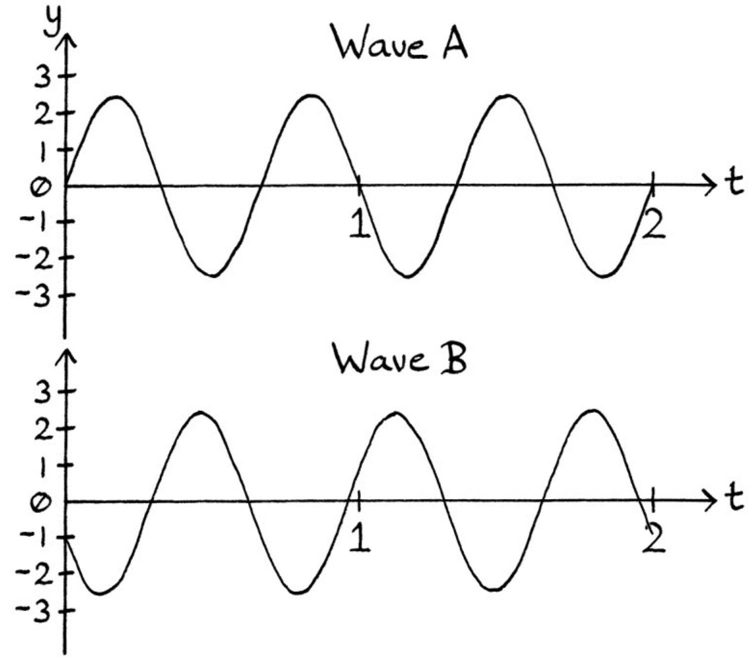
Note how this resulting wave is entirely below y = 0. The phases of the two original waves are 180 degrees apart, and the result is entirely in the bottom half of the graph. The phase of the resulting wave is +90 degrees.

The formula of the resulting wave is:
“y = −3.125 + 3.125 sin ((360 * 3t) + 90)”

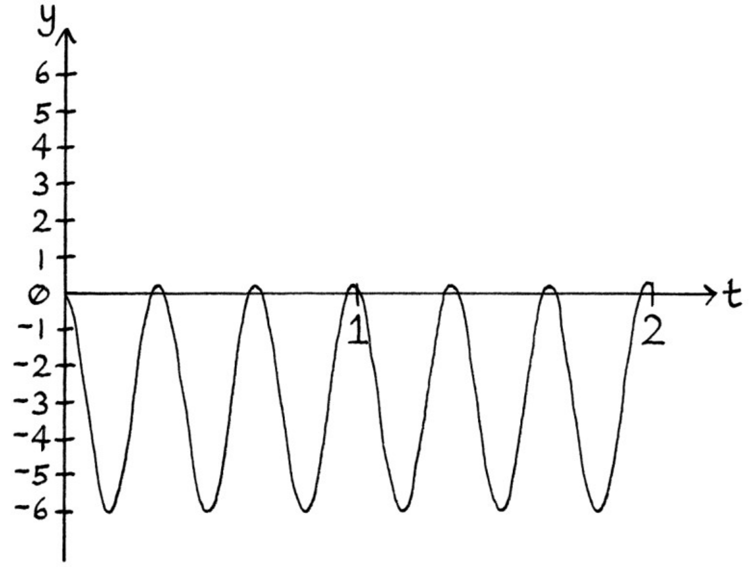
Wave B with 202.5 degree phase

Next, we will give Wave B a phase of 202.5 degrees. We will be multiplying the following two waves:
Wave A: “y = 2.5 sin (360 * 1.5t)”
... and:
Wave B: “y = 2.5 sin ((360 * 1.5t) + 202.5)”

The two waves look like this:



The result looks like this:

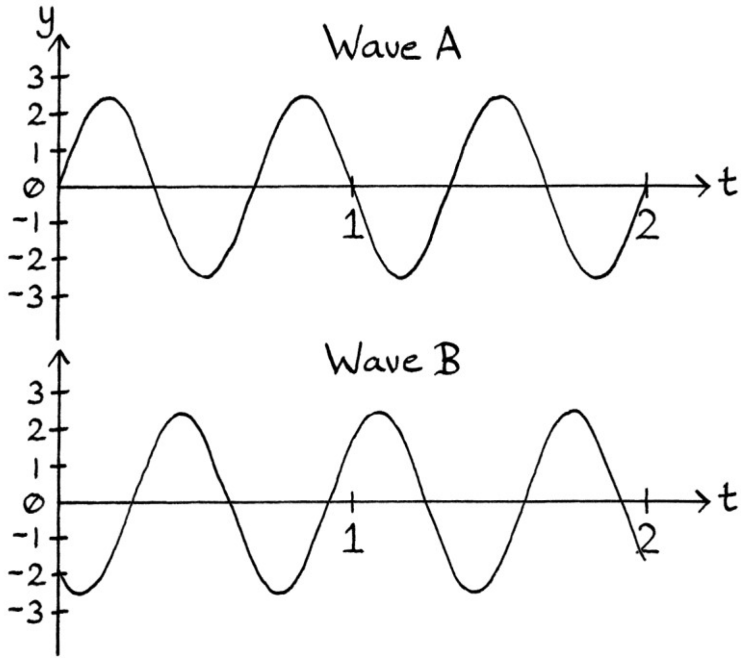


Note how the resulting wave is moving back up the y-axis. The mean level is now −2.8871 units, and is the same as when Wave B's phase was 157.5 degrees. Note also that the phase of this resulting wave is different from when Wave B's phase was 157.5 degrees – the wave is moving back upwards at the same rate as it moved downwards, but its phase is not the same.

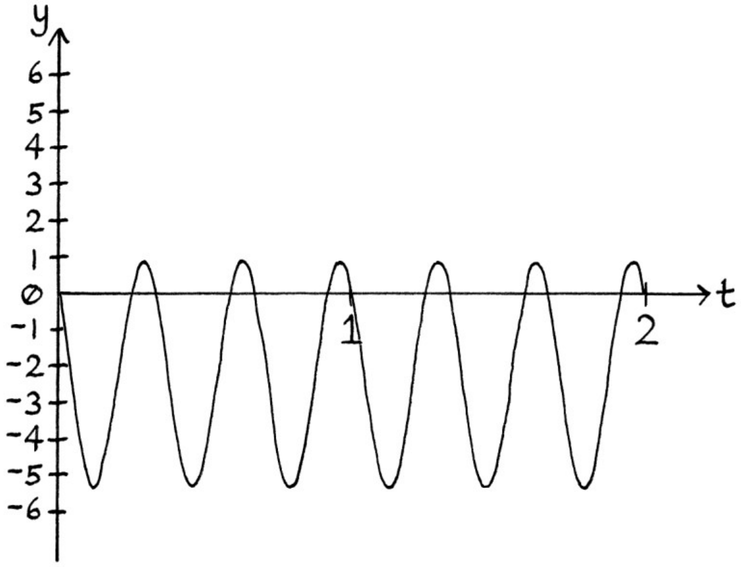
The formula of the resulting wave is:
"y = −2.8871 + 3.125 sin ((360 * 3t) + 112.5)"

Wave B with 225 degree phase

Next, we will give Wave B a phase of 225 degrees. We will be multiplying the following two waves:
Wave A: "y = 2.5 sin (360 * 1.5t)"
... and:
Wave B: "y = 2.5 sin ((360 * 1.5t) + 225)"



The result looks like this:

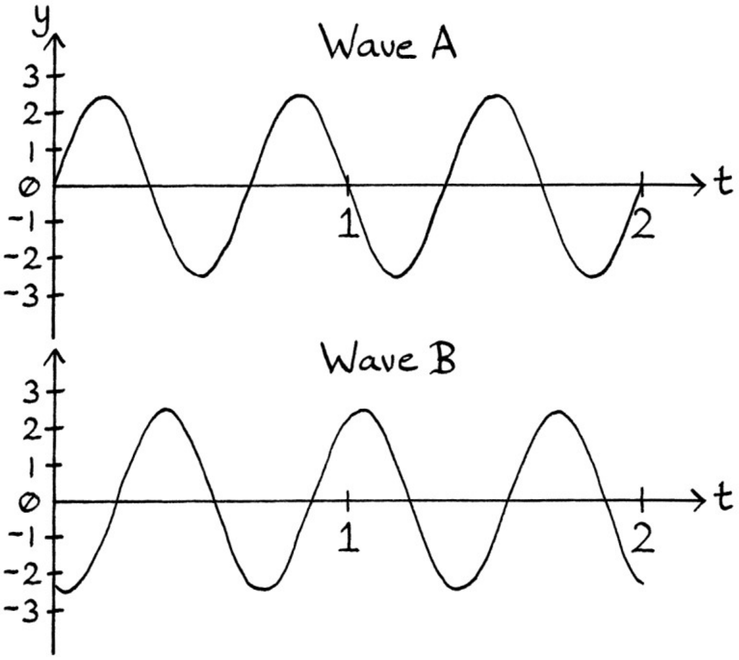


The resulting wave is continuing to move up the y-axis. Its mean level is −2.2097, which is the same as when Wave B had a phase of 135 degrees.

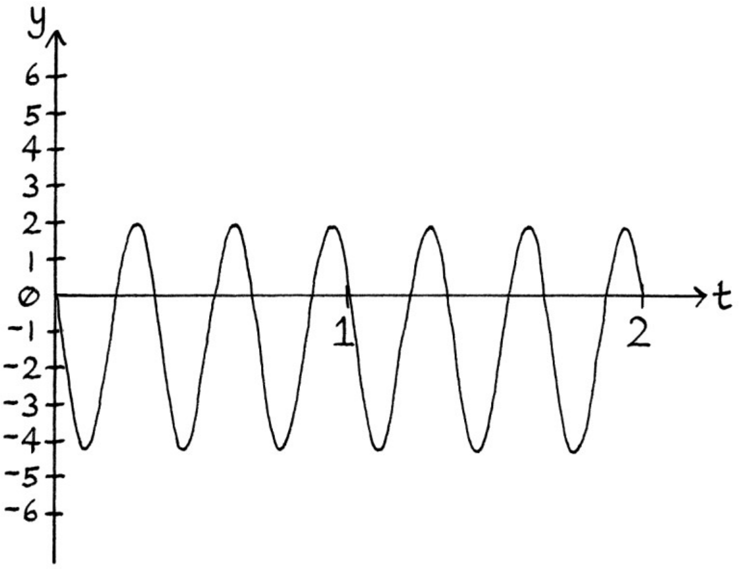
The formula of the resulting wave is:
"y = −2.2097 + 3.125 sin ((360 * 3t) + 135)"

Wave B with 247.5 degree phase

Next, we will give Wave B a phase of 247.5 degrees. We will be multiplying the following two waves:
Wave A: "y = 2.5 sin (360 * 1.5t)"
... and:
Wave B: "y = 2.5 sin ((360 * 1.5t) + 247.5)"



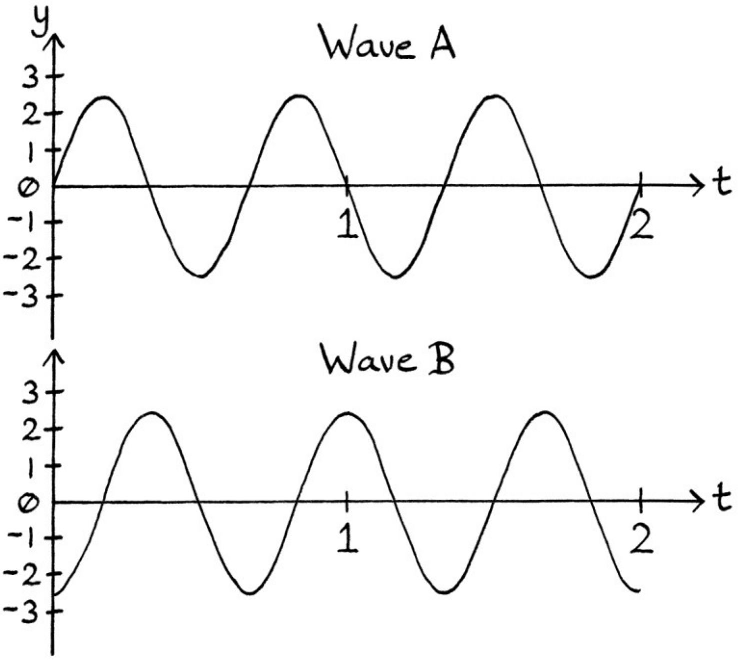
The result looks like this:



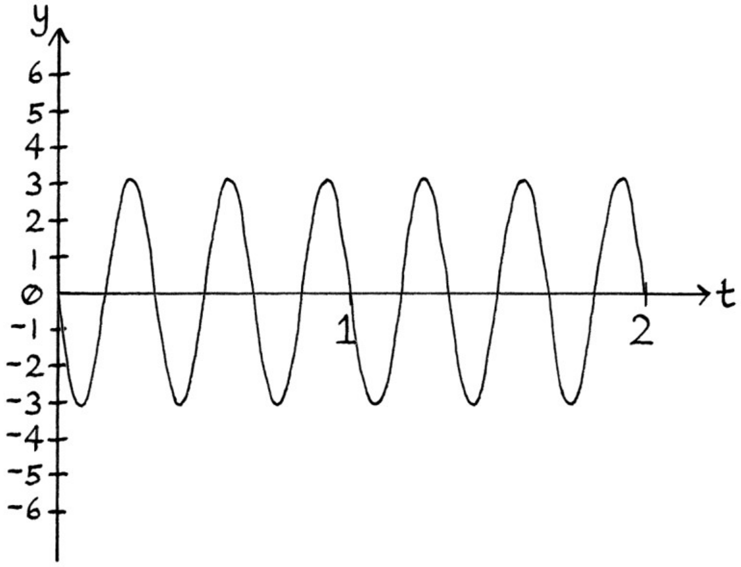
The formula of the resulting wave is:
“y = −1.1959 + 3.125 sin ((360 * 3t) + 157.5)”

Wave B with 270 degree phase

Next, we will give Wave B a phase of 270 degrees:
Wave A: “y = 2.5 sin (360 * 1.5t)”
Wave B: “y = 2.5 sin ((360 * 1.5t) + 270)”



The result looks like this:



This is an important result to examine. The mean level of this result is zero – the wave is centred on y = 0. The phase of this resulting wave is 180 degrees. When Wave B had a phase of 90 degrees, the mean level was also zero, but the phase in that case was zero. We can say that when Wave B has a phase of 270 degrees, the result is the same as when Wave B has a phase of 90 degrees, except the resulting wave is upside down.

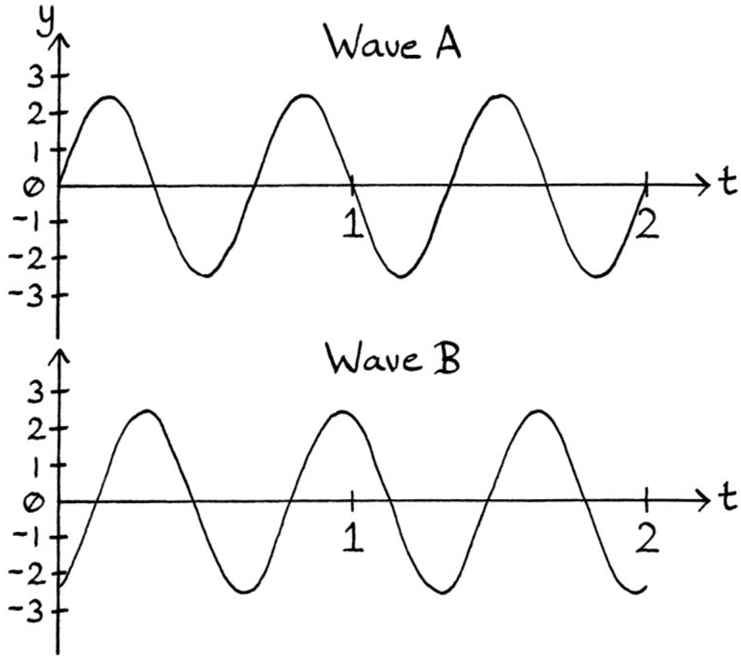
Here, we can also see the rule from when Wave B had a +90 degree phase – the two waves have phases that are 90 degrees apart (because 270 degrees one way around is 90 degrees the other way around), and the mean level of the result is zero.

The formula of this resulting wave is:
"y = 3.125 sin ((360 * 3t) + 180)"

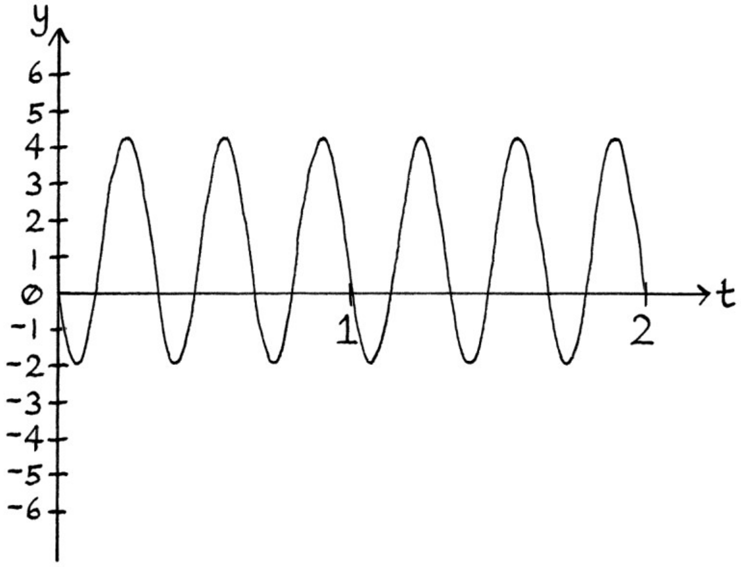
Wave B with 292.5 degree phase

If Wave B has a phase of 292.5 degrees, then we are going to be multiplying the following two waves:
Wave A: “y = 2.5 sin (360 * 1.5t)”
... and:
Wave B: “y = 2.5 sin ((360 * 1.5t) + 292.5)”

The two waves look like this:



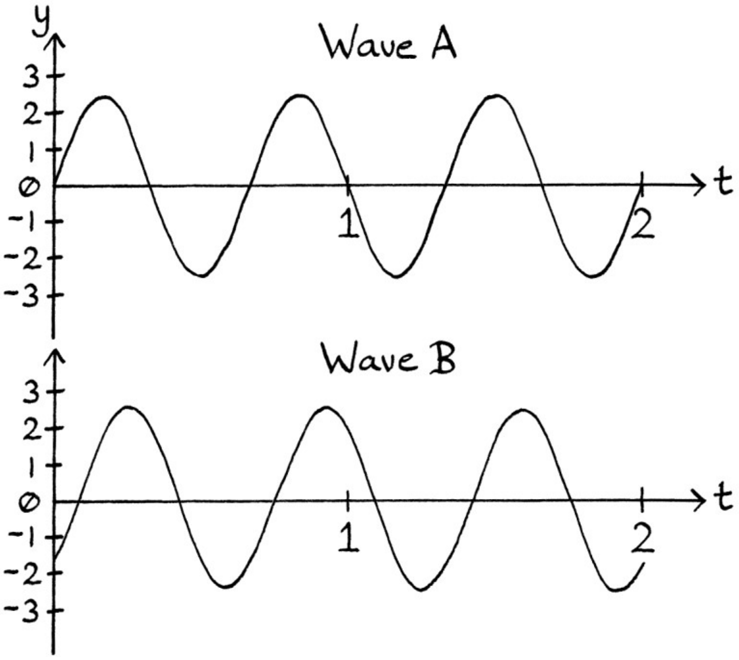
The result looks like this:



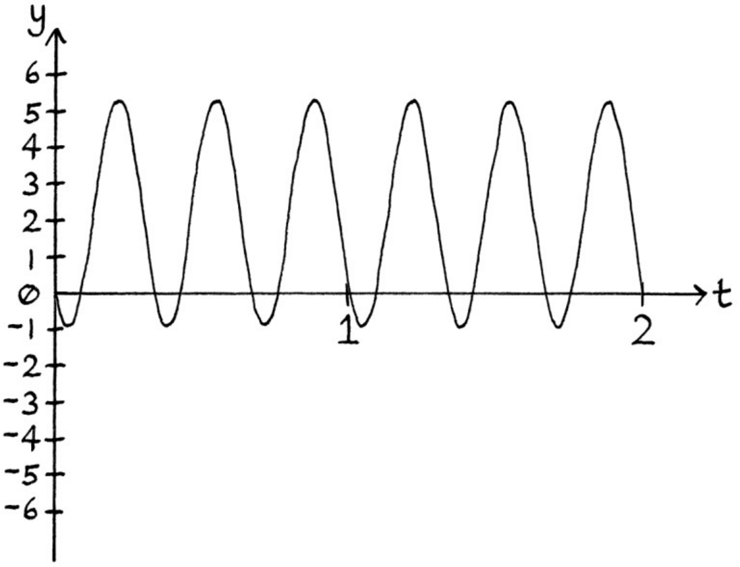
The formula of the resulting wave is:
"y = 1.1959 + 3.125 sin ((360 * 3t) + 202.5)"

Wave B with 315 degree phase

Next, we will give Wave B a phase of 315 degrees:
Wave A: "y = 2.5 sin (360 * 1.5t)"
Wave B: "y = 2.5 sin ((360 * 1.5t) + 315)"



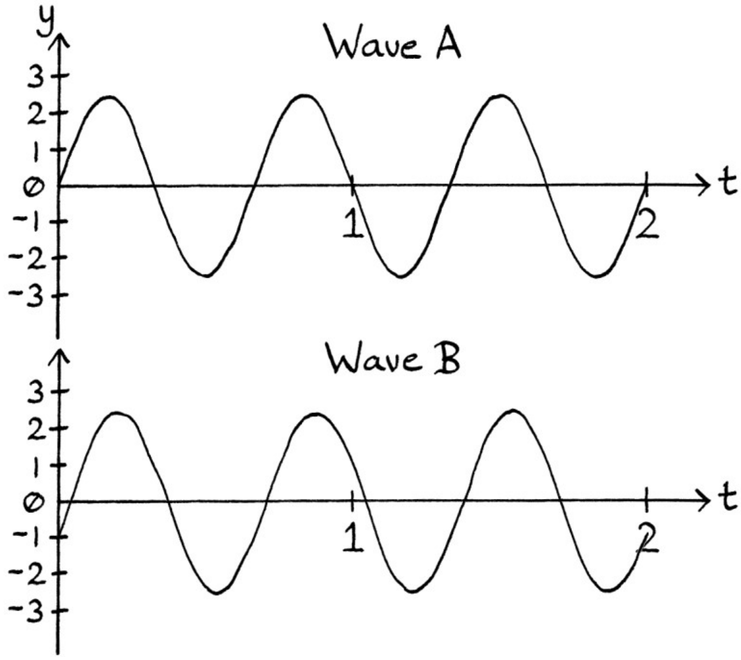
The result looks like this:



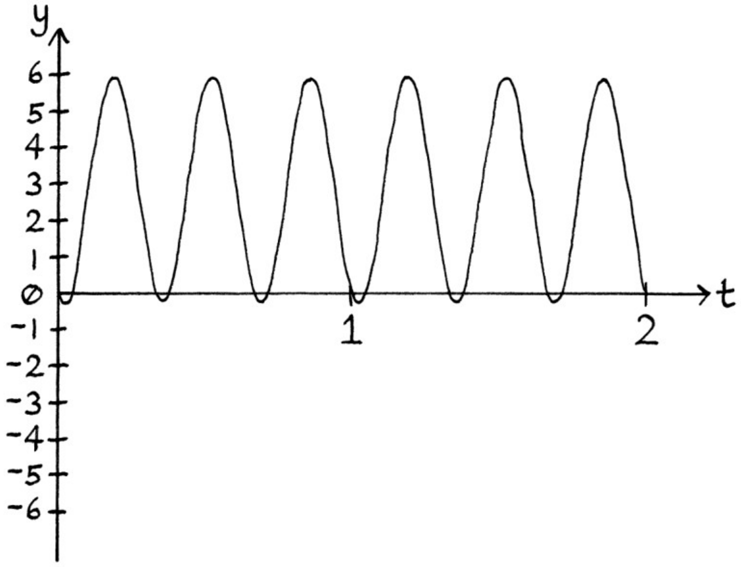
The formula of the resulting wave is:
“y = 2.2097 + 3.125 sin ((360 * 3t) + 225)”

Wave B with 337.5 degree phase

Next, we will give Wave B a phase of 337.5 degrees:
Wave A: “y = 2.5 sin (360 * 1.5t)”
Wave B: “y = 2.5 sin ((360 * 1.5t) + 337.5)”



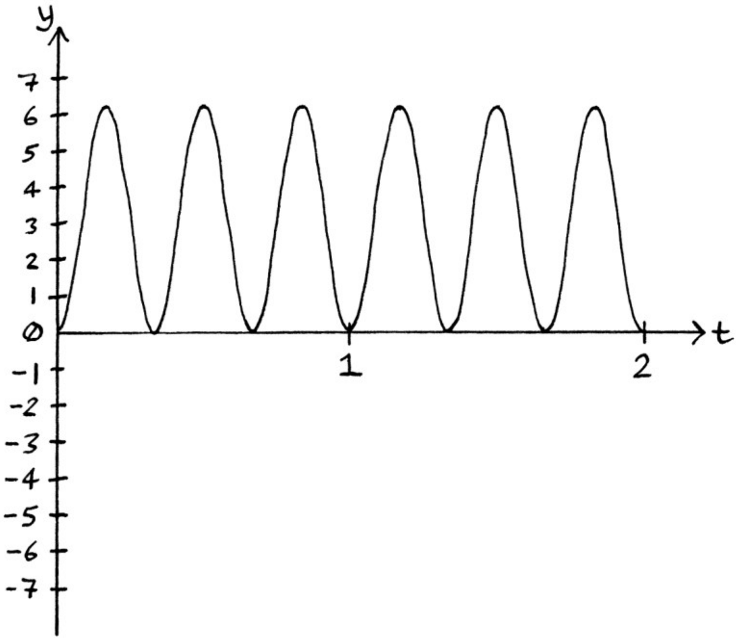
The result looks like this:



The formula of the resulting wave is:
"y = 2.8871 + 3.125 sin ((360 * 3t) + 247.5)"

Wave B with 360 degree phase

If we give Wave B a phase of 360 degrees, the result will (obviously) be the same as when Wave B had a phase of 0 degrees.

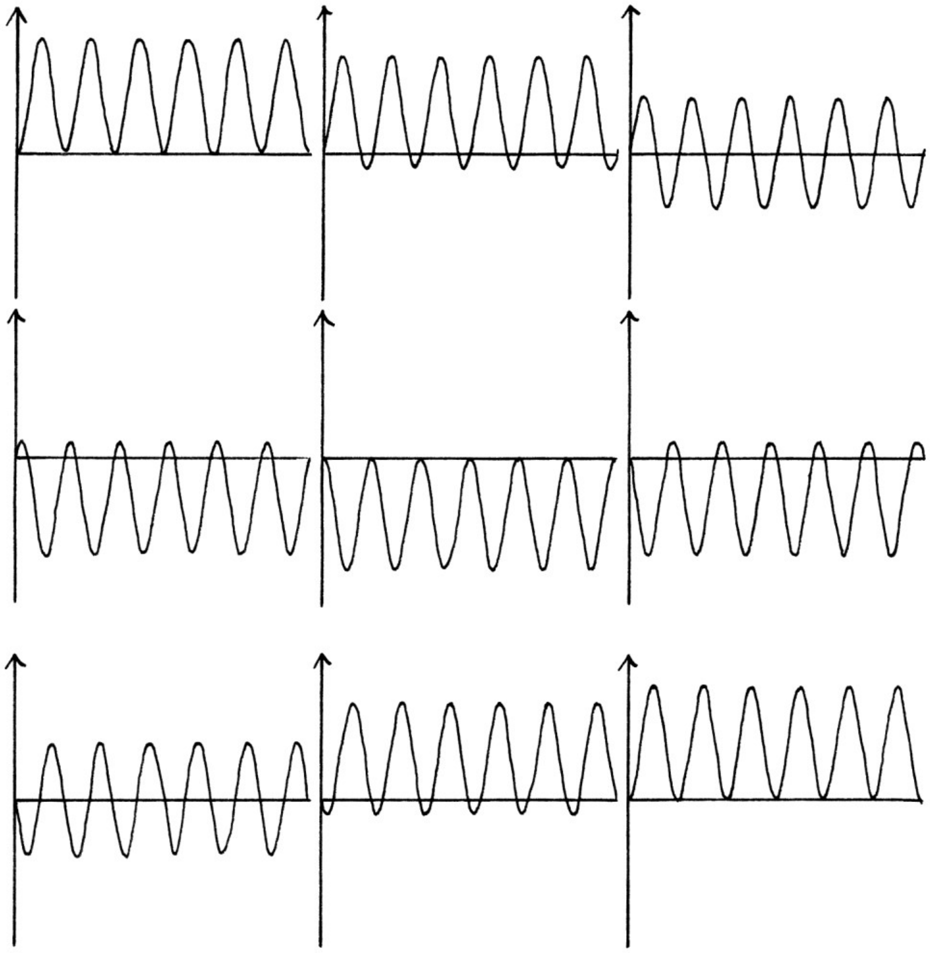


Analysis: mean levels

The general principle seen in this example is that the basic shape of the resulting wave is identical in every case. In fact, the shape is based on the amplitude and frequency, and nothing else. The differences in the result of each multiplication are that the mean level of the resulting wave is different and the phase of the resulting wave is different.

When it comes to mean level, increasing the phase of Wave B produces results with *lower* mean levels up until the moment when Wave B's phase is 180 degrees or higher, when increasing the phase further produces results with *higher* mean levels.

For ever-increasing phases in Wave B, the resulting wave moves down the y-axis from being totally above to being totally below, and then it moves back up the y-axis. This is clearer when we show some of the resulting waves next to each other [drawn at intervals of 45 degrees for Wave B, going from left to right, line by line]:



Some interesting and important observations of all this are:

- When Wave B's phase is zero, the resulting mean level is at its maximum.
- When Wave B's phase is 90 degrees, the resulting mean level is zero.
- When Wave B's phase is 180 degrees, the resulting mean level is at its minimum.
- When Wave B's phase is 270 degrees, the resulting mean level is zero.

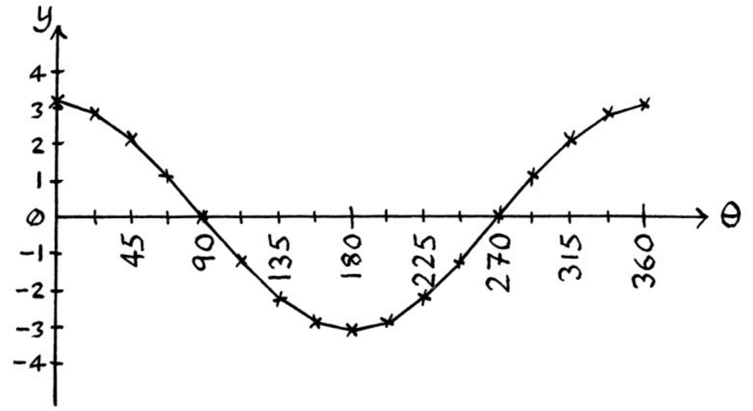
Restating these observations in a more general way:

- When the phases of both original waves are the same, the resulting mean level is at its maximum.
- When the phases of both original waves are 90 degrees apart, the resulting mean level is zero. This also means that if the phases are 270 degrees apart, the resulting mean level is zero too, because if they are 270 degrees apart, then they are also 90 degrees apart.
- When the original phases are 180 degrees apart, the resulting mean level is at its minimum.

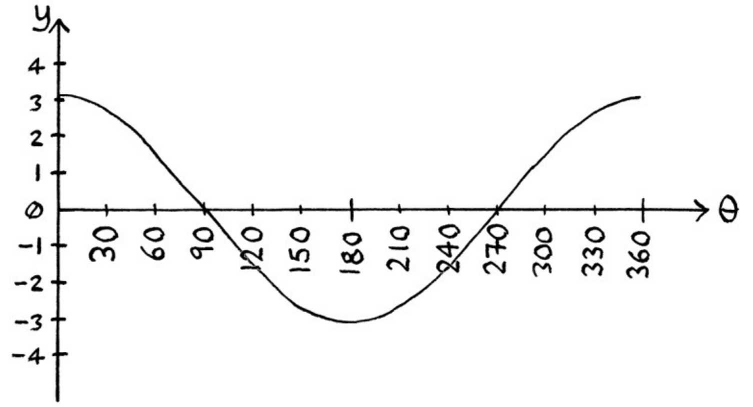
To examine the resulting mean levels more, we will put the values in a table:

Phase of Wave A	**Phase of Wave B**	**Resulting Mean Level**
0	0	3.125
0	22.5	2.8871
0	45	2.2097
0	67.5	1.1959
0	90	0
0	112.5	−1.1959
0	135	−2.2097
0	157.5	−2.8871
0	180	−3.125
0	202.5	−2.8871
0	225	−2.2097
0	247.5	−1.1959
0	270	0
0	292.5	1.1959
0	315	2.2097
0	337.5	2.8871
0	360	3.125

To understand the resulting mean levels better, we can plot the information from this table on a graph. The x-axis shows the possible phases of Wave B; the y-axis shows the mean levels on the resulting wave created by multiplying Wave A and Wave B. The x-axis goes from 0 to 360 degrees. As the x-axis is showing angles, we will call it the “θ-axis”.



If we had more values in the table, then the graph would turn out smoother:



The most important thing that we can see from this graph is that it is a pure wave. This is quite significant. It means that the mean level of a wave that results from multiplying two waves varies according to the Sine function (or in this case the Sine function with a 90-degree phase, which is the Cosine function with no phase). Whereas the resulting *phases* from multiplying two *identical* waves with the same phase increased at an even rate, when the original phases are different, the resulting *mean levels* increase according to the Sine function.

This particular "resulting mean-levels from a given phase" Sine wave graph has the following qualities:

- It has a mean level of zero units.
- It has an amplitude of 3.125 units.
- It has a phase of 90 degrees.
- It cannot be said to have a frequency as it relates to angles and not time.
- This wave's formula is "y = 3.125 sin (θ + 90)"

[Irrelevant side note: as this wave is a pure wave, we could recreate the circle from which it could be said to be derived by reading just 2 points from the graph or the table. We take the point at θ = 0, which is +3.125, and the point at θ = 90, which is 0. We then use arctan and Pythagoras's theorem to work out the amplitude and phase point of the circle that would have created that wave.]

We can use the mean level graph to calculate other resulting mean levels. If Wave B were "y = 2.5 sin ((360 * 1.5t) + 211)", we would read off the y-axis value on the graph at the point where "θ" is 211 degrees. That point has a y-axis value of −2.6786 units. Therefore, the wave resulting from multiplying Wave A and this Wave B will have a mean level of −2.6786 units.

The graph does not tell us the phase or frequency of the resulting wave – it only tells us the mean level. The graph has the same amplitude as the resulting wave.

We can ignore the graph, and instead use the formula that represents the graph:
"y = 3.125 sin (θ + 90)".
In which case, "θ" will be the phase of Wave B, so we calculate "3.125 sin (211 + 90)", which will produce the same result. The mean level resulting from multiplying Wave A by a Wave B with any phase can be calculated using the formula.

Given what we know about multiplying waves so far, we can create a wider ranging formula for calculating the mean level of the result of multiplying two waves:

- The amplitude of the "mean level wave" will be the same as the amplitude of the resulting wave. We calculate that by multiplying the original amplitudes and dividing by two. We can phrase this as:
  $(a_1 * a_2) \div 2$
  ... or as:
  $0.5 * (a_1 * a_2)$
  ... where a_1 and a_2 are the amplitudes of Wave A and Wave B.

- The mean level wave's mean level will be zero.

- The mean level wave's phase will be 90 degrees.

We end up with:
"resulting mean level = $0.5 * (a_1 * a_2) * \sin (\phi + 90)$"
... where ϕ is Wave B's phase, and Wave A has zero phase.

We could make this more succinct by saying that the wave is a Cosine wave with zero phase, as that is exactly the same thing as a Sine wave with a +90 degree phase. We end up with:
"resulting mean level = 0.5 * (a_1 * a_2) * cos (ϕ)"
... where ϕ is Wave B's phase, and Wave A has zero phase.

However, for consistency with later formulas in this chapter, I will keep the formula in terms of Sine.

The formula works when Wave A has zero phase and Wave B has any phase. If we experimented more with the multiplication of waves, we would find out that we can adjust it so that it works for when Wave A has any phase too. It is the case that it is the *difference* in phase that is important for the resulting mean level. The relevant formula is:
"resulting mean level = 0.5 * (a_1 * a_2) * sin ((ϕ_2 − ϕ_1) + 90)"
... where ϕ_1 is Wave B's phase, and ϕ_2 is Wave A's phase.

When it comes to the subtraction of the phases, it could be either way around and it would still mean the same thing. Two points on a circle can be said to be a positive number of degrees away from each other, or a negative number of degrees away from each other, but both values mean the same thing on a circle. Therefore, the equation could just as easily be:
"resulting mean level = 0.5 * (a_1 * a_2) * sin ((ϕ_1 − ϕ_2) + 90)"

The above formula will calculate the resulting mean level for the multiplication of any two waves, no matter what their amplitude, frequency or phase, as long as the frequencies are the same and the original waves both have zero mean level.

As an example, suppose we are multiplying:
"y = 2.52 sin ((360 * 12t) + 111)"
... and:
"y = 4.72 sin ((360 * 12t) + 78)"
... then:

- The resulting amplitude would be: 0.5 * (2.52 * 4.72) = 5.9472 units.
- The resulting frequency would be: 2 * 12 = 24 cycles per second.
- The resulting mean level would be:
  0.5 * (2.52 * 4.72) * sin (111 – 78 + 90) = 4.9877 units.
  [If we subtracted the phases in the other order, we would have the same result:
  0.5 * (2.52 * 4.72) * sin (78 – 111 + 90) = 4.9877 units].

- Using the knowledge we have so far, we cannot calculate the resulting phase yet.

As a summary of this section on mean level: the resulting mean level is completely dependent on solely the resulting amplitude and the difference in phase of the original waves.

Analysis: phases

In the above examples with Wave A and Wave B, we can have a better idea of how the phases of the resulting waves relate to the phases of the original waves if we list them in a table:

Phase of Wave A	**Phase of Wave B**	**Phase of resulting wave**
0	0	270
0	22.5	292.5
0	45	315
0	67.5	337.5
0	90	0 or 360
0	112.5	22.5
0	135	45
0	157.5	67.5
0	180	90
0	202.5	112.5
0	225	135
0	247.5	157.5
0	270	180
0	292.5	202.5
0	315	225
0	337.5	247.5
0	360	270

In the table, the phases of the resulting waves increase at an even rate, and also at the *same* rate as the phases of Wave B. The rule for our particular Wave A and the ranges of Wave B is that the resulting wave's phase will be equal to the phase of Wave B minus 90 degrees (or plus 270 degrees as that is the same thing).

From this, we can make up a formula that will work when Wave A has a phase of zero degrees, and for when Wave B has any phase:
“resulting phase = phase of Wave B − 90”

... or we could put it more mathematically, and say the same thing with:
“$\phi_R = \phi_B - 90$”
... where ϕ_R is the phase of the resulting wave and ϕ_B is the phase of Wave B.

From this, we can see that the resulting phase is based entirely on the original phases, and is independent of the size of the amplitudes or frequency of the original waves.

The situation surrounding resulting phases turns out to be more complicated than this though, as we find out if we give Wave A a non-zero phase. As an example, we will give Wave A the formula “y = 2.5 sin ((360 * 1.5t) + 25)” and we will run through various phase values for Wave B, which starts off as “y = 2.5 sin ((360 * 1.5t) + 0). In other words, Waves A and B are the same as before, but Wave A has a phase of 25 degrees. If we calculate the phases resulting from multiplying Wave A against a range of Wave Bs, we will end up with the following table:

Phase of Wave A	**Phase of Wave B**	**Phase of resulting wave**
25	0	295
25	22.5	317.5
25	45	340
25	67.5	2.5 (or 362.5)
25	90	25 (or 385)
25	112.5	47.5
25	135	70
25	157.5	92.5
25	180	115
25	202.5	137.5
25	225	160
25	247.5	182.5
25	270	205
25	292.5	227.5
25	315	250
25	337.5	272.5
25	360	295

The previous formula does not work here. The resulting phase still increases at a constant rate, but it is not equal to the phase of Wave B − 90 degrees. In fact, it is equal to the phase of Wave B – 65 degrees. [−65 degrees is also +295 degrees.] The number −65 is "−90 + 25". [The number 295 is "270 + 25".] It turns out that the resulting phase is actually equal to the phase of Wave A added to the phase of Wave B, with 90 degrees subtracted [or 270 degrees added.] When we look at the previous examples when Wave B had a phase of zero degrees, we can see that the result was also the two phases with 90 degrees subtracted [or 270 degrees added], but it was impossible to tell.

We can say that the formula for calculating the resulting phase for the multiplication of any two waves that have any amplitude, any phase, the same frequency (of any value), and zero mean level is:
"resulting phase = phase of Wave A + phase of Wave B − 90"

A more mathematical and concise formula is:
"$\phi_R = \phi_A + \phi_B - 90$"
... where ϕ_R is the resulting phase, ϕ_A is the phase of Wave A and ϕ_B is the phase of Wave B.

As an example of the "$\phi_R = \phi_A + \phi_B - 90$" formula working, we will multiply the waves:
"y = 3.89 sin ((360 * 11.3t) + 201.71)"
... and:
"y = 12.56 sin ((360 * 11.3t) + 341.99)"

The phase of the resulting wave will be:
201.71 + 341.99 − 90 = 453.7, which, after subtracting 360 degrees, is 93.7 degrees.

An overall formula

Now we have worked out formulas for the resulting mean level and phase, we can have a formula for the entire result. For the multiplication of two waves of the same frequency, any amplitude, any phase, and with zero mean levels, we know that:

- The resulting amplitude will always be the two original amplitudes multiplied by each other and divided by two.

- The resulting frequency will always be twice the original frequency. We can also phrase this as being the two frequencies added together. Since the two frequencies are the same, this means the same thing.

- The resulting mean level will be the resulting amplitude multiplied by the Sine of the difference in phase between the original waves added to 90 degrees. [It does not matter which way around the difference is calculated]. This can also be phrased as the resulting amplitude multiplied by the *Cosine* of the difference in phase between the original waves.

- The resulting phase will be the sum of the original phases minus 90 degrees.

The formula that gives the resulting wave for all of this is as follows: [split into three lines to make it clearer, with the mean level on the first line]
$0.5 * (a_1 * a_2) \sin (\phi_1 - \phi_2 + 90)$
+
$0.5 * (a_1 * a_2) \sin ((360 * 2f * t) + (\phi_1 + \phi_2 - 90))$
... where:
“a_1” and “a_2” are the amplitudes of each wave.
“ϕ_1” and “ϕ_2” are the phases of each wave.
“f” is the frequency of each wave (the frequencies are the same)
... and the mean level of each original wave is zero.

We can rephrase this formula so that instead of subtracting 90 degrees from the phase of the second half, we add 270 degrees. This means the same thing, but loses the pattern of having 90 degrees added in the first half, and 90 degrees subtracted in the second half:
$0.5 * (a_1 * a_2) \sin (\phi_1 - \phi_2 + 90)$
+
$0.5 * (a_1 * a_2) \sin ((360 * 2f * t) + (\phi_1 + \phi_2 + 270))$

Sine multiplied by Cosine

Multiplying a Sine wave with no phase by a Cosine wave with no phase is the same as multiplying a Sine wave by another Sine wave with a +90 degree phase. Because the phases are 90 degrees apart, the resulting signal will have a zero mean level. We can tell this is true because the mean level will be either:
$0.5 * (a_1 * a_2) \sin (90 - 0 + 90) = 0.5 * (a_1 * a_2) \sin (180)$
... or:
$0.5 * (a_1 * a_2) \sin (0 - 90 + 90) = 0.5 * (a_1 * a_2) \sin (0)$
... both of which end up as:
$0.5 * (a_1 * a_2) * 0$
$= 0$.

The resulting phase will be:
$0 + 90 - 90 = 0$ degrees.

As an example, if we multiplied “y = sin 360t” by “y = cos 360t”, the result would be: “y = 0.5 sin (360 * 2t)”

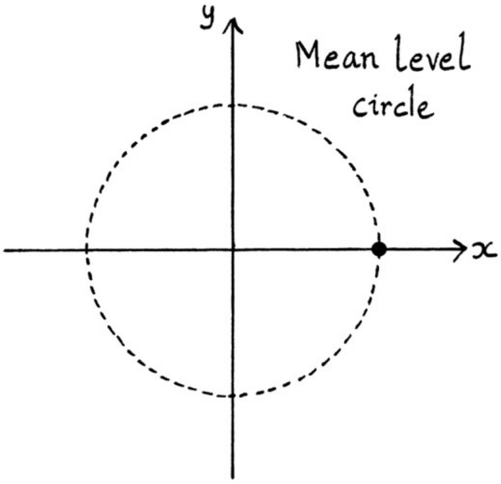
Visualising results with circles

We can understand the *meaning* of a result, but not necessarily how we obtained that result, by thinking of circles. We can think of the general formula as really being the sum of two circles. First, we will look at the first part:
$0.5 * (a_1 * a_2) \sin (\phi_1 - \phi_2 + 90)$

To portray this as an object rotating around a circle, we will make it into a wave that relates to time, but one that has a frequency of zero cycles per second:
$0.5 * (a_1 * a_2) \sin ((360 * 0t) + (\phi_1 - \phi_2 + 90))$

The circle, from which this wave derives, will be based on the two waves:
$0.5 * (a_1 * a_2) \sin ((360 * 0t) + (\phi_1 - \phi_2 + 90))$
... and:
$0.5 * (a_1 * a_2) \cos ((360 * 0t) + (\phi_1 - \phi_2 + 90))$

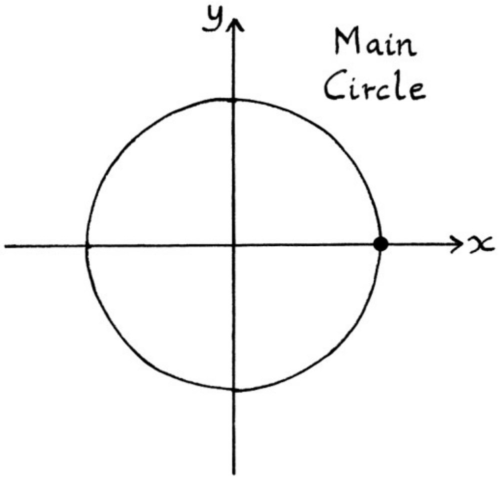
This circle shows the movement of an object that is “rotating” around a circle at 0 cycles per second. In other words, the object is stationary and the circle will act as a mean level for the other circle, with the other circle centred on its phase point. We will call this circle the “mean level circle”. It looks like this:



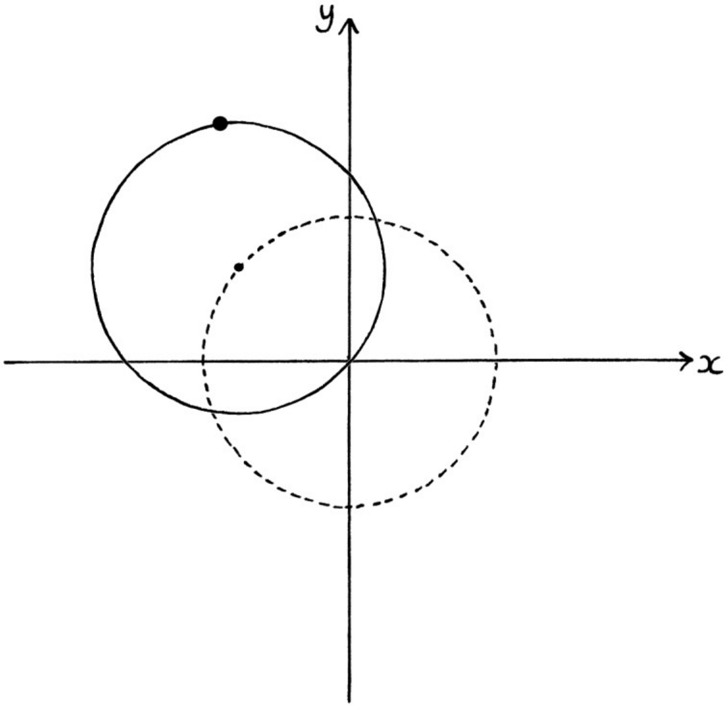
The second part of the formula is:
$0.5 * (a_1 * a_2) \sin ((360 * 2f * t) + (\phi_1 + \phi_2 - 90))$

The circle, from which this wave is derived, is the one based on:
$0.5 * (a_1 * a_2) \sin ((360 * 2f * t) + (\phi_1 + \phi_2 - 90))$
... and:
$0.5 * (a_1 * a_2) \cos ((360 * 2f * t) + (\phi_1 + \phi_2 - 90))$

We will call this circle the "main circle", as that is as good a name as anything else. It looks like this:

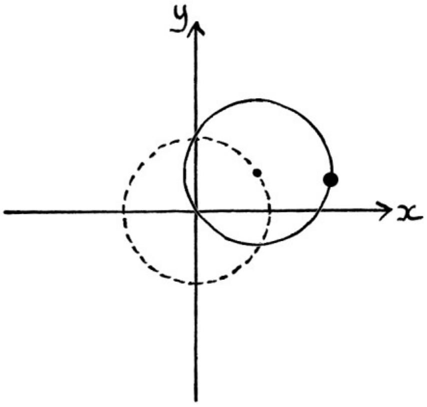


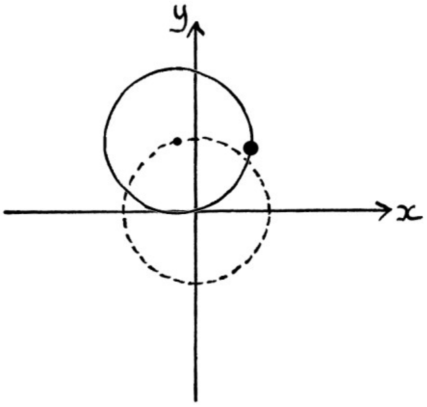
The two circles have the same radius. It is $0.5 * (a_1 * a_2)$, or to put it in English, half the two original amplitudes multiplied by each other. The "mean level" circle will be centred on the origin of the axes. The "main" circle will be placed on the phase point of the "mean level" circle.

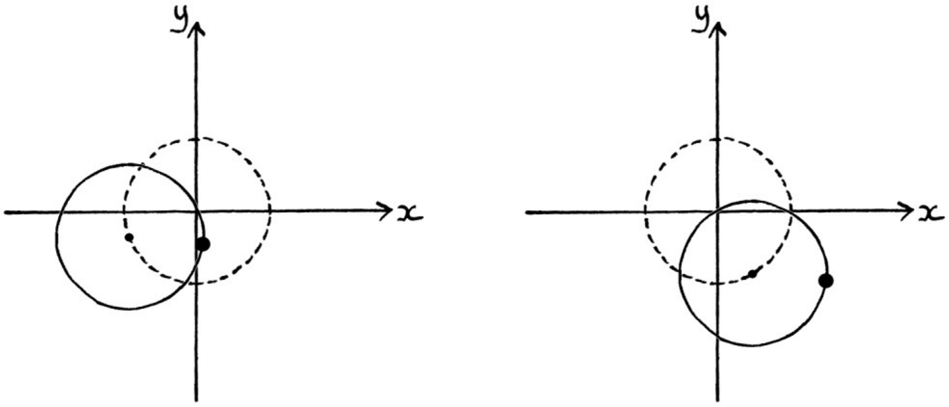


The outer object rotates around the outer "main" circle, yet the outer "main" circle is stationary on the "mean level" circle's phase point.

Depending on the characteristics of the two multiplied waves, the "mean level" circle's phase point might be anywhere on its edge, but the "main" circle will *always* be centred on that phase point.







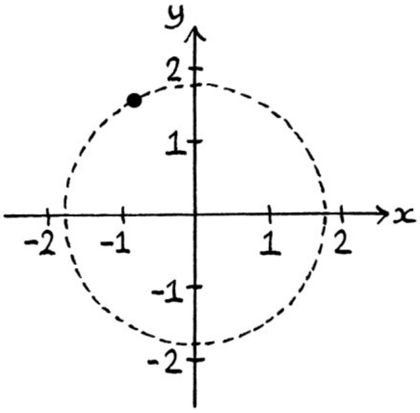
As an example of this idea in use, we will multiply:
“y = 1.7 sin ((360 * 2.2t) + 45)”
by:
“y = 2.1 sin ((360 * 2.2t) + 17)”

The result will be:
0.5 * (1.7 * 2.1) sin (45 – 17 + 90)
+
0.5 * (1.7 * 2.1) sin ((360 * 4.4t) + (17 + 45 – 90))

...which ends up as:
1.785 sin (118)
+
1.785 sin ((360 * 4.4t) + 332)

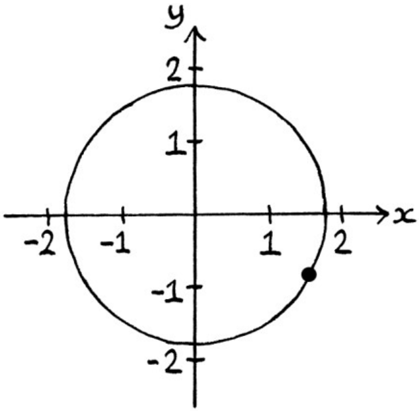
We rephrase the first part to be a time related Sine wave, and we have:
1.785 sin ((360 * 0t) + 118)
Turning these into circles, we have the “mean level” circle as that based on:
1.785 sin ((360 * 0t) + 118)
... and:
1.785 cos ((360 * 0t) + 118)

It looks like this:

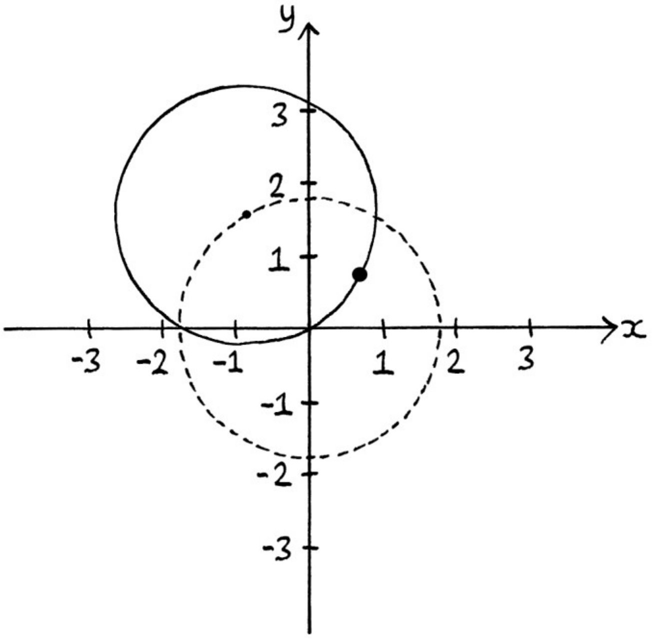


... and we have the “main” circle as that based on the waves:
1.785 sin ((360 * 4.4t) + 332)
... and:
1.785 cos ((360 * 4.4t) + 332)

It looks like this:



Joined together, the circles look like this:



Supposing the phases in the original multiplication were different, then the phase points of the “mean level” circle and the “main” circle would be different. However, the radiuses would still be the same, the “mean level” circle would still be centred on the origin of the axes, and the “main” circle would still be centred on the “mean level” circle’s phase point. No matter what happens, the “main” circle is trapped on the “mean level” circle’s edge.

Portraying the result of multiplication as circles helps in understanding the constraints of the results and gives clues as to why the results end up as they do. In practice, if we only want to find the result of a multiplication, there is probably not much point in doing it.

Note that we are not portraying the process of multiplication using circles – we are portraying the addition that is equivalent to the result with circles. We are showing which two circles when added together are equivalent to a multiplication.

The importance of mean level

One of the most important aspects of this section to remember is that the mean level fluctuates according to the phases of the multiplied waves. The rules relating to this are, as I have said before:

- When the phases of both original waves are the same, the resulting mean level is at its maximum.
- When the phases of both original waves are 90 degrees apart, the resulting mean level is zero. This also means that if the phases are 270 degrees apart, the resulting mean level is zero too, because if they are 270 degrees apart, then they are also 90 degrees apart.
- When the original phases are 180 degrees apart, the resulting mean level is at its minimum.

Given this knowledge, if we saw the result of a multiplication, we would be able to know something about the relationship between the phases of the original waves by paying attention to the resulting mean level.

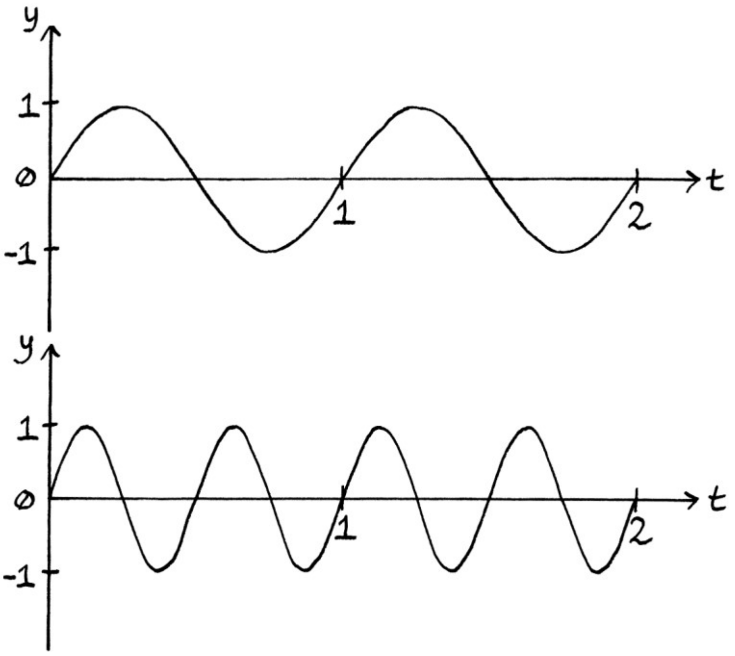
Later on in this book, these rules will have more importance.

Different frequencies

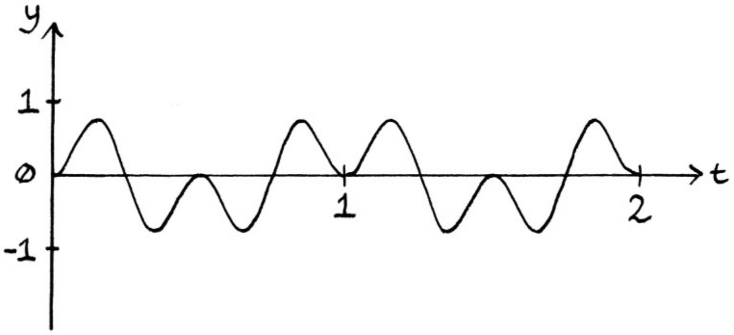
We will now look at multiplying waves that have different frequencies (and zero mean levels). When we multiplied waves of the same frequency, the result was a pure wave. Now we are multiplying waves of different frequencies, the results will *not* be pure waves.

Example 1

As a simple example, we will multiply "y = sin 360t" and "y = sin (360 * 2t)". These two waves look like this:



The resulting signal looks like this:



You might recognise that the shape of this signal is the same as if we had *added* two waves with a frequency ratio of 3 : 1. [It looks like the letters MWMWMW repeated forever].

The main observations about the resulting signal are:

- It is *not* a pure wave.

- The maximum y-axis value is +0.7698 units

- The minimum y-axis value is −0.7698 units.

- The centre point is at y = 0, or in other words, the signal's mean level is 0 units.

- The frequency of the resulting signal is 1 cycle per second. This is the frequency of the slower of the original frequencies.

- The resulting signal is the same as two waves added together that have a frequency ratio of 3 : 1, which is something we can tell by recognising the pattern. In fact, the resulting signal is identical to the result of adding these two waves together:
  "y = 0.5 sin ((360 * 3t) + 270)"
  ... and:
  "y = 0.5 sin (360t + 90)".

If we give the resulting signal's formula in terms of the multiplication that produced it, we would give it as:
"y = (sin (360 * 2t)) * (sin 360t)"

If we give the resulting signal's formula in terms of the *addition* that would produce an identical result, we could give it as:
"y = 0.5 sin ((360 * 3t) + 270) + 0.5 sin (360t + 90)"

From this, we can say that:
"y = (sin (360 * 2t)) * (sin 360t)"
... and:
"y = 0.5 sin ((360 * 3t) + 270) + 0.5 sin (360t + 90)"
... are identical.

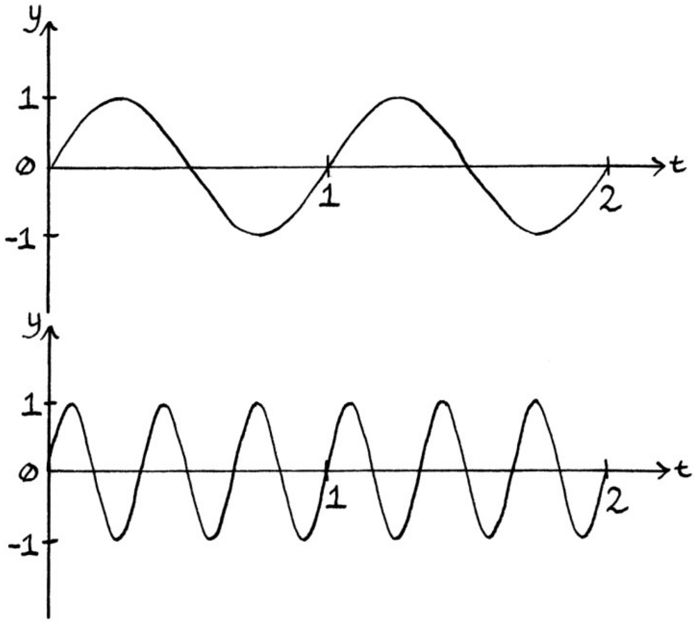
[Note that all periodic signals that are not pure waves can be said to be the sum, or approximately the sum, of two or more pure waves, but in this case, it is obvious straightaway that the signal is the sum of *exactly* two pure waves].

The maximum or minimum points are not more than the amplitudes of either of the original waves because in one cycle of the slower wave, there is never a time when both waves peak or dip to their full extent at the same time. The slower wave peaks at 0.25 seconds and dips at 0.75 seconds, but the faster wave peaks at 0.125 and 0.625 seconds, and dips at 0.375 and 0.875 seconds.

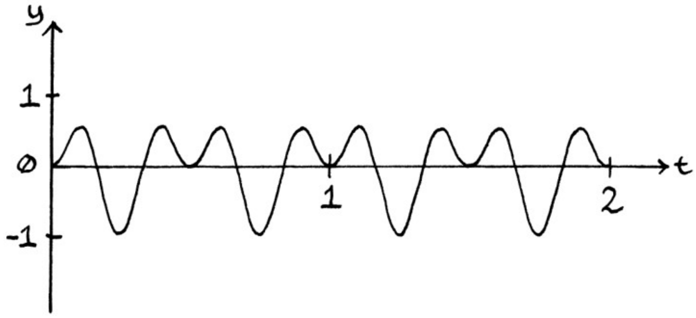
With the *addition* of waves, a resulting pattern repeats when the original waves' cycles align. With *multiplication*, this is also true, however, the resulting pattern might repeat twice as soon if negative values become multiplied together to mimic positive values being multiplied together. This was seen when we were multiplying two waves with the same frequency – the resulting frequency was twice that of the original frequency.

Example 2

We will multiply "y = sin 360t" and "y = sin (360 * 3t)". These two waves look like this:



The resulting signal looks like this:



Significant observations about this signal are:

- It has a frequency of 1 cycle per second.

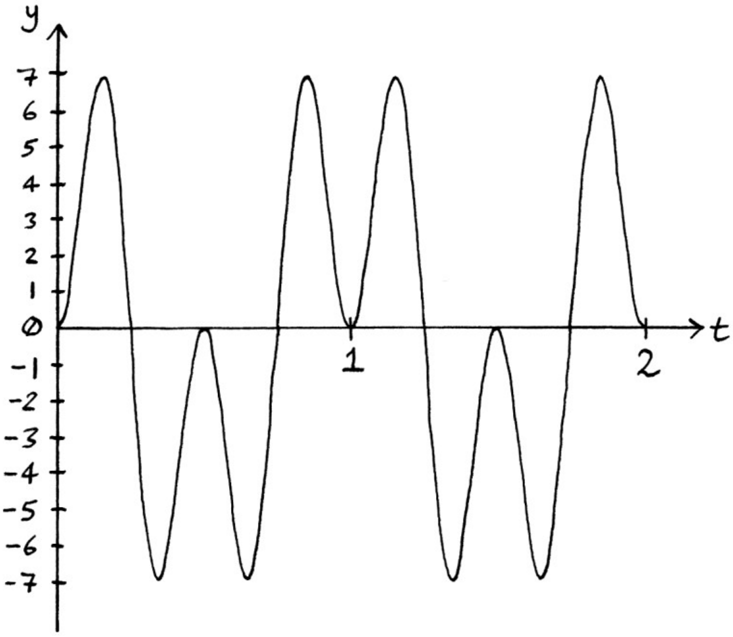
- The signal's shape is not symmetrical around y = 0. The maximum y-axis value is 0.5625, yet the minimum y-axis value is −1. However, the mean level of the signal is still zero. The single dip in each cycle is the size of the two peaks in each cycle, so it all averages out to zero.

In the previous example, the signal was recognisable as the sum of two waves. In this case, this signal is also the same as the sum of two waves. These are:
"y = 0.5 sin ((360 * 2t) + 90)"
... and:
"y = 0.5 sin (360 * 4t) + 270)".

[This is an addition with a 2 : 1 frequency ratio, but with phases that skew it and thus disguise its appearance.]

Example 3

We will multiply “y = 3 sin (360t)” and “y = 3 sin (360 * 2t)”. This is the same as the first example, but the amplitudes are both 3 units. The resulting signal looks like this:



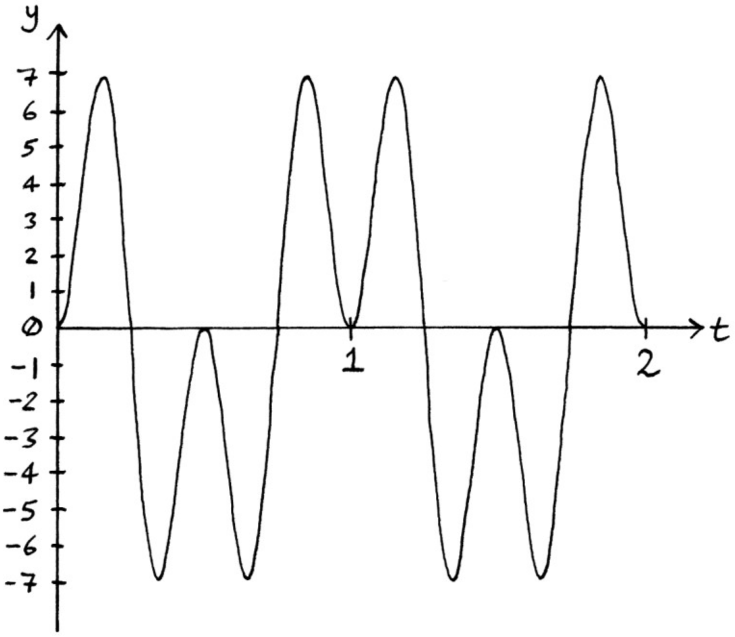
Observations about the resulting signal are:

- The signal has a frequency of 1 cycle per second.
- The maximum y-axis value is +6.9281 units.
- The minimum y-axis value is −6.9281 units.
- The mean level is zero.
- The signal is identical to:
  “y = 4.5 sin ((360 * 1t) + 90)”
  ... added to:
  “y = 4.5 sin ((360 * 3t) + 270)”.

  Notice how these two added waves have amplitudes that are half the result of multiplying the original multiplied waves. In other words, (3 * 3) ÷ 2 = 4.5. This happened to be true in Examples 1 and 2, but it was not obvious because then the original amplitudes were all 1 unit.

Example 4

Next, we will multiply “y = 4.5 sin (360t)” and “y = 2 sin (360 * 2t)”. These are the same two waves as in the previous example, except the amplitudes are different. However, multiplying the amplitudes together still results in the number 9, as was true in the last example. The resulting signal is identical to that in the last example:



Therefore, the resulting signal is *still* equal to the sum of these waves:
“y = 4.5 sin ((360 * 1t) + 90)”
... and:
“y = 4.5 sin ((360 * 3t) + 270)”.

In fact, it is the *result* of multiplying the two original amplitudes that is relevant to a resulting signal, and not what the actual individual amplitudes are. For example, if everything else is the same, then two original amplitudes of 3 and 3, or 9 and 1, or 18 and 0.5, or 4.5 and 2, will all produce exactly the same result. The reason for this becomes clear if we think about multiplication in general. If we have the multiplication:
4.5 * a * 2 * b
... then it is the same as:
4.5 * 2 * a * b
... which is:
9 * a * b

With waves, if we have the multiplication:
4.5 sin (360t) * 2 sin (360 * 2t)
... then it is the same as:
4.5 * 2 * sin (360t) * sin (360 * 2t)
... which is:
9 * sin (360t) * sin (360 * 2t).

Any two amplitudes for these two waves, that, when multiplied together, result in 9, will produce the same signal. That signal will be equivalent to two added waves each with an amplitude of 4.5 units.

Thoughts so far

Given that each multiplication we have seen so far results in a signal that can be said to be the sum of two waves, the best way of describing the result of a multiplication is in terms of those two added waves. It is much easier to visualise a signal as being the sum of two waves, than it is to visualise it as the result of multiplying two waves (the "product" of two waves). The sum of two waves can also be portrayed on a circle, and therefore, any strange characteristics can be understood by drawing the circles. The reason for the maximum and minimum values of the resulting signals so far is not particularly obvious, but when we think of the circles from which the two equivalent added waves can be said to be derived, the reasons will be clearer. [It is due to the way that the cycles line up, and how the movement of an outer object on the circle chart does not necessarily follow a simple or symmetrical path.]

We will call the original waves that are multiplied together, "the Multiplication Waves" and the waves that when added together produce the same signal, "the Addition Waves". So far, in our examples, the signals that were created by multiplying two Multiplication Waves were identical to the sum of exactly two Addition Waves. In later examples in this chapter, the signal created by the two Multiplication Waves might be equal to the sum of more than two Addition Waves.

For the examples so far, we can say that the multiplication of two waves results in a signal that is also result of the addition of two other waves, or we can say that the multiplication of two waves is *equivalent* to the addition of two other waves. These two ways of thinking amount to exactly the same thing.

The resulting amplitude of each Addition Wave will be half the original amplitudes multiplied together. The amplitudes are multiplied together and then shared between each Addition Wave.

The formula for this is:
“amplitude of each Addition Wave = 0.5 * (amplitude$_1$ * amplitude$_2$)”
... or to put this slightly more mathematically:
“$a_{a1} = a_{a2} = 0.5\ (a_{m1} * a_{m2})$”
... where a_{a1} and a_{a2} are the amplitudes of the Addition Waves, and a_{m1} and a_{m2} are the amplitudes of the Multiplication Waves.

Negative amplitudes

If we wanted to multiply two waves with negative amplitudes, we could rephrase them to be waves with positive amplitudes and phases of 180 degrees. However, we will look at what happens if we keep the amplitudes negative. We will multiply:
“y = −3 sin (360t)”
... and:
“y = 3 sin (360 * 2t)”.

Following the rules we know so far, the amplitudes of the resulting Addition Waves should both be 0.5 * (−3 * 3) = −4.5 units. In fact, the formulas of the Addition Waves in full are:
“y = −4.5 sin ((360 * t) + 90)”
... and:
“y = −4.5 sin ((360 * 3t) + 270)”

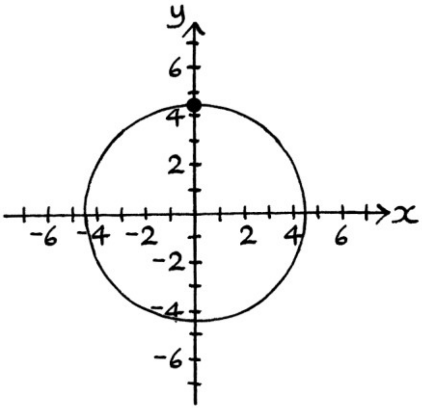
It does not matter whether the amplitudes are positive or negative – the rule for calculating the amplitudes will always work. This is one of the situations where it is simpler to use negative amplitudes than it is to make them positive and add 180 degrees to their phases.

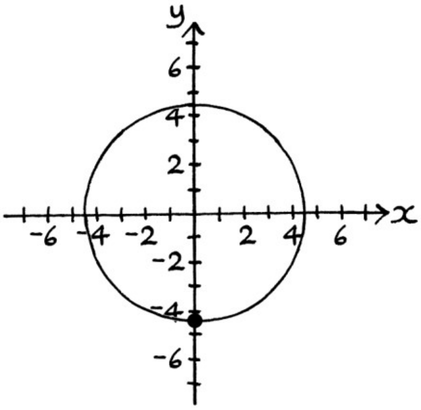
Using circles to visualise the Addition Waves

When multiplying two waves, we can portray the equivalent Addition Waves as circles on a circle chart. This helps visualise the result. We will use the previous example of:
"y = 3 sin (360t)"
... multiplied by:
"y = 3 sin (360 * 2t)"

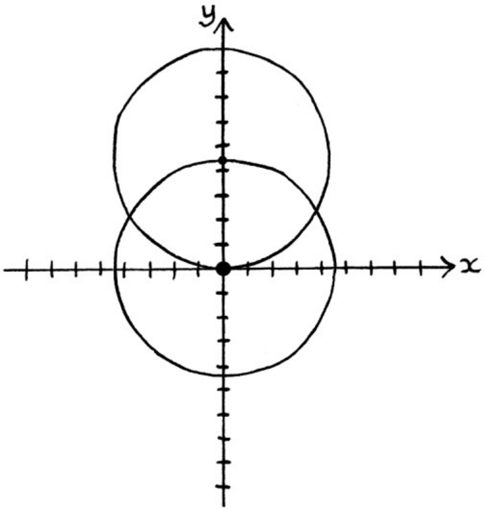
The equivalent Addition Waves are:
"y = 4.5 sin ((360 * 1t) + 90)"
... added to:
"y = 4.5 sin ((360 * 3t) + 270)"

These waves drawn as circles look like this:

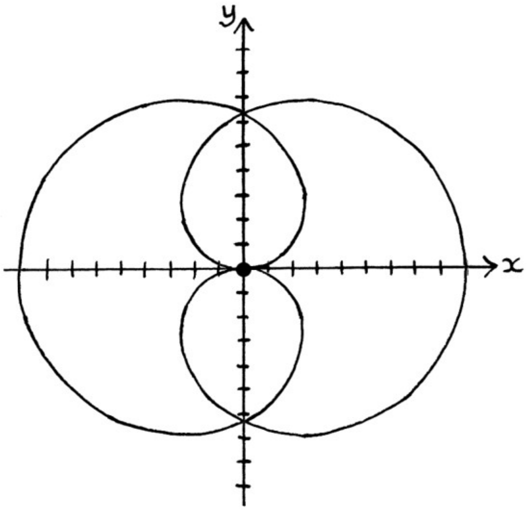




The circles arranged for adding look like this:



The resulting shape looks like this:



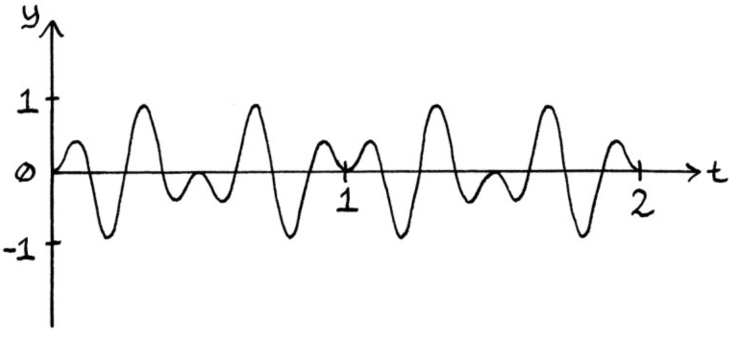
[Remember that we are not portraying the process of multiplication using circles. Instead, we are portraying the equivalent addition with circles.]

Seeing the waves as the sum of two circles helps in better understanding the result. Earlier in this chapter, when we were multiplying two waves with the same frequency, we saw how the object "rotating" around the first circle was stationary, and thus acting as a mean level. Now that the frequencies are different, the objects are both moving around their circles.

Example 5

Given that we now know how the amplitudes of the Addition Waves are calculated, future examples will stick to amplitudes of 1 to simplify matters. Now, we will multiply "y = sin 360t" by "y = sin (360 * 4t)". These have a frequency ratio of 4 : 1.

The result looks like this:



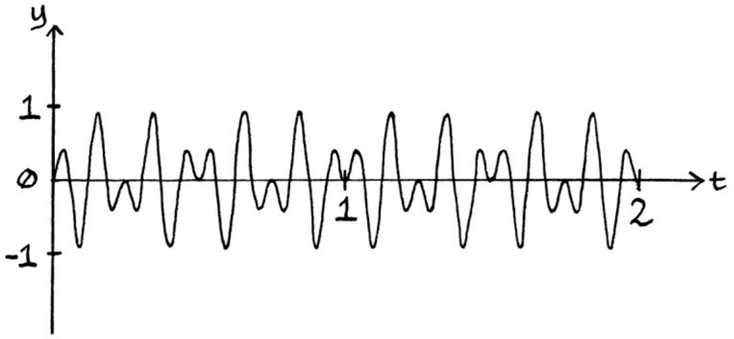
This signal is the same as the sum of:
“y = 0.5 sin ((360 * 3t) + 90)”
... and:
“y = 0.5 sin ((360 * 5t) + 270)”

The original Multiplication Waves had a frequency ratio of 4 : 1. The Addition Waves that produce the same signal have a frequency ratio of 5 : 3.

Example 6

We will multiply “y = sin (360 * 2t)” by “y = sin (360 * 8t)”. These waves also have a frequency ratio of 4 : 1, as in the last example, but the actual values of the frequencies are twice what they were before.

The result looks like this:



The result is equal to the sum of these two waves:
“y = 0.5 sin ((360 * 6t) + 90)”
... and:
“y = 0.5 sin ((360 * 10t) + 270)”

These Addition Waves have a frequency ratio of 10 : 6. This is twice the ratio in Example 5, where the ratio was 5 : 3. The Addition Waves and the Multiplication Waves in this example have twice the frequency of those in Example 5. From this, we can say that scaling the frequencies of two multiplied waves will produce two added waves that are scaled in the same way in their frequencies.

Frequency ratios

Here is a table showing the relationship between the original Multiplication Wave frequency ratio, and the resulting Addition Wave frequency ratio. Note that I have not reduced the ratios to a number to 1. By not doing this, the relationships are easier to see.

Frequency ratio of Multiplication Waves	**Frequency ratio of equivalent Addition Waves**
2 : 1	3 : 1
3 : 1	4 : 2
4 : 1	5 : 3
5 : 1	6 : 4
6 : 1	7 : 5
7 : 1	8 : 6

With each step down the table, both halves of the Addition Wave ratio increase by 1. We can also see that at every step down the table, the second half of the Addition Wave ratio is 2 units less than the first half – the first is 3 : 1, the second is 4 : 2, the third is 5 : 3 and so on.

Given all this information, we can formulate a rule *for the values in this table*:
“If the frequency ratio of the Multiplication Waves is z : 1, then the ratio for the equivalent “Addition Waves” will be (z + 1) : (z − 1).”

We can test this. We will multiply:
“y = sin (360 * 1t)”
... and:
“y = sin (360 * 15t)”

These have a frequency ratio of 15 : 1. Therefore, the resulting Addition Waves should have a ratio of 16 : 14. It turns out the resulting Addition Waves are:
“y = 0.5 sin ((360 * 14t) + 90)”
... and:
“y = 0.5 sin ((360 * 16t) + 270)”

This means the rule is correct. [I will explain how we can calculate which waves were added to create a signal in Chapter 18.]

The above rule is fine for when the multiplication ratio is an integer to one, but it does not work if the ratio is an integer to two, such as 5 : 2. Here is a table showing multiplication ratios to two:

Frequency ratio of Multiplication Waves	**Frequency ratio of equivalent Addition Waves**
3 : 2	5 : 1
4 : 2	6 : 2
5 : 2	7 : 3
6 : 2	8 : 4
7 : 2	9 : 5
8 : 2	10 : 6

The rule here is:
“If the frequency ratio of the Multiplication Waves is z : 2, then the ratio for the resulting Addition Waves will be (z + 2) : (z – 2).”

Therefore, if we have a multiplication frequency ratio of 15 : 2, that will be equivalent to an addition frequency ratio of:
15 + 2 : 15 – 2
... which is:
17 : 13

For ratios to 3, we can find a rule using this table:

Frequency ratio of Multiplication Waves	**Frequency ratio of equivalent Addition Waves**
4 : 3	7 : 1
5 : 3	8 : 2
6 : 3	9 : 3
7 : 3	10 : 4
8 : 3	11 : 5
9 : 3	12 : 6

The rule here is:
"If the frequency ratio of the Multiplication Waves is z : 3, then the ratio for the equivalent Addition Waves will be (z + 3) : (z − 3)."

Similar rules apply for frequency ratios of z : 4, z : 5, and so on.

If we examine these three tables and their rules, we can see a pattern that will lead on to a general rule for *any* ratio:
"If a multiplication frequency ratio is a : b, then the equivalent addition frequency ratio will be: (a + b) : (a – b)"

We can stop thinking about ratios, and have a general rule as this:
"Given the two multiplication frequencies, one of the equivalent addition frequencies will be the sum of the two multiplication frequencies; the other will be the faster minus the slower of the two multiplication frequencies."

To make a standard rule for this, we will say that:

- The frequency of the *first* Addition Wave will be equal to the frequency of the faster Multiplication Wave minus the frequency of the slower Multiplication Wave.
- The frequency of the *second* Addition Wave will be equal to the sum of the frequencies of the Multiplication Waves.

In this way, we are classifying each of the Addition Waves as being either "first" or "second". Strictly speaking, the idea of ordering them is unnecessary, and which is considered "first" or "second" is a completely arbitrary choice. However, thinking of the two Addition Waves in this way will make future calculations more straightforward, especially as the calculations become more complicated.

We can put the rule slightly more mathematically as:
The first Addition Wave frequency = $f_1 - f_2$
The second Addition Wave frequency = $f_1 + f_2$
... where f_1 and f_2 are the frequencies of the Multiplication Waves, and f_1 is faster than f_2.

We can put this even more mathematically as:
$f_{a1} = f_{m1} - f_{m2}$
$f_{a2} = f_{m1} + f_{m2}$
... where f_{a1} is the frequency of the first Addition Wave, f_{a2} is the frequency of the second Addition Wave, f_{m1} is the *faster* of the two frequencies of the Multiplication Waves, and f_{m2} is the *slower* of the two frequencies of the Multiplication Waves.

The above rule works for all frequencies.

Note how I use the words “faster” and “slower” instead of words such as “larger” and “smaller” or “higher” and “lower”. This is because the rule works for negative frequencies too. If we have a frequency of −10 cycles per second, then it is *faster* than a frequency of −2 cycles per second. An object rotating around a circle the “wrong” way at 10 cycles per second will be completing more revolutions per second than one going the “wrong” way at 2 cycles per second. The number −10 is a lower number than −2, but it is a faster frequency. Similarly, supposing we had frequencies of −5 cycles per second and +3 cycles per second, then −5 is the faster of the two frequencies, despite being a lower number. An object rotating at −5 cycles per second is completing more revolutions per second than one completing +3 cycles per second.

As an example of the rule working, we will multiply a wave with a frequency of 8 cycles per second against a wave with a frequency of 5 cycles per second. The frequency of the first Addition Wave will be: 8 – 5 = 3 cycles per second, and the frequency of the second Addition Wave will be: 8 + 5 = 13 cycles per second.

If we multiply waves with frequencies of 123 cycles per second and 77 cycles per second, then the frequency of the first Addition Wave will be: 123 – 77 = 46 cycles per second, and the frequency of the second Addition Wave will be: 123 + 77 = 200 cycles per second.

If we multiply waves with frequencies of 3 cycles per second and 6 cycles per second, then, *paying attention to which is the faster original frequency*, the frequency of the first Addition Wave will be: 6 – 3 = 3 cycles per second, and the frequency of the second Addition Wave will be: 6 + 3 = 9 cycles per second.

The rule also works for non-integer frequencies. For example, the multiplication frequencies of 7.5 and 1 cycles per second are equivalent to the addition frequencies of: 7.5 – 1 = 6.5 cycles per second, and: 7.5 + 1 = 8.5 cycles per second.

The rule even works for irrational frequencies and irrational ratios. If the Multiplication Wave frequencies are π cycles per second (in other words, 3.141592653589793238462643383... cycles per second) and 1 cycle per second, then the Addition Wave frequencies will be: π − 1 = 2.14159265 cycles per second and π + 1 = 4.14159265 cycles per second.

The rule also works for identical frequencies. In such cases, there is no "faster frequency", so we have to choose one or the other to act as if it had a faster frequency. It does not actually matter which way around the calculation is done in these situations. For example, the Multiplication Wave frequencies of 5 and 5 cycles per second are equivalent to the Addition Wave frequencies of: 5 – 5 = 0 cycles per second, and 5 + 5 = 10 cycles per second. We saw how multiplying two waves with the same frequency works earlier in this chapter, but we can use the new rule to do the same thing. Supposing we were multiplying:
"y = sin (360 * 5t)"
... and:
"y = sin (360 * 5t)"
... then the Addition Wave frequencies would be 0 cycles per second and 10 cycles per second.

No matter which way around we calculated the frequencies, the resulting Addition Waves, in full, would be:
"y = 0.5 sin ((360 * 0t) + 90)"
... and:
"y = 0.5 sin ((360 * 10t) + 270)".

This result is interesting because the first wave has a frequency of zero cycles per second. It represents an object "moving" around a circle at zero cycles per second, or, in other words, an object fixed at the same point on the circle at all times. We could rephrase that wave's formula to be: "y = 0.5 sin (90)", or "y = 0.5". This first wave ends up acting as the mean level for the second wave. Earlier in this chapter, when we were looking at multiplying waves with the same frequencies, we saw how the resulting mean level could be phrased as a time-based wave, and now we are seeing exactly the same thing. Given that, when multiplying waves of the same frequency, it pays to always think of the mean level in the result as a zero-frequency time-based wave in an addition of two waves.

Thoughts on the resulting phases

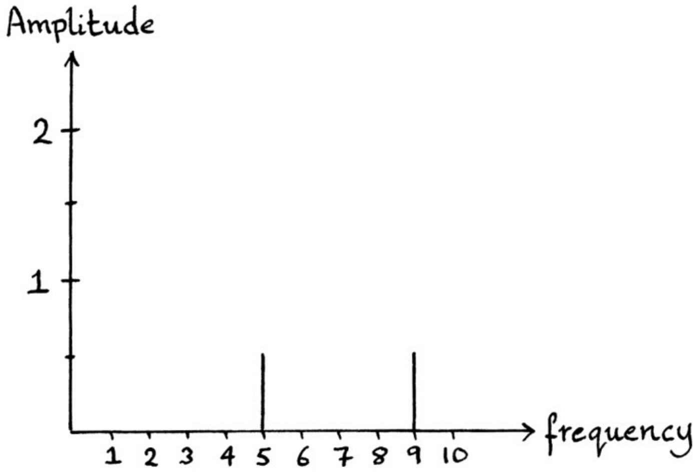
When we multiply two waves with different frequencies, but *zero* phases, a phase of 90 degrees is assigned to the Addition Wave that has the frequency of "$f_{m1} - f_{m2}$", and a phase of −90 degrees (270 degrees) is assigned to the wave that has the frequency of "$f_{m1} + f_{m2}$". In other words, the first wave receives a phase of +90 degrees, and the second wave receives a phase of −90 degrees (which we will usually write as +270 degrees).

Frequency domain graphs

The resulting Addition Waves from a multiplication can be drawn on a wave frequency domain graph. This allows us to see more clearly what is happening. We will multiply:
“y = sin (360 * 7t)”
... and:
“y = sin (360 * 2t)”

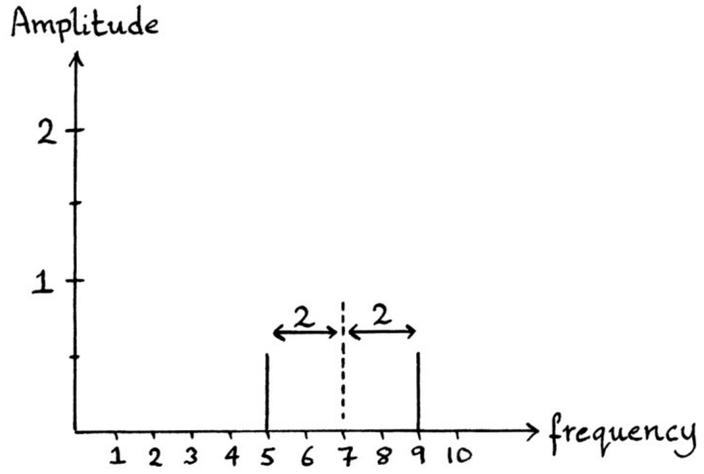
The equivalent Addition Waves are:
“y = 0.5 sin ((360 * 5t) + 90)”
... and:
“y = 0.5 sin ((360 * 9t) + 270)”.

Drawn on a frequency domain graph, they look like this:

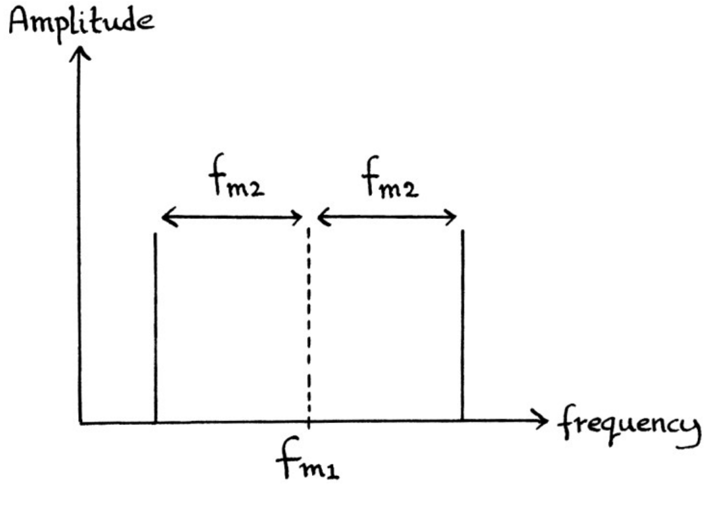


[Remember that as this is a wave frequency domain graph, it cannot portray the phases.]

The two waves appearing on the graph are equidistant from the frequency of 7 cycles per second. They are both 2 cycles per second away:



If we call the faster frequency of a multiplication "f_{m1}" and the slower frequency "f_{m2}", then for the multiplication of waves with zero mean levels, it will always be the case that the two lines on the frequency domain graph showing the addition result will be centred around "f_{m1}", and will be "f_{m2}" cycles per second either side of it:

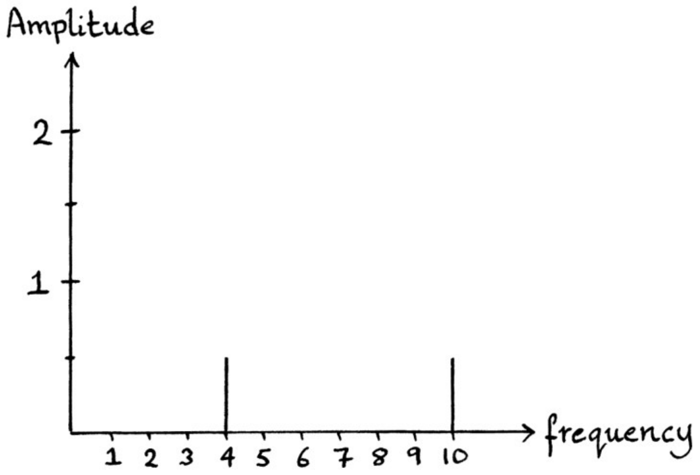


In other words, if we multiply two waves with zero mean levels that have frequencies of 4 and 10 cycles per second, the resulting Addition Wave frequencies will be 4 cycles per second either side of 10. They will be 6 and 14 cycles per second.

Doing the calculation the wrong way round

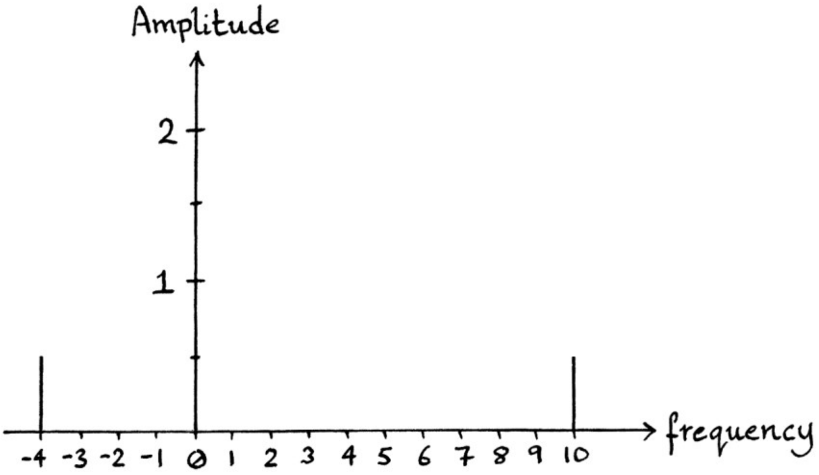
We will look at what happens if we calculate the Addition Wave frequencies the wrong way round. If we multiply the frequencies 7 and 3 cycles per second, then we know that our rule works if we calculate the addition waves as (7 – 3) and (7 + 3), with the slower frequency being subtracted from the faster frequency. If we did the calculation the wrong way round, we would calculate them as (3 – 7) and (3 + 7). As an addition is the same whichever way around it is done, the only difference in the calculations is which way around the subtraction is done.

If our waves are "y = sin (360 * 3t)" and "y = sin (360 * 7t)", then if we do the subtraction the usual way, we end up with:
"y = 0.5 sin ((360 * 4t) + 90)"
... added to:
"y = 0.5 sin ((360 * 10t) + 270)"
... which is the correct result. On a frequency domain graph, these two frequencies appear either side of 7 cycles per second. In fact, they are 3 cycles per second either side of 7 cycles per second. [Remember, as always, that the graph does not show their phases].

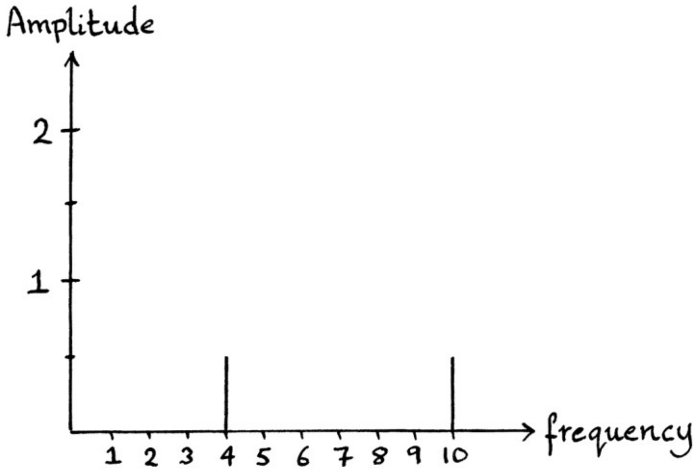


If we do the subtraction the other way around, we end up with:
"y = 0.5 sin ((360 * −4t) + 90)"
... added to:
"y = 0.5 sin ((360 * 10t) + 270)"

These two frequencies are either side of +3 cycles per second. In fact, they are 7 cycles per second either side of 3 cycles per second:



However, the wave “y = 0.5 sin ((360 * −4t) + 90)” is actually the same as the wave “y = 0.5 sin ((360 * 4t) + 90)”. This is because a Sine wave with a negative frequency and a phase of +90 degrees is the same as that Sine wave with the frequency made positive. [For converting a Sine wave formula with a negative frequency to one with a positive frequency, think of the circle that represents that wave, and mirror it left-to-right or right-to-left.] The above frequency domain graph could be shown as in the following picture too, and the phases (which are not shown) would still be the same:



Therefore, we can say that it does not matter which way around we do the subtraction in this particular case, because both possible resulting formulas will represent the same wave.

If we were multiplying waves with non-zero phases, then it would probably not be the case that we end up with a Sine wave with a 90-degree phase, in which case the negative-frequency wave would not have the same phase as that same wave with the frequency made positive. However, it would still be possible to convert the negative frequency to a positive frequency, and that positive frequency would have the same value as if we had done the original subtraction the normal way around.

To give an example of this, we will temporarily jump ahead to show multiplication of different frequencies and non-zero phases. We will multiply:
"y = 3 sin ((360 * 11t) + 327)"
... by:
"y = 2 sin ((360 * 5t) + 45)"

We will do this the normal way first. The amplitudes of both resulting Addition Waves will be 0.5 * (2 * 3) = 3 units. The frequency of the first wave will be 11 – 5 = 6 cycles per second. The frequency of the second wave will be 11 + 5 = 16 cycles per second. The phases will be 12 degrees and 282 degrees (I will explain how to calculate these later in the chapter). The two resulting Addition Waves will be:
"y = 3 sin ((360 * 6t) + 12)"
... and:
"y = 3 sin ((360 *16t) + 282)"

Now, we will do this the reverse way. The amplitudes will still be 3 units each. The frequency of the first wave will be 5 – 11 = −6 cycles per second. The frequency of the second wave will be 5 + 11 = 16 cycles per second. The phases will be 168 degrees and 282 degrees. The two resulting Addition Waves will be:
"y = 3 sin ((360 * −6t) + 168)"
... and:
"y = 3 sin ((360 * 16t) + 282)"

From this, we can see that the second Addition Wave will be the same no matter in which order we do the calculation. However, if we do the calculation in the "wrong" order, the first Addition Wave will end up as the negative frequency version of the "correct calculation way" wave. If we imagine:
"y = 3 sin ((360 * 6t) + 12)"
... on a circle, then we will be able to see that:
"y = 3 sin ((360 * −6t) + 168)"
... will be its mirror image. The phase points of 12 degrees and 168 degrees are both equidistant from 90 degrees [and from 270 degrees]. This means that the formulas of "y = 3 sin ((360 * 6t) + 12)" and "y = 3 sin ((360 * −6t) + 168)" refer to *exactly* the same curve. [See Chapter 11 to refresh your memory on negative

frequencies]. They are, in essence, the same thing. If they are the same thing, then the sum of the two Addition Waves will also refer to exactly the same curve too.

This all means that it does not actually make any difference as to which way around the calculation is done. The resulting formulas might be different, but the curves those formulas represent will be the same. However, in my opinion, it makes everything a lot easier if we end up with either two positive-frequency Addition Wave formulas or two negative-frequency Addition Wave formulas after doing the calculation. First, if the frequencies are of the same type, we can tell much more easily how the frequencies of the Addition Waves relate to each other. Second, it is simpler not to have to convert negative frequencies to positive frequencies or vice versa, so there is an advantage in subtracting the slower frequency from the faster frequency. Therefore, I recommend that you always do the calculation paying attention to which is the faster frequency wave. It makes everything generally easier.

One zero frequency

If we multiply:
“y = sin (360 * 2t)”
... by:
“y = sin (360 * 0t)”
... then the resulting Addition Waves will have frequencies of (2 – 0) cycles per second and (2 + 0) cycles per second, which are the same as 2 cycles per second and 2 cycles per second.
The resulting Addition Waves are:
“y = 0.5 sin ((360 * 2t) + 90)”
... and:
“y = 0.5 sin ((360 * 2t) + 270)”

If you visualise what these waves look like, you might realise that they are the same but 180 degrees apart. Therefore, when adding them together, they cancel each other out, and the result is just a straight line at y = 0.

We could have guessed this result by thinking about how “y = sin (360 * 0t)” is really the same as “y = 0t” or “y = 0”, where “y” is zero for all time. We are, therefore, multiplying one wave by zero, which is the same as multiplying every single y-axis value of that wave by zero, which in turn results in “y = 0” for all time.

Negative frequencies

We will multiply these two waves:
“y = sin (360 * −2t)”
... and:
“y = sin (360 * −5t)”

They both have negative frequencies. [We could rephrase these to have positive frequencies and phases of 180 degrees, but we will look at multiplying waves with different frequencies and non-zero phases later in this chapter.]

We will use our rule to calculate the resulting Addition Waves. The *faster* frequency is −5. Therefore, the addition frequencies will be: −5 − −2 = −3 cycles per second, and: −5 + −2 = −7 cycles per second. [Remember that when it comes to negative frequencies, the *faster* frequency is not necessarily the *higher* frequency. The number −2 is a *higher* number than −5, but −5 cycles per second is a *faster* frequency – the cycles repeat more quickly. An object rotating around a circle at 5 cycles per second the wrong way around is completing more cycles than one doing 2 cycles per second the wrong way (or the right way) around. We could say we are looking for the *absolute* larger value – as in the value made positive regardless of whether it was positive or negative to start with. Another way of thinking about this is that we are looking for the number that is furthest away from zero, whether that number is negative or positive.]

The resulting Addition Waves are:
“y = 0.5 sin ((360 * −3t) + 90)”
... and:
“y = 0.5 sin ((360 * −7t) + 270)”

As another example, we will multiply:
“y = sin (360 * −7.8t)”
... and:
“y = sin (360 * −2.5t)”

The faster frequency is −7.8, so the Addition Waves will have frequencies of:
−7.8 − −2.5 = −5.3 cycles per second, and: −7.8 + −2.5 = −10.3 cycles per second. These are both equidistant from −7.8.

The Addition Waves are, therefore:
"y = 0.5 sin ((360 * −5.3t) + 90)"
... and:
"y = 0.5 sin ((360 * −10.3t) + 270)"

As another example, we will multiply these two waves:
"y = sin (360 * −2t)"
... and:
"y = sin (360 * −2t)"

These have the same frequency, so it does not matter in which order we do the calculation. The resulting Addition Waves will have frequencies of: −2 − −2 = 0 cycles per second, and: −2 + −2 = −4 cycles per second. The resulting Addition Waves will be:
"y = 0.5 sin ((360 * 0t) + 90)"
... and:
"y = 0.5 sin ((360 * −4t) + 270)"

The first wave ends up as "y = 0.5 sin (90)" or "y = 0.5", and acts as a mean level to the second wave. This results in one single wave:
"y = 0.5 + 0.5 sin ((360 * −4t) + 270)"

One negative frequency; one positive frequency

Things become slightly different when we multiply one negative-frequency wave with a positive-frequency wave. As an example, we will multiply these two waves:
"y = sin (360 * −2t)"
... and:
"y = sin (360 * +3t)"

The faster frequency is +3 cycles per second. Therefore, the resulting frequencies will be: 3 − −2 = 5 cycles per second, and: 3 + −2 = 1 cycle per second. The resulting Addition Waves are:
"y = 0.5 sin ((360 * 5t) + 90)"
... and:
"y = 0.5 sin ((360 * 1t) + 270)"

Now, we will multiply these two waves:
“y = sin (360 * −10.5t)”
... and:
“y = sin (360 * +7t)”

The faster frequency is −10.5. Therefore, the resulting Addition Wave frequencies will be:
−10.5 – 7 = −17.5 cycles per second
... and:
−10.5t + 7 = −3.5 cycles per second.

The resulting Addition Waves are:
“y = 0.5 sin ((360 * −17.5t) + 90)”
... and:
“y = 0.5 sin ((360 *−3.5t) + 270)”

Now, we will multiply these two waves:
“y = sin (360 * −2t)”
... and:
“y = sin (360 * +2t)”

The two frequencies are the negative of each other, but they are each equally fast. An object rotating around a circle at 2 cycles per second anticlockwise completes as many cycles as one rotating at 2 cycles per second clockwise. Therefore, it is hard to know which way around to perform the calculation.

Depending on whether we say −2 is the faster frequency, or whether we say +2 is the faster frequency, the resulting addition frequencies will be either:
(−2 – 2) and (−2 + 2), which are −4 and 0
... or:
(2 – −2) and (2 + −2), which are +4 and 0.

This means that the resulting Addition Waves will be either:
“y = 0.5 sin ((360 * −4t) + 90)” and “y = 0.5 sin ((360 * 0t) + 270)”
... or:
“y = 0.5 sin ((360 * +4t) + 90)” and “y = 0.5 sin ((360 * 0t) + 270)”

In the first pair of Addition Waves, the second wave ends up as “y = 0.5 sin (270)”, which becomes: “y = −0.5”. Therefore, it acts as a mean level to the first wave of that pair, and we can rephrase the pair as one wave, which is:
“y = −0.5 + 0.5 sin ((360 * −4t) + 90)”

In the second pair of Addition Waves, the second wave also ends up as “y = −0.5”, and so also acts as a mean level to the other wave. We can give that pair as one wave in the form:
“y = −0.5 + 0.5 sin ((360 * +4t) + 90)”

Therefore, if we consider −2 the faster frequency, we have the result as:
“y = −0.5 + 0.5 sin ((360 * −4t) + 90)”
... and if we consider +2 the faster frequency, we have the result as:
“y = −0.5 + 0.5 sin ((360 * +4t) + 90)”

These formulas refer to exactly the same wave curve. [If you imagine each wave portrayed as a circle, the circles will be mirror images of each other.] These two waves being identical means that it does not matter which of −2 or +2 we treat as the “faster” frequency. This also confirms that the rule for calculating the resulting frequencies works in all situations. However, it is important to note that this would not have worked if the Multiplication Waves had had non-zero phases, as we will see later in this chapter.

Rules so far

So far, we know that if we multiply two pure waves of different, or the same, frequencies, and that have any amplitude, zero phases and zero mean levels, then:

- The result will be equivalent to the sum of two pure waves, which we are calling the "Addition Waves".

- Those two resulting summed waves will have amplitudes equal to half the original amplitudes multiplied together.

- The first of the resulting Addition Waves will have a frequency equal to the faster of the two original frequencies minus the slower of the two original frequencies. The second of the resulting Addition Waves will have a frequency equal to the sum of the original frequencies. Remember that it is the speed of the frequencies that matters, and not whether one is higher than the other – this distinction is important if we are dealing with negative frequencies, where, for example, –5 cycles per second is faster than +2 cycles per second.

- The first Addition Wave will have a phase of 90 degrees; the second Addition Wave will have a phase of −90 degrees (which we will generally write as +270 degrees).

- The mean level of the resulting signal and of both the resulting Addition Waves will be zero. [Note, that if we are multiplying waves of the same frequency (whether both positive, both negative, or positive and negative), it is still the case that we have two Addition Waves with zero mean levels. It is just that one of those waves will have zero frequency, and therefore *act* as the mean level to the other wave. The zero-frequency wave, itself, still has a mean level of zero, but its having a zero frequency means that it behaves as a mean level.]

From all of the above, we can create a general rule for calculating the two resulting waves that when added together are equivalent to two other waves multiplied together.

If we multiply:
“$y = a_1 \sin (360 * f_1 * t)$”
... by:
“$y = a_2 \sin (360 * f_2 * t)$”
... where f_1 is faster than, the negative of, or equal to f_2, then the equivalent sum of waves will be:
“$y = 0.5 * (a_1 * a_2) * \sin ((360 * (f_1 - f_2)) + 90)$”
... added to:
“$y = 0.5 * (a_1 * a_2) * \sin ((360 * (f_1 + f_2)) - 90)$”

Notes:

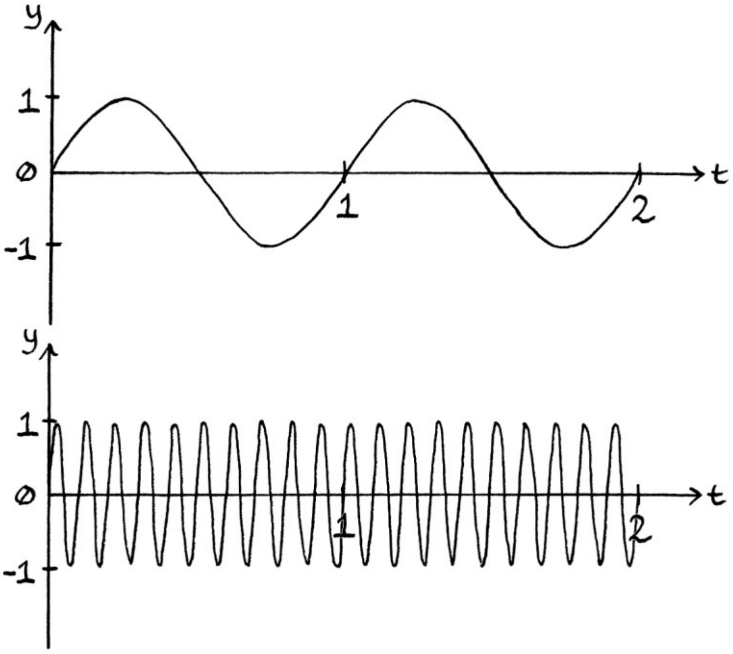
- If f_1 and f_2 are different and both positive *or* both negative, it will always be the case that $(f_1 + f_2)$ will be faster than $(f_1 - f_2)$.
- If f_1 and f_2 are different and positive *and* negative, then the opposite will be true.
- If f_1 and f_2 are the same (whether both positive *or* negative), one of the Addition Waves will have zero frequency and act as a mean level to the other.
- If f_1 and f_2 are the negative of each other, one of the Addition Waves will have zero frequency and act as a mean level to the other, and there will be two possible results for the *formula* of the other wave, each of which will represent exactly the same wave curve.
- If $(f_1 + f_2)$ and $(f_1 - f_2)$ result in the same number, and the original amplitudes are the same, then the Addition Waves will be identical but with phases 180 degrees different from each other, which means that they will cancel each other out and result in “$y = 0$” for all time.

More thoughts on multiplication

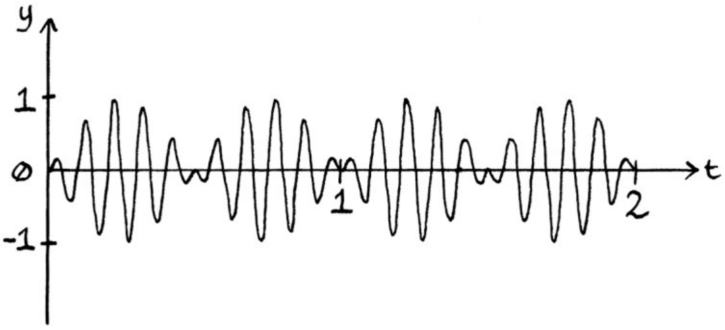
Before we look at the effects of multiplying different frequencies and non-zero phases, we will look at some interesting aspects of multiplication.

High frequency ratios

When we multiply two waves with significantly different frequencies, we produce noteworthy results. We will start by multiplying:
"y = sin 360t"
... by:
"y = sin (360 * 10t)"
These two waves look like this:

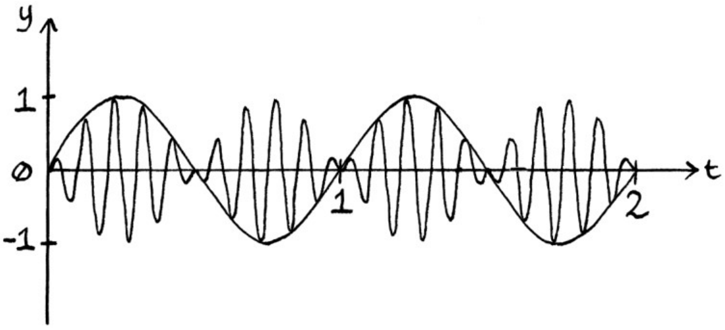


The resulting signal looks like this:



The resulting signal has a frequency of 1 cycle per second. [At first glance, it looks as if it repeats once every 0.5 seconds, but it just has a similar shape for the second half second]. The waves that if added together would reproduce this result are:
"y = 0.5 sin ((360 * 9t) + 90)" and "y = 0.5 sin ((360 * 11t) + 270)".

What is interesting about this result is that the shape of the first Multiplication Wave is apparent in the shape of the resulting signal. It is as if the fast moving curve is filling in the first wave's curve. This is more obvious if we overlay the first Multiplication Wave on to the result:



This effect will still be apparent if we increase the frequency of the second wave more. Similarly, if we increase the amplitude of the first wave, then the resulting signal will still look as if it is mimicking the shape of the first wave. A variation of this idea that incorporates mean levels is used in amplitude modulation, which I will look at later in this chapter.

Modulation

The word "modulation" is generally used in the academic subject of waves to mean the altering of a wave to convey a message. I explain amplitude modulation later in this chapter, and I explain modulation in more detail in a later chapter. Modulation is a way in which a message can be encoded into a wave to produce a signal that contains that message. If the wave is a radio wave, then the resulting signal will travel over a distance in the way that radio waves do, and carry the message with it. Someone receiving the radio signal can then decode the message from the signal by examining the received signal.

As a very simple and quick example of the basis of modulation, we will imagine that someone is either going to send a wave of 2 cycles per second to indicate the word "yes", or they are going to send a wave of 4 cycles per second to indicate the word "no". They want to send the message via radio to someone a long way away, and they only have equipment that can transmit between 6.9 MHz and 7.1 MHz

[which is another way of saying between 6,900,000 and 7,100,000 cycles per second.]

Supposing they want to send the wave that represents the word “yes”, then they could multiply their wave of 2 cycles per second by a wave of 7,000,000 cycles per second. This would produce a signal consisting of two waves added together. These would have frequencies of 7,000,002 cycles per second and 6,999,998 cycles per second. If they wanted to send the wave that represents the word “no”, then they could multiply their wave of 4 cycles per second by a wave of 7,000,000 cycles per second. This would produce a signal consisting of two waves added together, which would have frequencies of 7,000,004 cycles per second and 6,999,996 cycles per second.

The signal resulting from the multiplication would be transmitted by the transmitting equipment. Someone awaiting the message would be listening for signals around 7 MHz. If they received a signal that, when analysed, turned out to be the sum of waves with the frequencies 7,000,002 cycles per second and 6,999,998 cycles per second, they would know that the message was “yes”. If they received a signal that, when analysed, turned out to be the sum of waves with the frequencies 7,000,004 cycles per second and 6,999,996 cycles per second, they would know that the message was “no”.

By multiplying the “yes” and “no” waves by a wave of 7,000,000 cycles per second, the messages can be transmitted as radio waves. [A radio wave with a frequency of just 2 or 4 cycles per second would, in practice, be extremely difficult or impossible to transmit, but waves of around 7,000,000 cycles per second would be straightforward to transmit]. The multiplication allows “the essence” of a single slow frequency wave to be transmitted around the 7,000,000 cycles per second frequency, but with the unavoidable consequence that it becomes split into two separate waves either side of 7,000,000 cycles per second.

This example is very contrived, and the person could have saved time by just sending waves around 7 MHz without bothering with multiplication. However, the underlying concept is important in the everyday world of radio transmissions. It is also the reason why amplitude modulated radio broadcasts are generally mirrored either side of a central frequency.

One thing to realise here is that the same result could have been achieved with an addition and a subtraction, instead of with a multiplication.

Different frequencies and phases

Now, we will look at multiplying waves with different frequencies and non-zero phases.

When we were multiplying two waves with the same frequency, the rule for the result was that the first wave (the mean level wave) had a phase equal to the difference in phases added to 90 degrees, and the second wave had a phase equal to the two phases added together with 90 degrees subtracted (which is the same as 270 degrees added). The formula for multiplying two waves of the same frequency was this (if both halves are phrased as time waves):
$0.5 * (a_1 * a_2) \sin ((360 * 0t) + (\phi_1 - \phi_2 + 90))$
+
$0.5 * (a_1 * a_2) \sin ((360 * 2f * t) + (\phi_1 + \phi_2 - 90))$

This formula always produces a single pure wave with a mean level because the first part has zero frequency.

When it comes to multiplying waves with different frequencies and different phases, the rule is pretty much the same. To find the rule, we will see what happens when we multiply various waves with phases.

Example 1

We will multiply:
“y = sin (360t)”
... by:
“y = sin ((360 * 2t) + 35)”

Using the method explained in Chapter 18, we would find out that the resulting Addition Waves are:
“y = 0.5 sin (360t + 125)”
... added to:
“y = 0.5 sin ((360 * 3t) + 305)”

The first of these waves has a phase of 125 degrees, and the second has a phase of 305 degrees. If we remember how multiplication worked before, we might wonder if the first wave’s phase might be including a phase of +90 degrees, and the second wave’s phase might be including a phase of −90 degrees (which is also 270 degrees). This idea can be tested if we rephrase the Addition Waves as so:

"y = 0.5 sin ((360t) + (35 + 90))"
... added to:
"y = 0.5 sin ((360 * 3t) + (35 – 90))"
[... which is also "y = 0.5 sin ((360 * 3t) + (35 + 270))"]

The first wave has a phase of 35 + 90 degrees, and the second wave has a phase of 35 – 90 degrees (which is also 35 + 270 degrees). The number 35 happens to be both the difference in phase and the sum of the phases from the original multiplied waves.

Example 2

We will multiply these two waves:
"y = sin (360t + 20)"
... by:
"y = sin ((360 * 2t) + 35)"

The resulting Addition Waves are:
"y = 0.5 sin (360t + 105)"
... added to:
"y = 0.5 sin ((360 * 3t) + 325)"

We can rephrase these to show the underlying +90 degree and −90 degree phases, and we have:
"y = 0.5 sin ((360t) + (15 + 90))"
... added to:
"y = 0.5 sin ((360 * 3t) + (55 – 90))"

The original phases were 20 and 35 degrees. The number 15 is the difference between those two phases, and the number 55 is the sum of those two phases. We are getting closer to a pattern that we can turn into a rule.

Example 3

We will try the same phases as in Example 2, but give the waves different frequencies. We will multiply:
"y = sin ((360 * 3t) + 20)"
... by:
"y = sin ((360 * 7t) + 35)"

The resulting Addition Waves are:
“y = 0.5 sin ((360 *4t) + 105)”
... added to:
“y = 0.5 sin ((360 * 10t) + 325)”

We can rephrase these to be:
“y = 0.5 sin ((360 * 4t) + (15 + 90))”
... added to
“y = 0.5 sin ((360 * 10t) + (55 – 90))”

Despite giving the original waves different frequencies, the resulting phases are the same as they were in Example 2. This suggests that the value of the frequencies does not affect the resulting phases.

Example 4

Now we will multiply the same waves from Example 3, but we will swap the frequencies around to see what happens. We will multiply:
“y = sin ((360 * 7t) + 20)”
... by:
“y = sin ((360 * 3t) + 35)”

The resulting Addition Waves are:
“y = 0.5 sin ((360 * 4t) + 75)”
... added to:
“y = 0.5 sin ((360 * 10t) + 325)”

If we take into account the +90 degree and −90 degree phases that always appear in the results, the formulas can be rephrased as:
“y = 0.5 sin ((360 * 4t) + (−15 + 90))”
... added to:
“y = 0.5 sin ((360 * 10t) + (55 – 90))”

In this case, the first Addition Wave has a phase of −15 + 90 degrees, instead of the +15 + 90 degrees that was produced in Example 3. This shows that the *order* of the frequencies makes a difference to the result, even if the actual *values* of the frequencies do not.

Rule for phase

Given everything we know so far, we can make rules for multiplication that take into account phase:

"The phase of the first Addition Wave will be the phase of the faster frequency wave minus the phase of the slower frequency wave, added to 90 degrees."

"The phase of the second Addition Wave will be the sum of the two original phases with 90 degrees subtracted."

To put these slightly more mathematically:
"phase of first Addition Wave = $\phi_1 - \phi_2 + 90$"
"phase of second Addition Wave = $\phi_1 + \phi_2 - 90$"
... where ϕ_1 is the phase of the original wave with the faster frequency, and ϕ_2 is the phase of the original wave with the slower frequency.

This rule works when the frequencies of each wave being multiplied are both positive, both negative, or positive and negative.

Rules so far

A general rule for multiplying two waves that have any amplitude, any frequency, any phase and zero mean level is:

If we multiply:
"$y = a_1 \sin ((360 * f_1 * t) + \phi_1)$"
... by:
"$y = a_2 \sin ((360 * f_2 * t) + \phi_2)$"
... where f_1 is faster than, the negative of, or equal to f_2, and f_1 and f_2 can be both positive, both negative, or positive and negative, then the equivalent sum of waves will be:

"$y = 0.5 * (a_1 * a_2) * \sin ((360 * (f_1 - f_2)) + (\phi_1 - \phi_2 + 90)$"
... added to:
"$y = 0.5 * (a_1 * a_2) * \sin ((360 * (f_1 + f_2)) + (\phi_1 + \phi_2 - 90)$"

We will test this new rule with various phases.

Example 5

We will multiply:
“y = 1.5 sin ((360 * 4.7t) + 77.5)”
... by:
“y = 3.4 sin ((360 * 1.2t) + 334)”

From our knowledge of the rules so far, we know that the resulting Addition Waves will have the following properties:

- The amplitude of each wave will be 0.5 * (1.5 * 3.4) = 2.55 units.

- As 4.7 cycles per second is faster than 1.2 cycles per second, the frequency of the first Addition Wave will be 4.7 − 1.2 = 3.5 cycles per second. The frequency of the second Addition Wave will be 4.7 + 1.2 = 5.9 cycles per second.

- The phase of the *faster* original wave is 77.5 degrees. Therefore, the first Addition Wave will have a phase of 77.5 − 334 + 90 = −166.5 = 193.5 degrees. The second Addition Wave will have a phase of 77.5 + 334 − 90 = 321.5 degrees.

- The mean levels of each Addition Wave will be zero. In fact, they will always be zero as long as the two multiplied waves have zero mean levels.

The full formulas of the resulting Addition Waves are therefore:
“y = 2.55 sin ((360 * 3.5t) + 193.5)”
... and:
“y = 2.55 sin ((360 * 5.9t) + 321.5)”

Example 6

We will use the values from Example 5, but swap the frequencies around. We will multiply:
“y = 1.5 sin ((360 * 1.2t) + 77.5)”
... by:
“y = 3.4 sin ((360 * 4.7t) + 334)”

- The amplitudes and mean levels of the resulting Addition Waves will be the same as before.

- As 4.7 cycles per second is faster than 1.2 cycles per second, the frequency of the first Addition Wave will be 4.7 – 1.2 = 3.5 cycles per second. The frequency of the second Addition Wave will be 4.7 + 1.2 = 5.9 cycles per second. These are the same frequencies as in Example 5.

- In this example, the phase of the faster original wave is 334 degrees. Therefore, the first Addition Wave will have a phase of 334 – 77.5 + 90 = 346.5 degrees. The second Addition Wave will have a phase of 77.5 + 334 – 90 = 321.5 degrees, which is the same as before.

The full formulas of the resulting Addition Waves are therefore:
“y = 2.55 sin ((360 * 3.5t) + 346.5)”
... and:
“y = 2.55 sin ((360 * 5.9t) + 321.5)”

The result is the same as that in Example 5, except that the first Addition Wave now has a phase of 346.5 degrees, when in Example 5 it had a phase of 193.5 degrees.

Example 7

We will now multiply two waves with the same frequency:
“y = 3 sin ((360 * 2.5t) + 220)”
... by:
“y = 3 sin ((360 * 2.5t) + 110)”

- The amplitude of both resulting Addition Waves will be 0.5 * (3 * 3) = 4.5.

- As each wave has the same frequency, it does not matter which way around we calculate the resulting frequencies. The resulting frequencies will be:
  2.5 – 2.5 = 0 cycles per second for the first wave, and:
  2.5 + 2.5 = 5 cycles per second for the second wave.

- As the original frequencies are the same, there are two possible ways we can do the first Addition Wave. We can give it a phase of:
  220 – 110 + 90 = +200 degrees, or we can give it a phase of:
  110 – 220 + 90 = –20 = +340 degrees.

- The second Addition Wave will have a phase of:
  110 + 220 − 90 = 600 = 240 degrees.

- The mean level of each Addition Wave will be zero.

The formulas for the two Addition Waves are *either*:
"y = 4.5 sin ((360 * 0t) + 200)"
... added to:
"y = 4.5 sin ((360 * 5t) + 240)"

... *or*:

"y = 4.5 sin ((360 * 0t) + 340)"
... added to:
"y = 4.5 sin ((360 * 5t) + 240)"

The zero-frequency wave for the first result ends up as:
4.5 sin ((360 * 0t) + 200) = 4.5 sin (200) = −1.5391.

The zero-frequency wave for the second result ends up as:
4.5 sin ((360 * 0t) + 340) = 4.5 sin (340) = −1.5391.

Therefore, it does not matter which way around we do the subtraction of phases when the frequencies are the same – the result will be the same. [Note that it *does* matter if the frequencies are different, in the sense that whichever Addition Wave is given the frequency of "$f_1 - f_2$" must also be given the phase of "$\phi_1 - \phi_2 + 90$".]

As both the first waves of each possible Addition Wave pair result in the same value of −1.5391, we can give the resulting Addition Waves in the form of one wave with a mean level. It will be:
"y = −1.5391 + 4.5 sin ((360 * 5t) + 240)"

Example 8

We will now multiply two waves with the same *negative* frequency:
"y = 3 sin ((360 * −2.5t) + 27)"
... by:
"y = 3 sin ((360 * −2.5t) + 200)"

We know that the amplitude of both resulting Addition Waves will be 0.5 * (3 * 3) = 4.5. We also know that the mean level of each Addition Wave will be zero.

There are two ways we can calculate the multiplication. We can rephrase the wave formulas to have positive frequencies and adjust the phases accordingly, or we can keep the formulas as they are.

If we keep the formulas as they are, then we would proceed as follows:

As the original frequencies are the same, it does not matter which way around we calculate the resulting frequencies. The first resulting Addition Wave frequency will be −2.5 – −2.5 = 0 cycles per second. The second will be −2.5 + −2.5 = −5 cycles per second.

The first Addition Wave's phase either will be 27 – 200 + 90 = −83 = +277 degrees, or it will be 200 – 27 + 90 = 263 degrees. You might notice that these are both equidistant from 90 degrees [and, more obviously from 270 degrees], and on the circle chart, these angles indicate the same y-axis value on a circle's edge.

The first Addition Wave's formula will be: "y = 4.5 sin ((360 * 0t) + 277)", which is "y = 4.5 sin (277)", which is −4.4665 units. Or, if we use the other phase calculation, it will be "y = 4.5 sin ((360 * 0t) + 263)", which is "y = 4.5 sin (263), which is also −4.4665 units. As a rule, if the original frequencies are the same, the first wave will have a frequency of zero, and result in a "wave" that is just the amplitude multiplied by the Sine of the phase. In that case, it does not matter which way around the phase of the first resulting wave is calculated because both ways will result in the same number.

The second Addition Wave's phase will be 27 + 200 − 90 = 137 degrees.

The full result will be:
"y = −4.4665 + 4.5 sin ((360 * −5t) + 137)"

If we rephrased the original formulas to have positive frequencies, then we would need to change the phases accordingly. We can do this by seeing how many degrees above or below 90 degrees each phase is, and then making them that same number of degrees below or above 90 degrees. Or, we can imagine the circles from which the waves are derived and mirror them left-to-right or right-to-left to see where the phase point will end up. The phase of 27 degrees is 63 degrees below 90 degrees, so the positive frequency equivalent will be 63 degrees *above* 90 degrees – it will be 153 degrees. The phase of 200 degrees is 110 degrees above 90 degrees, so the positive frequency equivalent will be 110 degrees below 90 degrees, which is −20, which is 340 degrees. [In this case, it would have been easier to see how many degrees below 270 degrees it is – it is 70 degrees below 270 degrees, and therefore, the equivalent positive frequency phase would be 70 degrees *above* 270 degrees, which is 340 degrees.]

The original multiplication, when changed to have positive frequencies, would look like this:
"y = 3 sin ((360 * 2.5t) + 153)"
... multiplied by:
"y = 3 sin ((360 * 2.5t) + 340)"
Calculating this uses the same method. The resulting Addition Waves will each have amplitudes of 4.5 units and mean levels of zero units.
The first Addition Wave will have a frequency of 0 cycles per second; the second will have a frequency of 5 cycles per second.

The first Addition Wave will have a phase of either 153 − 340 + 90 = −97 = 263 degrees, or 340 – 153 + 90 = 277 degrees. Therefore, either its formula will be:
"y = 4.5 sin ((360 * 0t) + 263)", which is "y = 4.5 sin (263)" = −4.4665
... or it will be:
"y = 4.5 sin ((360 * 0t) + 277)", which is "y = 4.5 sin (277)" = −4.4665 as well.

The second Addition Wave will have a phase of 153 + 340 − 90 = 403 = 43 degrees.

The resulting formula will be:
"y = −4.4665 + 4.5 sin ((360 * 5t) + 43)"
... which refers to exactly the same curve as the previous negative frequency result, which was:
"y = −4.4665 + 4.5 sin ((360 * −5t) + 137)"

The circles, from which these two waves are derived, are mirror images of each other. The phases 43 degrees and 137 degrees are either side of 90 degrees.

Example 9: Part A

We will now multiply waves where one has a frequency that is the negative of the other. We can do this in three different ways. We can convert both frequencies to positive frequencies (and adjust the phases accordingly), we can convert both frequencies to negative frequencies (and adjust the phases accordingly), or we can just perform the calculations as they are. We will do all three to see how the calculations differ.

We will multiply these two waves:
“y = 3 sin ((360 * +7.5t) + 111)”
... by:
“y = 3 sin ((360 * −7.5t) + 352)”

First, we will find the result by making the second wave formula have a positive frequency. In this way, both frequencies will be positive. To convert a Sine wave formula with a negative frequency into one with a positive frequency, we imagine the circle from which the wave derived, and mirror it left or right across the y-axis, and then observe the new position of the phase point. To put it another way, we see how many degrees *above* or *below* 90 degrees the phase is, and then make it that number of degrees *below* or *above* 90 degrees. [We could also use 270 degrees instead of 90 degrees.] Therefore, to turn the −7.5 cycles per second frequency wave formula into a positive-frequency wave formula, we turn the phase of +352 degrees into a phase of +188 degrees [because 352 is 262 degrees above 90 degrees, so we want the number 262 degrees below 90 degrees, which is −172, which is +188]. In this way, our negative-frequency wave formula, “y = 3 sin ((360 * −7.5t) + 352)”, becomes “y = 3 sin ((360 * +7.5t) + 188)”.

Therefore, we will be multiplying:
“y = 3 sin ((360 * 7.5t) + 111)”
... by:
“y = 3 sin ((360 * 7.5t) + 188)”

As the frequencies are the same, we will arbitrarily choose the first Multiplication Wave to act as the one with the faster frequency. We could just as easily have picked the other wave. Whichever wave we pick, we have to be consistent for the whole calculation.

- The amplitudes of the resulting Addition Waves will be: 0.5 * (3 * 3) = 4.5 units.

- The mean levels of both waves will be zero.

- The frequency of the first Addition Wave will be 7.5 – 7.5 = 0 cycles per second. The frequency of the second Addition Wave will be 7.5 + 7.5 = 15 cycles per second.

- The phase of the first Addition Wave will be 111 – 188 + 90 = 13 degrees.

- The phase of the second Addition Wave will be 111 + 188 – 90 = 209 degrees.

The formula of the first Addition Wave will be "y = 4.5 sin ((360 * 0t) + 13)". This ends up as "y = 1.01228". [If we had picked the second Multiplication Wave as the "faster frequency wave", we would have ended up with "y = 4.5 sin ((360 * 0t) + 167), which also ends up as "y = 1.01228".]

No matter in which order we put the multiplication waves, the formula of the second Addition Wave will be "y = 4.5 sin ((360 * 15t) + 209)".
The complete result of the multiplication will be:
"y = 1.01228 + 4.5 sin ((360 * 15t) + 209)"

Example 9: Part B

Now, we will give each Multiplication Wave formula negative frequencies. We will convert the first wave to a negative frequency, so "y = 3 sin ((360 * 7.5t) + 111)" becomes "y = 3 sin ((360 * −7.5t) + 69)". We will then multiply:
"y = 3 sin ((360 * −7.5t) + 69)"
... by:
"y = 3 sin ((360 * −7.5t) + 352)"

- The amplitudes of both Addition Waves will be 0.5 * (3 * 3) = 4.5 units.

- The first Addition Wave's frequency is: −7.5 – −7.5 = 0 cycles per second.

- The second Addition Wave's frequency is: −7.5 + −7.5 = −15 cycles per second.

- The first Addition Wave's phase is: 69 – 352 + 90 = −193 = 167 degrees. [If we had put the Multiplication Waves the other way around, it would be 352 – 69 + 90 = 373 degrees].

- The second Addition Wave's phase is: 69 + 352 – 90 = 331 degrees.

The first Addition Wave's formula is "y = 4.5 sin ((360 * 0t) + 167)", which ends up as "y = 1.01228". [If we had put the Multiplication Waves the other way around, we would have had "y = 4.5 sin ((360 * 0t) + 373)", which ends up as "y = 1.01288" as well.

The second Addition Wave's formula is "y = 4.5 sin (360 * −15t) + 331)".

The full result of the multiplication is:
"y = 1.01228 + 4.5 sin ((360 * −15t) + 331)"

This result is a negative frequency result, but it refers to exactly the same curve as the result of Part A, which was:
"y = 1.01228 + 4.5 sin ((360 * 15t) + 209)"
[We can tell this is true because 331 and 209 are equidistant from 90 degrees, which is the same thing as being equidistant from 270 degrees. The circles for each wave would be mirror images of each other.]

Example 9: Part C

Now we will multiply the two waves while leaving their frequencies as one positive and one negative. We will multiply:
"y = 3 sin ((360 * +7.5t) + 111)"
... by:
"y = 3 sin ((360 * −7.5t) + 352)"

To start with, we will arbitrarily choose the first Multiplication Wave as the "faster frequency".

- The amplitudes of the two Addition Waves will be: 0.5 * (3 * 3) = 4.5 units.

- The frequency of the first Addition Wave will be: +7.5 − −7.5 = +15 cycles per second.

- The frequency of the second Addition Wave will be: +7.5 + −7.5 = 0 cycles per second.

- The phase of the first Addition Wave will be: 111 – 352 + 90 = −151 = 209 degrees.

- The phase of the second Addition Wave will be: 111 + 352 – 90 = 373 degrees.

The resulting Addition Waves will be:
“y = 4.5 sin ((360 * 15t) + 209)”
... and:
“y = 4.5 sin ((360 * 0t) + 373)”

The second of these waves ends up as “y = 1.01228”, so the two waves end up as:
“y = 1.01228 + 4.5 sin ((360 * 15t) + 209)”
... which is the same as the result we calculated when we converted the original waves to have positive frequencies.

Now we will choose the second Multiplication Wave as the “faster frequency”, and do the calculation again:

- The amplitudes will be: 0.5 * (3 * 3) = 4.5 units.

- The frequency of the first Addition Wave will be: −7.5 − 7.5 = −15 cycles per second.

- The frequency of the second Addition Wave will be: −7.5 + 7.5 = 0 cycles per second.

- The phase of the first Addition Wave will be: 352 – 111 + 90 = 331 degrees.

- The phase of the second Addition Wave will be: 352 + 111 – 90 = 373 degrees.

The resulting Addition Waves will be:
“y = 4.5 sin ((360 * −15t) + 331)”
... and:
“y = 4.5 sin ((360 * 0t) + 373)”

The second wave ends up as “y = 1.01228”, so the pair of waves ends up as:
“y = 1.01228 + 4.5 sin ((360 * −15t) + 331)”
... which is the same result as when we converted both original waves to have negative frequencies.

The rule works no matter whether we change the frequencies to be either both positive or both negative, or leave them as they are, and no matter which wave we consider having the "faster frequency".

Cosine waves

We can make a rule for Cosine waves based on the rule for Sine waves. The rule for Sine waves was:

If we multiply:
"$y = a_1 \sin((360 * f_1 * t) + \phi_1)$"
... by:
"$y = a_2 \sin((360 * f_2 * t) + \phi_2)$"
... where f_1 is faster than, the negative of, or equal to f_2, then the equivalent sum of waves will be:
"$y = 0.5 * (a_1 * a_2) * \sin((360 * (f_1 - f_2)) + (\phi_1 - \phi_2 + 90)$"
... added to:
"$y = 0.5 * (a_1 * a_2) * \sin((360 * (f_1 + f_2)) + (\phi_1 + \phi_2 - 90)$"

If we converted the two multiplied Cosine waves to Sine waves, then the rule would work as normal – we would first add 90 degrees to each phase before we used the rule to turn them into Sine waves. Then, when the calculations were complete, we would subtract 90 degrees from the phase of the resulting Addition Waves to turn them back into Cosine waves.

Therefore, we would start with phases of:
$\phi_1 + 90$
... and:
$\phi_2 + 90$

... and we would end up with:
$((\phi_1 + 90) - (\phi_2 + 90) + 90) - 90$ as the phase for the first Addition Wave,
... and:
$((\phi_1 + 90) + (\phi_2 + 90) - 90) - 90$ as the phase for the second Addition Wave.

The phase of the first Addition Wave can be made more concise:
$((\phi_1 + 90) - (\phi_2 + 90) + 90) - 90$
... becomes:
$\phi_1 + 90 - \phi_2 - 90 + 90 - 90$
... which is:

$\phi_1 - \phi_2$.

The phase of the second Addition Wave can also be made more concise:
$((\phi_1 + 90) + (\phi_2 + 90) - 90) - 90$
... becomes:
$\phi_1 + 90 + \phi_2 + 90 - 90 - 90$
... which is:
$\phi_1 + \phi_2$.

The first resulting phase is 90 degrees less than if we had been working with Sine waves, and the second resulting phase is 90 degrees more than if we had been working with Sine waves.

Given all that, we can make a rule for the multiplication of Cosine waves, which is:
If we multiply:
"$y = a_1 \cos((360 * f_1 * t) + \phi_1)$"
... by:
"$y = a_2 \cos((360 * f_2 * t) + \phi_2)$"
... where f_1 is faster than, the negative of, or equal to f_2, then the equivalent sum of waves will be:
"$y = 0.5 * (a_1 * a_2) * \cos((360 * (f_1 - f_2)) + (\phi_1 - \phi_2)$"
... added to:
"$y = 0.5 * (a_1 * a_2) * \cos((360 * (f_1 + f_2)) + (\phi_1 + \phi_2)$"

As an example of this rule in use, we will multiply:
"$y = 9 \cos((360 * 2.5t) + 18)$"
... by:
"$y = 1 \cos((360 * 3t) + 126)$"

- The amplitudes of both the resulting Addition Waves will be: 0.5 * (9 * 1) = 4.5 units.

- The mean levels will both be zero.

- The faster frequency is 3 cycles per second. Therefore, the first Addition Wave will have a frequency of 3 – 2.5 = 0.5 cycles per second. The second wave will have a frequency of 3 + 2.5 = 5.5 cycles per second.

- The phase of the first Addition Wave will be 126 – 18 = 108 degrees. The phase of the second Addition Wave will be 18 + 126 = 144 degrees.

The resulting Addition Waves in full will be:
“y = 4.5 cos ((360 * 0.5t) + 108)”
... and:
“y = 4.5 cos ((360 * 5.5t) + 144)”

As a second example, we will multiply:
“y = 2.2 cos ((360 * −1.25t) + 349)”
... by:
“y = 1.7 cos ((360 * −6.45t) + 17.8)”

- The amplitudes of both the resulting Addition Waves will be: 0.5 * (2.2 * 1.7) = 1.87 units.

- The mean levels will both be zero.

- The faster frequency is −6.45 cycles per second. Therefore, the first Addition Wave will have a frequency of −6.45 − −1.25 = −5.2 cycles per second. The second wave will have a frequency of −6.45 + −1.25 = −7.7 cycles per second.

- The phase of the first Addition Wave will be 17.8 – 349 = −331.2 = 28.8 degrees. The phase of the second Addition Wave will be 17.8 + 349 = 366.8 = 6.8 degrees.

The resulting Addition Waves in full will be:
“y = 1.87 cos ((360 * −5.2t) + 28.8)”
... and:
“y = 1.87 cos ((360 * −7.7t) + 6.8)”

Although it is irrelevant to this chapter, for the sake of refreshing what was explained in Chapter 11 (on the more complicated characteristics of frequency), we will convert these negative-frequency Addition Wave formulas into positive-frequency formulas. Remember that it is only Cosine waves *with no phase* for which a positive-frequency and negative-frequency formula are equivalent. To convert these waves, we imagine where the phase points appear on the circle, and then flip the circle upwards or downwards. Alternatively, we can see how far these phase angles are from 0 degrees and find the corresponding angle on the other side of 0 degrees. The first phase is 28.8 degrees above 0 degrees, so the equivalent positive frequency phase will be 28.8 degrees below 0 degrees. This is −28.8 degrees, which is +331.2 degrees. The second phase is 6.8 degrees above 0

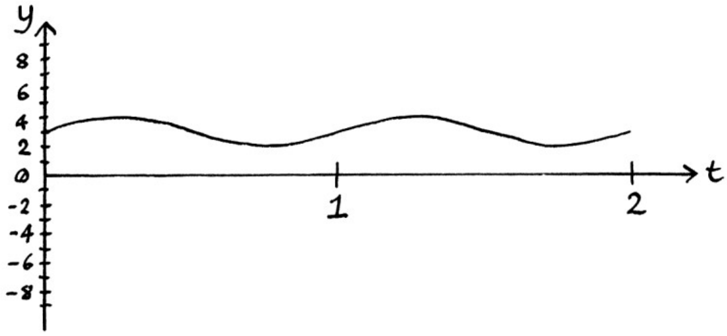
degrees, so the equivalent positive frequency phase will be 6.8 degrees below 0 degrees. This is −6.8, which is +353.2 degrees. Therefore, those two negative-frequency Addition Waves rephrased to be positive-frequency waves are:
“y = 1.87 cos ((360 * 5.2t) + 331.2)”
... and:
“y = 1.87 cos ((360 * 7.7t) + 353.2)”

Waves with non-zero mean levels

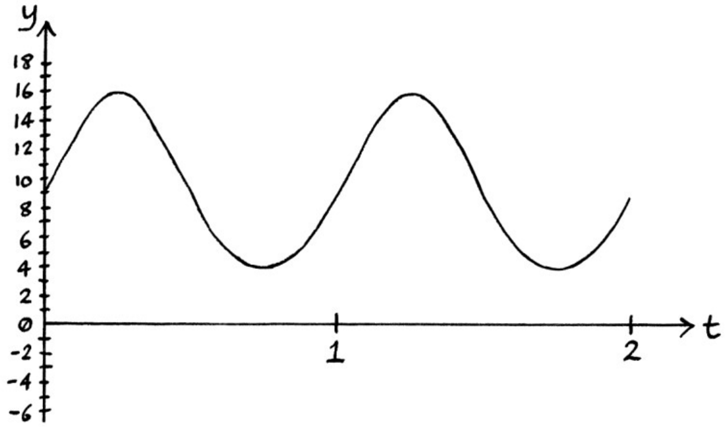
With the *addition* of two waves, a non-zero mean level is easy to deal with. With the *multiplication* of two waves, a non-zero mean level complicates things more than any other attribute of a wave. However, we can still find a rule for this.

Example 1

We will multiply “y = 3 + sin 360t” by itself. The wave looks like this:



The wave multiplied by itself looks like this:



Significant observations about this result are:

- The resulting signal's mean level is much higher than the mean level of the original wave. This is because the original wave is centred around 3 on the y-axis, and therefore, its y-axis values will go from 4 down to 2. The squares of these numbers are from 16 down to 4. The mean level is halfway between 16 and 4 at y = 10.

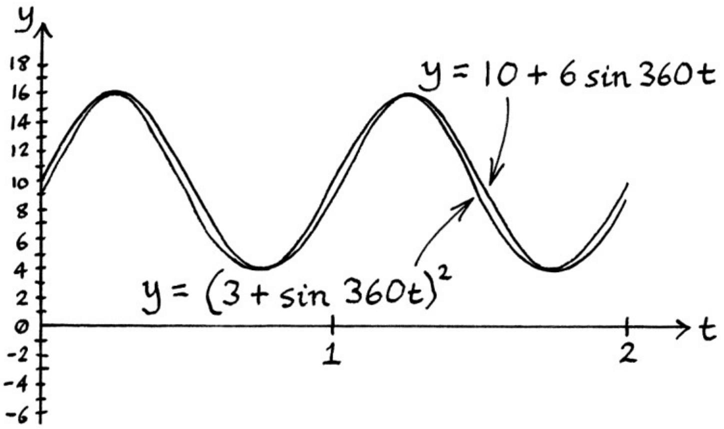
- The frequency of this resulting signal is the same as the frequency of the original wave. This is because all the original y-axis values were above zero. This means that there were no negative dips to end up as positive peaks, which is what would have happened if the original waves had had zero mean levels.

- The maximum y-axis value is 16; the minimum y-axis value is 4.

- The most important thing to notice about the resulting signal is that it is *not* a pure wave, even though it resembles one.

The resulting signal very closely *resembles* the wave "y = 10 + 6 sin 360t", but it is slightly off. At first glance, there is no obvious difference between the result and that formula – they both have the same frequency, and they *appear* to have the same amplitude, the same mean level and the same phase. However, the curve of the resulting signal does not exactly match the curve of "y = 10 + 6 sin 360t". Apart from at the peaks and dips the resulting signal is always slightly lower. The dips of the signal are shallower than the wave it resembles, and the peaks are pointier.

The square of "y = 3 + sin 360t" and the wave "y = 10 + 6 sin 360t" drawn on the same graph appear as in the following picture. The curve of "y = 10 + 6 sin 360t" is the higher of the two curves at all times except at the peaks and dips.



This difference is most obvious in the middle of the curves between the peaks and dips. In these places, the resulting signal's values are clearly lower than those of "y = 10 + 6 sin 360t".

The difference is best seen by using a calculator to find particular y-axis values. On the squared wave, at 0.5 seconds, the y-axis value is 9. We can calculate this on a calculator, or do the maths mentally, as we can know from thinking about a circle that the Sine of 180 is 0:
(3 + sin (360 * 0.5)) * (3 + sin (360 * 0.5))
... which is:
(3 + sin (180)) * (3 + sin (180))
... which is:
(3 + 0) * (3 + 0)
... which is:
9.

If we use a calculator, or our heads, to find the result of the formula, "y = 10 + 6 sin 360t", at t = 0.5 seconds, we would have:
10 + 6 sin (360 * 0.5)
... which is:
10 + 6 sin (180)
... which is:
10 + (6 * 0)
... which is:
10 + 0
which is:
10.

Therefore, at 0.5 seconds, there is a y-axis difference of 1 unit on the two curves. However, at the peaks and dips there is no difference at all. For example at t = 0.25, both the resulting signal and the "y = 10 + 6 sin 360t" wave have a y-axis value of 16.

The presence of a non-zero mean level causes the resulting signal to be ever so slightly distorted in comparison to a pure wave.

The fact that a non-zero mean level distorts the resulting signal means that a wave with a *zero* mean level squared and then squared again will not be a pure wave. Multiplying a wave with a zero mean level by itself once produces a resulting pure wave with a non-zero mean level. Multiplying that result by itself will produce an impure signal.

Any periodic signal that is not a pure wave is the sum, or approximately the sum, of pure waves. We will use the method that will be explained in Chapter 18 to find out which waves make up this sum. The method only finds the waves in a sum after any of the same frequency have been added together. Given that, we can calculate that the sum of the following waves would be equal to the signal:
“y = 6 sin 360t”
... added to:
“y = 0.5 sin (360 * 2t) + 270)”
... added to a mean level of:
9.5 units.

Note how the sum of these waves is similar to “y = 10 + 6 sin 360t”. It is really “y = 6 sin 360t” added to a wave with a small amplitude and faster frequency, and then added to a mean level of 9.5 units.

If there had been no mean level in the original squared wave, we would have ended up with:
“y = 0.5 + 0.5 sin ((360 * 2t) + 270)”
... but as we have learnt in this chapter, the 0.5 unit mean level in that case is really a wave with zero frequency. There are two Addition Waves, and they are:
“y = 0.5 sin ((360 * 0t) + 90)”
... and:
“y = 0.5 sin ((360 * 2t) + 270)”

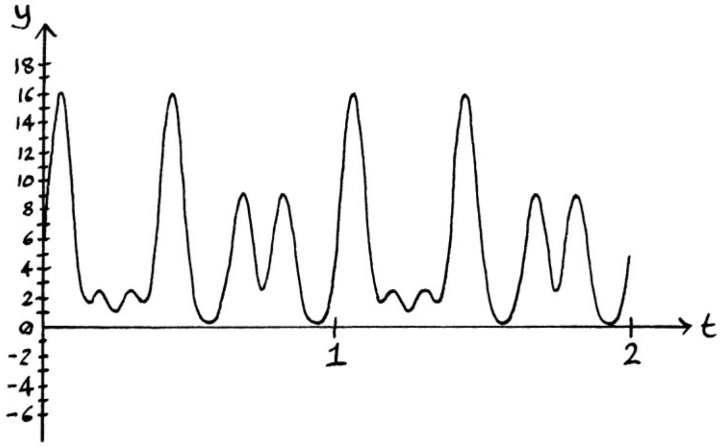
Therefore, in this example with an original wave with a mean level, we might guess that the sum consists of more than two Addition Waves, which are possibly these ones:
“y = 9 + 6 sin (360 * 1t)”
“y = 0.5 sin (360 * 0t) + 90)”
“y = 0.5 sin ((360 * 2t) + 270)”
The second of these waves acts as a mean level for the first and third.

We will discover what is happening here as we look at more examples.

Example 2

We will multiply:
“y = 3.1 + 2.3 sin (360 * 5t)”
... by:
“y = 1.7 + 1.5 sin (360 * 3t)”

Even without mean levels to complicate things, as the frequencies are different, the result of this will not be a pure wave. The resulting signal looks like this:



Using the methods that will be explained in Chapter 18, we can discover that the resulting Addition Waves are:
“y = 1.725 sin ((360 * 2t) + 90)”
“y = 1.725 sin ((360 * 8t) + 270)”
“y = 3.91 sin (360 * 5t)”
“y = 4.65 sin (360 * 3t)”
... and a mean level of 5.27 units.

If we look at each of these in turn, we can see patterns that relate to the original multiplied waves.

The first two waves are what we would have expected to see if there had been no mean level. We could have calculated these using the standard rule for multiplying two waves with zero mean levels. They follow this rule:
“y = 0.5 * (a_1 * a_2) * sin ((360 * (f_1 – f_2)) + (ϕ_1 – ϕ_2 + 90)”
... added to:
“y = 0.5 * (a_1 * a_2) * sin ((360 * (f_1 + f_2)) + (ϕ_1 + ϕ_2 – 90)”

- Their amplitudes are both: 0.5 * (1.5 * 2.3) = 1.725 units.
- The first wave's frequency is: 5 – 3 = 2 cycles per second.
- The second wave's frequency is: 5 + 3 = 8 cycles per second.
- The phase of the first wave is: 0 – 0 + 90 = 90 degrees.
- The phase of the second wave is: 0 + 0 – 90 = −90 = +270 degrees.

The third wave has a frequency equal to that of the Multiplication Wave with the faster frequency – it is 5 cycles per second.

The fourth wave has a frequency equal to that of the Multiplication Wave with the slower frequency – it is 3 cycles per second.

The third wave has an amplitude equal to the amplitude of the faster Multiplication Wave multiplied by the mean level of the slower Multiplication Wave. It is: 2.3 * 1.7 = 3.91 units.

The fourth wave has an amplitude equal to the amplitude of the slower Multiplication wave multiplied by the mean level of the faster Multiplication Wave. It is: 1.5 * 3.1 = 4.65 units.

The mean level that is added on to the four waves is the mean level of the faster Multiplication Wave multiplied by the mean level of the slower Multiplication Wave. It is: 3.1 * 1.7 = 5.27 units.

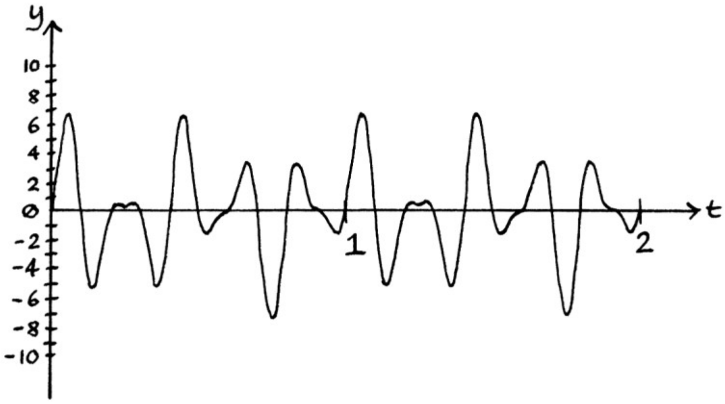
We will keep looking for patterns with more examples.

Example 3

We will multiply:
"y = 2.3 sin (360 * 5t)"
... by:
"y = 1.7 + 1.5 sin (360 * 3t)"

These are the same waves as in Example 2, yet the first wave has zero mean level.

The result looks like this:



Using the methods that will be explained in Chapter 18, we can discover that the resulting Addition Waves are:
“y = 1.725 sin ((360 * 2t) + 90)”
“y = 1.725 sin ((360 * 8t) + 270)”
“y = 3.91 sin (360 * 5t)”

Rules so far

Given what we have just seen, we can create a rule for multiplying two waves with non-zero mean levels, any amplitudes, any frequencies, *and zero phases*.

If we multiply:
“y = h_1 + a_1 sin (360 * f_1 * t)”
... by:
“y = h_2 + a_2 sin (360 * f_2 * t)”
... where f_1 is faster than, the negative of, or equal to f_2, then the equivalent sum of waves and a mean level will be:
“y = 0.5 * (a_1 * a_2) * sin ((360 * (f_1 – f_2)) + 90)”
“y = 0.5 * (a_1 * a_2) * sin ((360 * (f_1 + f_2)) – 90)”
“y = (h_2 * a_1) sin (360 * f_1 * t)”
“y = (h_1 * a_2) sin (360 * f_2 * t)”
“h_1 * h_2”

In this way, there are four added waves and a mean level.

It would be possible to phrase the mean level in the result as a wave with no frequency, and in that way, there would be five waves, and a more consistent sum. This would be:
"$y = 0.5 * (a_1 * a_2) * \sin ((360 * (f_1 - f_2)) + (\phi_1 - \phi_2 + 90)$"
"$y = 0.5 * (a_1 * a_2) * \sin ((360 * (f_1 + f_2)) + (\phi_1 + \phi_2 - 90)$"
"$y = (h_2 * a_1) \sin (360 * f_1 * t)$"
"$y = (h_1 * a_2) \sin (360 * f_2 * t)$"
"$y = (h_1 * h_2) \sin ((360 * 0 * t) + 90)$"

It would also be possible to split the mean level into two pieces and add it on to each of the second pair of waves, as so:
"$y = 0.5 * (a_1 * a_2) * \sin ((360 * (f_1 - f_2)) + (\phi_1 - \phi_2 + 90)$"
"$y = 0.5 * (a_1 * a_2) * \sin ((360 * (f_1 + f_2)) + (\phi_1 + \phi_2 - 90)$"
"$y = (0.5 * h_1 * h_2) + (h_2 * a_1) \sin (360 * f_1 * t)$"
"$y = (0.5 * h_1 * h_2) + (h_1 * a_2) \sin (360 * f_2 * t)$"

Although turning the mean level into a zero-frequency wave, or splitting the mean level and applying it to the third and fourth waves, make the list of Addition Waves more succinct, they also reduce the simplicity of the list. It is a matter of choice, but personally, I think it is easiest to remember how the list of Addition Waves is calculated if the mean level is kept separate. We can say that there are 4 Addition Waves and a mean level.

If both the original waves have zero mean levels, then the third and fourth waves will end up as "y = 0", and the resulting mean level will be zero.

If just one of the original waves has a zero mean level, then either the third or fourth wave will end up as "y = 0", and the resulting mean level will be zero. This raises an interesting point about multiplying waves with mean levels – if only one of the original waves has a non-zero mean level, then it means that the resulting signal will have a zero mean level (presuming the frequencies are different).

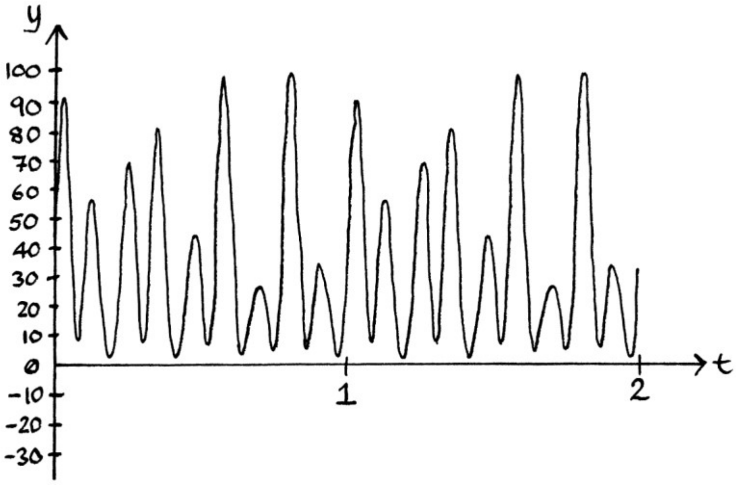
If the frequencies of the original waves are the same, then the first wave will end up acting as a mean level, and that mean level will end up being added to the resulting mean level in the final signal.

The rule works for all situations where the phase of each multiplied wave is zero.

Example 4

As an example of the rule working, we will multiply:
“y = 7 + 6 sin (360 * 9t)”
... by:
“y = 5 + 3 sin (360 * 4t)”

The resulting signal looks like this:



We will use the rules to calculate the four Addition Waves and the mean level.

For the first pair of Addition Waves:

- They will both have an amplitude of 0.5 * (6 * 3) = 9 units.
- The first wave will have a frequency of 9 – 4 = 5 cycles per second.
- The second wave will have a frequency of 9 + 4 = 13 cycles per second.
- The first wave will have a phase of +90 degrees; the second will have a phase of −90 degrees, which is +270 degrees.
- Both waves will have zero mean levels.

For the second pair of Addition Waves:

- The first wave will have an amplitude of 5 * 6 = 30; the second wave will have an amplitude of 7 * 3 = 21.
- The first wave of this pair will have a frequency of 9 cycles per second; the second will have a frequency of 4 cycles per second.
- Both waves will have phases of zero degrees.
- Both waves will have zero mean levels.

For the mean level:

- It will be 5 * 7 = 35 units.

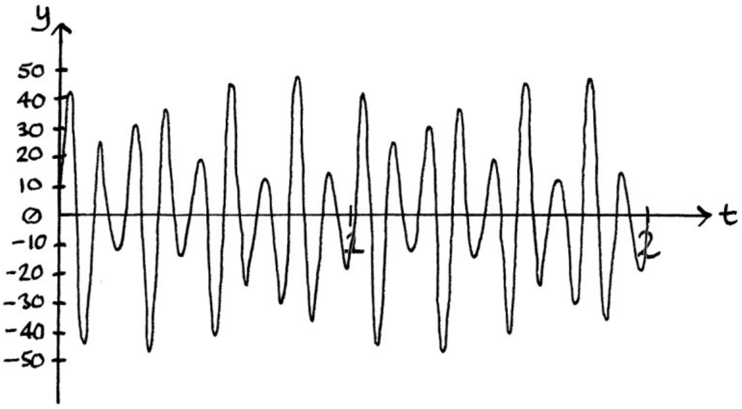
All four Addition Waves given in full, and the mean level are:
“y = 9 sin ((360 * 5t) + 90)”
“y = 9 sin ((360 * 13t) + 270)”
“y = 30 sin (360 * 9t)”
“y = 21 sin (360 * 4t)”
The mean level is 35 units.

[Note that when calculating the resulting formulas, it is very easy to mix up the mean levels and the amplitudes by mistake, especially since we have not used mean levels in formulas that much until now].

Example 5

In this example, we will use the waves from Example 4, but give one of the waves a zero mean level. We will multiply:
“y = 6 sin (360 * 9t)”
... and:
“y = 5 + 3 sin (360 * 4t)”

The resulting signal looks like this:



Notice how it is centred around y = 0. If only one wave has a mean level, and the frequencies are different, then this will always be the case.

The first pair of Addition Waves will be the same as in Example 4.

For the second pair of Addition Waves:

- The first wave will have an amplitude of 5 * 6 = 30 units.
- The second wave will have an amplitude of 0 * 3 = 0 units. This is because we are multiplying by the mean level of the first wave, which is zero.
- The first wave of this pair will have a frequency of 9 cycles per second; the second will have a frequency of 4 cycles per second.
- Both waves will have phases of zero degrees.
- Both waves will have zero mean levels.

For the mean level, it will be 0 * 5 = 0 units.

The Addition Waves in full will be:
“y = 9 sin ((360 * 5t) + 90)”
“y = 9 sin ((360 * 13t) + 270)”
“y = 30 sin (360 * 9t)”
“y = 0 sin (360 * 4t)”
The mean level is 0 units.

As the amplitude of the fourth wave is zero, it ends up as “y = 0” for all time, so can be ignored. As the mean level is zero, that can be ignored too. We are left with the following three Addition Waves:
“y = 9 sin ((360 * 5t) + 90)”
“y = 9 sin ((360 * 13t) + 270)”
“y = 30 sin (360 * 9t)”

Example 6

As another example, we will multiply:
“y = −5.6 + 11 sin (360 * 5t)”
... by:
“y = −3.2 + 13 sin (360 * 3t)”

We will calculate the four resulting Addition Waves and the mean level using the rule from before.

For the first pair of waves:

- They will both have an amplitude of 0.5 * (11 * 13) = 71.5 units.
- The first wave will have a frequency of 5 – 3 = 2 cycles per second.
- The second wave will have a frequency of 5 + 3 = 8 cycles per second.
- The first wave will have a phase of +90 degrees; the second will have a phase of −90 degrees, which is +270 degrees.
- Both waves will have zero mean levels.

For the second pair of waves:

- The first wave will have an amplitude of −3.2 * 11 = −35.2 units. The second wave will have an amplitude of −5.6 * 13 = −72.8 units.
- The first wave of this pair will have a frequency of 5 cycles per second; the second will have a frequency of 3 cycles per second.
- Both waves will have phases of zero degrees (although, if we give both waves positive amplitudes, we might change the phases to be 180 degrees).
- Both waves will have zero mean levels.

For the mean level:

- It will be −5.6 * −3.2 = 17.92 units.

The resulting Addition Waves in full are:
“y = 71.5 sin ((360 * 2t) + 90)”
“y = 71.5 sin ((360 * 8t) + 270)”
“y = −35.2 sin (360 * 5t)”
“y = −72.8 sin (360 * 3t)”
... and a mean level of 17.92 units.

As we have negative amplitudes, we could rephrase these waves to be:
“y = 71.5 sin ((360 * 2t) + 90)”
“y = 71.5 sin ((360 * 8t) + 270)”
“y = 35.2 sin ((360 * 5t) + 180)”
“y = 72.8 sin ((360 * 3t) + 180)”
... and a mean level of 17.92 units.

Example 1 revisited

In Example 1 of this section, we multiplied “y = 3 + sin 360t” by itself. We ended up with the result being the sum of two waves and a mean level:
“y = 6 sin 360t”
“y = 0.5 sin (360 * 2t) + 270)”
... and a mean level of 9.5 units.

Now we know the rules for calculating the result, we can know that the result is really the sum of four waves and a mean level. These waves are:
“y = 0.5 sin (360 * 0t) + 90)”
“y = 0.5 sin ((360 * 2t) + 270)”
“y = 3 sin 360t”
“y = 3 sin 360t”
... and a mean level of 9 units.

The first wave has zero frequency and a phase of 90 degrees. Therefore, it becomes a mean level of 0.5 units. In the result, this becomes added to the mean level of 9 units to create an overall mean level of 9.5 units.

The third and fourth waves have the same frequency, so become added together to be “y = 6 sin 360t”.

This means that the previous result was correct, but it did not show the result in as much detail as that gained from working out the result using maths. It is a good example of how adding waves is similar to pouring different liquids into a bucket: adding waves of the same frequency is similar to pouring different containers of water into a bucket – afterwards, we can see how much water there is in the bucket, but we cannot tell how many containers were used. In this case, each of the “y = 3 sin 360t” waves in the sum are similar to a container of water. After the sum, we only know that there are six lots of “sin 360t”, and not two lots of three.

Example 7: phases

We will try a multiplication where one wave has a non-zero phase. We will multiply:
"y = 2 + 3 sin ((360 * 4t) + 50)"
... by:
"y = 4 + 5 sin (360 * 2t)"

We will use the method that will be explained in Chapter 18 to analyse the signal. [The method only finds the waves that were added after waves of the same frequency have been added together]. The sum of waves turns out to be:
"y = 6.4299 sin ((360 * 2t) + 48.5702)"
"y = 12 sin ((360 * 4t) + 50)"
"y = 7.5 sin ((360 * 6t) + 320)"
... and a mean level of 8 units.

These are only three waves, instead of four.

If we had calculated the results for the first pair of Addition Waves, we would have found that:

- They should both have an amplitude of 0.5 * (3 * 5) = 7.5 units.
- The first wave should have a frequency of 4 – 2 = 2 cycles per second.
- The second wave should have a frequency of 4 + 2 = 6 cycles per second.
- The first wave should have a phase of: 50 − 0 + 90 = 140 degrees.
- The second wave should have a phase of: 50 + 0 − 90 = −40 = 320 degrees.
- Both waves will have zero mean levels.

If the rule still works, then the first pair of waves *should* be:
"y = 7.5 sin ((360 * 2t) + 140)"
"y = 7.5 sin ((360 * 6t) + 320)"

In the analysis of the signal, the first of these is missing, however, there is "y = 6.4299 sin ((360 * 2t) + 48.5702)", which has a matching frequency, and we could guess that it is a sum of the missing first wave and another wave with the same frequency. [If we thought about circles and phase points, we could calculate what the other wave was in this case].

Using the rules for calculating the second pair of Addition Waves, we would calculate that:

- The first wave should have an amplitude of: 4 * 3 = 12 units.
- The second wave should have an amplitude of: 2 * 5 = 10 units.
- The first wave of this pair will have a frequency of 4 cycles per second; the second will have a frequency of 2 cycles per second.

If the rule works for non-zero phases, then this pair of waves *should* be:
"y = 12 sin (360 * 4t)"
"y = 10 sin (360 * 2t)"

The analysis of the signal found "y = 12 sin ((360 * 4t) + 50)", which is the first of the pair, but with a phase of 50 degrees. Therefore, the rule must be incomplete.

The second of the pair must be part of "y = 6.4299 sin ((360 * 2t) + 48.5702)". In fact, if we add:
"y = 7.5 sin ((360 * 2t) + 140)"
... and:
"y = 10 sin (360 * 2t)"
... we end up with:
"y = 6.4299 sin ((360 * 2t) + 48.5702)". Therefore, the second wave of this pair is correct, and the first wave of the first pair of waves is also correct. [As always, to add two waves of the same frequency, imagine the waves as circles arranged for adding. Then find the coordinates of the outer phase point using Sine and Cosine, then use Pythagoras's theorem to find the distance of that phase point from the centre of the inner circle, and use arctan to find its angle].

For the mean level, we multiply 2 by 4 to produce 8 units.

This all means that the rule works, except for the third wave, which should have a phase of 50 degrees. This is the phase of the first multiplied wave. Although the phase of the fourth wave in this case is zero, one might guess that the fourth wave actually has the phase of the second wave being multiplied.

The four Addition Waves and the mean level are as so:
"y = 7.5 sin ((360 * 2t) + 140)"
"y = 7.5 sin ((360 * 6t) + 320)"
"y = 12 sin ((360 * 4t) + 50)"
"y = 10 sin (360 * 2t)"
... and a mean level of 8.

An all-encompassing rule

A new rule for the multiplication of two waves, where each wave has any amplitude, any frequency, any phase, and any mean level is:

If we multiply:
“$y = h_1 + a_1 \sin((360 * f_1 * t) + \phi_1)$”
... by:
“$y = h_2 + a_2 \sin((360 * f_2 * t) + \phi_2)$”
... where f_1 is faster than, the negative of, or equal to f_2, then the equivalent sum of waves and a mean level will be:
“$y = 0.5 * (a_1 * a_2) * \sin((360 * (f_1 - f_2)) + (\phi_1 - \phi_2 + 90)$”
“$y = 0.5 * (a_1 * a_2) * \sin((360 * (f_1 + f_2)) + (\phi_1 + \phi_2 - 90)$”
“$y = (h_2 * a_1) \sin((360 * f_1 * t) + \phi_1)$”
“$y = (h_1 * a_2) \sin((360 * f_2 * t) + \phi_2)$”
“$h_1 * h_2$”

Example 8

To test the new rule, we will multiply:
“y = 1 + 3.5 sin ((360 * 3.7t) + 176)”
... by:
“y = −5 + 8.3 sin ((360 * 3t) + 225)”

For the first pair of Addition Waves:

- They will both have an amplitude of 0.5 * (3.5 * 8.3) = 14.525 units.
- The first wave will have a frequency of 3.7 − 3 = 0.7 cycles per second.
- The second wave will have a frequency of 3.7 + 3 = 6.7 cycles per second.
- The first wave will have a phase of: 176 − 225 + 90 = 41 degrees.
- The second wave will have a phase of: 176 + 225 − 90 = 311 degrees.
- Both waves will have zero mean levels (as they always will).

For the second pair of Addition Waves:

- The first wave will have an amplitude of −5 * 3.5 = −17.5 units.
- The second wave will have an amplitude of 1 * 8.3 = 8.3 units.
- The first wave of this pair will have a frequency of 3.7 cycles per second; the second will have a frequency of 3 cycles per second.
- The first wave will have a phase of 176 degrees.
- The second wave will have a phase of 225 degrees.
- Both waves will have zero mean levels (as they always will).

For the mean level:

- It will be 1 * −5 = −5 units.

The full list of Addition Waves and the mean level will be as so:
“y = 14.525 sin ((360 * 0.7t) + 41)”
“y = 14.525 sin ((360 * 6.7t) + 311)”
“y = −17.5 sin ((360 * 3.7t) + 176)”
“y = 8.3 sin ((360 * 3t) + 225)”
... and a mean level of −5 units.

If we do not like the third wave having a negative amplitude, we could rephrase it to have a positive amplitude and add 180 degrees to its phase. The phase would become 176 + 180 = 356 degrees. The list of waves and the mean level, in that case, would become:
“y = 14.525 sin ((360 * 0.7t) + 41)”
“y = 14.525 sin ((360 * 6.7t) + 311)”
“y = 17.5 sin ((360 * 3.7t) + 356)”
“y = 8.3 sin ((360 * 3t) + 225)”
... and a mean level of −5 units.

The result of the sum would be identical.

For interest’s sake, we will calculate the *overall* frequency of the resulting signal by using the methods in Chapter 13 on the addition of waves. We calculate what the frequency would be from adding the first two of the Addition Waves. Then, we find out what the frequency would be if we added the third wave to that, then we find out what the frequency would be if we added the fourth wave to that. For each addition, we check if one frequency is an integer multiple of the other, or if they are both integers per second. If they are not, then we use the ratio method. That method involves finding the lowest integer multiple of the ratio between the frequencies, then dividing either of the original frequencies by the corresponding ratio value.

To start with, we will look at the first two Addition Waves. These are:
“y = 14.525 sin ((360 * 0.7t) + 41)”
... and:
“y = 14.525 sin ((360 * 6.7t) + 311)”

The frequencies are 0.7 cycles per second and 6.7 cycles per second.

The ratio is 6.7 : 0.7. This is not an integer to one, and neither of these are integers. Therefore, we scale up the ratio until we have an integer to an integer:
Multiplying both sides by 1 gives us: 6.7 : 0.7
Multiplying by 2 produces: 13.4 : 1.4
By 3 produces: 20.1 : 2.1
By 4 produces: 26.8 : 2.8
By 5 produces: 33.5 : 3.5
By 6 produces: 40.2 : 4.2
By 7 produces: 46.9 : 4.9
By 8 produces: 53.6 : 5.6
By 9 produces: 60.3 : 6.3
By 10 produces: 67 : 7. These are both integers.

Therefore, the frequency of the signal created by adding the first two waves will be: 6.7 ÷ 67 = 0.1 cycles per second. We could also have calculated it as 0.7 ÷ 7 = 0.1 cycles per second.

Next, we find the frequency of the signal resulting from adding a 0.1-cycle-per-second wave to the frequency of the third wave (3.7 cycles per second). We have the ratio 3.7 : 0.1. It might be obvious that to turn this into a ratio with the two lowest possible integers in it, it will need to be multiplied by 10, thus ending up with: 37 : 1.

Therefore, the frequency of the result of adding the first three waves will be: 3.7 ÷ 37 = 0.1. We could also have calculated this as 0.1 ÷ 1 = 0.1.

We then find the resulting frequency of a signal created by adding a wave with a frequency of 0.1 to a wave with the frequency of the fourth wave (3 cycles per second). The ratio is 3 : 0.1. The lowest multiple of this ratio that has integers on both sides will be 30 : 1. Therefore, the frequency of adding the four Addition Waves will be: 3 ÷ 30 = 0.1 cycles per second. We could also have calculated it as 0.1 ÷ 1 = 0.1.

From this, we now know that the frequency of the resulting signal of adding those four Addition Waves is 0.1 cycles per second. As the addition of the four Addition Waves produces an identical signal to the multiplication of the two original waves, it means the result of the multiplication also has a frequency of 0.1 cycles per second. The period of the signal is 1 ÷ 0.1 = 10 seconds. If we look at the graph on a graphing calculator, we can confirm this is correct. (For this particular wave, there are moments before ten seconds, where the wave's shape very nearly repeats sooner. At first glance, we might mistakenly think it repeats more quickly.)

Rule for Cosine waves

We can turn the rule for Sine waves into a rule for Cosine waves by adjusting the phases in the first pair of waves (in the same was as when we were looking at waves with non-zero phases and zero mean levels). The rest of the rule is the same as for Sine waves. The rule becomes:

If we multiply:
"$y = h_1 + a_1 \cos ((360 * f_1 * t) + \phi_1)$"
... by:
"$y = h_2 + a_2 \cos ((360 * f_2 * t) + \phi_2)$"
... where f_1 is faster than, the negative of, or equal to f_2, then the equivalent sum of waves and a mean level will be:
"$y = 0.5 * (a_1 * a_2) * \cos ((360 * (f_1 - f_2)) + (\phi_1 - \phi_2)$"
"$y = 0.5 * (a_1 * a_2) * \cos ((360 * (f_1 + f_2)) + (\phi_1 + \phi_2)$"
"$y = (h_2 * a_1) \cos ((360 * f_1 * t) + \phi_1)$"
"$y = (h_1 * a_2) \cos ((360 * f_2 * t) + \phi_2)$"
"$h_1 * h_2$"

Final formulas for multiplication

For the multiplication of two waves, no matter what their amplitude, frequency, phase or mean level, it will be the case that the resulting signal will be equal to the sum of 4 waves with zero mean levels, added to a separate mean level.

If the two multiplied waves have zero mean levels, then two of the waves in the sum will end up as “y = 0”, and the mean level will also be zero, in which case they can be ignored.

If the two multiplied waves have the same frequency and zero mean level, then one of the two non-zero waves in the sum will have zero frequency, and end up acting as a mean level to the other non-zero wave. This will give the appearance of the resulting sum only having one wave. It will also mean that the resulting signal will be a pure wave.

If the two multiplied waves have the same frequency and zero mean level, then the resulting signal will be a pure wave. If they have different frequencies and any mean level, then the resulting signal will not be a pure wave. If they have the same frequency and non-zero mean levels, then the resulting signal will not be a pure wave.

The rule for finding the sum of four Sine waves and the mean level that are equivalent to the multiplication of two Sine waves is as follows:

If we multiply:
“$y = h_1 + a_1 \sin((360 * f_1 * t) + \phi_1)$”
... by:
“$y = h_2 + a_2 \sin((360 * f_2 * t) + \phi_2)$”
... where f_1 is faster than, the negative of, or equal to f_2, then the equivalent sum of waves and a mean level will be:
“$y = 0.5 * (a_1 * a_2) * \sin((360 * (f_1 - f_2)) + (\phi_1 - \phi_2 + 90)$”
“$y = 0.5 * (a_1 * a_2) * \sin((360 * (f_1 + f_2)) + (\phi_1 + \phi_2 - 90)$”
“$y = (h_2 * a_1) \sin((360 * f_1 * t) + \phi_1)$”
“$y = (h_1 * a_2) \sin((360 * f_2 * t) + \phi_2)$”
“$h_1 * h_2$”

If f_1 is the slower frequency, the rule will still work but, if the frequencies are different, one of the first pair of resulting Addition Waves will have a negative frequency.

If the first Addition Wave is calculated as "$f_1 - f_2$", but its phase is calculated as "$\phi_2 - \phi_1 + 90$" [with the phases the wrong way around], then the calculation will still work, but *only* if the original waves both have positive frequencies, or if they both have negative frequencies. If one has a positive frequency and one has a negative frequency, then it will not work.

The rule for finding the sum of four Cosine waves and the mean level that are equivalent to the multiplication of two Cosine waves is as follows:

If we multiply:
"$y = h_1 + a_1 \cos((360 * f_1 * t) + \phi_1)$"
... by:
"$y = h_2 + a_2 \cos((360 * f_2 * t) + \phi_2)$"
... where f_1 is faster than, the negative of, or equal to f_2, then the equivalent sum of waves and a mean level will be:
"$y = 0.5 * (a_1 * a_2) * \cos((360 * (f_1 - f_2)) + (\phi_1 - \phi_2)$"
"$y = 0.5 * (a_1 * a_2) * \cos((360 * (f_1 + f_2)) + (\phi_1 + \phi_2)$"
"$y = (h_2 * a_1) \cos((360 * f_1 * t) + \phi_1)$"
"$y = (h_1 * a_2) \cos((360 * f_2 * t) + \phi_2)$"
"$h_1 * h_2$"

Multiplying signals by waves

It is important to have a basic understanding of what happens when we multiply a pure wave by a signal that is not a pure wave.

We can do this in two ways. The first way is to multiply every y-axis value from the pure wave with every y-axis value from the signal at corresponding moments in time. This method is simple and does not require us to know anything about the signal. In practice, this method could only be done by multiplying a series of evenly spaced y-axis values, and joining up the results. It would be best done by a computer than by hand. We could still achieve a good approximation doing it by hand, but it would generally take a long time.

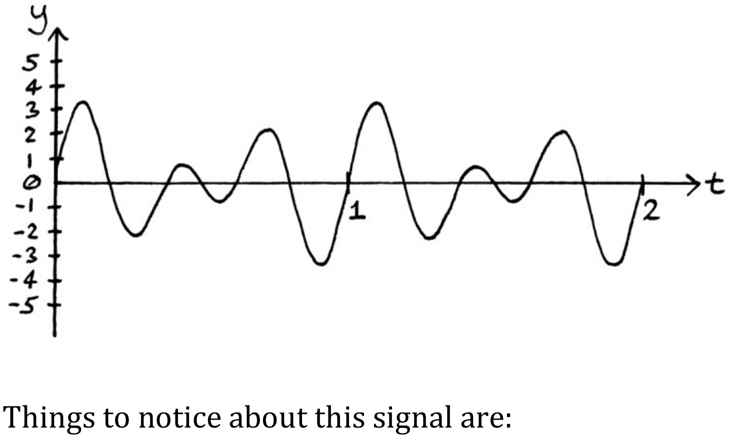
The second way requires that we know the signal is made up of a sum of pure waves, and what those waves are. It relies on the fact that different frequencies do not mix together in a signal. A signal made up of adding waves of different frequencies can still be treated as being the sum of those individual waves (as opposed to being treated as "just a signal"). To calculate the result of multiplying a wave by a signal, we multiply the wave by each of the waves that make up that signal in turn, and then add up the results. To put this another way, we break the signal up into the constituent waves that were added to create it. Then we perform the multiplication of our wave against each of the constituent waves in turn. Then we add up the results of those multiplications. Although this way requires more work, the resulting signal will be *identical* to if we had multiplied every y-axis value, but with the advantage that we will have the exact formula of the result.

[Previously in this book, I have given the rule: "most periodic signals that are not pure waves are equal to, or approximately equal to, the sum of two or more pure waves". When it comes to the second method, we can pretend that this rule is: "*all* periodic signals that are not pure waves *are equal* to the sum of two or more pure waves", and it will make things easier for us, but at the expense of sometimes having inaccurate results, depending on the nature of the signal. In the examples here, I will only be dealing with signals that are definitely the sum of pure waves.]

Example 1

As an example, we will use a signal that is based on the sum of these two waves:
"y = 1.5 sin (360 * 2t)"
... added to:
"y = 2 sin (360 * 3t)"

This signal looks like this:

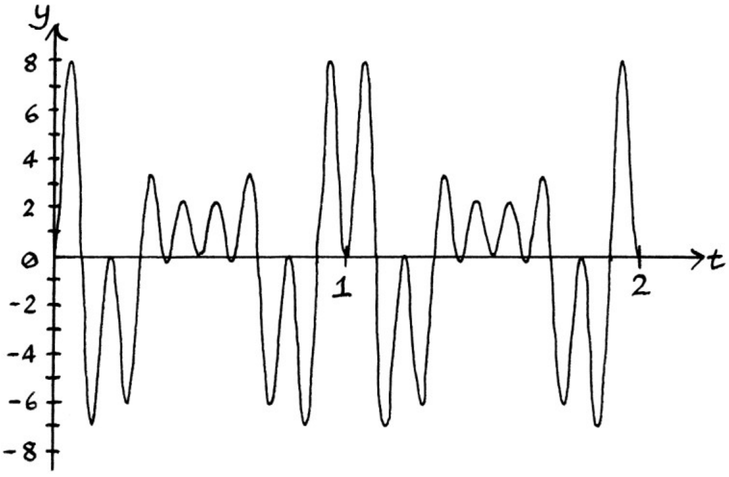


Things to notice about this signal are:

- It has a mean level of zero.
- It has a frequency of 1 cycle per second.

We will multiply the signal by this wave:
“y = 3 sin (360 * 5t)”

If we used the method of multiplying every y-axis value of the signal by every y-axis value of the wave for corresponding moments of time, we would end up with the following signal, but we would not really know much about the signal except for its shape:



Things to notice about this signal are:

- It has a mean level of zero.
- It has a frequency of 1 cycle per second.

Now we will find out the same resulting signal using maths. The original signal before the multiplication was made up of “y = 1.5 sin (360 * 2t)” added to “y = 2 sin (360 * 3t)”. Therefore, we multiply our wave by each of these, and add up the results. We need to calculate:
“y = 1.5 sin (360 * 2t)” multiplied by “y = 3 sin (360 * 5t)”
... added to:
“y = 2 sin (360 * 3t)” multiplied by “y = 3 sin (360 * 5t)”

First, we will look at the multiplication of:
“y = 1.5 sin (360 * 2t)”
... and:
“y = 3 sin (360 * 5t)”

We will use our multiplication formula. The faster frequency is 5 cycles per second, so we will arrange the waves this way round to make it clearer:
"y = 3 sin (360 * 5t)"
... and:
"y = 1.5 sin (360 * 2t)"

For the first pair of Addition Waves for this multiplication:

- The amplitude of each wave will be 0.5 * (3 * 1.5) = 2.25 units.
- The frequency of the first wave will be: 5 – 2 = 3 cycles per second.
- The frequency of the second wave will be: 5 + 2 = 7 cycles per second.
- The phase of the first wave will be: 0 – 0 + 90 = 90 degrees.
- The phase of the second wave will be: 0 + 0 – 90 = –90 = +270 degrees.
- The mean levels of each wave will be zero, as always.

For the second pair of Addition Waves for this multiplication:

- The amplitude of the first wave will be 0 * 3 = 0 units.
- The amplitude of the second wave will be 0 * 1.5 = 0 units.
- Given that the amplitudes are both zero, both these waves will end up as "y = 0" for all time, so there is no point in calculating the other characteristics.

The mean level will be 0 * 0 = 0 units.

Therefore, the result of the first multiplication will be the sum of:
"y = 2.25 sin ((360 * 3t) + 90)"
"y = 2.25 sin ((360 * 7t) + 270)"
"y = 0"
"y = 0"
... and the mean level of 0 units.

[As these waves have zero mean levels, there was no point in trying to calculate the second pair of Addition Waves or the mean level.]

For the second multiplication, we are multiplying:
"y = 2 sin (360 * 3t)"
... by:
"y = 3 sin (360 * 5t)"

As 5 cycles per second is the faster wave, we will arrange them as so to make things simpler:
“y = 3 sin (360 * 5t)”
... by:
“y = 2 sin (360 * 3t)”

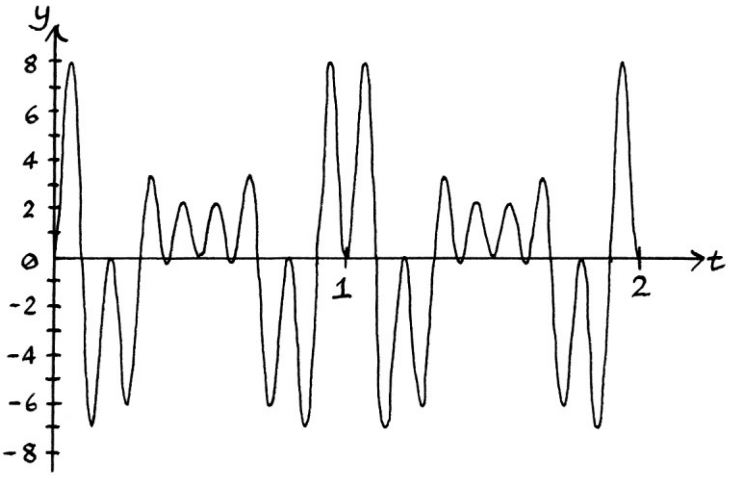
As the mean levels are zero, there will only be the first pair of waves in the resulting sum of Addition Waves. For that pair of waves, we know that:

- The amplitudes will both be: 0.5 * (3 * 2) = 3 units.
- The frequency of the first wave will be: 5 – 3 = 2 cycles per second.
- The frequency of the second wave will be: 5 + 3 = 8 cycles per second.
- The phase of the first wave will be: 0 – 0 + 90 = 90 degrees.
- The phase of the second wave will be: 0 + 0 – 90 = −90 = 270 degrees.

Therefore, the result of the second multiplication will be equal to the sum of:
“y = 3 sin ((360 * 2t) + 90)”
... and:
“y = 3 sin ((360 * 8t) + 270)”

We have multiplied our wave by each of the waves that make up the signal, and ended up with four waves. We add them all up to find the full resulting signal. The sum of waves will be:
“y = 2.25 sin ((360 * 3t) + 90)”
“y = 2.25 sin ((360 * 7t) + 270)”
“y = 3 sin ((360 * 2t) + 90)”
“y = 3 sin ((360 * 8t) + 270)”

The sum of these waves looks like this:



The signal from the sum is identical to the signal created by multiplying our original signal's y-axis values by the corresponding (by time) y-axis values of our original wave.

From all this, we can say that:
"y = 1.5 sin (360 * 2t) + 2 sin (360 * 3t)" multiplied by "y = 3 sin (360 * 5t)"
... is equal to the sum of:
"y = 2.25 sin ((360 * 3t) + 90)"
"y = 2.25 sin ((360 * 7t) + 270)"
"y = 3 sin ((360 * 2t) + 90)"
"y = 3 sin ((360 * 8t) + 270)"

It makes no difference which approach we use to find the result, as the result will always be the same. This is a good example of how different frequencies do not mix – our original signal, although it appears as just a signal, is always inherently a sum of waves of different frequencies.

When we break a signal up into the waves that were added together to create it, then multiply each constituent wave by the original wave, and then add up the results, we end up with a formula for the result. When we just multiply the corresponding y-axis values of a signal by those of a wave, we end up with the exact same result, but we do not end up with a formula to describe the result. Therefore, the "constituent" method can be more useful, even if it takes longer to do.

Example 2

In this example, we will multiply a wave by a signal again, but that signal will contain a wave of the same frequency as the wave.

We will use the signal created by adding:
"y = 2.2 sin (360 * 5t)"
... and:
"y = 1.1 sin (360 * 3t)"

We will multiply that signal by the wave:
"y = 3 sin (360 * 3t)"

To do the calculation, we will multiply each of the signal's constituent waves by our original wave, and then add up the results. Therefore, we will calculate:
"y = 2.2 sin (360 * 5t)" multiplied by "y = 3 sin (360 * 3t)".
... and add it to:
"y = 1.1 sin (360 * 3t)" multiplied by "y = 3 sin (360 * 3t)"

The first multiplication is:
"y = 2.2 sin (360 * 5t)"
... multiplied by:
"y = 3 sin (360 * 3t)".

As the mean levels are zero, there will only be the first pair of waves in the result, and no mean level. For this first and only pair:

- Both waves will have amplitudes of: 0.5 (2.2 * 3) = 3.3 units.
- The frequency of the first wave will be: 5 – 3 = 2 cycles per second.
- The frequency of the second wave will be: 5 + 3 = 8 cycles per second.
- The phase of the first wave will be: 0 – 0 + 90 = 90 degrees.
- The phase of the second wave will be: 0 + 0 – 90 = −90 = 270 degrees.

The waves will be:
"y = 3.3 sin ((360 * 2t) + 90)"
... added to:
"y = 3.3 sin ((360 * 8t) + 270)".

The second multiplication will be:
"y = 1.1 sin (360 * 3t)"
... multiplied by:
"y = 3 sin (360 * 3t)"

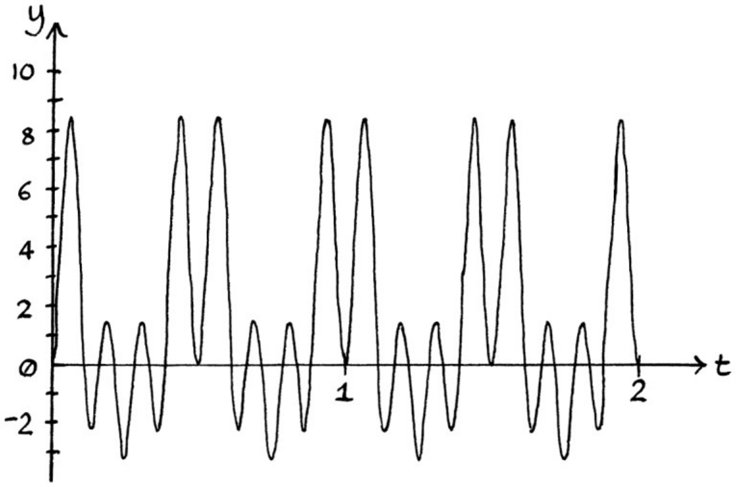
Again, as the mean levels are zero, there will only be the first pair of waves in the result, and no mean level. For this first and only pair of waves:

- The amplitude of both waves will be: 0.5 * (1.1 * 3) = 1.65 units.
- The frequency of the first wave will be: 3 – 3 = 0 cycles per second.
- The frequency of the second wave will be: 3 + 3 = 6 cycles per second.
- The phase of the first wave will be: 0 – 0 + 90 = 90 degrees.
- The phase of the second wave will be: 0 + 0 – 90 = −90 = 270 degrees.

The two waves will be:
"y = 1.65 sin ((360 * 0t) + 90)"
... added to:
"y = 1.65 sin ((360 * 6t) + 270)"
As the first wave has a zero frequency, it can be tidied up:
"y = 1.65 sin ((360 * 0t) + 90)"
... becomes:
"y = 1.65 sin (90)"
... which is:
"y = 1.65 * 1"
"y = 1.65"

This wave ends up as a mean level of 1.65 units.
The four waves, or actually the three waves and a mean level, that when added together are equal to the original multiplication of our signal and wave are:
"y = 3.3 sin ((360 * 2t) + 90)"
"y = 3.3 sin ((360 * 8t) + 270)".
"y = 1.65"
"y = 1.65 sin ((360 * 6t) + 270)"

The graph of this signal looks like this:



We could also have done the calculation by multiplying every y-axis value (or, more realistically, a series of evenly spaced y-axis values) from the signal by every y-axis value in the wave for corresponding moments in time, and we would have achieved *exactly* the same result. The difference being that we would not have known the waves that make up that resulting signal.

From the graph, we can tell that the overall signal has a non-zero mean level. From the waves that are added to make up this signal, we know that this mean level is 1.65 units.

Example 2: significance

What has happened in Example 2 is very significant, and it is one of the most important aspects of maths with waves.

In this chapter, we have seen that if we multiply two waves of the same frequency with zero mean level and zero phase, the result will be one wave with zero frequency added to one wave with double the original frequency. The wave with zero frequency ends up as a mean level to the other wave in the sum. This means that whenever two waves of the same frequency with zero mean level and zero phase are multiplied together, the result will always be a single pure wave and a *non-zero* mean level.

When we multiply a wave against a signal that is made up of added waves, we perform the multiplication against each of the waves that were added to make up that signal. Therefore, if one of the waves that was added to make up that signal has the same frequency as the wave against which it is being multiplied, and both waves have zero mean levels and zero phases, the result of that particular multiplication will be a pure wave with a non-zero mean level. That, in turn, means that if we multiply one signal by one wave, and that wave and the waves that make up that signal have zero mean levels and zero phases, and that wave has a frequency that matches one of the waves that was in the set of waves that made up that signal, then the resulting signal will *always* have a non-zero mean level.

As an example, supposing a signal is made up of the sum of:
“y = sin (360 * 2t)”
“y = sin (360 * 3t)”
“y = sin (360 * 4t)”
... and we multiply that signal by:
“y = sin (360 * 3t)

... then the result will really be the sum of these multiplications:
“y = sin (360 * 2t)” multiplied by “y = sin (360 * 3t)
“y = sin (360 * 3t)” multiplied by “y = sin (360 * 3t)
“y = sin (360 * 4t)” multiplied by “y = sin (360 * 3t)

In this case, because a wave of 3 cycles per second is one of the constituent waves of the original signal, and because we are multiplying it by a wave of 3 cycles per second, that particular multiplication will produce a pure wave and a mean level. Therefore, the resulting signal will contain a non-zero mean level.

Conversely, supposing a signal is made up of the sum of these waves:
"y = sin (360 * 2t)"
"y = sin (360 * 3t)"
"y = sin (360 * 4t)"
... and we multiply that signal by:
"y = sin (360 * 5t)

... then the result will really be the sum of these multiplications:
"y = sin (360 * 2t)" multiplied by "y = sin (360 * 5t)
"y = sin (360 * 3t)" multiplied by "y = sin (360 * 5t)
"y = sin (360 * 4t)" multiplied by "y = sin (360 * 5t)

In this case, none of the frequencies in each multiplication match, and therefore, none of the multiplications will produce a single pure wave with a mean level (they will each produce two pure waves with zero mean levels), and therefore, the resulting signal will have a zero mean level.

From all of this, we have a way of testing if a signal (that is made up of a sum of waves with zero mean levels and zero phases) contains a wave of a particular frequency. To test, we just multiply it by a test wave of a particular frequency – if the result has a *non-zero* mean level, then we know that that frequency is in the signal; if the result has a *zero* mean level then we know that the frequency is not in the signal.

The beauty of this is that we do not need to reduce a signal to its constituent summed waves to perform the test. We can just multiply every y-axis value of the signal against every y-axis value from the wave for corresponding moments in time. If the result has a non-zero mean level, then we know that that frequency is in the original signal; if the result has a zero mean level, then we know that it is not.

It is this aspect of waves that led me in previous chapters to say that frequencies do not mix. Although a signal might look complicated, it is still possible to discern the frequencies of the waves that were added together to make it. Adding waves of different frequencies together is analogous to mixing quantities of unmixable liquids together.

Given all of this, if we are presented with a periodic signal that we know is made up of waves with zero mean levels and zero phases, then we can test a range of frequencies against it to reveal the frequency of every single wave that was added to make that signal. [We know that all the waves that were added together to make up a signal must have frequencies that are integer multiples of the frequency of the signal (as explained in Chapter 13). Therefore, we only have to test frequencies that are integer multiples of the signal's frequency. If the signal has a frequency of 2 cycles per second, then we only need to test waves of 2 cycles per second, 4 cycles per second, 6 cycles per second, 8 cycles per second and so on. I explain this process in Chapter 18].

Non-zero phases

If our signal were made up of waves that did not have zero phases, this idea would be slightly more complicated.

To demonstrate the difference, we will first look at an example of multiplying two waves of the same frequency with zero mean levels and zero phases. We will multiply:
"y = 2 sin (360 * 5t)"
... by:
"y = 3 sin (360 * 5t)"

The result is:
"y = 3 sin ((360 * 0t) + 90)"
... added to:
"y = 3 sin ((360 * 10t) + 270)"

The first of these waves ends up as:
"y = 3 sin (90)", which is "y = 3 * 1", which is "y = 3", or a mean level of 3 for the second wave.

The significant thing to realise here is that the first wave ends up as the amplitude multiplied by the Sine of the phase. Because the phase is 90 degrees in this situation, the result will be just the amplitude multiplied by 1, which is just the amplitude.

If the waves being multiplied had had non-zero phases, then the resulting phase for the first Addition Wave would probably not have been 90. It could have been any angle at all. Therefore, the first Addition Wave would have ended up as the amplitude multiplied by the Sine of whatever the phase was. If the phase had been zero degrees, then the first Addition "Wave" would have been the amplitude multiplied by zero, which is zero. If the phase had been 45 degrees, then the first Addition "Wave" would have been the amplitude multiplied by the Sine of 45 degrees, which is the amplitude multiplied by 0.7071. The point I am trying to make is that when the first Addition Wave is reduced to being a mean level for the second Addition Wave, it might end up as any number from 0 up to the amplitude, or 0 down to the negative of the amplitude. The resulting mean level of the particular multiplication varies, depending on the phase in the first Addition Wave's formula. Therefore, so does the mean level of the resulting signal.

All of this means that if we multiplied a signal containing a particular frequency against a wave of that same frequency, and the phases were not both zero, then we might end up not giving the overall signal a non-zero mean level. In such a situation, we would not know if the original signal contained our frequency of interest or not.

As we know, the phases of the Addition Waves (for zero-mean-level Sine waves) are calculated as:
$(\phi_1 - \phi_2 + 90)$ for the first wave,
... and:
$(\phi_1 + \phi_2 - 90)$ for the second wave.

If the frequencies of the waves being multiplied are the same, then the first wave will have zero frequency, and it will end up being:
"$y = 0.5 * (a_1 * a_2) * \sin(\phi_1 - \phi_2 + 90)$

It will act as a mean level for the second wave, and, if the waves are part of a signal, it will act as the mean level for that signal.

If the frequencies are the same, and the phases of the two waves being multiplied are the same, then the first Addition Wave ends up as:
"$y = 0.5 * (a_1 * a_2) * \sin(90)$
... which ends up as "$y = 0.5 * (a_1 * a_2)$". This is the maximum possible mean level value.

If $\phi_1 - \phi_2$ results in 270 degrees or −90 degrees, then the first Addition Wave will end up as:
"$y = 0.5 * (a_1 * a_2) * \sin(0)$
... which is "$y = 0$". In such a situation, there is no mean level for the resulting signal.

If $\phi_1 - \phi_2$ results in 180 degrees, then the first Addition Wave will end up as:
"$y = 0.5 * (a_1 * a_2) * \sin(270)$
... which is the same as:
"$y = 0.5 * (a_1 * a_2) * \sin(-90)$
... which ends up as "$y = 0.5 * (a_1 * a_2) * -1$", which is "$y = -0.5 * (a_1 * a_2)$". This is the *minimum* possible mean level value.

From these, we can see that the resulting mean level will fluctuate from the negative of the amplitude up to the amplitude, depending on the relationship between the phases being multiplied. We saw this happening in the section "Different Phases" earlier in this chapter, although back then we did not know the rules for what was happening. In that section, we discovered that:

- When the phases of both original waves are the same, the resulting mean level is at its maximum.
- When the phases of both original waves are 90 degrees apart, the resulting mean level is zero. This also means that if the phases are 270 degrees apart, the resulting mean level is zero too, because if they are 270 degrees apart, then they are also 90 degrees apart.
- When the original phases are 180 degrees apart, the resulting mean level is at its minimum.

These observations can be rephrased with our current knowledge to be:

- When $\phi_1 = \phi_2$, the resulting mean level is at its maximum.
- When $\phi_1 - \phi_2 = 90$ or $\phi_1 - \phi_2 = 270$, the resulting mean level is zero.
- When $\phi_2 - \phi_1 = 180$, the resulting mean level is at its minimum.

If we were multiplying waves against a signal to test if that frequency was in it, we might, by bad luck, test with a wave with a phase 90 degrees away from that of the wave in the signal. If that happened, the resulting mean level would be zero, so there would be no evidence of a match. Therefore, to test for waves, it is necessary to try at least two waves and to ensure that the phase of one of them cannot be exactly 90 degrees away from that of the wave in the signal. To do this, we pick any two phases that are exactly 90 degrees apart. If one of them happens to be 90 degrees away from the phase of the wave in the signal, then the other one cannot be 90 degrees away from the phase of the wave in the signal. If we have a higher or lower mean level from multiplying one or both of the test waves, then that frequency is definitely in the signal. If the mean levels are zero for the multiplications of *both* test waves, then the frequency is definitely not in the signal.

We will look more at this idea in Chapter 18.

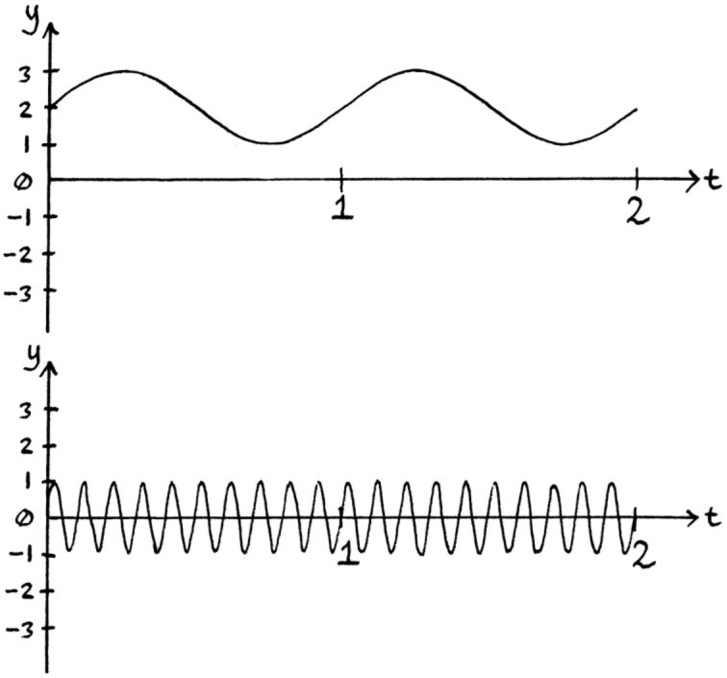
More aspects of multiplication

Amplitude modulation

If we multiply one wave with a non-zero mean level by another with a zero mean level and a much faster frequency, we can achieve interesting effects.

We will multiply:
“y = sin (360 * 10t)”
... by:
“y = 2 + sin (360 * 1t)”

These two waves look like this:



For the sake of practising multiplication, we will work out the result in terms of Addition Waves. As there is a mean level involved, there will be 4 Addition Waves and a mean level.

For the first pair of waves:

- They will both have an amplitude of 0.5 * (1 * 1) = 0.5 units.
- The first wave will have a frequency of 10 – 1 = 9 cycles per second.
- The second wave will have a frequency of 10 + 1 = 11 cycles per second.
- The first wave will have a phase of 0 – 0 + 90 = 90 degrees.
- The second wave will have a phase of 0 + 0 – 90 = 270 degrees.
- Both waves will have mean levels of zero, as always.

For the second pair of waves:

- The first wave will have an amplitude of 2 * 1 = 2 units.
- The second wave will have an amplitude of 0 * 1 = 0 units.
- The first wave will have a frequency of 10 cycles per second.
- The second wave will have a frequency of 1 cycle per second.
- Both waves will have zero phases and zero mean levels.

For the mean level:

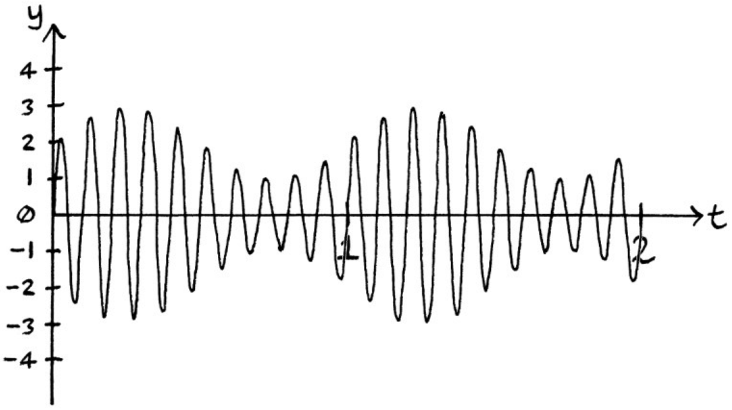
- It will be: 0 * 2 = 0 units.

The full sum of waves and mean level will be:
“y = 0.5 sin ((360 * 9t) + 90)”
“y = 0.5 sin ((360 * 11t) + 270)”
“y = 2 sin (360 * 10t)”
“y = 0 sin (360 * 1t)”
... and a mean level of 0.

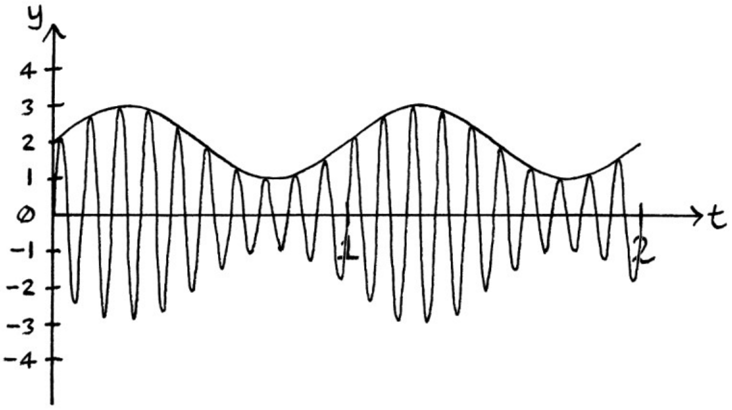
Ignoring the zero-amplitude wave and the zero mean level, and putting the waves in order of frequency, we have:
“y = 0.5 sin ((360 * 9t) + 90)”
“y = 2 sin (360 * 10t)”
“y = 0.5 sin ((360 * 11t) + 270)”

We, therefore, know that this signal will be centred around y = 0.

The signal, whether we treat it as a multiplication or as an addition, looks like this:



What is interesting about this resulting signal is that its fluctuations draw out the shape of the slower original wave:



The signal does not just draw out the wave – the wave appears at the same height as it did originally, too. The signal is carrying a version of the slower wave in its fluctuations. [Technically, it is carrying two versions – the signal is symmetrical, which means there is an upside down version of the wave below y = 0.]

In the above picture, it is as if the slower wave is riding on top of the resulting signal. One could also say that the wave's curve is surrounding the resulting signal. Another way to say this is that the curve is *enveloping* the signal– its curve is enclosing it. In the world of waves, people are likely to say that the slower wave appears in the "envelope" of the signal. In this sense, the word "envelope" refers to the line made from joining up all the peaks of a signal. Although the English noun "envelope" is more commonly used to mean an enclosure for a letter, here the word is being used in a way more in keeping with the verb "envelop" meaning to cover or wrap. It is not a perfect word for the situation, but it is the one that has become generally accepted. If someone refers to the "envelope of a signal", they are referring to the line that is drawn when we join up all the peaks of a signal.

What is happening in this example is important because it enables a signal to incorporate all the information of a slower frequency wave. The idea is used in the transmitting of radio broadcasts when it is called “amplitude modulation”. With amplitude modulation, a wave with a slower frequency is given a mean level and then multiplied by a wave with a faster frequency to produce a signal that carries the slower frequency wave. If that signal is then transmitted as a radio signal, the characteristics of the slower frequency wave become transmitted with it.

[Note that the mean level has to be added to the slower frequency wave or else it will not work.]

Supposing we had an audio tone that was a wave of 440 Hz, and we wanted to transmit it to someone else, we would not be able to convert that 440 Hz audio tone to a radio wave at that frequency. It would be very difficult to transmit such a low frequency radio wave, and if it were easy, other people would be doing the same thing, so there would be a lot of interference. It is easier and better to transmit the tone at a faster frequency by letting a radio signal of a faster frequency carry the audio tone. We will use the example of a radio signal around 252 KHz. We multiply our 440 Hz tone by a wave of 252,000 Hz, and then transmit the resulting signal as a radio signal. We have to give our tone a non-zero mean level for there to be a proper “envelope”. We will give it a mean level of 2 units. The calculation looks like this:
“y = sin (360 * 252,000t)”
... multiplied by:
“y = 2 + sin (360 * 440t)”

We will calculate the resulting Addition Waves.

For the first pair of waves:

- The amplitude of both waves will be 0.5 * (1 * 1) = 0.5 units.
- The frequency of the first wave will be 252,000 – 440 = 251,560 cycles per second.
- The frequency of the second wave will be 252,000 + 440 = 252,440 cycles per second.
- The phase of the first wave will be 0 – 0 + 90 = 90 degrees.
- The phase of the second wave will be 0 + 0 – 90 = 270 degrees.
- The mean level of the two waves will be zero, as always.

For the second pair of waves:

- The amplitude of the first wave will be 2 * 1 = 2 units.
- The amplitude of the second wave will be 0 * 2 = 0 units.
- The frequency of the first wave will be 252,000 cycles per second.
- The frequency of the second wave will be 440 cycles per second.
- The phase of both waves will be zero.
- The mean level of both waves will be zero, as always.

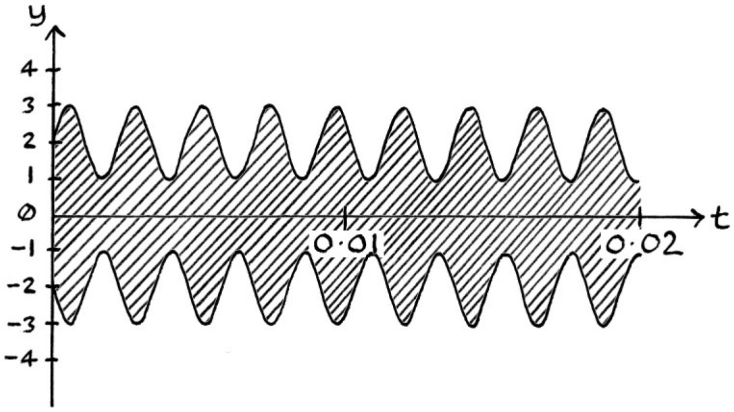
For the mean level:

- The mean level will be 0 * 2 = 0 units.

The full list of resulting Addition Waves will be:
"y = 0.5 sin ((360 * 251,560t) + 90)"
"y = 0.5 sin ((360 * 252,440t) + 270)"
"y = 2 sin (360 * 252,000t)"
"y = 0 sin (360 * 440t)"
... and a mean level of 0.

If we ignore the zero-amplitude wave and the mean level of zero, and we put the waves in order of frequency, we end up with:
"y = 0.5 sin ((360 * 251,560t) + 90)"
"y = 2 sin (360 * 252,000t)"
"y = 0.5 sin ((360 * 252,440t) + 270)"

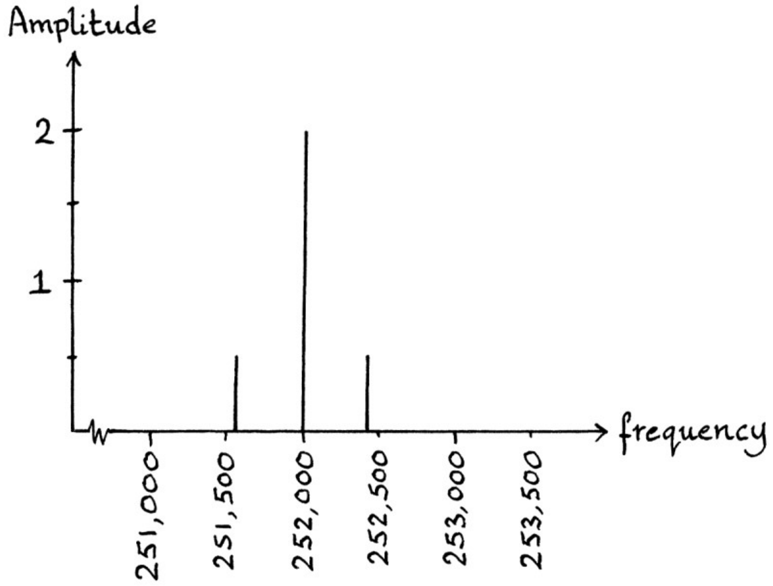
The resulting signal appears in the following picture. Note that it would be impossible to draw the curve of the signal accurately because the lines would be so close together. Therefore, I have indicated the curves of the signal with hatching.



We can see how the 440 Hz wave is carried in the envelope of the signal. In the picture, we can see 0.02 seconds' worth of the signal, which is enough time to see 8.8 cycles of the 440 Hz wave in the envelope.

Anyone receiving the signal on an AM radio would hear the tone because the radio would decode it. Someone receiving radio waves around 252,000 Hz in some other way would have to decode the signal themselves.

As we know the Addition Waves, we can see what the signal looks like in the frequency domain without much effort:



In this case, and in every case when a single wave *with* a mean level is multiplied against a faster wave with *no* mean level, there will be three resulting waves. There will be one in the centre at the frequency of the faster frequency wave, and there will be two either side. The distance between the centre wave and each of the other waves will be equal to the frequency of the original slower wave. In this example, the outer waves are both 440 Hz away from the centre wave.

The very nature of using multiplication to let a signal carry a slower wave in this way means that the result will always be symmetrical around the centre frequency.

Normally with amplitude modulation, we would not be transmitting one single continuous frequency wave, but instead a huge number of waves of ever changing frequencies, making up music, speech or other sounds. In such a case, the sum of the waves making up the sound must be multiplied by the faster frequency wave, and there will be countless waves transmitted around the central frequency. However, it will still be the case that all the frequencies of all the resulting waves will be symmetrical around the central frequency.

Although it is the resulting *signal* that carries the characteristics of the slower frequency wave, generally, the faster frequency *wave* in the multiplication is called the “carrier wave”. The carrier wave is the wave that is used to create a signal carrying the slower wave.

The slower frequency wave that is carried by the signal is generally called the “modulating wave” or “modulator”. It “modulates” the faster frequency wave to produce the resulting signal. In this sense, “to modulate” means to change or to alter. If the slower frequency wave were not there, the faster frequency wave would be left alone and appear as normal. As it is, the slower frequency wave alters or changes or “*modulates*” the faster frequency wave. The modulating wave is essentially the message. It is, or it contains, the information that we want to convey.

Putting all the jargon in a sentence, we can say, “The modulating wave modulates the carrier wave to produce a signal that carries a copy of the modulating wave in its envelope.”

In most situations, we would not be sending just one wave, but instead a signal made up of many waves. In such cases, there would not be a modulating *wave* but a modulating *signal*. There would still be a carrier *wave* though. The characteristics of the modulating signal would appear in the envelope of the resulting signal.

If we were sending a modulating *signal*, then sometimes, it might be referred to as the “baseband signal”. “Band” in this phrase refers to the bandwidth of frequencies taken up by a signal. It is a “band of frequencies”. “Base” can be thought of as referring to how these frequencies are relatively low. Any modulating signal will be made up of lower frequencies – for amplitude modulation to work, it is necessary that they are significantly lower than the frequency of the wave with which they will be multiplied. The “baseband signal” is the signal made up of the band of lower frequency waves. The “baseband signal” is the message that we wish to convey.

Note that in this example, instead of performing a multiplication against our original tone [the modulating wave], it would have been just as easy to transmit the results of an addition instead – it would have been an identical signal. However, when it comes to transmitting complicated modulating signals, doing an addition requires a lot more effort because the signal would need to be reduced to its constituent waves first.

Demodulation

For the amplitude modulation signal that we just created, there are several possible ways to recover the original 440 Hz tone. If we had a drawing of the signal, we could just join up its peaks, and we could see the encoded wave. If we were receiving the signal in real time as a radio signal, we could note the maximum points reached by the signal's fluctuations and reconstruct the wave from that. One method that is appropriate to this chapter is to multiply the received signal by the faster wave that created it in the first place, and then filter the results.
To do this, we would multiply the received signal by "y = sin (360 * 252,000t)". In practice, this would be done by multiplying every y-axis value of the signal by the corresponding (in time) y-axis value from "y = sin (360 * 252,000t)". As we actually know the waves that, when added, make up our signal, we can do the multiplication against each individual wave instead. This allows us to see exactly what is happening.

The waves that make up our signal are:
"y = 0.5 sin ((360 * 251,560t) + 90)"
"y = 0.5 sin ((360 * 252,440t) + 270)"
"y = 2 sin (360 * 252,000t)"

Therefore, we need to do three multiplications:
"y = 0.5 sin ((360 * 251,560t) + 90)" multiplied by "y = sin (360 * 252,000t)"
"y = 0.5 sin ((360 * 252,440t) + 270)" multiplied by "y = sin (360 * 252,000t)"
"y = 2 sin (360 * 252,000t)" multiplied by "y = sin (360 * 252,000t)"

For the first multiplication, we will end up with:
"y = 0.25 sin (360 * 440t)"
"y = 0.25 sin (360 * 503,560t)"

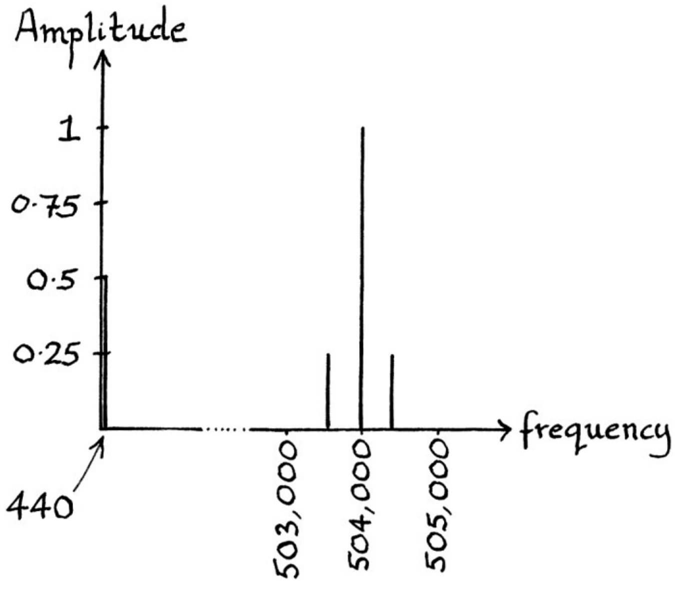
For the second multiplication, we will end up with:
"y = 0.25 sin (360 * 440t)"
"y = 0.25 sin ((360 * 504,440t) + 180)"

For the third multiplication, we will end up with:
"y = sin ((360 * 0t) + 90)"
"y = sin ((360 * 504,000t) + 270)"
... which, as the first wave has a zero frequency, are the same as a single wave with a mean level:
"y = 1 + sin ((360 * 504,000t) + 270"

This means that our final signal is made up of the sum of all these waves:
“y = 0.25 sin (360 * 440t)”
“y = 0.25 sin (360 * 503,560t)”
“y = 0.25 sin (360 * 440t)”
“y = 0.25 sin ((360 * 504,440t) + 180)”
“y = 1 + sin ((360 * 504,000t) + 270”

If we add the two 440 Hz waves, and put the waves in order of frequency, we have:
“y = 0.5 sin (360 * 440t)”
“y = 0.25 sin (360 * 503,560t)”
“y = sin ((360 * 504,000t) + 270”
“y = 0.25 sin ((360 * 504,440t) + 180)”
... and a mean level of 1.

As we can see, our original 440 Hz tone is one of these waves, except its amplitude is half what it was originally. There is a mean level of 1, which again is half the mean level that we originally gave our tone so that it could be amplitude modulated in the first place. The other three waves have frequencies of 503,560 Hz, 504,000 Hz and 504,440 Hz. Interestingly, 503,560 and 504,440 are both 440 Hz away from 504,000. By multiplying the signal by a wave with a 252,000 Hz frequency, we have not only isolated our 440 Hz tone and half its mean level, but we have also created another version of the signal at twice the frequency, and with half the amplitudes. The frequency domain graph of our new signal appears in the following picture. [I have removed the mean level to make things simpler, and I have given the frequency axis a huge gap so that it is possible to distinguish between the frequencies.]



We now filter out the unwanted frequencies. In the real world, the method used to filter out these frequencies will vary according to the situation. To avoid becoming sidetracked with the details of filtering, we will ignore any actual method, and just say that we will filter out frequencies below 252,000 cycles per second. The technical term for such a filter would be "a low pass filter at 252,000 Hz". The filter would allow frequencies lower than 252,000 Hz to pass on to the next stage of whatever it is we are doing.

After the filtering, we would be left with just:
"y = 0.5 sin (360 * 440t)"
... and a mean level of 1.

... or more concisely:
"y = 1 + 0.5 sin (360 * 440t)"

We would need to remove the mean level from this wave, and then double the amplitude of the wave, and we would end up with our 440 Hz tone that we started with:
"y = sin (360 * 440t)"

[In the real world with radio waves, we would be more interested in the tone itself, and not its exact original amplitude. For one thing, any evidence of its original amplitude would have been lost as the signal travelled from the transmitter to the receiver – the perceived amplitude would diminish as it travelled. Also, the amplitude of the tone when it was transmitted might not suit that required by the person receiving it – in other words, someone receiving it might want to increase or decrease the volume of the tone to their own taste, rather than rely on exactly the volume at which it was transmitted. Given all that, it would be unlikely that someone would bother specifically *doubling* the amplitude of the wave. Instead, they would scale the amplitude to whatever they wanted it to be.]

We will look at modulation in more depth in Chapters 35 to 38.

Finding the multiplied waves from some added waves

It is possible to work backwards from a set of Addition Waves to find the two waves that could have been multiplied to create them. As different multiplications can result in the same Addition Waves, we are really finding one of the countless possible pairs of Multiplication Waves.

Just to demonstrate how this is possible, we will look at one pair of Addition Waves. We will look at these two Addition Waves:
"y = 4.5 sin ((360 * 7t) + 90)"
... and:
"y = 4.5 sin ((360 * 5t) + 270)"

To calculate the amplitudes of the Multiplication Waves, we can use the rule for finding the Addition Wave's amplitudes. It is that both Addition Wave amplitudes will be: 0.5 * (a_1 * a_2). Working backwards from this, we have:
0.5 * (a_1 * a_2) = 4.5
(a_1 * a_2) = 9

There are countless answers to this, so we will just pick one and say that a_1 and a_2 are equal to each other, and say they are 3. They could just as easily be 9 and 1, or 4.5 and 2, or 18 and 0.5, and so on.

To calculate the Multiplication Wave frequencies we use the rule for finding the Addition Wave frequencies. The first Addition Wave is calculated with $f_1 - f_2$, and the second Addition Wave is calculated with $f_1 + f_2$. The faster of our two given waves is 7 cycles per second, so that must be the result of "$f_1 + f_2$", and 5 cycles per second must, therefore, be the result of "$f_1 - f_2$". We need to find "f_1" and "f_2" in these calculations:
$f_1 + f_2$ =7
... and:
$f_1 - f_2 = 5$.
We can work out that f_1 must be 6 cycles per second, and f_2 must be 1 cycle per second.

For the phases, the Addition Waves would have originally been found with ($\phi_1 - \phi_2$ + 90) and ($\phi_1 + \phi_2$ − 90). As the phases in our given waves are +90 and +270, we know that $\phi_1 - \phi_2$ is zero, and $\phi_1 + \phi_2$ is also zero. Therefore, the phases must both have been zero.

From all of this, we can tell that the two Multiplication Waves that when multiplied together would produce our two Addition Waves *could* have been:
"y = 3 sin (360 * 7t)"
... and:
"y = 3 sin (360 * 5t)"
... although there are other waves that would have produced the same results.

This was just a simple example to give an idea of what is possible.

Conclusion

In this chapter, we have seen much about the multiplication of waves. The most important thing to know is that the multiplication of two waves will always result in a signal that is equivalent to the addition of four waves and a mean level. Depending on the characteristics of the two multiplied waves, two of those four waves might end up as "y = 0", the mean level might be zero, and one of the waves might, itself, end up as a mean level.

There is more to know about multiplication, and there are countless ways in which multiplication can be used to assist in the analysis of waves.

Chapter 17: Multiplication with circles

In this chapter, we will look at multiplication involving circles. As before, I use the term “circle” as shorthand to refer to an object rotating around a circle. The basic idea behind multiplication with circles is similar to addition with circles.

If we multiply two circles, we will actually be multiplying corresponding (by time) coordinates from one object’s path around its circle with the coordinates from a second object’s path around its own circle. This is the same as multiplying the corresponding (by time) y-axis values from the derived Sine waves of two objects rotating around their circles, then multiplying the corresponding (by time) y-axis values from the derived Cosine waves of those two objects, and then using the results as coordinates for a new path.

When we perform multiplication with circles, either we can think of the multiplication as involving the actual coordinates of the object rotating around the circle, or we can think of the Cosine and Sine wave that make up the coordinates of that circle. They amount to the same thing, but sometimes one is easier to visualise than the other.

Multiplication involving circles might not be as obviously useful a concept as multiplication involving waves, but for the sake of completeness, it helps to have a basic understanding of the idea.

There are three types of multiplication when it comes to circles:

- Multiplication of a circle by a number. By this, I mean multiplying all the coordinates of an object rotating around a circle by the same number. This is identical to multiplying every value from the two waves that make up the circle by the same number. The effect of this is scaling how far away each point is from the origin of the axes, and thus how big the circle is.

- Multiplication of a circle by another circle. By this, I mean multiplying the x-axis coordinate from every point of one circle by the corresponding x-axis coordinate from every point of the other circle, and then doing the same for the y-axis coordinates. [By “corresponding”, I mean “corresponding by *time*” rather than *angle*. Therefore, the two sets of coordinates are the coordinates of two objects rotating around their circles at particular frequencies, phases, mean levels and amplitudes]. This form of multiplication is identical to multiplying every value from the Sine wave that makes up one circle by the corresponding values from the Sine wave that makes up the other circle, then doing the same for the values from the Cosine waves, and then using the results as coordinates.

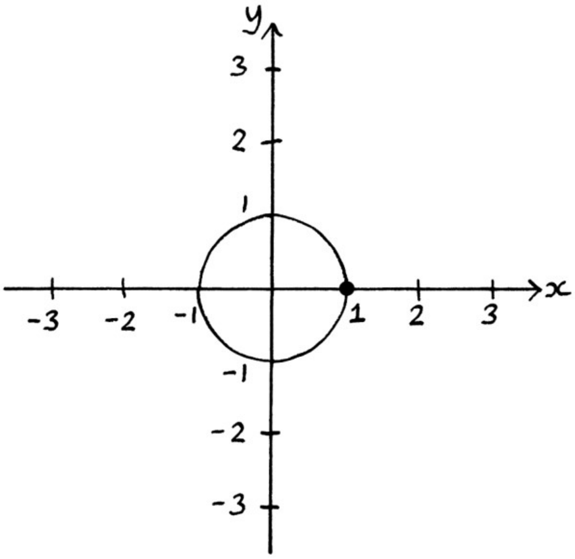
- Multiplication of a circle by a wave. By this, I mean multiplying both the x and y-axis coordinates of every point of an object rotating around a circle by the corresponding (in time) value from a wave. This is identical to multiplying each of the pair of waves that make up a circle by the same particular wave, and then using the results as coordinates.

Multiplying by a number

Multiplication of a circle by a number means that every x-axis and y-axis coordinate of the object rotating around a circle is multiplied by that number.

Circles centred on the origin of the axes

If we have an object rotating around a unit-radius circle at 1 cycle per second, and it started at 0 degrees, and that circle is centred on the origin of the axes, then:
At t = 0, its coordinates will be (1, 0)
At t = 0.125, its coordinates will be (0.7071, 0.7071)
At t = 0.5, its coordinates will be (0, 1)
... and so on. The circle will look like this:

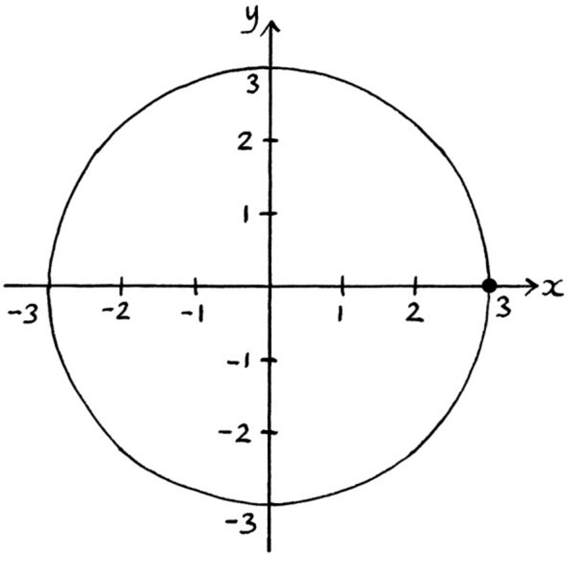


The circle is made up of the two waves “y = sin 360t” and “y = cos 360t”, which is another way of saying that the two derived waves from the circle are “y = sin 360t” and “y = cos 360t”.

If we multiply every coordinate of the object as it moves around the circle by the number 3, we will be scaling its position to be 3 times further away from the origin at every moment in time:
At t = 0, its coordinates will be (3, 0)
At t = 0.125, its coordinates will be (2.1213, 2.1213)
At t = 0.5, its coordinates will be (0, 3)
... and so on.

We are really multiplying each part of every coordinate of the circle by 3. If we gave every coordinate of the first circle as (x, y), then the after the multiplication, every coordinate would be (3x, 3y). If we gave every coordinate in terms of Sine and Cosine, then the original circle would have coordinates that for any moment in time would be (cos 360t, sin 360t), and the resulting circle would have coordinates that for any moment of time would be (3 cos 360t, 3 sin 360t).

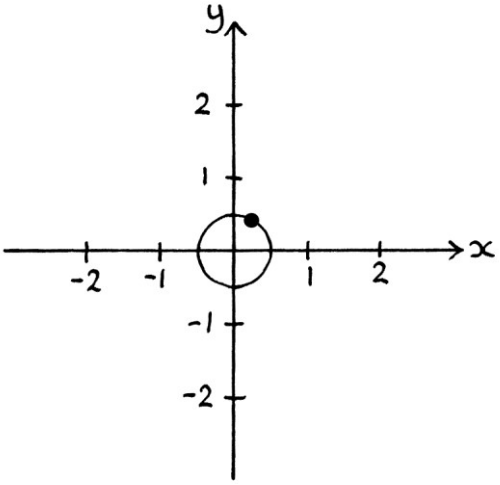
The multiplication of the coordinates, in this case, is the same as multiplying the radius of the circle by 3, or the amplitudes of the derived waves by 3. The rotating object moves around at 3 times the distance from the origin as before. The path it takes is a circle with a radius three times that of the original circle. The derived waves will have the formulas “y = 3 sin 360t” and “y = 3 cos 360t”.



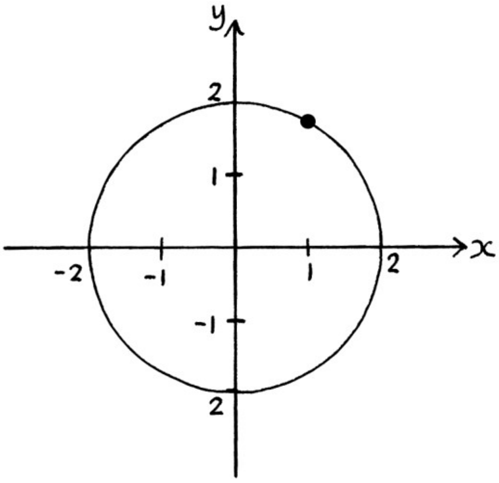
We already knew all of this from understanding amplitude in the formulas for waves, but it is good to think about the details of such a multiplication.

The start point of the circle will be at the same angle from the circle's centre as it was before. The frequency of the object rotating around the circle will also be the same as before. We can tell these are true because changing the amplitude of a wave does not change its phase or frequency.

As another example of multiplication of a circle by a number, we will look at the circle made up of these two waves:
"y = 0.5 sin ((360 * 4t) + 60)"
... and:
"y = 0.5 cos ((360 * 4t) + 60)"
The circle looks like this:



If we multiply the coordinates of the circle by the number 4, we end up with this circle:



We can calculate the waves that make up this circle. They will be:
“y = 4 * (0.5 sin ((360 * 4t) + 60))”
... and:
“y = 4 * (0.5 cos ((360 * 4t) + 60))”

... which end up as:

“y = 2 sin ((360 * 4t) + 60)”
... and:
“y = 2 cos ((360 * 4t) + 60)”

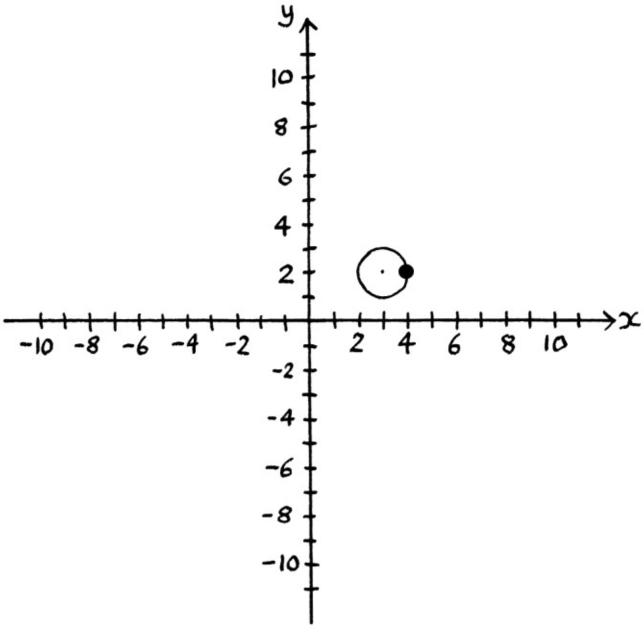
The phase is still 60 degrees; the frequency is still 4 cycles per second.

Circles not centred on the origin of the axes

If a circle is not centred on the origin of the axes, then at least one of the derived waves must have a non-zero mean level. The effects of multiplication by a number will be different from before.

We will look at the circle made up of these waves:
“y = 2 + sin 360t”
... and:
“y = 3 + cos 360t”

The circle looks like this:

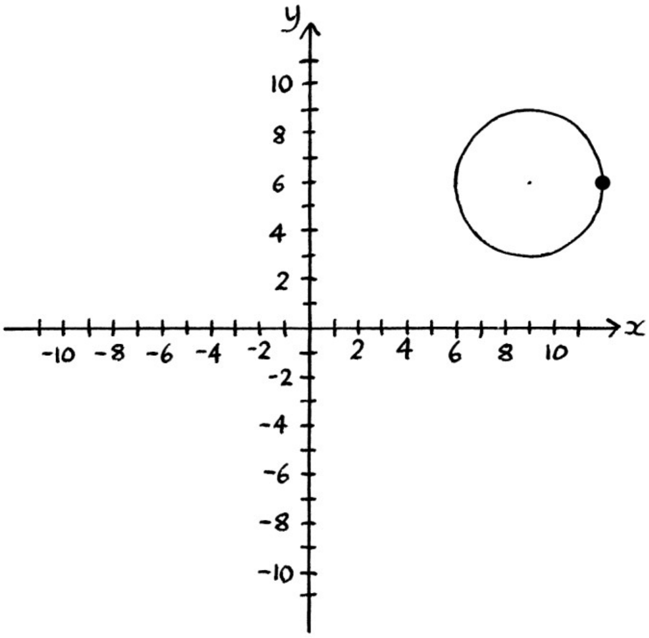


Supposing we multiplied the points around this circle by the number 3, then we know that the result would be made up of these waves:
“y = 3 * (2 + sin 360t)”
... and:
“y = 3 * (3 + cos 360t)”

... which end up as:

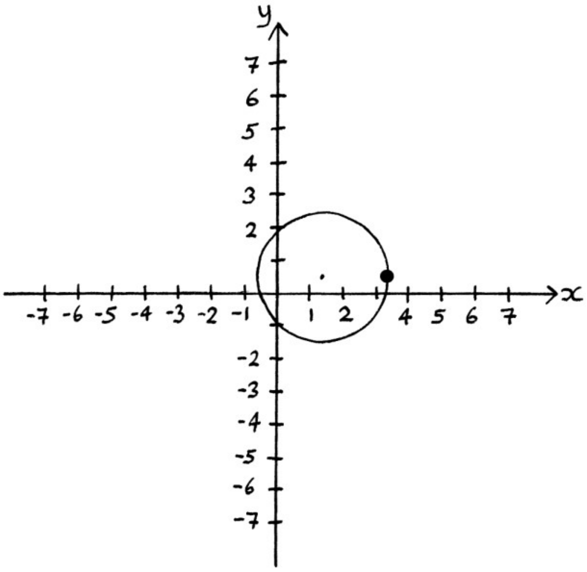
“y = 6 + 3 sin 360t”
... and:
“y = 9 + 3 cos 360t”

The resulting circle looks like this:

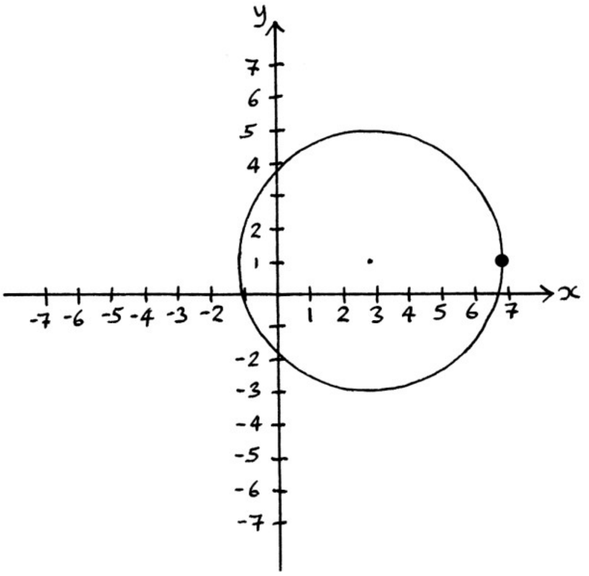


The resulting circle has a radius of 3 units, and it is much further away from the origin of the axes because the mean levels have become scaled by 3 as well as the radius. The frequency and phase are unaffected.

If we have a circle that borders the origin of the axes, but without being centred on it, multiplication will still work in the same way. A circle made up of the waves:
“y = 1.4 + 2 sin 360t”
... and:
“y = 0.5 + 2 cos 360t”
... will look like this:



If we multiply the circle by 2, we end up with this:



The result is still a circle. Its derived waves are:
“y = 2.8 + 4 sin 360t”
... and:
“y = 1 + 4 cos 360t”

Thoughts

In all cases of multiplication with circles, we can think of the process as multiplication of each coordinate in turn, or we can think of it as multiplication of the Sine and Cosine waves that make up the circle. As the coordinates of a circle can be given in terms of Sine and Cosine waves, these amount to the same thing.

Multiplying a circle by a circle

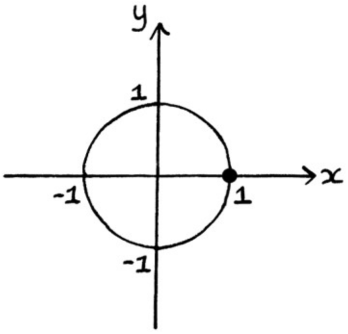
The effects of multiplication become more complicated when we multiply one circle by another circle. By “multiplying one circle by another circle”, I mean multiplying the x-axis coordinates of the objects rotating around each circle at corresponding moments in time, and multiplying the y-axis coordinates of the objects rotating around each circle at corresponding moments in time. This is the same as multiplying the Sine waves that make up the y-axis values of the two circles, and multiplying the Cosine waves that make up the x-axis coordinates of the two circles, and using the results as coordinates.

Example 1: A circle multiplied by itself

For the first example, we will multiply a circle by itself. This is the same as multiplying two identical circles. The results of doing this are that every x-axis value will be squared, and every y-axis value will be squared. An object rotating around the circle will have its coordinates squared for every moment in time.

We will use a circle based on these waves:
“y = sin 360t”
... and:
“y = cos 360t”

The coordinates of every point on this circle's edge are: (cos 360t, sin 360t). The circle looks like this:



The coordinates of the object rotating around this original circle at every sixteenth of a second are as follows:

Time in seconds	**x-axis value**	**y-axis value**
0	1	0
0.0625	0.9239	0.3827
0.125	0.7071	0.7071
0.1875	0.3827	0.9239
0.25	0	1
0.3125	−0.3827	0.9239
0.375	−0.7071	0.7071
0.4375	−0.9239	0.3827
0.5	−1	0
0.5625	−0.9239	−0.3827
0.625	−0.7071	−0.7071
0.6875	−0.3827	−0.9239
0.75	0	−1
0.8125	0.3827	−0.9239
0.875	0.7071	−0.7071
0.9375	0.9239	−0.3827
1	1	0

If we square each of the x and y-axis coordinates, we end up with this table:

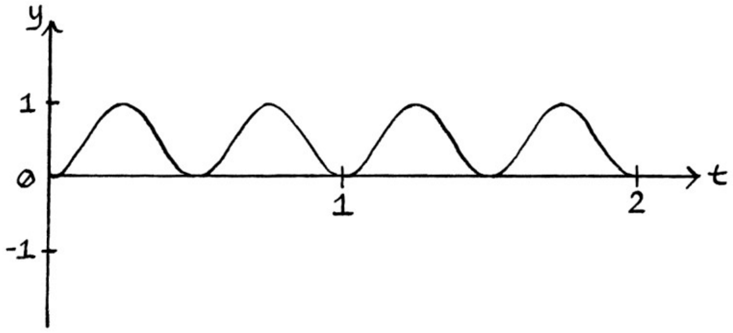
Time in seconds	**Squared x-axis value**	**Squared y-axis value**
0	1	0
0.0625	0.8536	0.1464
0.1250	0.5	0.5
0.1875	0.1464	0.8536
0.2500	0	1
0.3125	0.1464	0.8536
0.3750	0.5	0.5
0.4375	0.8536	0.1464
0.5	1	0
0.5625	0.8536	0.1464
0.6250	0.5	0.5
0.6875	0.1464	0.8536
0.7500	0	1
0.8125	0.1464	0.8536
0.8750	0.5	0.5
0.9375	0.8536	0.1464
1	1	0

From this table, we can see that every resulting value is positive – this should be expected as squaring a negative value will make it positive. If we were to use these as coordinates, every point would be in the top right quarter of the circle chart.

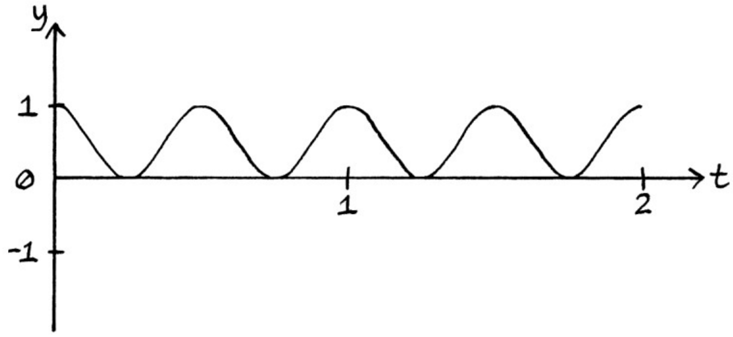
Before we see the shape drawn out by these coordinates, we will look at how we could have done the multiplication using wave formulas.

The circle's coordinates are (cos 360t, sin 360t) for every moment in time. We can use our knowledge of multiplication from the last chapter to calculate the squares of these waves.

The square of "y = sin 360t" is: "y = 0.5 + 0.5 sin ((360 * 2t) + 270)". It looks like this:

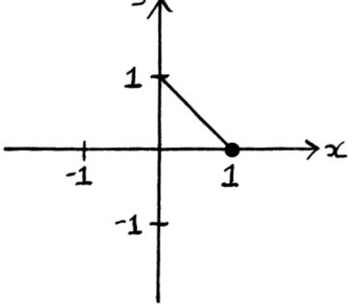


The square of "y = cos 360t" is: "y = 0.5 + 0.5 cos (360 * 2t)". It looks like this:



These two waves are no longer 90 degrees apart, but are, in fact, 180 degrees apart. [Remember that they are representing coordinates, and that they are not being added together]. This means that the shape they portray will definitely not be a circle. The coordinates of points along the edge of the shape will be:
(0.5 + 0.5 cos (360 * 2t), 0.5 + 0.5 sin ((360 * 2t) + 270))
... for all moments in time.

The shape, whether it is drawn from the table of coordinates, or whether it is drawn from the two waves, looks like this: [The black dot still marks the starting point of the object.]



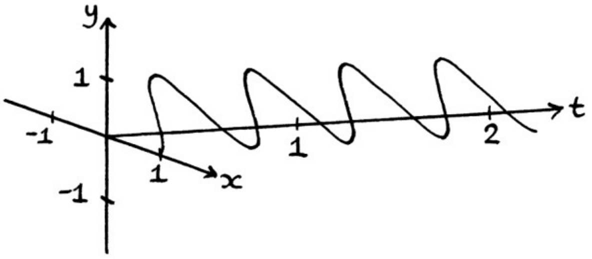
The resulting shape is just a straight line at 135 degrees going from (1, 0) up to (0, 1).

This line still represents the movement of an object. The object starts at the x-axis coordinate of: 0.5 + 0.5 cos (0) = 1, and at the y-axis coordinate of: 0.5 + 0.5 sin (0 + 270) = 0, or in other words at (1, 0). The object moves up the line until it reaches (0, 1) at 0.25 seconds, and then it moves back again, when it reaches (1, 0) at 0.5 seconds. It then moves up the line to reach (0, 1) at 0.75 seconds, and moves back to (1, 0) at 1 second. Its movement over time is easier to comprehend by looking at the table above that shows the results of the squared coordinates.

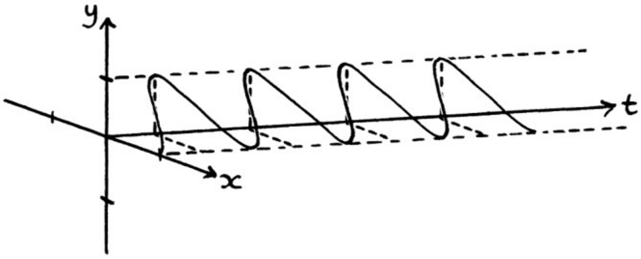
The object takes 0.5 seconds to do one trip up and down the line. This might be expected as the two waves that make up the coordinates have frequencies of 2 cycles per second, and so take 0.5 seconds to complete one cycle.

The helix chart

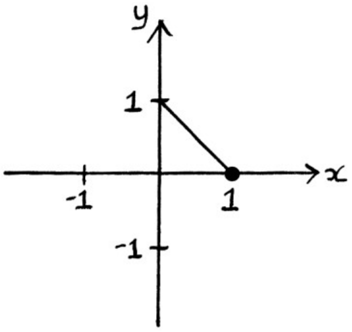
The resulting straight line is easier to understand when we look at the "helix" (which obviously is not literally a helix), on the helix chart:



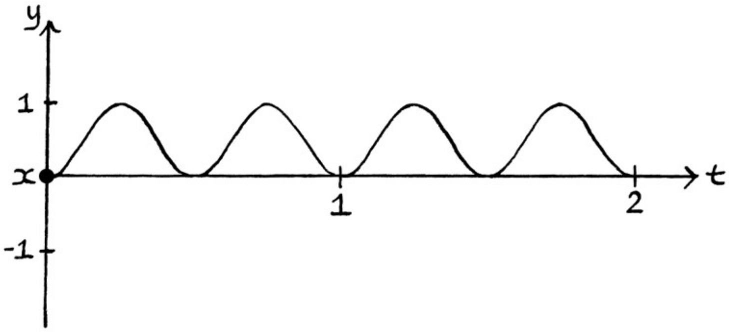
Here is the same picture drawn with dotted lines to clarify its dimensions:



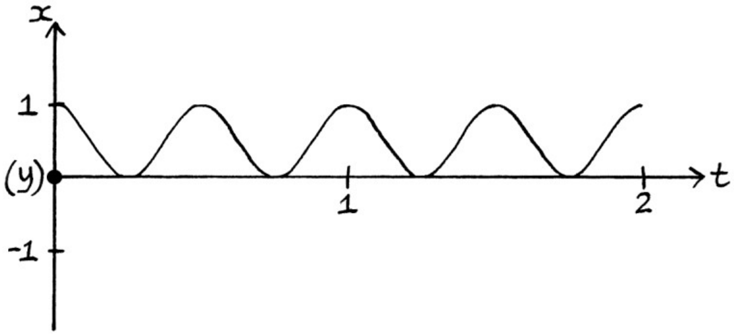
In the same way as if we were dealing with normal waves, if we view this "helix" end on, with the time axis pointing away from us, we see the straight line (or the circle chart view)



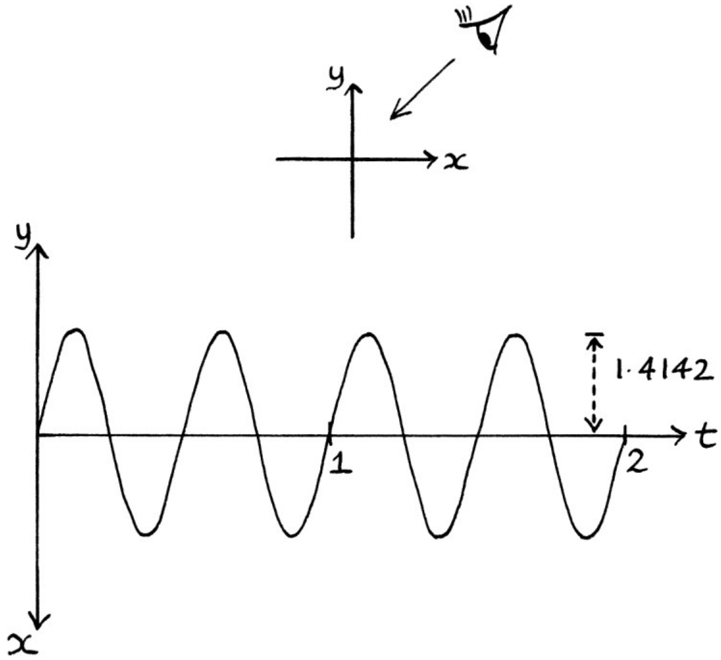
If we view the "helix" with the x-axis pointing directly at us, and the time axis pointing to the right, we see the result of the squared Sine wave:



If we view the "helix" from underneath, with the y-axis pointing directly away from us, and the time axis pointing to the right, we see the see the result of the squared Cosine wave:



If we viewed the helix chart at a 45-degree angle, with the y-axis pointing upwards, the x-axis pointing downwards, and the t-axis pointing to the right, we would see a Sine wave with an amplitude of 1.4142 units.



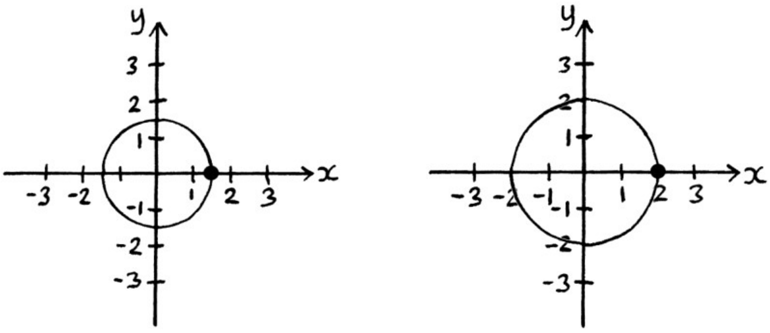
In this example, we have created a Sine wave placed at 45-degrees in the helix chart.

Example 2: Two different circles

Next, we will multiply two circles based on the waves:
“y = 1.5 sin (360 * 3t)” and “y = 1.5 cos (360 * 3t)”.
... and:
“y = 2 sin (360 * 5t)” and “y = 2 cos (360 * 5t)”.

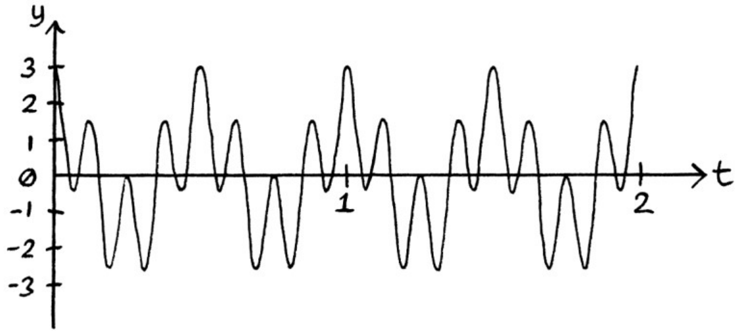
The first circle has a radius of 1.5 units and a frequency of 3 cycles per second; the second circle has a radius of 2 units and a frequency of 5 cycles per second. The frequencies of 3 cycles per second and 5 cycles per second will coincide every second. Therefore, the resulting waves and the resulting “shape” will have a frequency of either 1 cycle per second, or 2 cycles per second, depending on how the waves align with each other.

The two circles look like this:



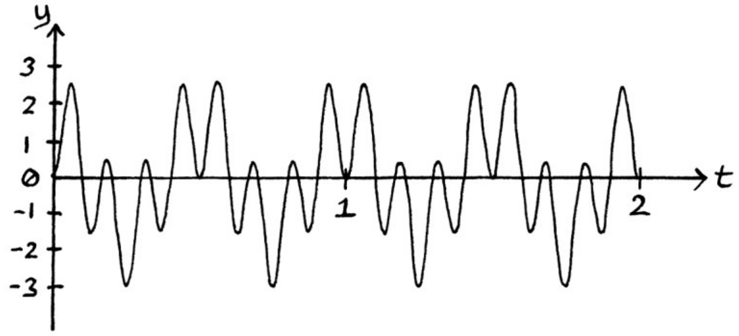
The x-axis coordinates of the resulting shape for every moment in time will be based on the signal:
“y = 1.5 cos (360 * 3t)” multiplied by “y = 2 cos (360 * 5t)”
... which is equivalent to:
“y = 1.5 cos (360 * 2t)” added to “y = 1.5 cos (360 * 8t)”

This signal looks like this:

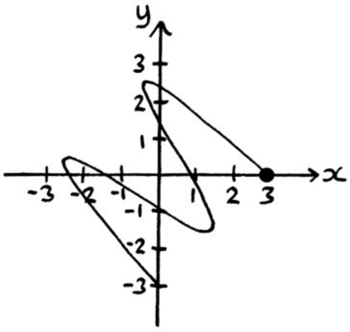


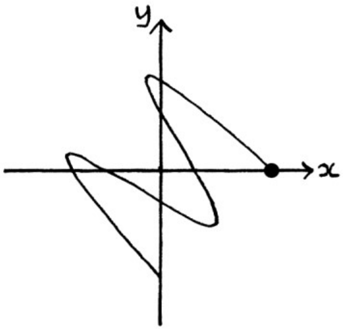
The y-axis coordinates of the resulting shape for every moment in time will be based on the signal:
“y = 1.5 sin (360 * 3t)” multiplied by “y = 2 sin (360 * 5t)”
... which is equivalent to:
“y = 1.5 sin ((360 * 2t) + 90)” added to “y = 1.5 sin ((360 * 8t) + 270)”

This signal looks like this:



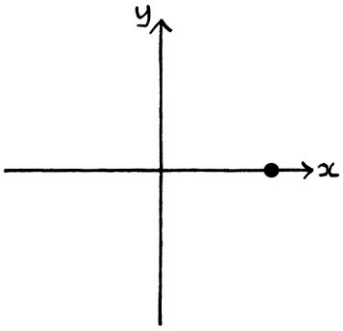
The resulting shape is really just a line, and it looks like this (drawn with axis numbering and without):



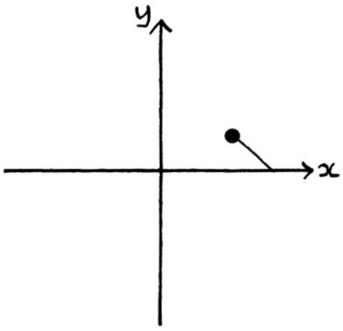


The object starts at the coordinates (3, 0) and moves along the line, until it reaches the coordinates (0, −3), which it does at 0.25 seconds. It then moves in the other direction and returns to its starting place at 0.5 seconds. Its position at every 0.025 seconds from t = 0 to t = 0.25 is as follows:

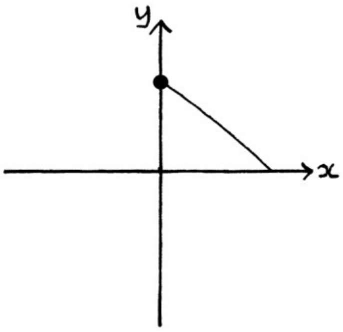
At 0 seconds:



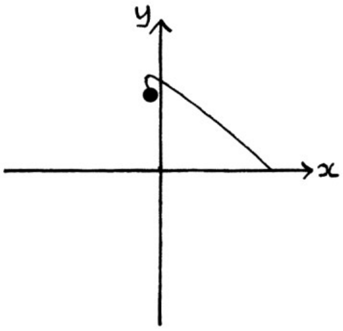
At 0.025 seconds:



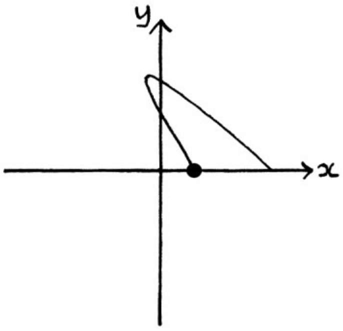
At 0.05 seconds:



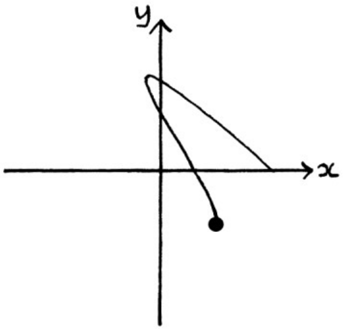
At 0.075 seconds:



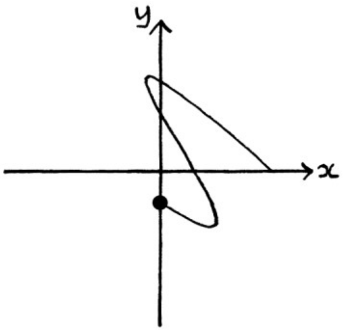
At 0.01 seconds:



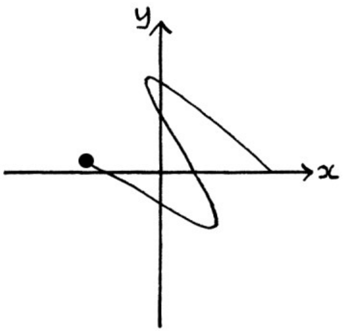
At 0.125 seconds:



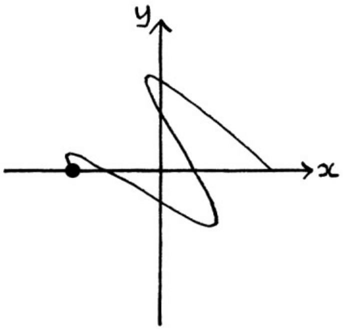
At 0.15 seconds:



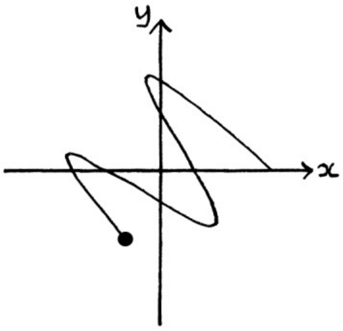
At 0.175 seconds:



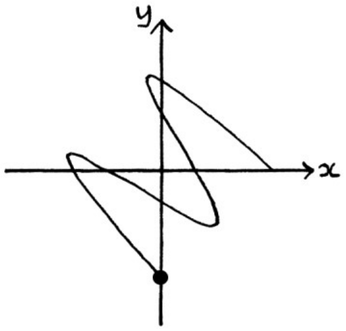
At 0.2 seconds:



At 0.225 seconds:



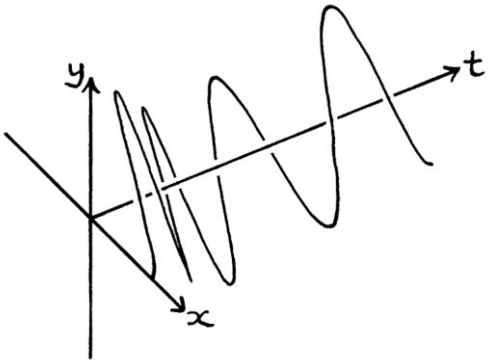
At 0.25 seconds:



The object then moves back to the start.

The helix chart

The movement of an object is usually easier to understand if we think about the helix chart. However, in this particular case, it is difficult to draw such a picture without it looking like a muddle of lines.



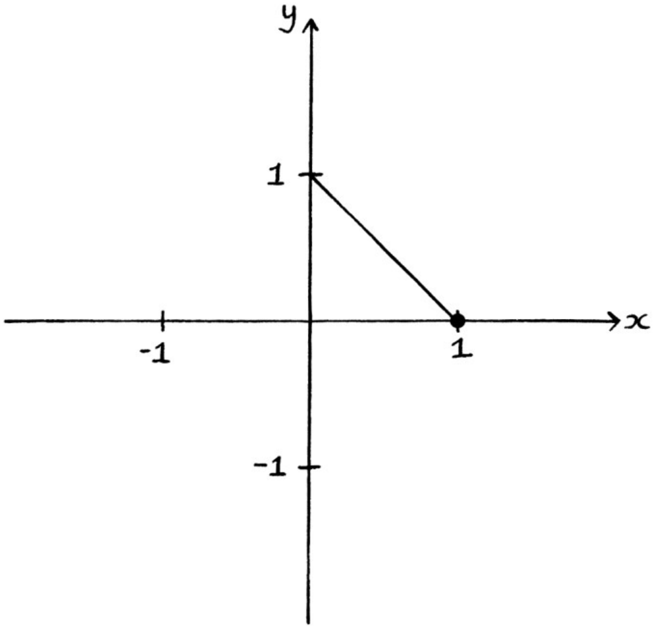
Circles by circles: consistent shapes

If we multiply two circles that differ only in frequency, we end up with particular shapes. The shapes are dependent on the ratio between the two frequencies.

Ratio of 1 : 1

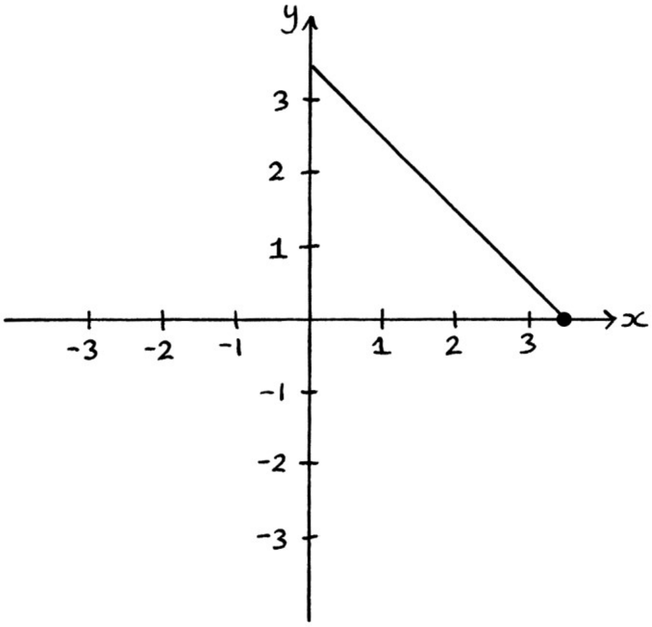
We will start with two identical circles each based on the waves:
“y = sin 360t” and “y = cos 360t”.

The frequency ratio between these circles is 1 : 1. The resulting shape on the circle chart looks like this, which is the same as we had in the first example:



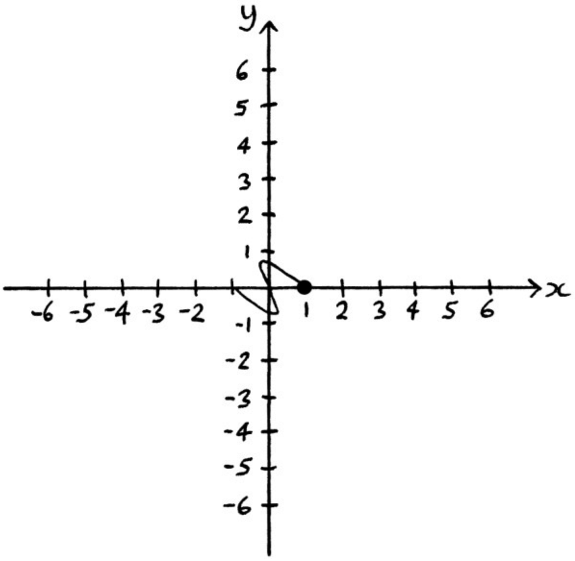
Any frequency ratio of 1 : 1, where the waves have zero mean levels, will produce a line *similar* to this.

For example, if we multiply the circle based on:
“y = 0.5 sin (360 * 5t)” and “y = 0.5 cos ((360 * 5t)”
... by the circle based on:
“y = 7 sin (360 * 5t)” and “y = 7 cos (360 * 5t)”
... then we will have the following graph:

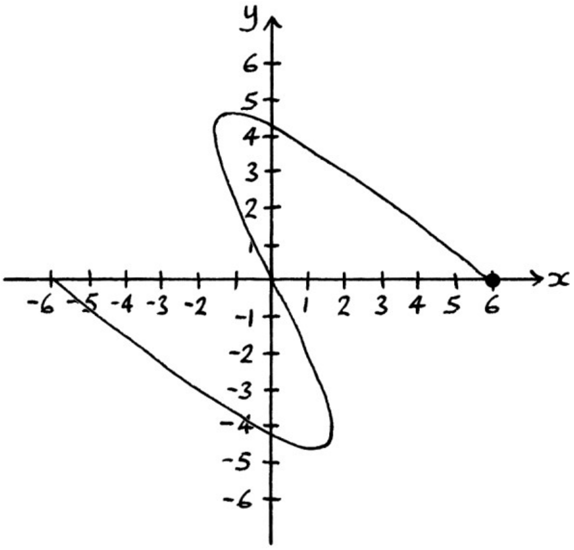


Ratio of 2 : 1

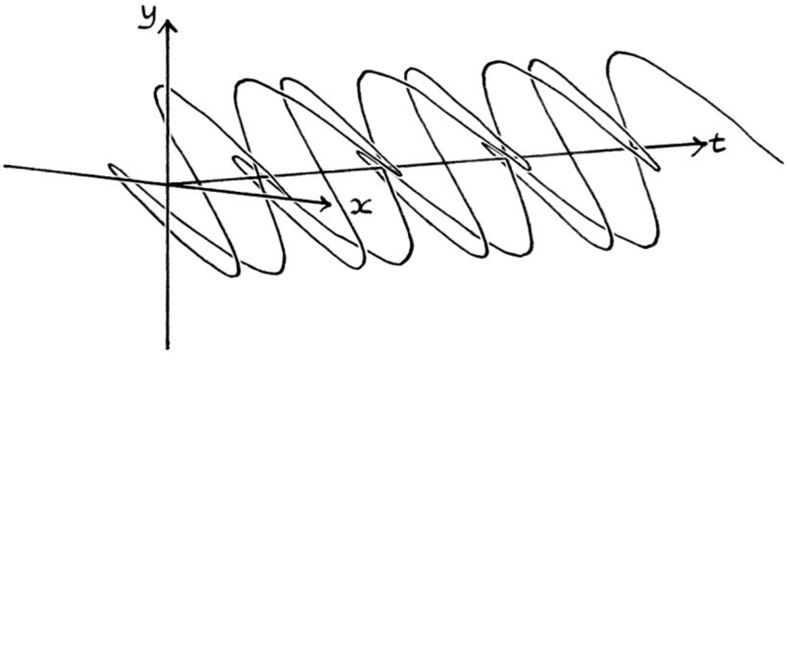
If we have a circle based on:
“y = sin (360 * 3t)” and “y = cos (360 * 3t)”
... and we multiply it by a circle based on:
“y = sin (360 * 6t)” and “y = cos (360 * 6t)”
... then we end up with this:



If the frequency ratio is 2 : 1, no matter what the actual frequencies are, the resulting shape will be the same. If we have a circle based on:
“y = 2 sin (360 * 3.5t)” and “y = 2 cos (360 * 3.5t)”
... and we multiply all its coordinates by those of the circle based on:
“y = 3 sin (360 * 7t)” and “y = 3 cos (360 * 7t)”
... then we end up with this:

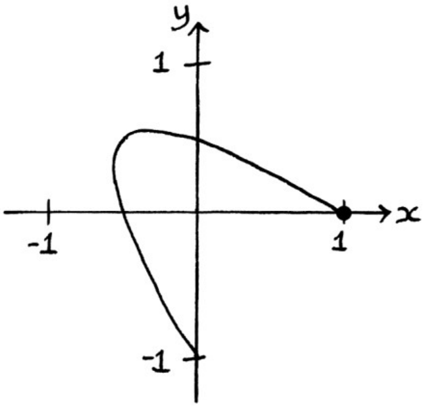


On the helix chart, it looks like this:

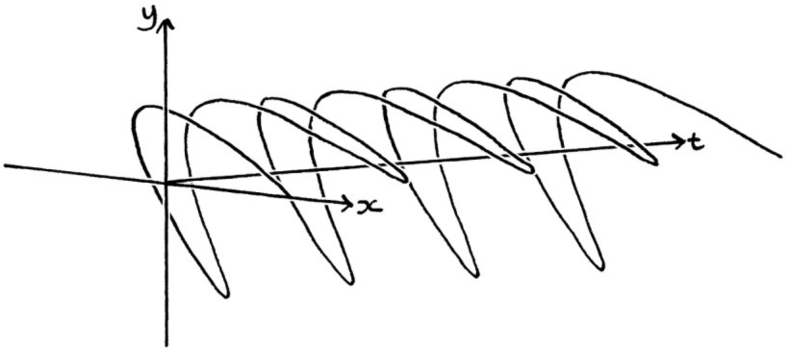


Ratio of 3 : 1

For the same circles, but with a frequency ratio of 3 : 1, we end up with this shape:

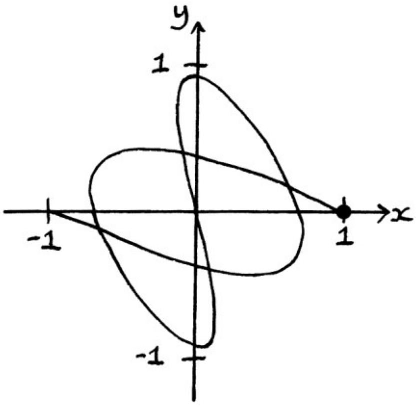


On the helix chart, it looks like this:



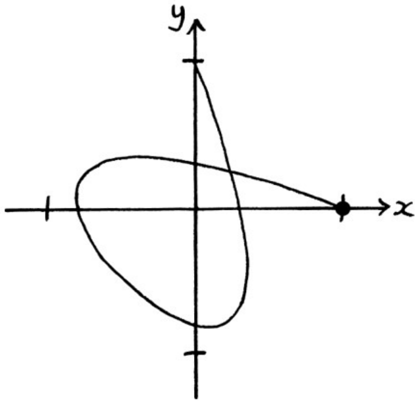
Ratio of 4 : 1

For the same circles with a frequency ratio of 4 : 1, we will have this shape:

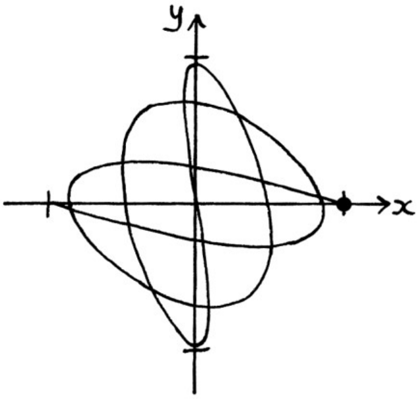


Other ratios:

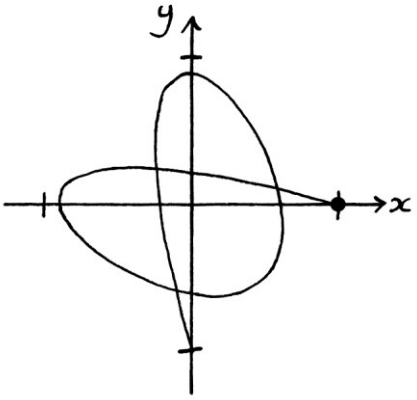
A ratio of 5 : 1 produces this shape:



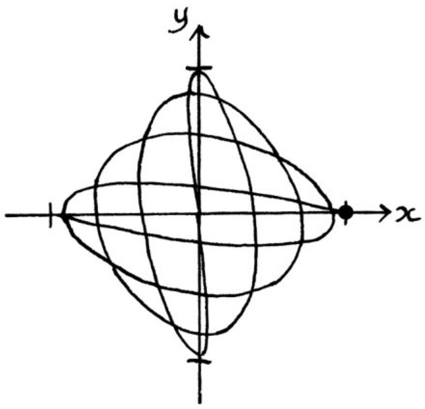
A ratio of 6 : 1 produces this shape:



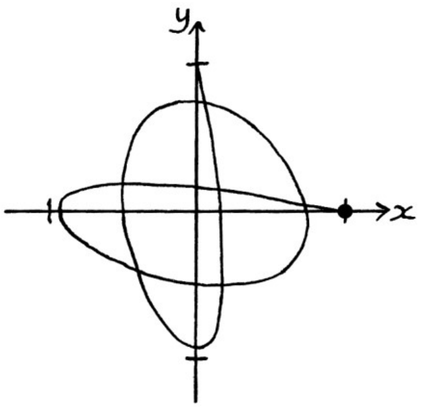
A ratio of 7 : 1 looks like this:



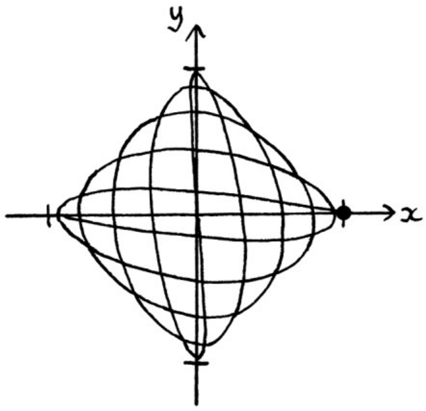
A ratio of 8 : 1 looks like this:



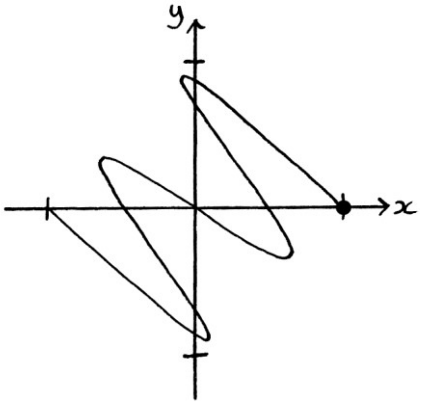
A ratio of 9 : 1 looks like this:



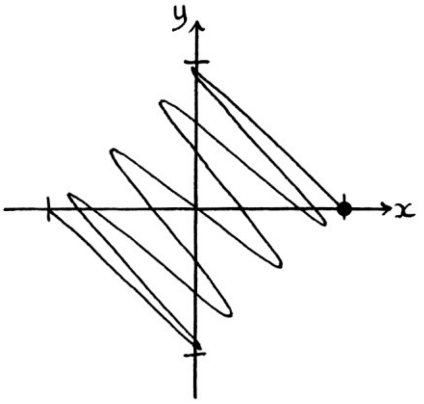
A ratio of 10 : 1 looks like this:



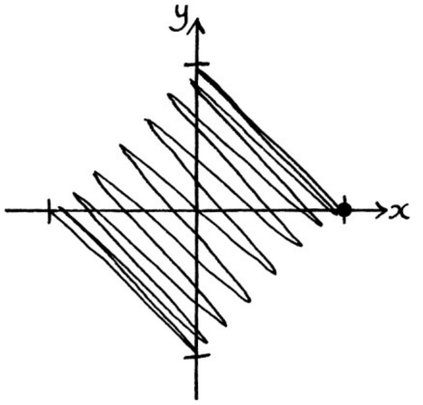
A ratio of 1.5 : 1 looks like this:



A ratio of 1.25 : 1 looks like this:

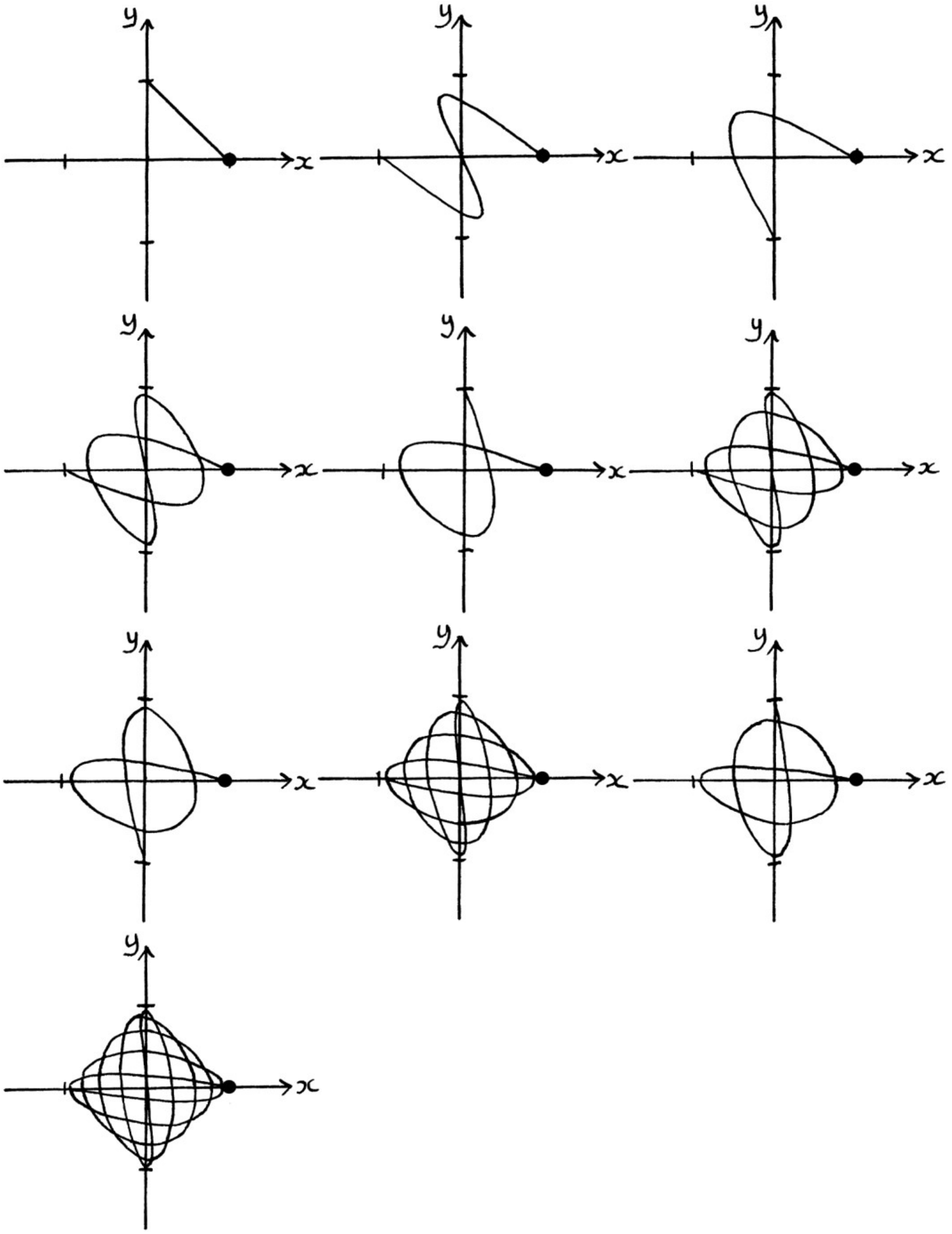


A ratio of 1.125 : 1 looks like this:



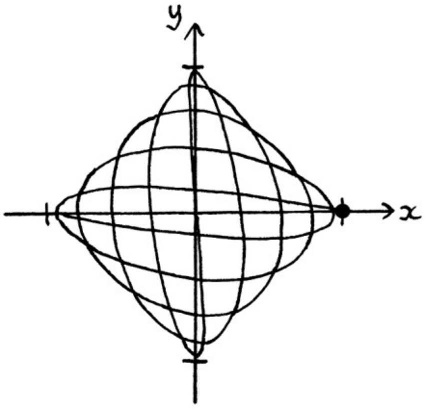
The different shapes together:

Here are the different shapes, from 1 : 1 to 10 : 1, laid out together:

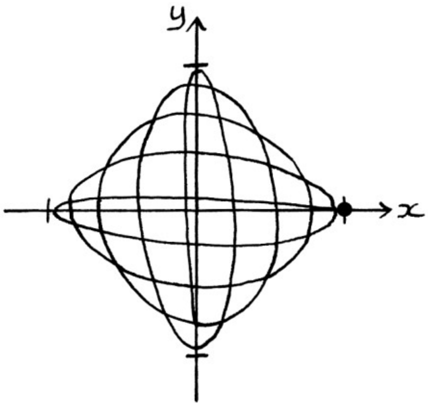


There are several interesting observations connected with these pictures. These are not necessarily useful to know, but they are interesting nonetheless. First, the odd frequency ratios have fewer lines in them than the even frequency ratios either side of them. If we redrew rough approximations of each picture with straight lines instead of curved lines, we would see how the first odd frequency ratio requires 1 line to draw, the second odd frequency ratio requires 2 lines to draw, the third odd frequency ratio requires 3 lines to draw, and so on. However, the first *even* frequency ratio requires 3 lines to draw, the second *even* frequency ratio requires 5 lines to draw, the third even frequency ratio requires 7 lines to draw and so on. The same is true for the number of changes in direction, which really amounts to the same thing. The fact that the odd ratios have fewer lines than the even ones, means that interesting events happen, such as how the shape for a frequency ratio of 10 : 1 is very similar to the shape for a frequency ratio of 19 : 1:

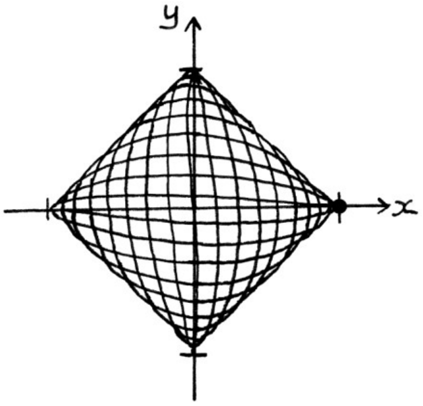
Ratio of 10 : 1:



Ratio of 19 : 1

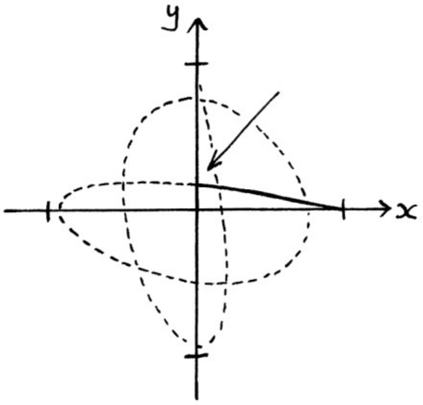


The second interesting aspect to these pictures is that the higher the frequency ratio, the squarer the resulting shape. If we used a frequency ratio of 20 : 1, with amplitudes of 1 for each wave, we would end up with this shape:



This is a square at a 45-degree angle. Its sides are $\sqrt{2}$ units long. Another way of thinking about this is that all the points for the result of the multiplication of two circles, where the radiuses are 1 unit long, will be enclosed within a square in this way. It does not matter what the phases of the two circles are as the points will always be within a square such as this – different phases just change the start point along the edge of the square. The distance from a corner of the enclosing square to the centre will be the two amplitudes multiplied by each other. The length of one side of the square will be the square root of twice the square of that distance. [In other words, we just use Pythagoras's theorem.]

Another interesting aspect of the shapes is how, in each shape, the object moves from its starting point on the x-axis upwards and to the left in each picture. What is of interest here is the y-axis value where the object first crosses the y-axis:



The y-axis value where the object first crosses the y-axis decreases as the frequency ratio increases. There is a distinct rule for where the object first crosses the y-axis. A table showing the first ten y-axis crossing places for when the circles have radiuses of 1 unit, zero phases and zero mean levels is as follows:

Frequency ratio	**y-axis value where the object first crosses the y-axis**
1: 1	1
2 : 1	0.7071
3 : 1	0.5
4 : 1	0.3827
5 : 1	0.3090
6 : 1	0.2588
7 : 1	0.2225
8 : 1	0.1951
9 : 1	0.1737
10 : 1	0.1564

The rule for calculating the y-axis value at which the object first crosses the y-axis is: "sin (90 ÷ the higher number in the ratio)".

Therefore, if we multiply two circles with frequency ratios of 11 : 1, the object on the resulting shape will first cross the y-axis at:
sin (90 ÷ 11) = sin (8.1818) = 0.1423 units.

Another interesting aspect is that the shapes for the ratios from 1 : 1 to 10 : 1 could be used as a slightly unwieldy universal number system. The tenth shape could be used to represent zero. The shapes are devoid of any social or cultural influence, and could be deduced by someone with too much time on their hands. In practice though, there are quicker and more intuitive ways to portray numbers in a universal way – for example, using patterns of dots such as appear on dominoes.

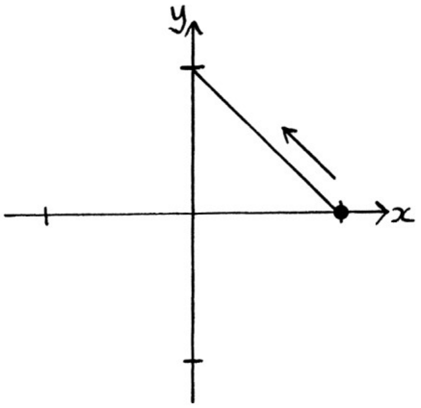
Circles by circles: phases

Now, we will look at the effects of multiplying two circles with non-zero phases. This is easiest to see if we repeatedly multiply one circle by a range of circles with a variety of phases.

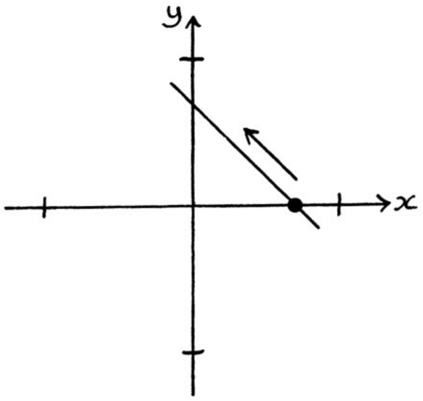
Ratio of 1 : 1

As an example, we will multiply the circle created from the waves:
“y = sin 360t” and “y = cos 360t”
... by the circle created from the waves:
“y = sin (360t + ϕ)” and “y = cos (360t + ϕ)”
... where ϕ is a range of phases from 0 to 360 degrees, spaced at intervals of 45 degrees.

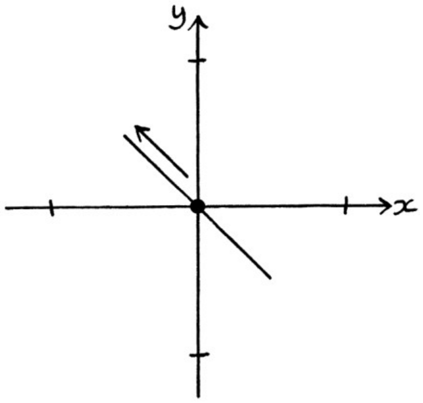
When the second wave has a phase of 0 degrees, the result is as follows. The object starts at (0, 1) and will move up the line. [The arrow indicates the direction in which the object will move to start with.]



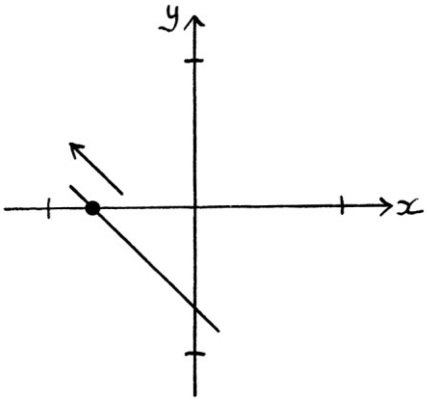
When the second wave has a phase of 45 degrees, the result is as follows. Notice how the whole line has been slid slightly towards the bottom left of the axes. The start point is no longer at the end of the line, but it is still on the x-axis.



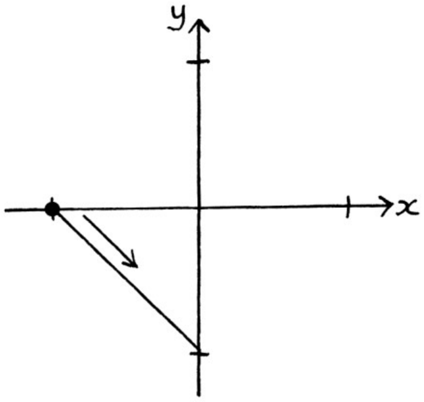
When the second wave has a phase of 90 degrees, the result appears as follows. The line is centred on the origin of the axes, and the start point is also on the origin of the axes. The object will still move up the line when it starts moving.



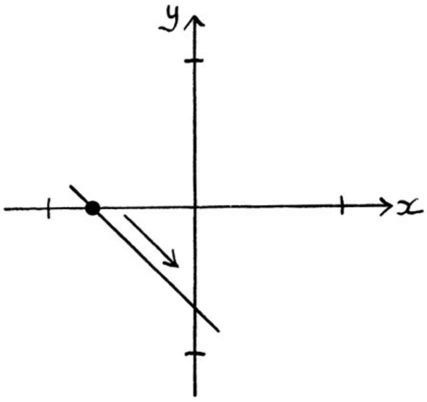
When the second wave has a phase of 135 degrees, the result is this:



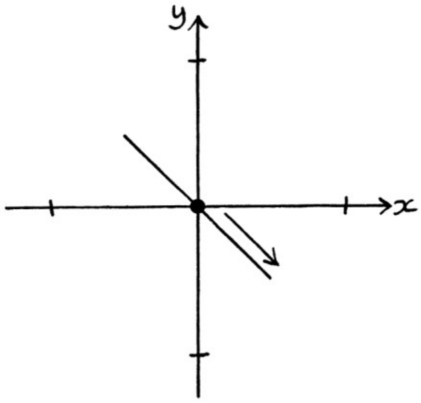
When the phase is 180 degrees, the result appears as follows. The line now extends from (0, −1) to (−1, 0), whereas when the second wave had a phase of zero degrees, it extended from (1, 0) to (0, 1). The starting point is now at (−1, 0). Given that the start point is at the top of the line, the object can only move down the line when it starts.



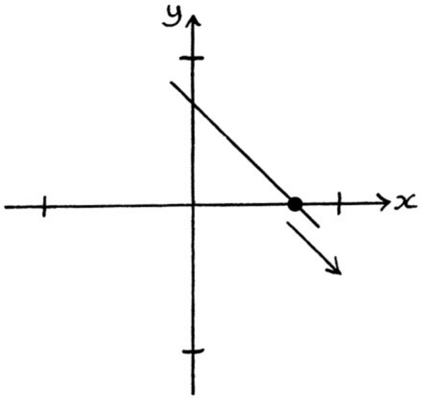
When the second wave has a phase of 225 degrees, the result is as follows. The line is moving back towards where it started. The object moves down the line when it starts.



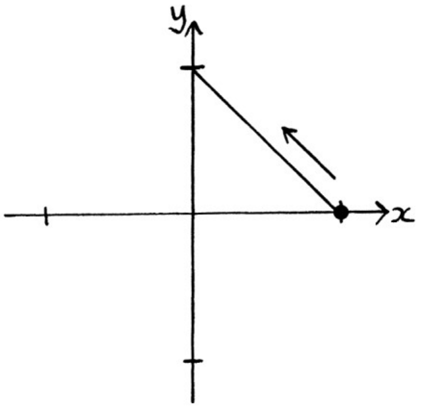
When the second wave has a phase of 270 degrees, the result is as follows. The line is centred over the origin of the axes, as is the start point. The object moves down the line when it starts.



When the second wave has a phase of 315 degrees, the result looks like this:



When the second wave has a phase of 360 degrees, the result is as in the following picture. The line and the object are where they would be if there were no phase difference, which should be expected as 360 degrees is the same as 0 degrees.

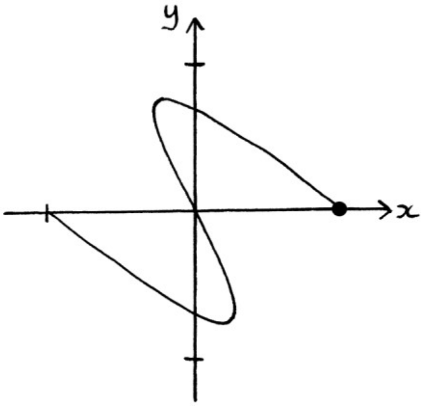


Ratio of 2 : 1

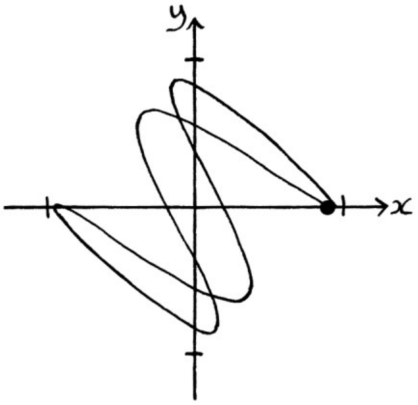
When the frequencies have a ratio of 2 : 1, and the phase changes from 0 to 360, the results are much more interesting. We will multiply the circle based on these waves:
"y = sin 360t" and "y = cos 360t)"
... by the circle based on these waves:
"y = sin ((360 * 2t) + ϕ)" and "y = cos ((360 * 2t + ϕ)"
... where ϕ is a range of phases from 0 to 360 degrees, spaced at intervals of 22.5 degrees.

What is interesting about the results is that at each phase increase, it is as if the shape is a three-dimensional picture that is being rotated.

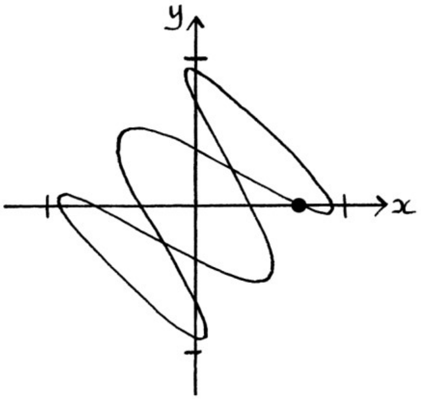
When the second wave's phase is zero degrees, the result looks like this:



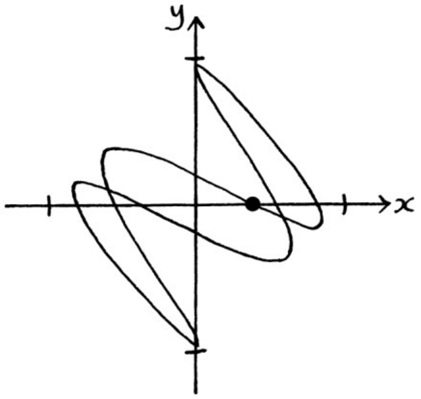
When the second wave's phase is 22.5 degrees:



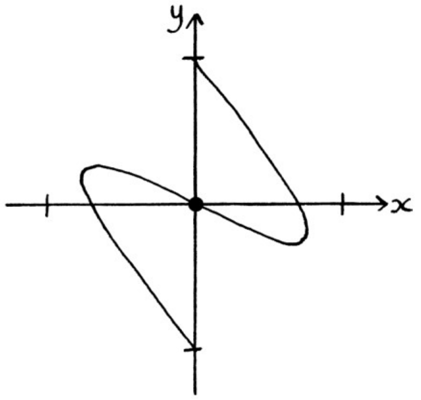
When the second wave's phase is 45 degrees, the result looks like this. Note how the start point of the object remains at y = 0 for all time.



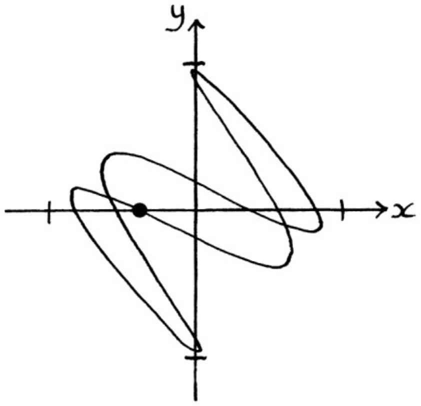
When the second wave's phase is 67.5 degrees:



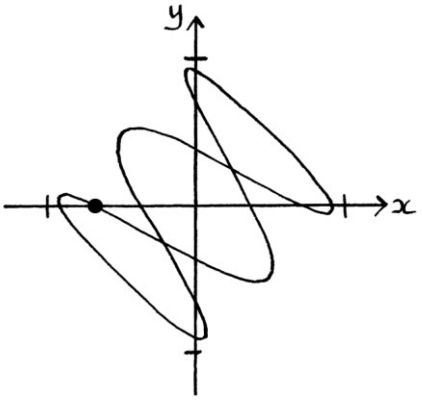
When the second wave's phase is 90 degrees, the result looks is as follows. The start point is directly on the origin of the axes. The shape looks like a backwards "S" drawn at 45 degrees, while when the second wave's phase was 0 degrees, the result looked like a forward "S" drawn at 45 degrees.



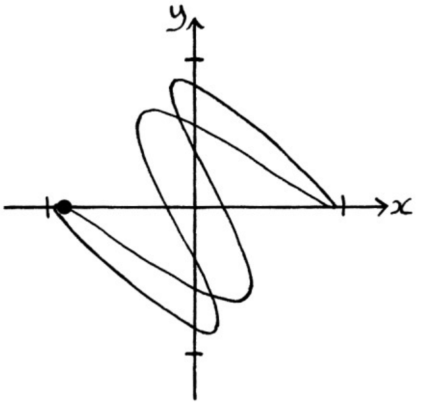
When the second wave's phase is 112.5 degrees:



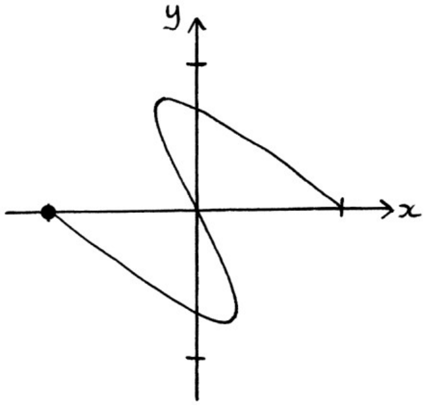
When the second wave's phase is 135 degrees:



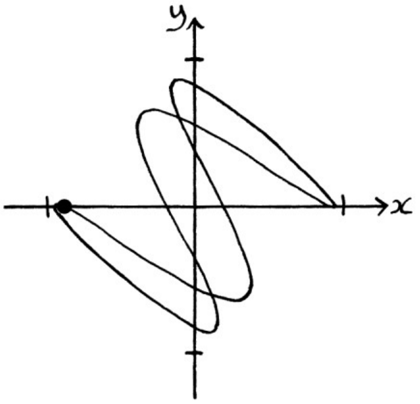
When the second wave's phase is 157.5 degrees:



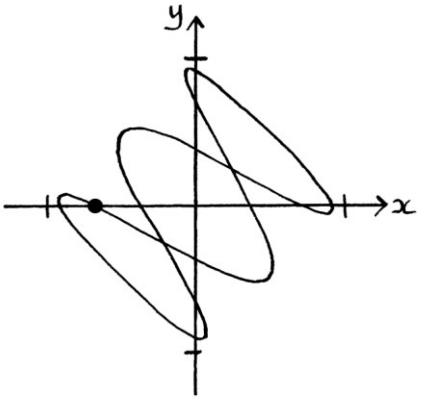
When the second wave's phase is 180 degrees, the result is as follows. Note how this shape is the same as when the phase was 0 degrees, however the start point is at the other end of the line:



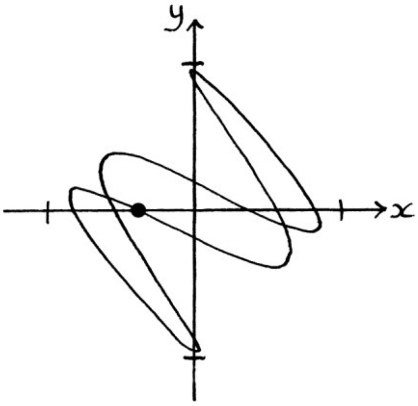
When the second wave's phase is 202.5 degrees:



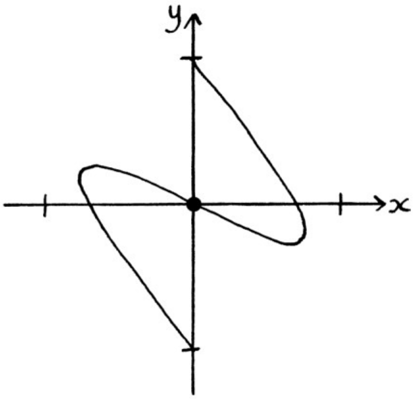
When the second wave's phase is 225 degrees:



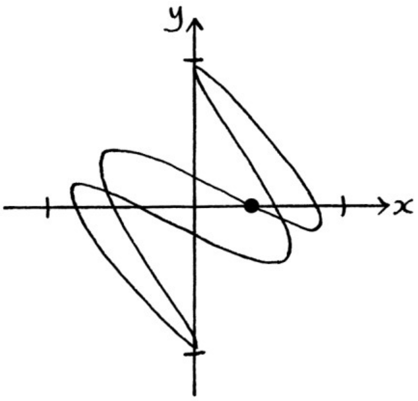
When the second wave’s phase is 247.5 degrees:



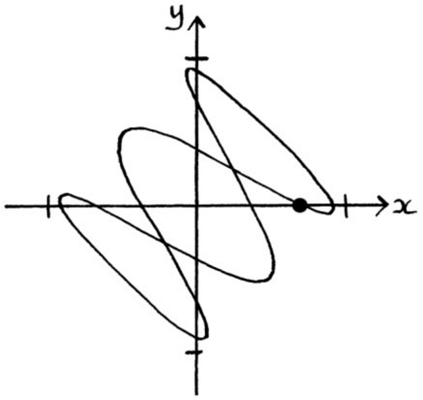
When the second wave’s phase is 270 degrees, the result looks as follows. This is the same shape as when the phase was 90 degrees. The start point is also over the origin of the axes. However, now the object will be moving down towards the right instead of up towards the left when it starts.



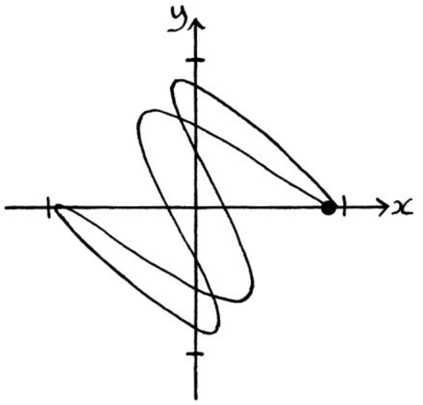
When the second wave’s phase is 292.5 degrees:



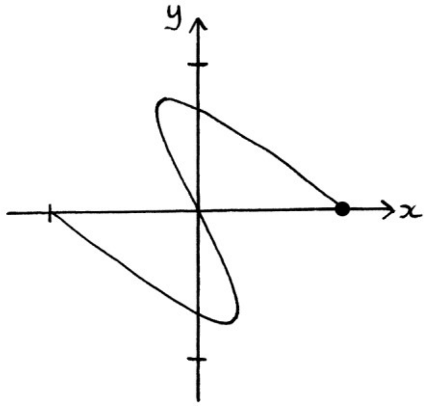
When the second wave's phase is 315 degrees:



When the second wave's phase is 337.5 degrees:



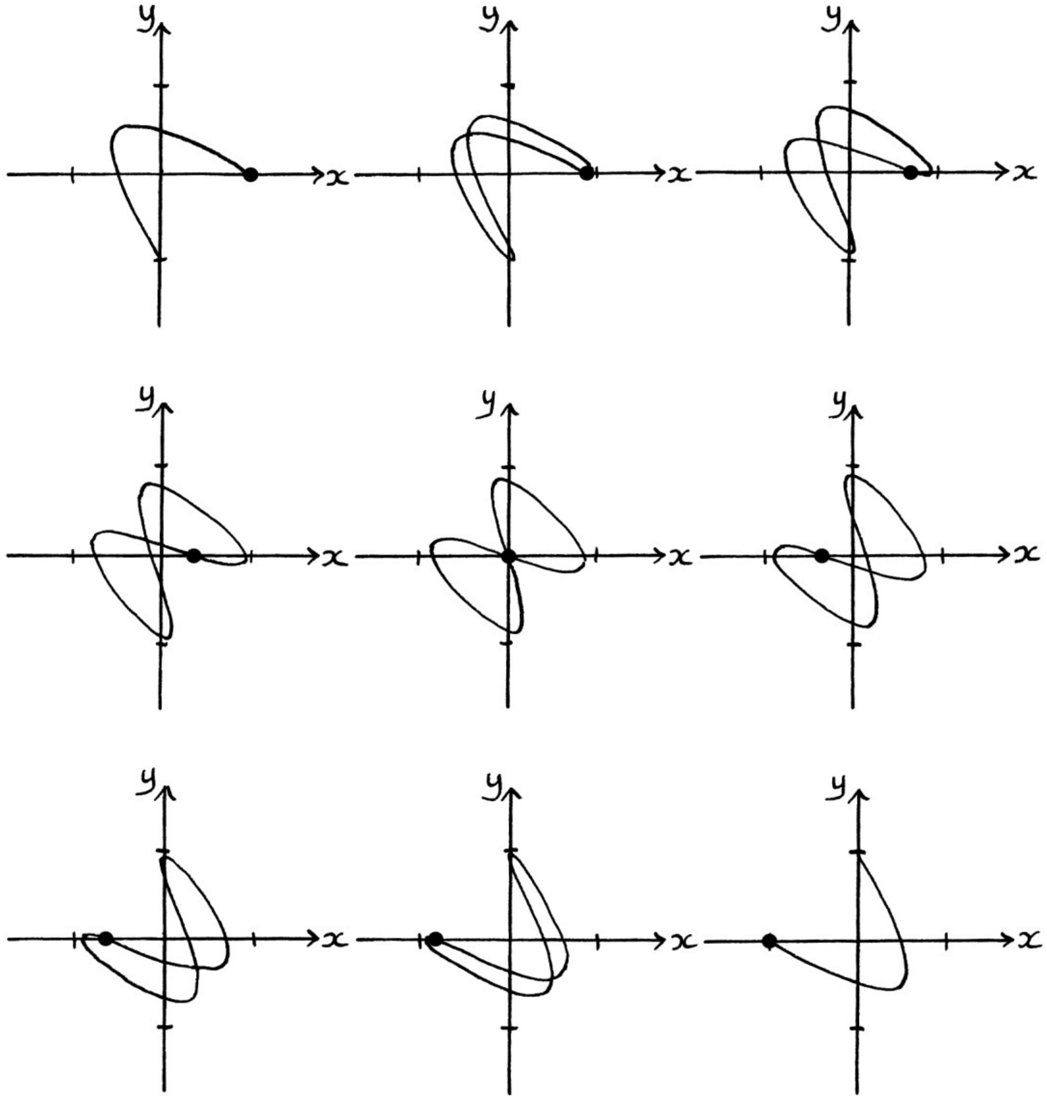
When the second wave's phase is 360 degrees:

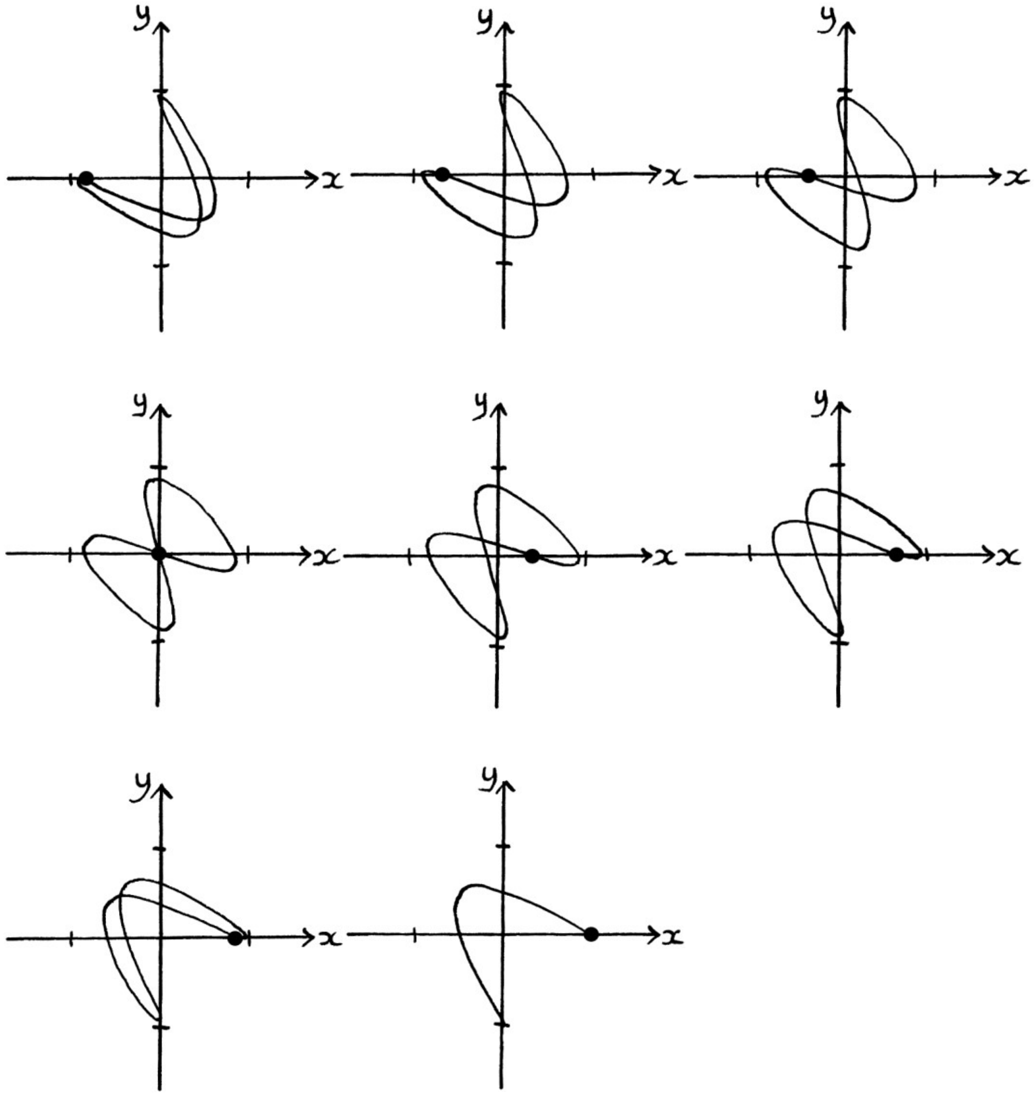


Ratio of 3 : 1

We will multiply the circle based on these waves:
“y = sin 360t” and “y = cos 360t)”
... by the circle based on these waves:
“y = sin ((360 * 3t) + ϕ)” and “y = cos ((360 * 3t) + ϕ)”
... where ϕ is a range of phases from 0 to 360 degrees, spaced at intervals of 22.5 degrees.

The results are as so, from left to right:




y
x
y
x
y
x
y
x
y
x
y
x
y
x
y
x

Circles by circles: mean levels

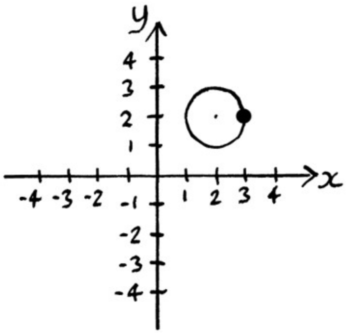
So far, the example calculations have used circles with zero mean levels – they have all been centred on the origin of the axes. Now we will experiment with non-zero mean levels.

Example 1

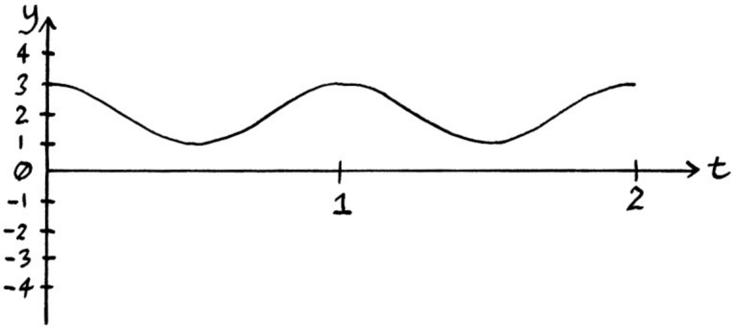
First, we will multiply a circle by itself. This time it will be one portrayed by the waves:
“y = 2 + cos 360t”
... and:
“y = 2 + sin 360t”

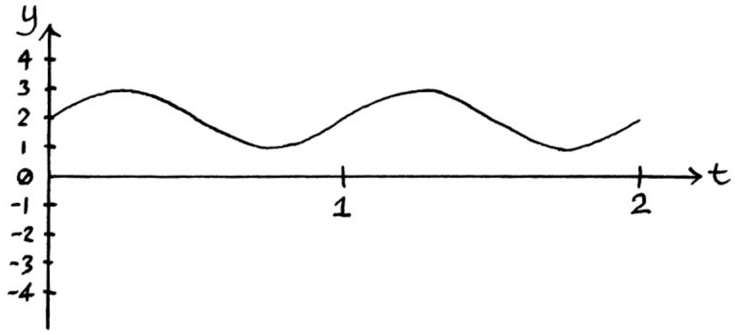
This circle has *two* mean levels – one of 2 units for the x-axis, and one of 2 units for the y-axis. The points on the circle’s edge are entirely positive, and therefore the resulting shape will be entirely in the top right quarter of the circle chart.

The original circle looks like this:

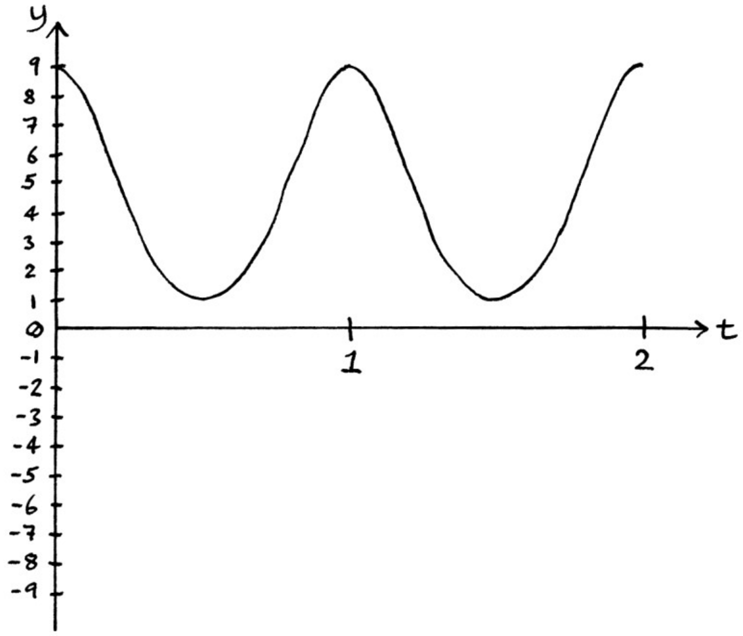


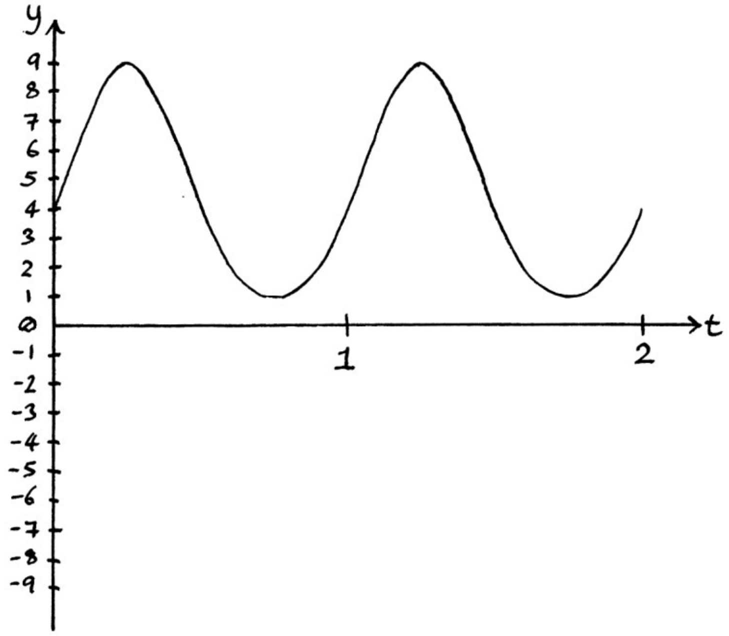
Its two derived waves look like this:



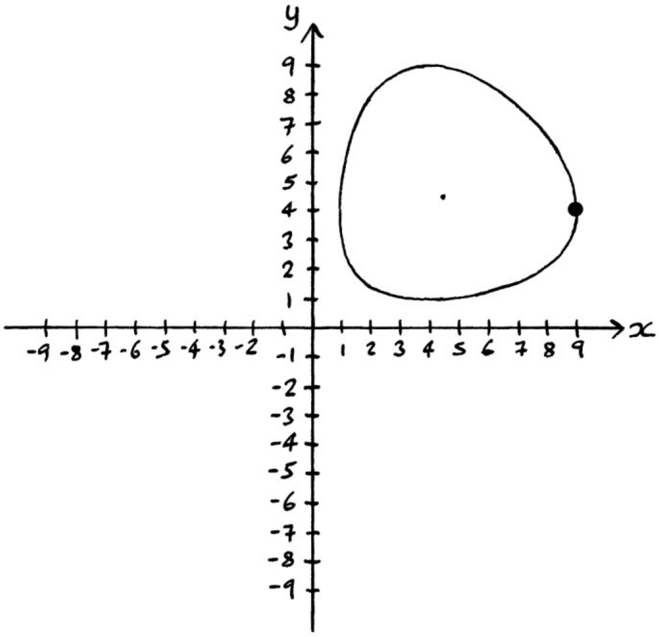


The graphs of each wave squared look like this:

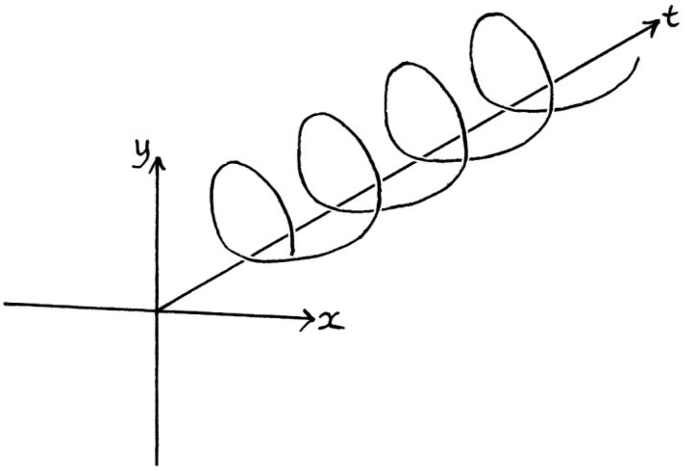




The resulting shape looks like this:



The resulting “helix” looks like this:



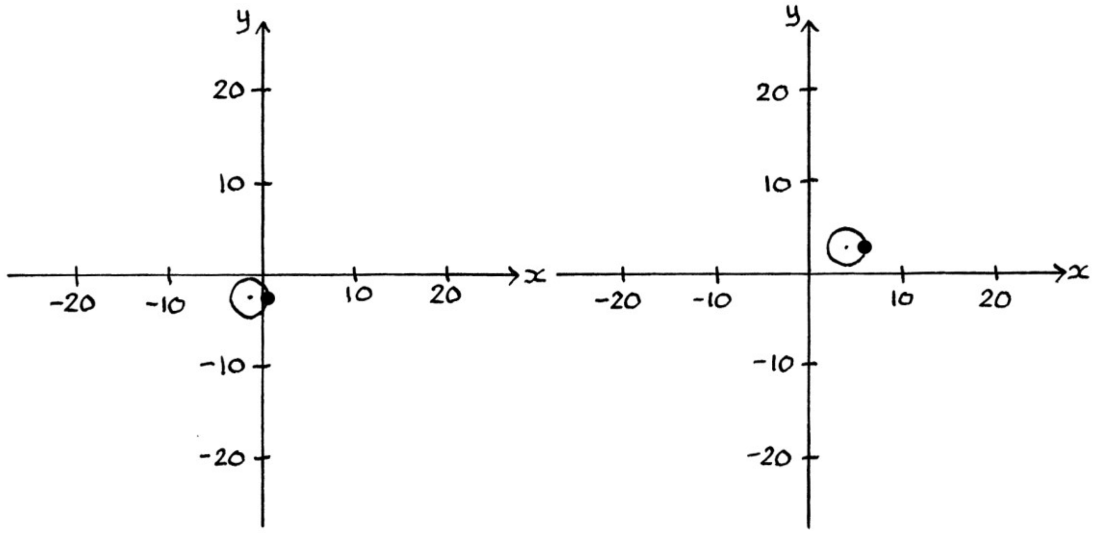
The resulting shape looks how it does because squaring coordinates that are further away from the origin results in points that are much further away from the origin. Squaring coordinates nearer the origin does not produce points that are as far away. Squaring the coordinates does not cause them to expand outwards in a uniform way.

Example 2

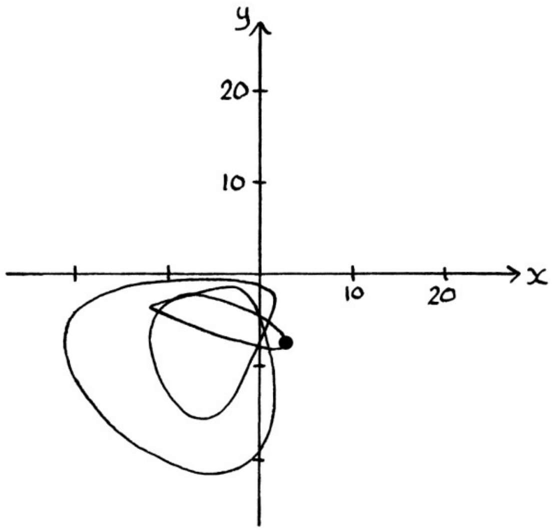
For this example, we will multiply two different circles with different attributes. The first circle will be that portrayed by the waves:
“y = −2.5 + 2 sin (360 * 3t)”
“y = −1.5 + 2 cos (360 * 3t)”

The second circle will be that portrayed by the waves:
“y = 3 + 2 sin (360 * 2t)
“y = 4 + 2 cos (360 * 2t)

The initial circles look like this:



The resulting shape looks like this:



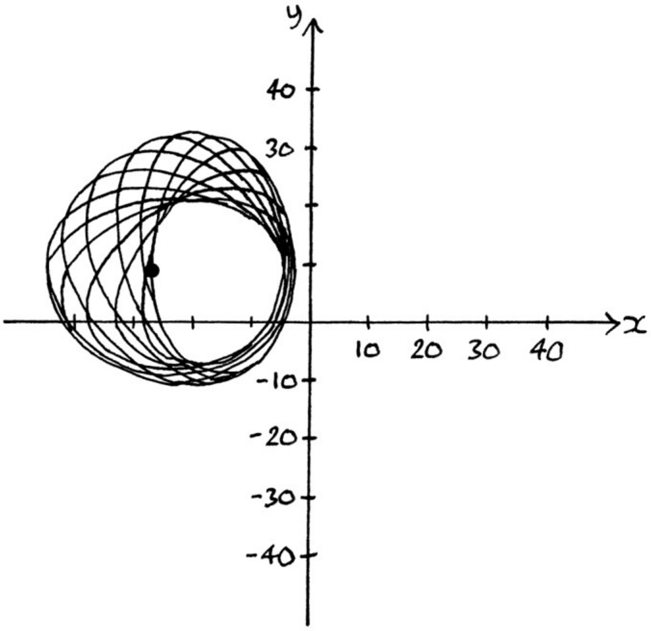
Circles by circles: various shapes

By altering the amplitude, frequency, phase and mean levels of two circles that are multiplied together, it is possible to create countless different shapes. One interesting thing about the shapes is that although many of them seem completely random, they are all a direct result of the attributes of the circles that created them. The exact same attributes will always create the same shape. It would be difficult to work out which circles were multiplied to create most of these shapes, but once we know the circles, it is extremely easy to make the shapes (if we can write a program that does such a thing, or if we can work out the maths on paper). This section might not be of much use, unless you need to draw decorative pictures.

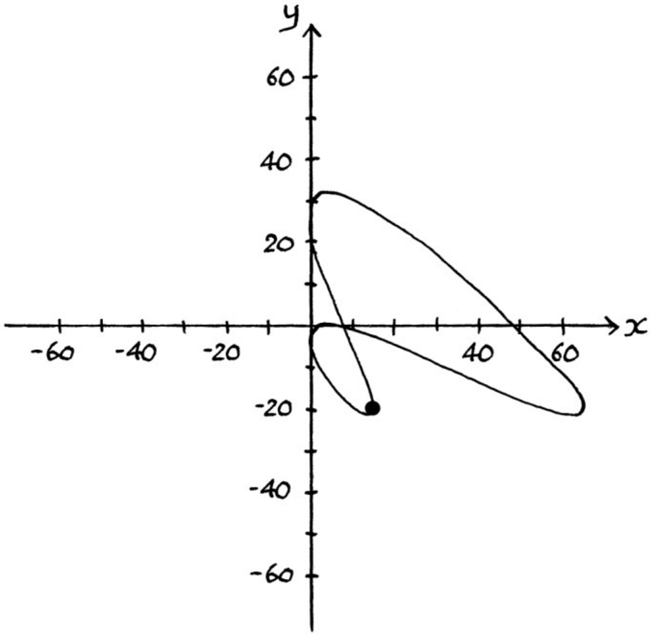
Amplitude, frequency and mean levels

Some example shapes from the multiplication of two circles that vary only in amplitude, frequency and mean level are as follows:

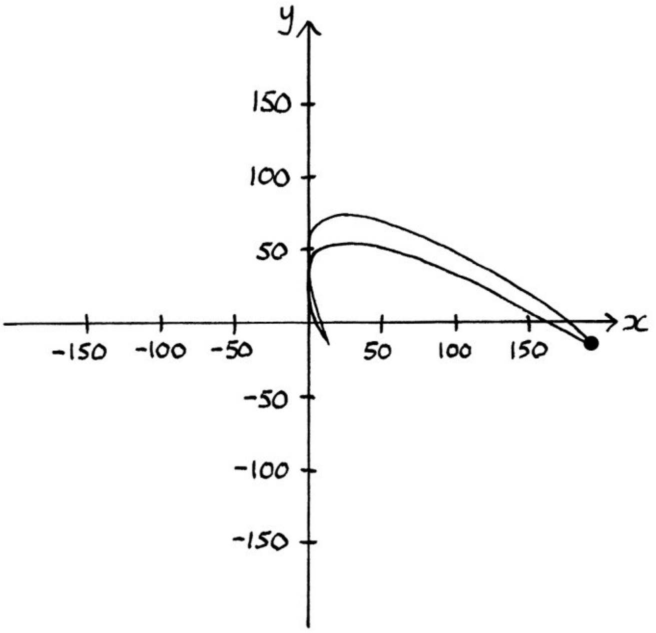
The following shape is the result of multiplying the circle made up of:
“y = 4.5 + 1 sin (360 * 3.5t)” and “y = −4 + 1 cos (360 * 3.5t)”
... by the circle made up of:
“y = 2 + 4 sin (360 * 5.5t)” and “y = 5 + 4 cos ((360 * 5.5t)”



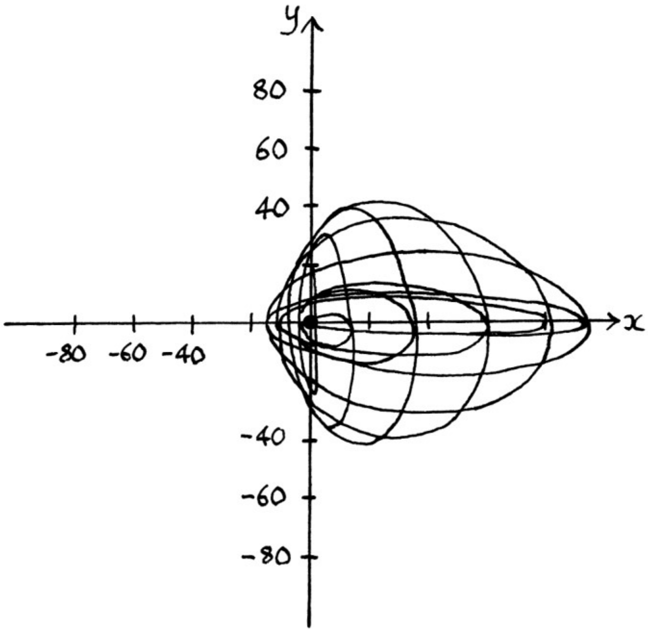
The following shape is the result of the circle made up of:
“y = 5 + 9 sin (360 * 6t) and “y = −4 + 9 cos (360 * 6t)”
... multiplied by the circle made up of:
“y = −4 + 4 sin (360 * 6t)” and “y = −1 + 4 cos (360 * 6t)”



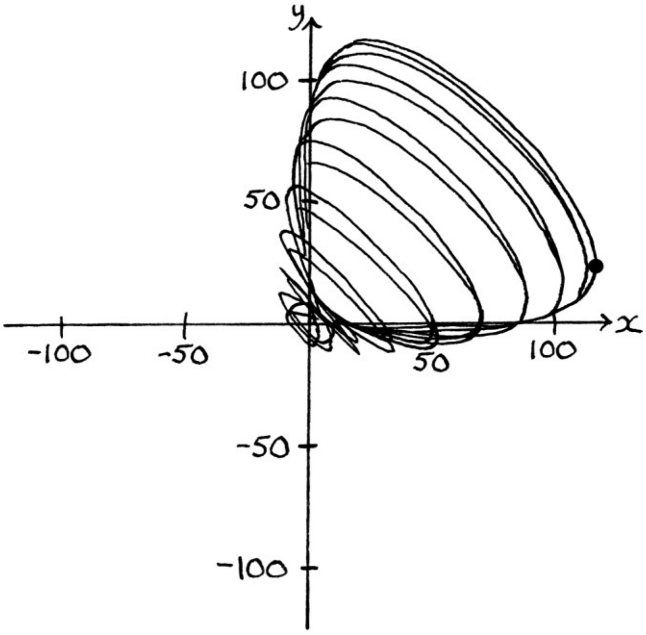
The following shape is the result of this circle:
“y = −3 + 8 sin (360 * 10t)” and “y = 5 + 8 cos (360 * 10t)”
... multiplied by this circle:
“y = 5 + 10 sin (360 * 10t)” and “y = 5 + 10 cos (360 * 10t)”



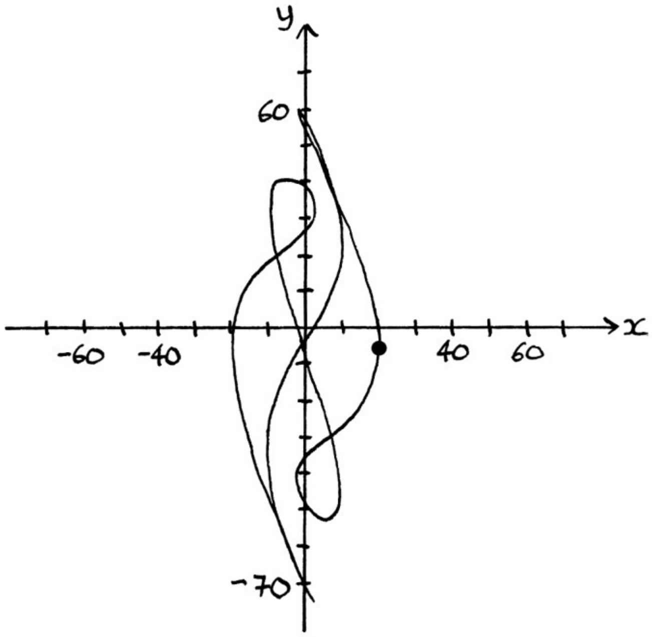
The following shape is the result of this circle:
“y = 7 sin (360 * 12t)” and “y = −5 + 7 cos (360 * 12t)”
... multiplied by this circle:
“y = −2 + 4 sin (360t)” and “y = −4 + 4 cos (360t)”



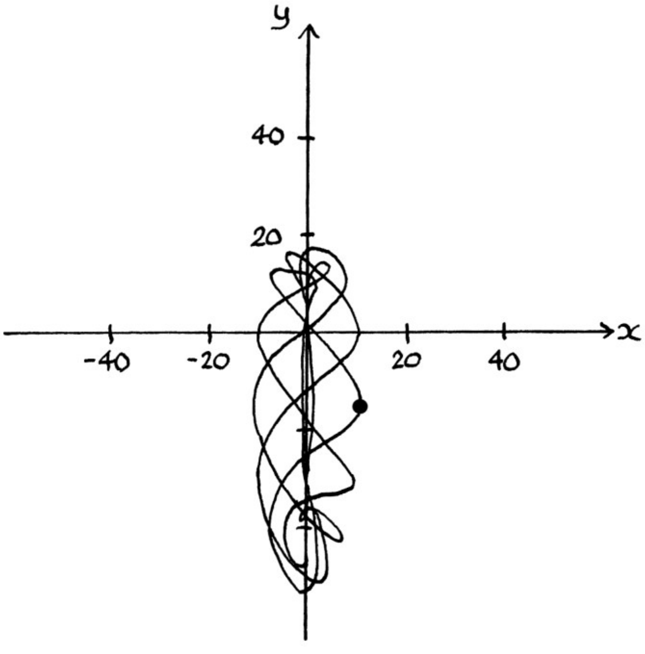
The following shape is the result of this circle:
“y = 6 + 7 sin (360 * 14t)” and “y = 6 + 7 cos (360 * 14t)”
... multiplied by this circle:
“y = 4 + 5 sin (360 * 13t)” and “y = 4 + 5 cos (360 * 13t)”



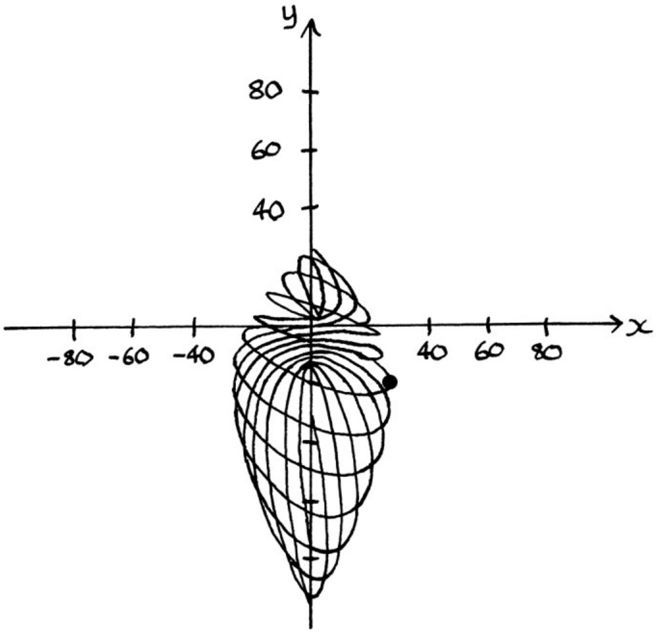
The following shape is the result of this circle:
“y = 6 + 2 sin (360 * 15t)” and “y = 2 cos (360 * 15t)”
... multiplied by this circle:
“y = −1 + 9 sin (360 * 10t)” and “y = 1 + 9 cos (360 * 10t)”



The following shape is the result of this circle:
“y = 3 + 6 sin (360 * 4t)” and “y = 5 + 6 cos (360 * 4t)”
... multiplied by this circle:
“y = −5 + sin (360 * 11t) and “y = 1 cos (360 * 11t)”



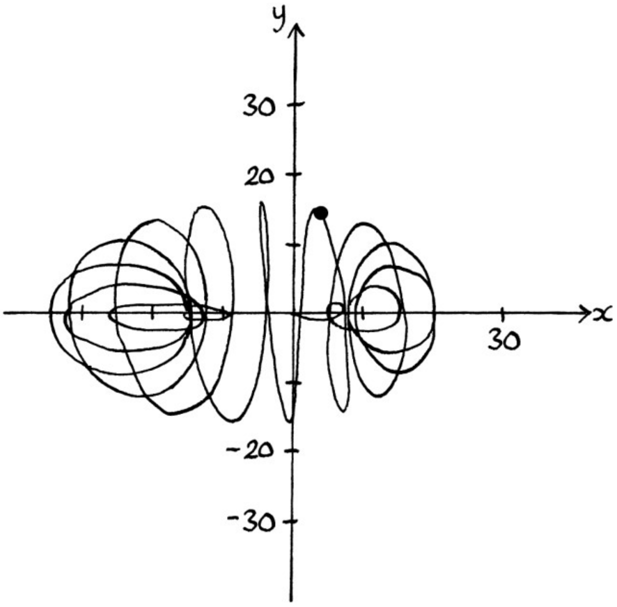
The following shape is the result of this circle:
"y = 4 + 3 sin (360 * 13t)" and "y = 3 cos (360 * 13t)"
... multiplied by this circle:
"y = −5 + 9 sin (360 * 2t)" and "y = 9 cos (360 * 2t)"



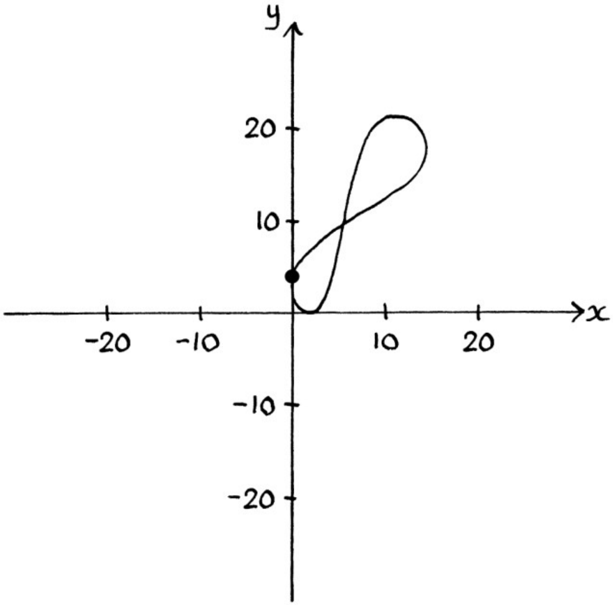
Amplitude, frequency, mean levels and phase

In this section, we will look at circles that vary in amplitude, frequency, mean level and phase. Again, there might not be much use in knowing any of this.

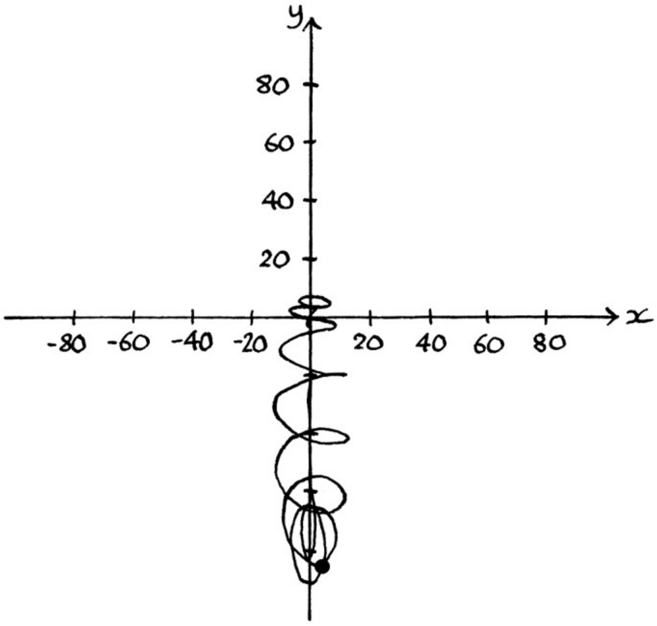
The following shape is the result of this circle:
“y = 5 + 2 cos ((360 * 15t) + 285)” and “y = 2 sin ((360 * 15t) + 285)”
... multiplied by this circle:
“y = −1 + 4 cos (360t + 296)” and “y = −4 + 4 sin (360t + 296)”



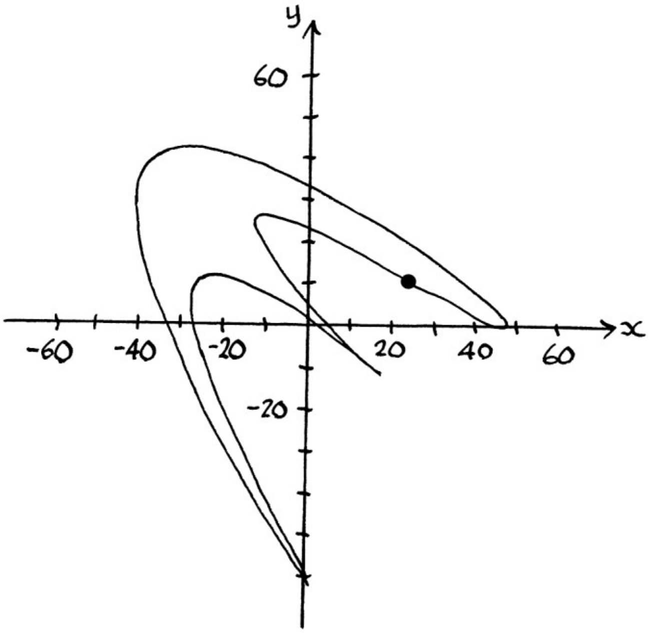
The following shape is this circle:
“y = 4 + 2 cos ((360 * 9t) + 41)” and “y = −2 + 2 sin ((360 * 9t) + 41)”
... multiplied by this circle:
“y = 2 + 2 cos ((360 * 9t) + 170)” and “y = −6 + 2 sin ((360 * 9t) + 170)”



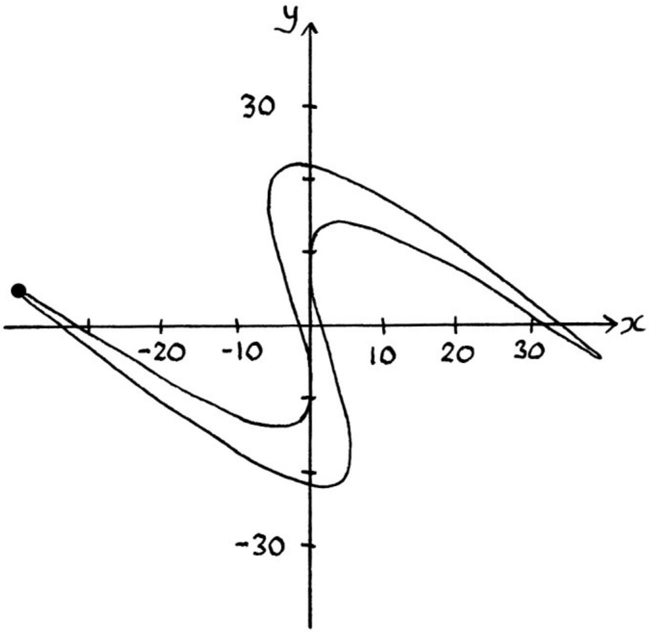
The following shape is this circle:
"y = 6 + 7 cos (360t + 265)" and "y = −6 + 7 sin (360t + 265)"
... multiplied by this circle:
"y = cos ((360 * 12t) + 41)" and "y = 6 + sin ((360 * 12t) + 41)"



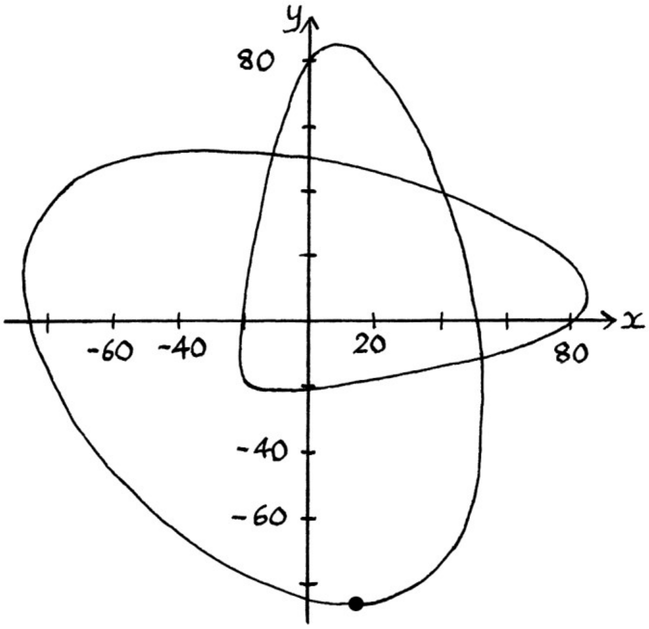
The following shape is this circle:
"y = 2 + 8 cos ((360 * 2t) + 316)" and "y = 3 + 8 sin ((360 * 2t) + 316)"
... multiplied by this circle:
"y = −1 + 6 cos ((360 * 4t) + 313)" and "y = 6 sin ((360 * 4t) + 313)"



The following shape is this circle:
“y = 3 cos ((360 * 2t) + 197)” and “y = 3 sin ((360 * 2t) + 197)”
... multiplied by this circle:
“y = 5 + 9 cos ((360 * 4t) + 349)” and “y = −4 + 9 sin ((360 * 4t) + 349)”



The following shape is this circle:
“y = −1 + 10 cos ((360 * 14t) + 96)” and "y = −1 + 10 sin ((360 * 14t) + 96)”
... multiplied by this circle:
“y = −6 + 4 cos ((360 * 7t) + 250)” and “y = −6 + 4 sin ((360 * 7t) + 250)”



Sum of two circles multiplied by a third

If a sum of two circles is multiplied by a third, more complicated shapes can be made.

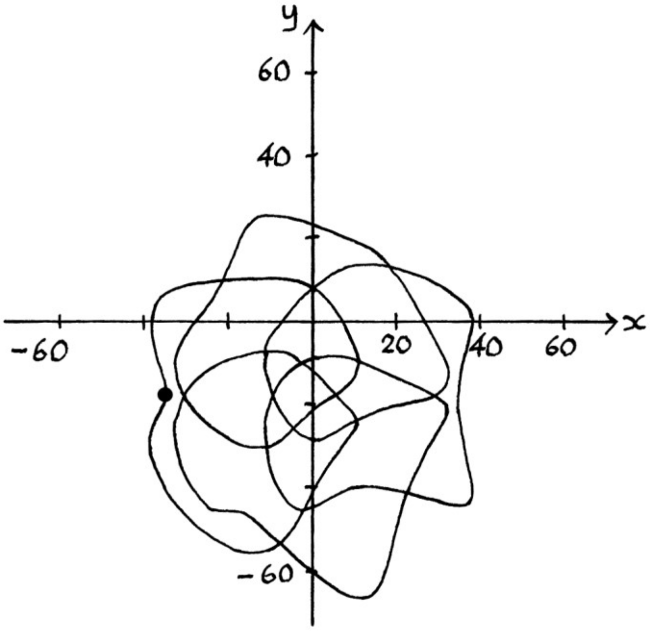
For example, if we add the circles based on these waves:

“y = −2 + 4 cos (360 * 5t)” and “y = 4 + 4 sin (360 * 5t)”
“y = 2 + 3 cos 360t” and “y = −1 + 3 sin 360t”

... and then multiply the result by the circle based on these waves:

“y = −6 + cos (360 * 14t)” and “y = −6 + sin (360 * 14t)”

... we end up with this shape:



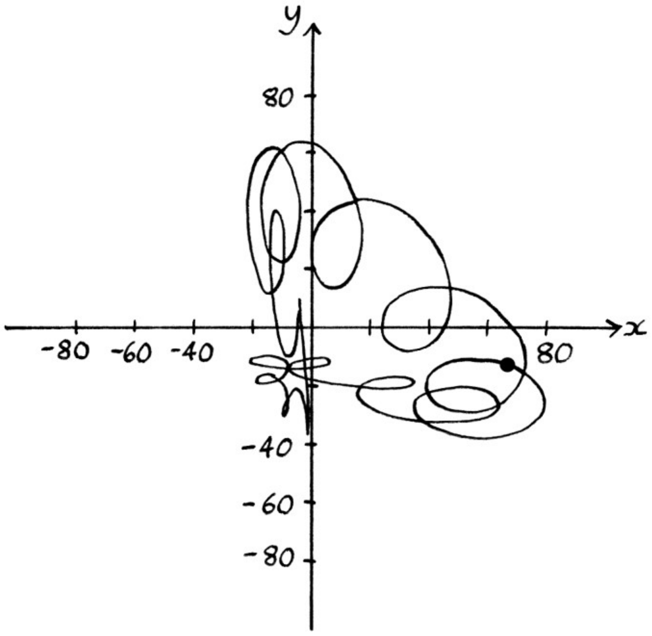
If we add the circles based on these waves:

“y = −3 + 8 cos 360t” and “y = 2 + 8 sin 360t”
“y = 1 + 2 cos ((360 * 14t) + 290)” and “y = 2 + 2 sin ((360 * 14t) + 290)”

... and multiply the result by the circle based on these waves:

“y = 5 + 5 cos 360t” and “y = −6 + 5 sin 360t”

... we end up with this shape:



Sum of three circles multiplied by a fourth

If we add three circles and multiply the result by a fourth circle, we will have more variations in the resulting shapes.

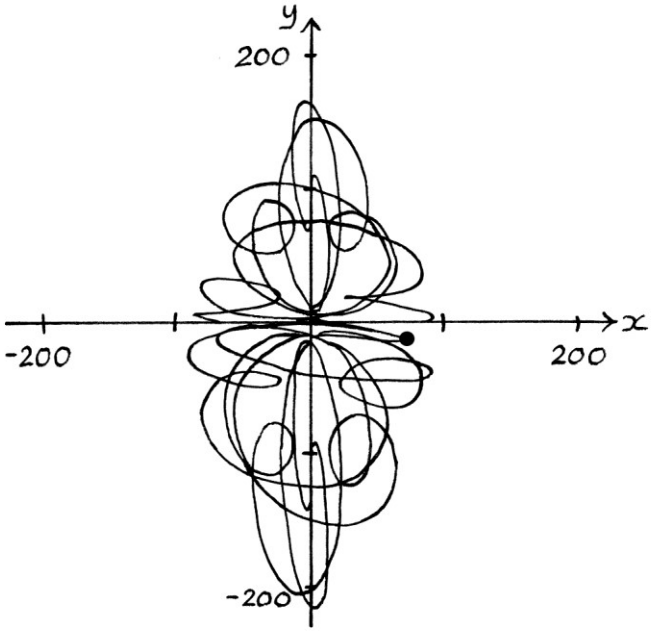
The sum of the circles created from the following pairs of waves:

"y = 6 cos (360 * 10t)" and "y = −4 + 6 sin (360 * 10t)"
"y = −5 + 2 cos (360 * 20t)" and "y = −6 + 2 sin (360 * 20t)"
"y = 1 + 5 cos (360 * 20t)" and "y = −3 + 5 sin (360 * 20t)"

... multiplied by the circle created from this pair of waves:

"y = 8 cos 360t" and "y = 1 + 8 sin 360t"

... results in this shape:



Multiplying a circle by a wave

Multiplying a circle by a wave involves multiplying the coordinates of a circle by the corresponding (in time) values of a wave. This is identical to multiplying the corresponding values (in time) of the two waves that make up a circle by a third wave.

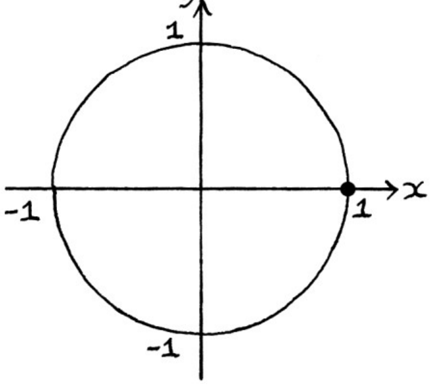
As with multiplying a circle by a circle, I am mainly explaining multiplying a circle by a wave for the sake of completeness. At this level of maths, it is mostly useful for making interesting pictures.

Example 1

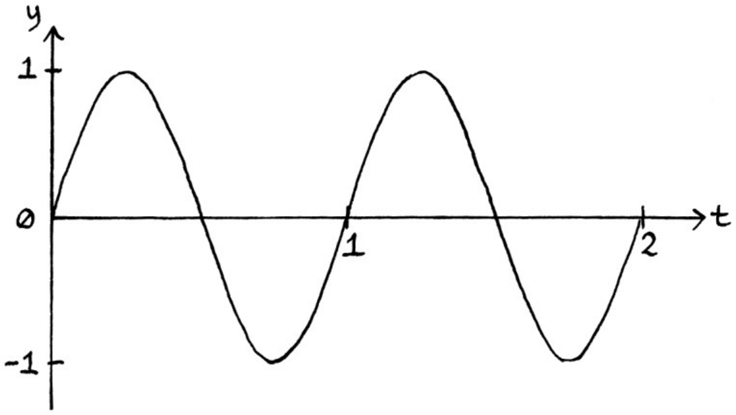
As an example, we will multiply a circle by its derived Sine wave. We will multiply every coordinate on the circle by the corresponding value from the Sine wave derived from that circle. Doing this is really taking the coordinates of the object at any moment in time as (x, y), and then multiplying both coordinates by "y". Therefore, we end up with (xy, y^2) for every moment in time. Another way of thinking about this is that we will be multiplying every point on the circle's edge by its y-axis value.

We will take the circle created from the waves "y = 3 sin 360t" and "y = 3 cos 360t", and multiply it by the "y = 3 sin 360t" wave.

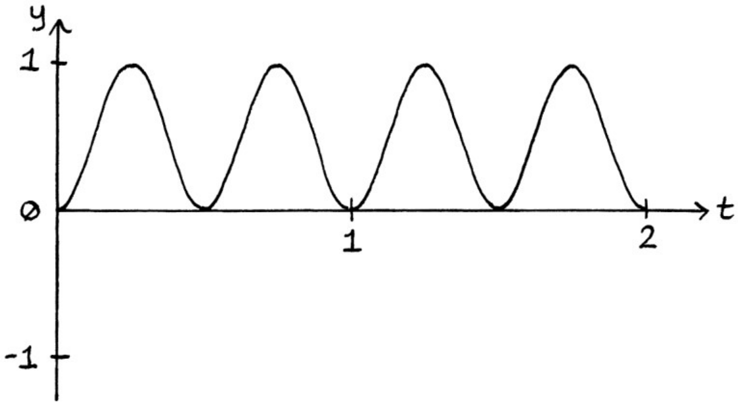
The initial circle looks like this:



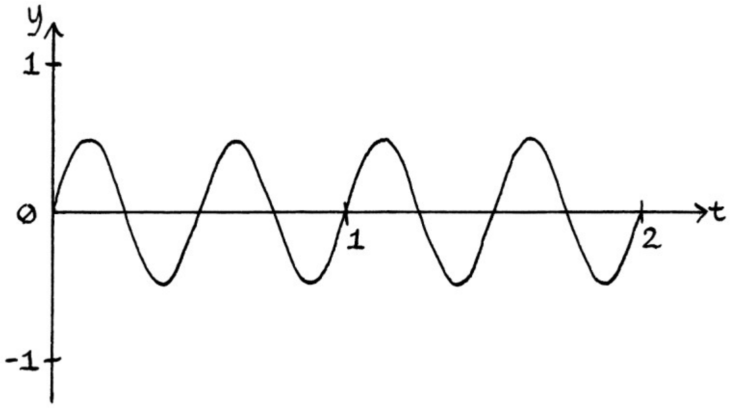
Its Sine wave looks like this:



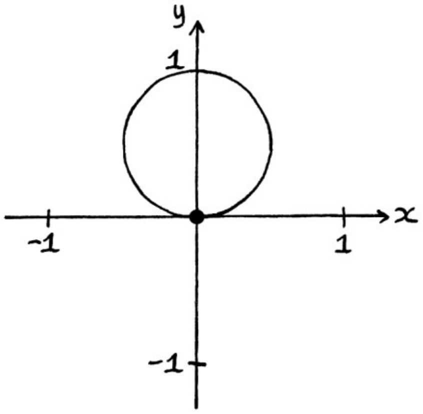
The result of multiplying the Sine wave from the circle by itself is a wave with the formula:
"y = 4.5 + 4.5 sin ((360 * 2t) + 270)":



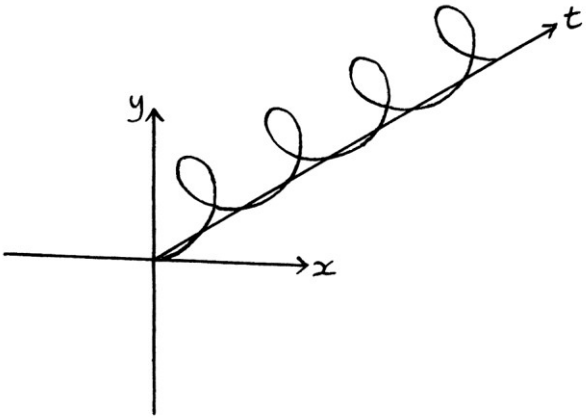
The result of multiplying the Cosine wave from the circle by the Sine wave is a wave with this formula:
"y = 4.5 sin (360 * 2t)"



The resulting shape made up from these two signals looks like this: [the start point is at the coordinates (0, 0)]



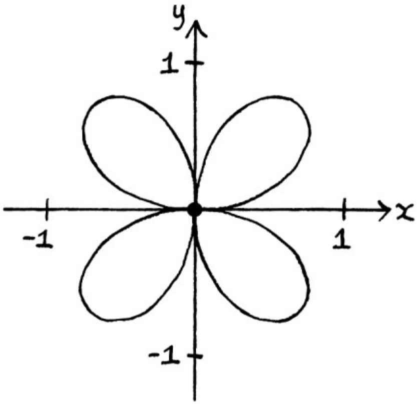
It is a perfect circle. The resulting helix looks like this:



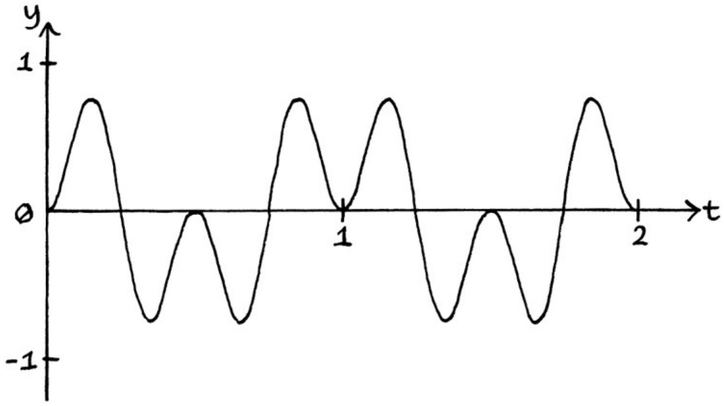
Example 2

As another example, we will multiply the circle made up of these two waves:
“y = sin 360t” and “y = cos 360t”
... by this wave:
“y = sin (360 * 2t)”.

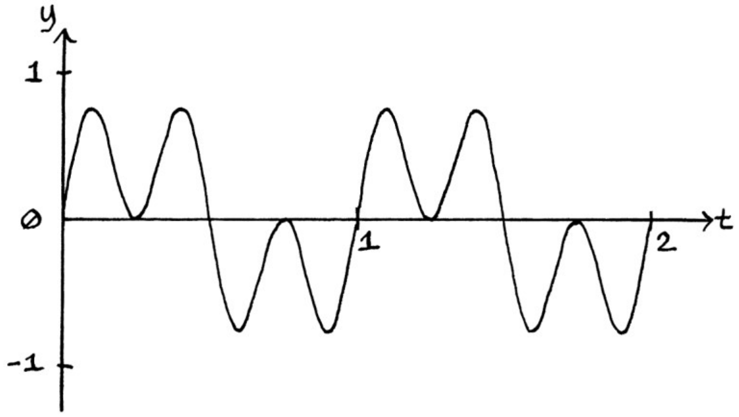
We end up with this shape, with the start point at (0, 0):



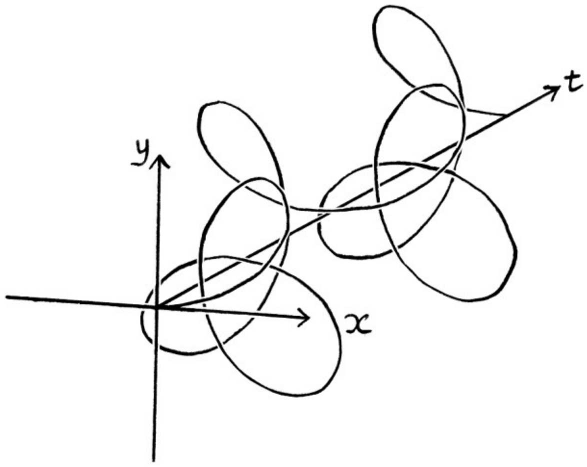
Its vertically derived signal looks like this:



Its horizontally derived signal looks like this:

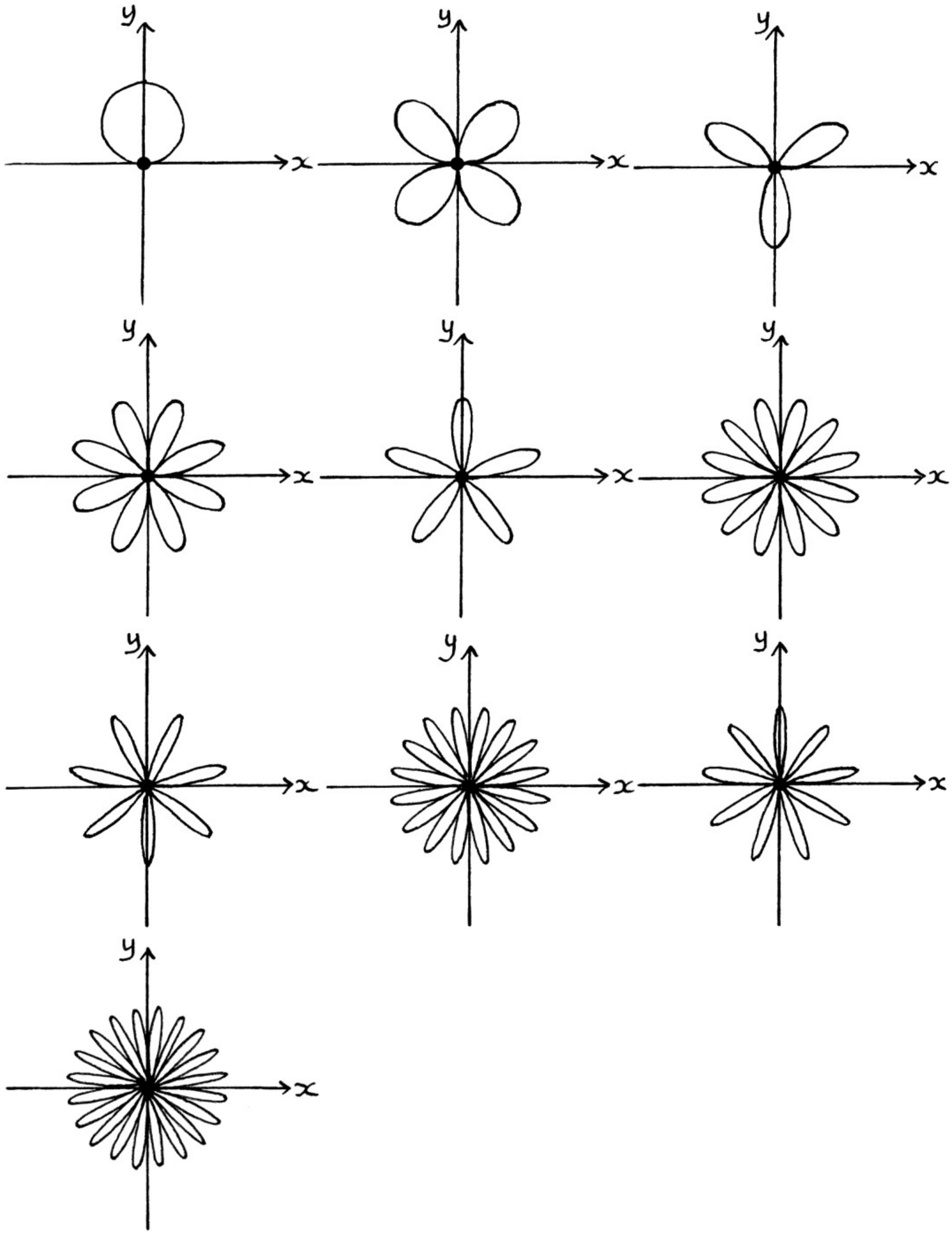


The helix looks like this:

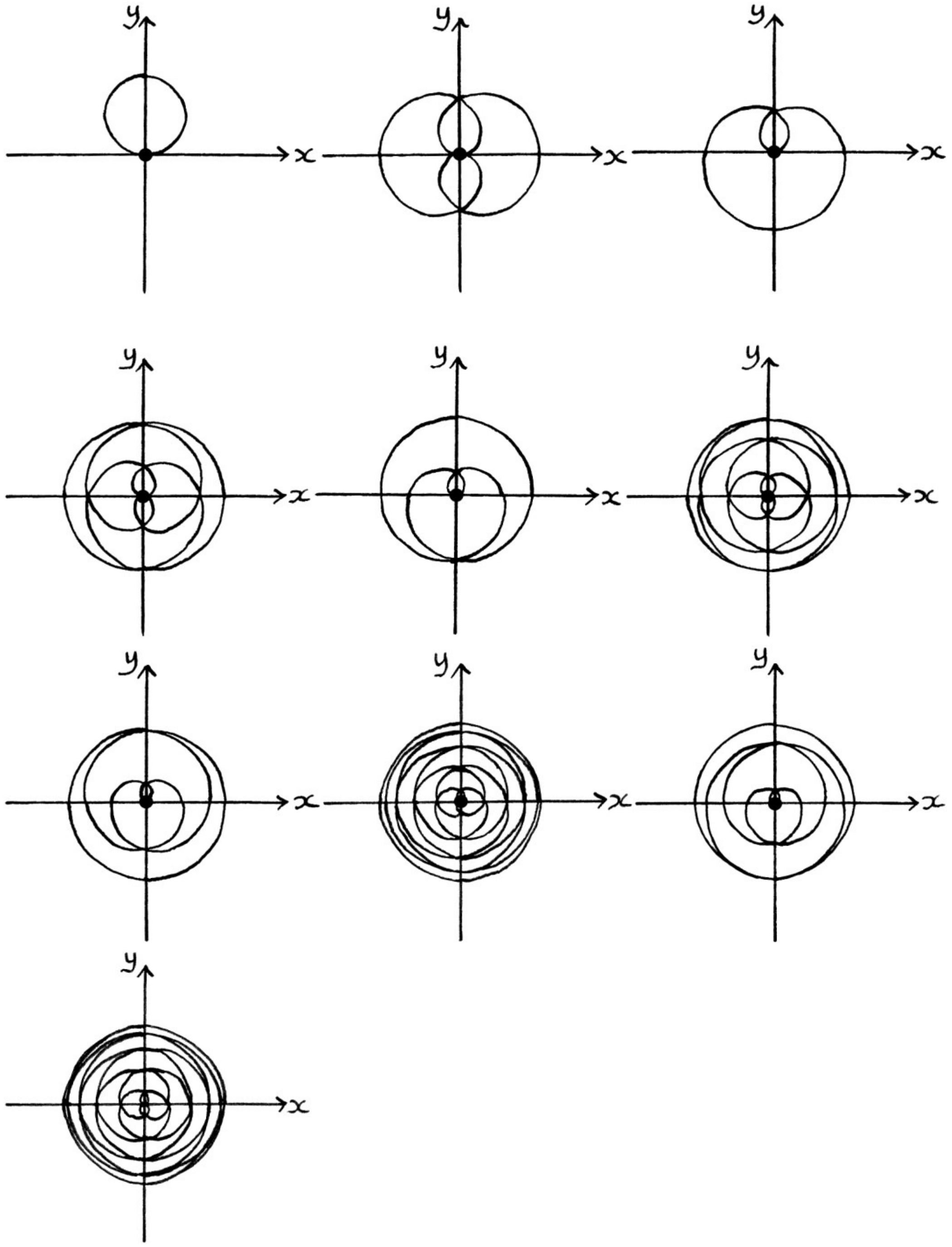


Circles by waves: frequency ratios

There is a pattern for shapes with particular frequency ratios. If we change the frequency of the *wave*, while keeping the frequency of the *circle* the same, the ratios of the frequencies of the wave to the circle from 1 : 1 up to 10 : 1 look like so:



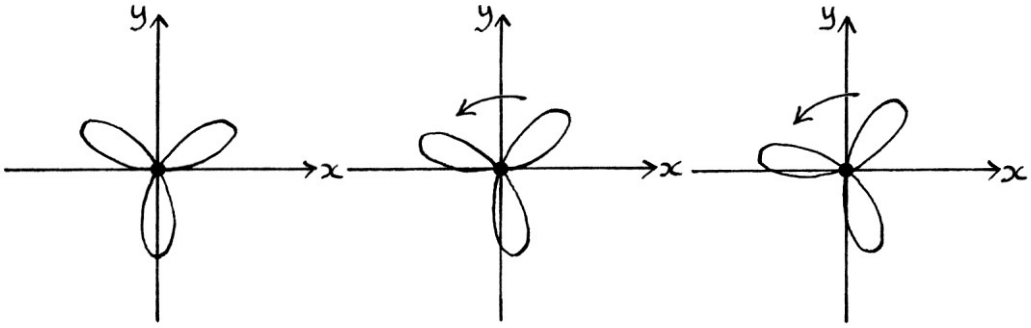
If we change the frequency of the *circle* while keeping the frequency of the *wave* the same, the shapes for the ratios of the frequencies of the circles to the waves from 1 : 1 to 10 : 1 are as follows:



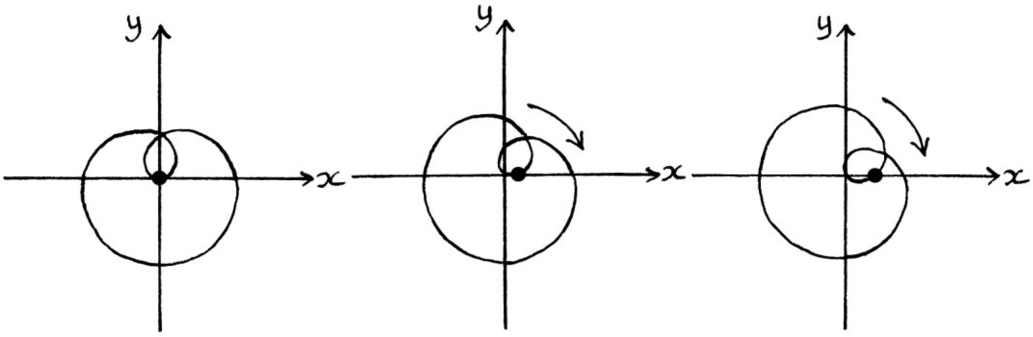
Circles by waves: varying phases

Changing the phase of either the wave or the circle will result in the entire shape becoming rotated.

For example, if we start with a frequency ratio of 3 : 1 for the wave to the circle, and increase the phase of the circle in 10 degree steps, the whole shape rotates anticlockwise:



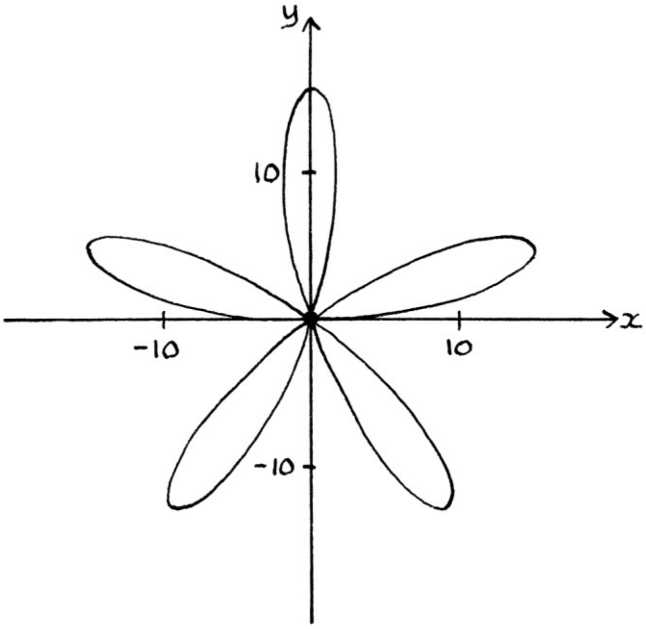
If we start with a frequency ratio of 3 : 1 for the *circle* to the *wave*, and increase the phase of the *wave* in 10 degree steps, the whole shape rotates clockwise:



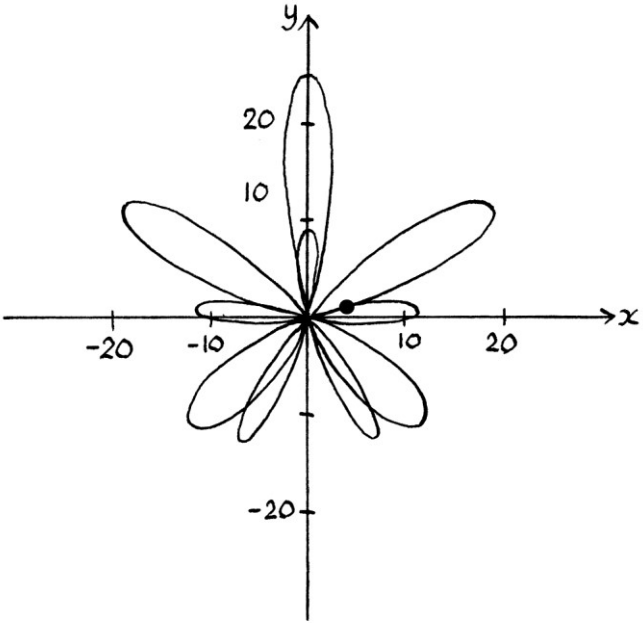
Circles by waves: varying mean levels

To see the effect of varying the mean levels, we will start with the circle based on these waves:
“4 sin 360t” and “4 cos 360t”
... and we will multiply it by this wave:
“4 sin (360 * 5t)”

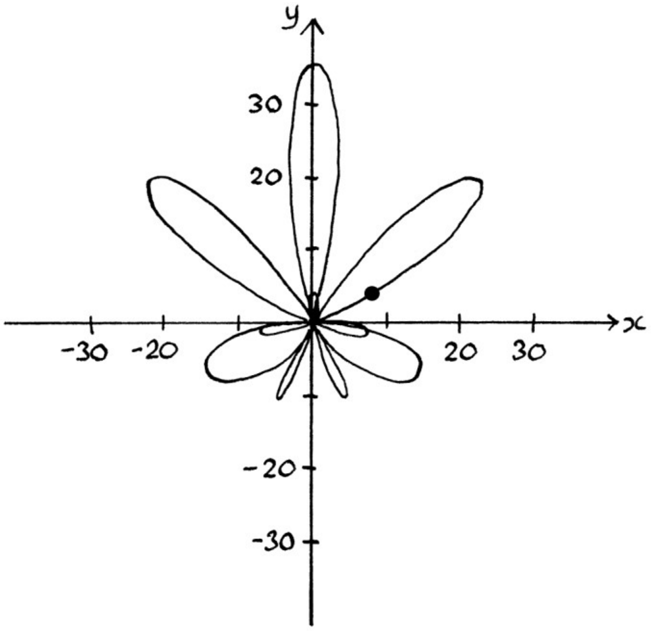
The shape looks like this:



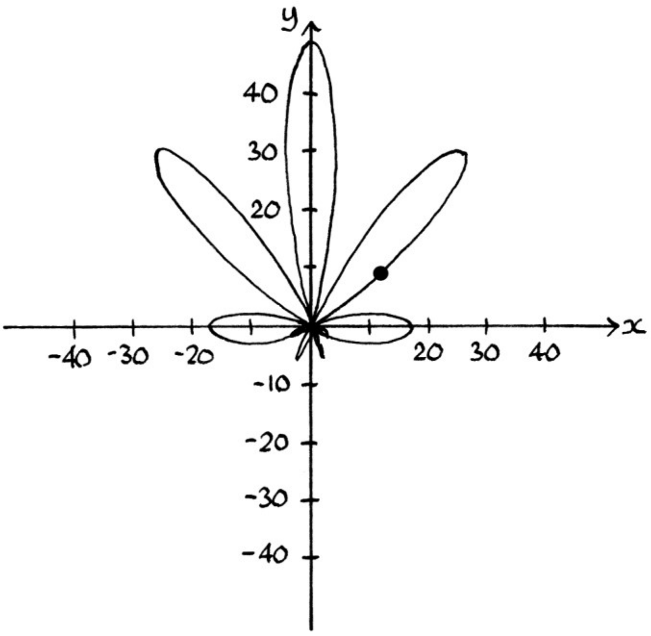
If we increase the mean level of both the circle and the wave by 1 y-axis unit, we end up with this shape:



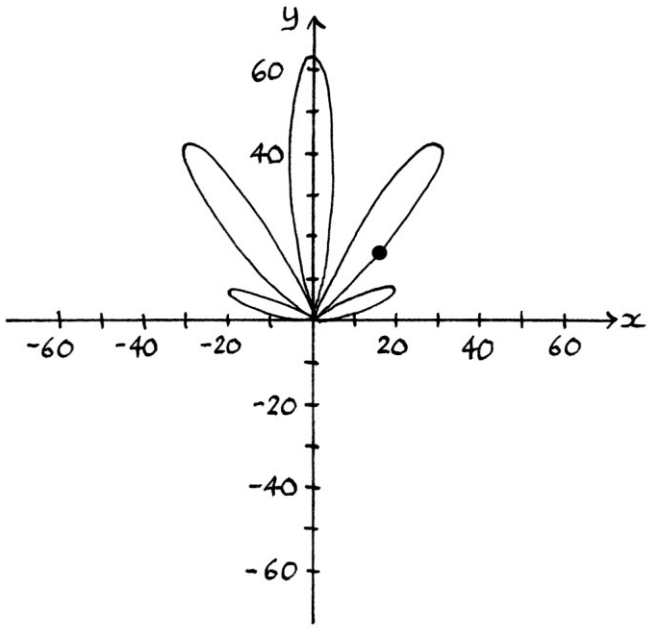
If we give the circle and the wave mean levels of 2 y-axis units, the shape looks like this:



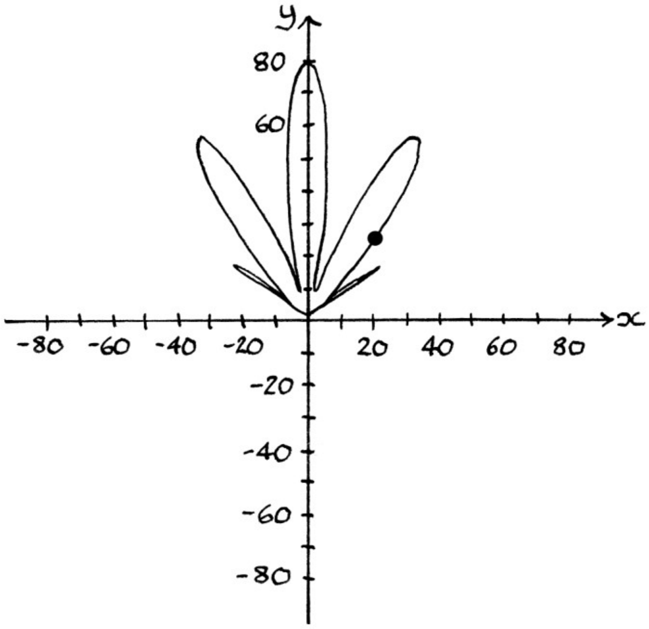
If we give the circle and the wave mean levels of 3 y-axis units, the shape looks like this:



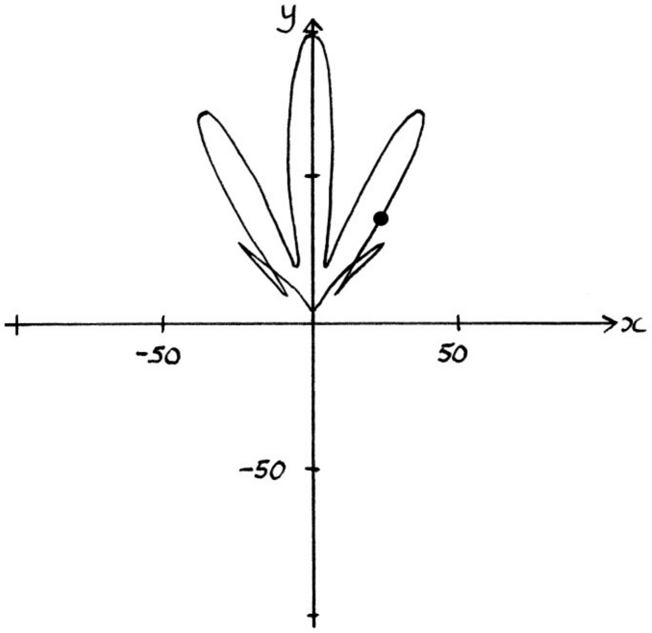
If we give the circle and the wave mean levels of 4 y-axis units, the shape looks like this:



If we give the circle and the wave mean levels of 5 y-axis units, the shape looks like this:



If we give the circle and the wave mean levels of 6 y-axis units, the shape looks like this:

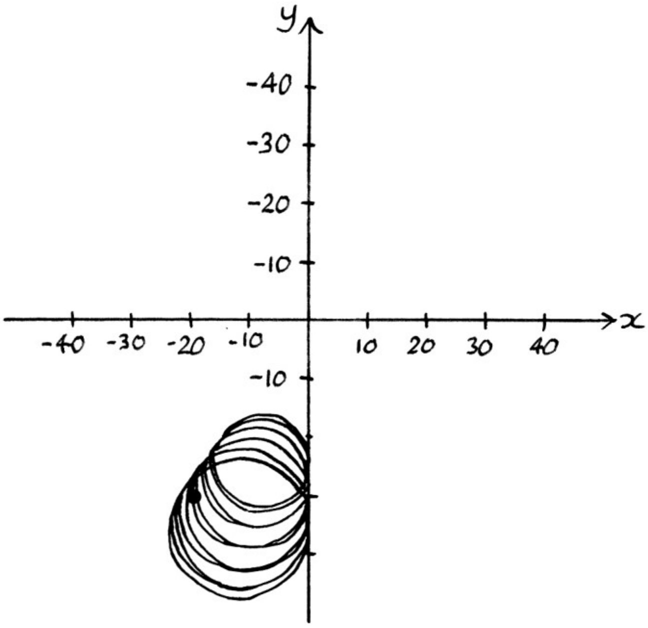


As the y-axis mean levels increase, the lobes of the resulting shape point further into the top half of the circle chart.

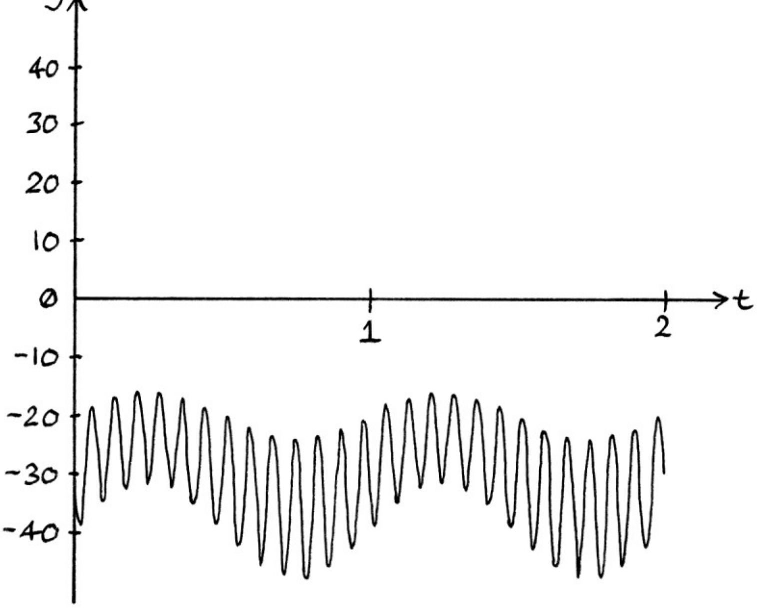
Circle by waves: various shapes

There can be interesting shapes resulting from multiplying circles by waves.

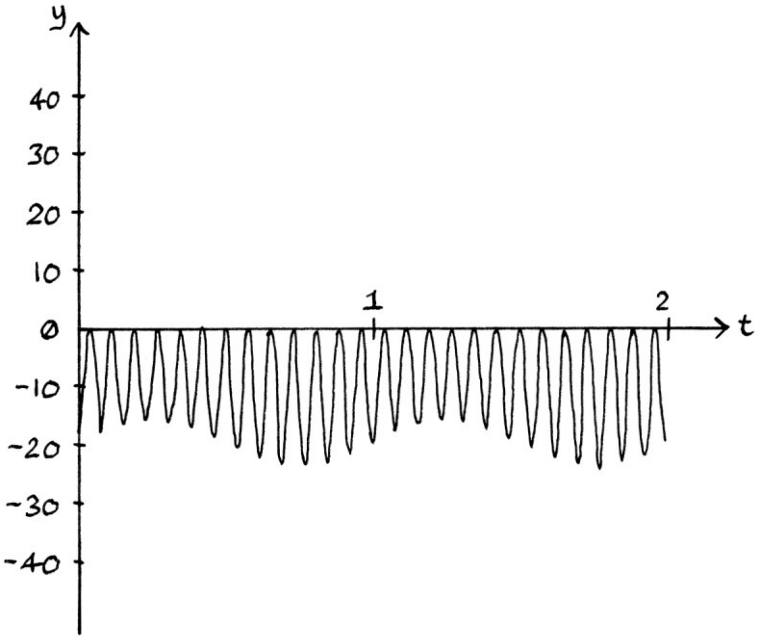
If we multiply the circle created from these waves:
“y = 2 + 2 cos (360 * 13t)” and “y = 6 + 2 sin (360 * 13t)”
... by this wave:
“y = −5 + sin 360t”
... we end up with this shape:



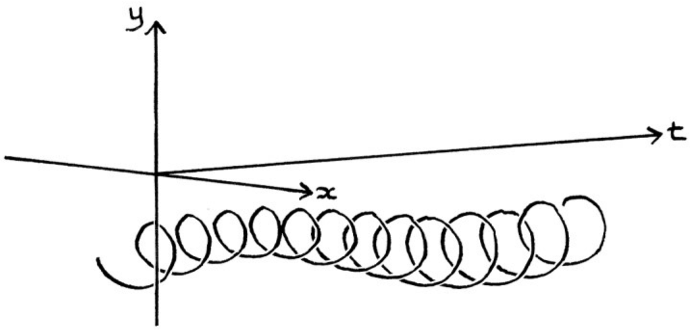
The vertically derived signal looks like this:



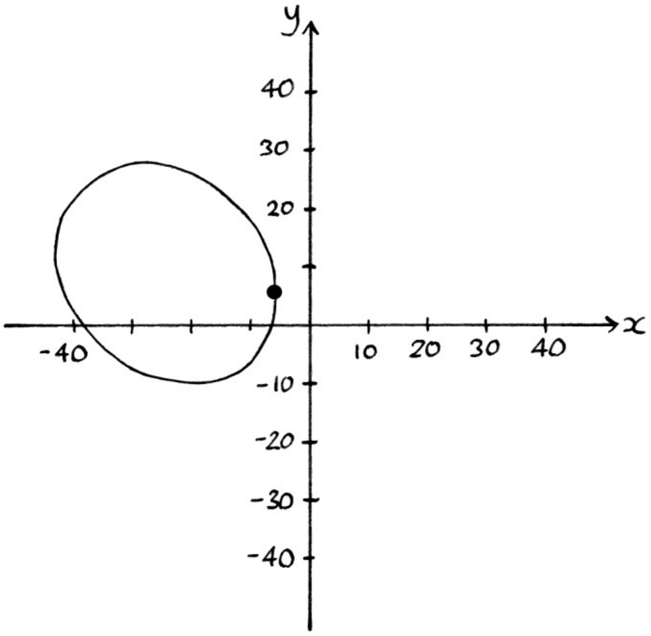
The horizontally derived signal looks like this:



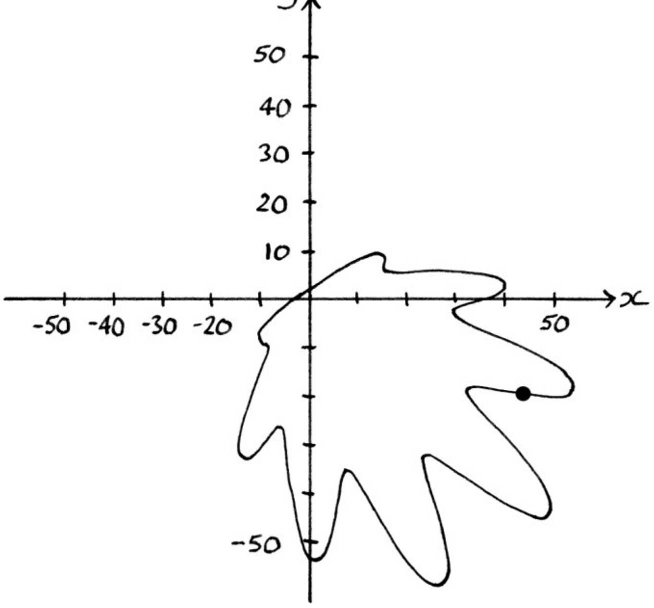
The helix looks like this, for one cycle:



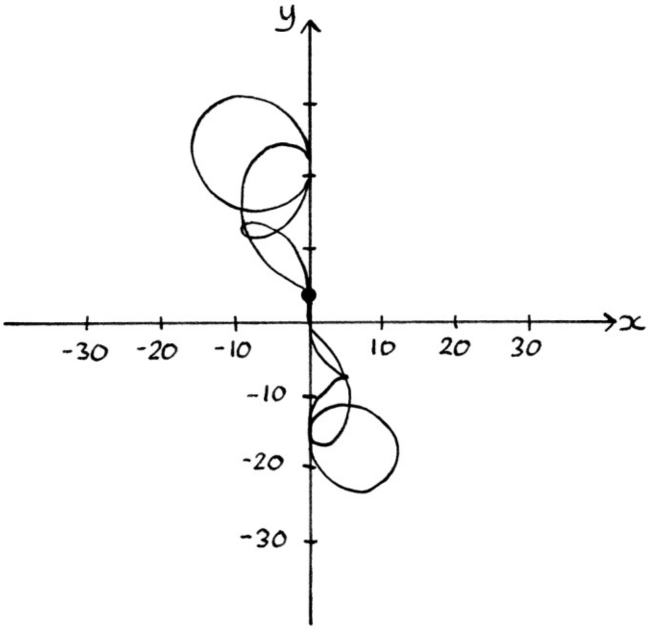
If we multiply the circle created from these waves:
“y = −4 + 3 cos (360 * 5t)” and “y = 1 + 3 sin (360 * 5t)”
... by this wave:
“y = 6 + sin (360 * 5t)”
... we end up with this shape, which is an ellipse:



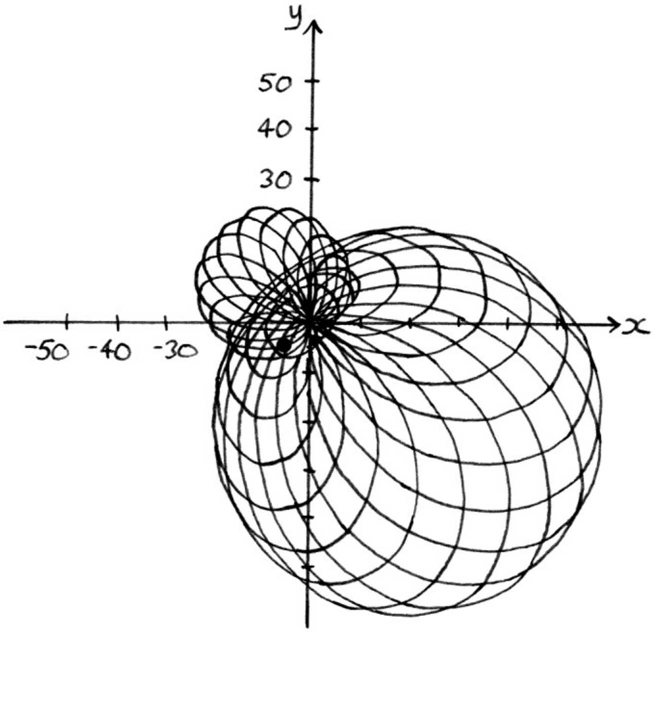
If we multiply the circle created from these waves:
“y = 4 + 7 cos 360t” and “y = −5 + 7 sin 360t”
... by this wave:
“y = 4 + sin (360 * 8t)”
... we end up with this shape:



If we multiply the circle created from these waves:
“y = −1 + cos (360 * 18t)” and “y = 3 + sin (360 * 18t)”
... by this wave:
“y = 1 + 7 sin (360 * 3t)”
... we produce this shape:



If we multiply the circle created from these waves:
“y = −2 + 4 cos (360 * 17t)” and “y = 2 + 4 sin (360 * 17t)”
... by this wave:
“y = −3 + 7 sin (360 * 15t)”
... we produce this shape:



Conclusion

There might not be much obvious use in knowing about multiplying circles by circles or circles by waves, unless we want interesting patterns, but this chapter has shown what happens. You might see patterns from multiplying circles by circles, or circles by waves, outside the world of waves, in which case, you can know how they were created.

There is a great deal more that we could discover about multiplying circles, but to keep this book from becoming even longer than it is, we will move on to other subjects.

Chapter 18: Fourier series analysis

In this chapter, we will learn how to analyse a periodic signal to see which waves were added to create it. The process is very important in the world of waves. The method is reasonably simple, but it is usually explained as if it were extremely complicated. Most explanations tend to concentrate on describing a mathematical summary of the meaning of the process, instead of actually explaining how to do it.

I will call the waves that were added together to make up a periodic signal, "the constituent waves". Note that even if a periodic signal were not literally created by adding waves, most of the time it will be possible to create it, or an approximation, by adding waves together. In this sense, the term "constituent waves" can refer to the waves that were literally added together to make up a signal, or it can refer to the waves that *could have been* added together to make up a signal.

The process involves making a list of waves with frequencies that could possibly be in the signal, and then multiplying each of those waves against the signal in turn to see if the multiplication produces a result with a non-zero mean level. A non-zero mean level in the result indicates that a wave of that frequency was in the original signal. It also gives us clues as to the other characteristics of the constituent wave for that frequency. There is slightly more to the process than this because if the phase of one of the tested waves is 90 degrees either side of an existing phase in the signal, it will not affect the mean level. Therefore, we have to take this into account.

The beauty of the process is that not only does it find the constituent waves from periodic signals that were literally created by adding waves, it also finds the constituent waves from most other periodic signals too. For example, we can find the waves that, when added together, are equivalent to the result of several multiplications, or equivalent to a signal that has been drawn on a piece of paper, or (nearly) equivalent to a square wave.

The process only works for *periodic* signals, and it only works for *most* periodic signals. If a signal is not periodic – if it does not repeat its pattern – then this method will not work. [There are other methods to use in such situations, but they require much longer explanations.] If a periodic signal has sudden jumps or straight edges, the method will at best find the waves that when added together produce an approximation to the signal. The basic rule is that if a periodic signal can be said to be exactly the sum of some waves, then the process will work; if a periodic signal cannot be said to be exactly the sum of some waves, then either the process will not work, or, more commonly, it will find an approximation. For example, there is no sum of waves that can exactly match a given square wave. Therefore, at best the process will find a sum of waves that will approximately match a square wave.

In this chapter, we will be paying attention to the mean levels of the results of multiplications. If we know the formulas of multiplied waves, we can calculate the mean level exactly using maths. If we do not know the formulas, we can calculate the mean level by taking an average of all the y-axis values for one cycle of the result – the mean level is the average level, hence its name. In practice, we have to take the average of a series of evenly spaced y-axis values because it would be impossible to take the average of *every* y-axis value. The more y-axis values for one cycle that we read, the more accurate the result will be.

What can be discovered

As I said in Chapter 13, adding waves of different frequencies is analogous to adding together liquids that do not mix. If we pour 3 litres of oil and 5 litres of water into a bucket, we will still be able to tell that we have two different liquids and how much we have of each because they do not mix together properly. Similarly, if we add a wave with a frequency of 3 cycles per second to a wave with a frequency of 5 cycles per second, we will still be able to tell (after doing some calculations) that we have a signal containing waves of 3 and 5 cycles per second, and how much of each (the amplitudes). The frequencies do not mix. The “unmixability” of frequencies is the property of waves that enables humans and animals to distinguish between colours and sounds in the real world. [The proper word for “unmixability” is “immiscibility”, but everyone would have to look that up in a dictionary.]

The analogy of unmixable liquids is also relevant to another aspect of adding waves of different frequencies. If we poured half a litre of water into a bucket already containing half a litre of water and one litre of oil, when examining the resulting contents, we would only be able to tell that there was one litre of water and one litre of oil. The two half litres of water *do* mix together, and are indistinguishable. Similarly, if we add two waves of a particular frequency to one wave of a different frequency, examining the resulting signal would only show that there were two different frequencies in the signal. Amplitudes of the same frequency *do* mix together, and cannot be distinguished in the summed signal.

Another useful fact to know is that the *mean levels* for each added wave, regardless of amplitude, frequency or phase, all become blended together, and are inseparable. We can know the total of all the mean levels (it will be the mean level of the signal being analysed), but we cannot know how that mean level was distributed between the originally added waves.

The phase of a wave can be recovered, but if there are two or more waves of different amplitudes with the same frequency, it will only be possible to recover the phase of the sum of those same-frequency waves.

The basic rule for what the analysis can find is that we can determine one wave of each frequency that was added together to make up a signal.

The summary of what attributes can be discerned from analysing a signal made up of added waves is:

- Amplitudes of the same frequency are indistinguishable for that frequency. In other words, if we are given a signal created from adding:
  “y = 2 sin (360 * 3t)”
  ... and:
  “y = 3 sin (360 * 3t)”
  ... the result of which is:
  “y = 5 sin (360 * 3t)”
  ... then we would only be able to detect one wave with an amplitude of 5 units, and not the amplitudes of the two waves that were added.

- Mean levels for all waves, regardless of frequency, amplitude or phase, become added together and cannot be distinguished. In other words, if we add:
  “y = 7 + 3 sin (360 * 6t)”
  ... and:
  “y = 9 + 6.5 sin ((360 * 4.4t) + 270)”
  ... then we would not be able to tell from the resulting signal which constituent wave had had which mean level. We would only be able to tell that the sum of the mean levels was 16. The sum of the mean levels of the waves that make up a signal will be the same as the mean level of the signal as a whole. If a signal had a mean level of, say, 5 units, then the mean levels of all the waves that were added to make it would have added up to 5 units too.

- Phases for the summed waves of each frequency can be distinguished, but if two or more waves of the same frequency but with different phases were added, we will only be able to recover the resulting phase of the sum for that frequency. In other words, if we added:
  "y = sin ((360 * 6t) + 90)"
  ... and:
  "y = sin ((360 * 6t) + 180)"
  ... the result of which is:
  "y = 1.4142 sin ((360 * 6t) + 135)"
  ... then we would only be able to recover the combined phase for that frequency, which is 135 degrees.

- Frequencies can always be distinguished because, as I have already said, they do not mix. As a negative-frequency wave formula refers to the same wave curve as that formula given a positive frequency and a different phase, it is easier to ignore the idea of negative frequencies in this chapter, and treat all frequencies as being positive. [Technically, we could recover negative frequencies, but only if we were treating all frequencies as being negative. Therefore, there is not much advantage in thinking about negative frequencies here.]

Step 1: Make a list of possible frequencies

We will now look at the three main steps for discovering which waves were (or could have been) added to make up a signal.

In Chapter 13, I showed that adding waves of different frequencies produces a signal with a frequency related to that of the added waves. [The frequency of the signal is often called "the *fundamental* frequency".] The frequency of the resulting signal will be the highest common divisor of the added frequencies. It will be the highest number for which each added frequency is an integer multiple. To put this another way, the frequencies of the added waves are *always* integer multiples of the resulting frequency.

For example, if we add waves with frequencies of 2, 4, 6 and 10 cycles per second, the resulting signal will have a frequency of 2 cycles per second. This is because 2, 4, 6 and 10 are all integer multiples of 2, and equally importantly, 2 is the highest number for which 2, 4, 6, and 10 are all integer multiples. The number 2 is the highest common divisor of 2, 4, 6, and 10.

If we add waves with frequencies of 6, 9, 15 and 33 cycles per second, the resulting signal will have a frequency of 3 cycles per second. The numbers 6, 9, 15 and 33 are all integer multiples of 3. The number 3 is the highest common divisor. It is the highest number for which 6, 9, 15 and 33 are all integer multiples.

If we add waves with frequencies of 1.5, 2, 2.5 and 4.5, the resulting signal will have a frequency of 0.5 cycles per second. This is the highest common divisor. It is the highest number for which 1.5, 2, 2.5 and 4.5 are all integer multiples. These frequencies are not all integers, but they are all integer *multiples*.

Another way of thinking about this is that the resulting signal from adding waves will always have a *period* that is an integer multiple of the periods of each of the added waves. For example, if we add waves with periods of 0.5, 0.75 and 2 seconds, the resulting signal will have a period of 6 seconds. The number 6 is the lowest integer multiple of 0.5, 0.75 and 2. (The frequencies of these added waves are 2, 1.3333 and 0.5 cycles per second. The frequency of the resulting signal is 0.1667 cycles per second.) Some people might find it easier to think of frequency; some people might find it easier to think of period.

Knowing all of this means that if we are presented with a signal, we can tell from which set of frequencies the added waves must have come. If we are given any signal that was created by adding waves, we know that those added waves must have had frequencies that were integer multiples of the frequency of the signal. From this, we can make a list of the possible frequencies in the signal. We will not be able to tell *exactly* which frequencies were added using just this information, but we can know from which set of frequencies those individual frequencies came.

For example, if a signal made by adding two or more waves has a frequency of 5 cycles per second, then the only waves that could have been added to make it *must* have had frequencies that were integer multiples of 5 cycles per second. Those frequencies must be two or more from the list: 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65 and so on. If there were added waves that had frequencies outside of this list, the resulting signal would not have a frequency of 5 cycles per second. This might be easier to understand if you think about how all the constituent waves will repeat an integer number of times in the time it takes a 5-cycle-per-second wave to repeat once, or else the resulting signal would not have the frequency of 5 cycles per second. If this is not clear, read the “Addition of different frequencies” section in Chapter 13 again.

As another example, if a given signal, created from adding two or more waves, has a frequency of 11 cycles per second, then we know for sure that the waves that were added to make it would all have had frequencies that were integer multiples of 11. Therefore, they would be two or more from the following list: 11, 22, 33, 44, 55, 66, 77 cycles per second and so on. The constituent waves *must* have frequencies in this list or else the given signal would not have had that frequency.

Later in this chapter, we will turn each of the frequencies in such a list into waves with which we will test the signal. For this reason, we will call the list of frequencies, a list of "test frequencies". The waves created from this list will be called "test waves".

Summary of step 1

The first step in discovering which frequencies were added to make a given signal is to make a list of the *possible* frequencies that could have been added to make it. This we can do because we know that they will all be integer multiples of the frequency of the given signal. This still gives us an endlessly long list of possible frequencies, but there is less work to do than if we had to take into account every conceivable frequency. The difference is between having to test the frequencies of, say, 4, 8, 12, 16 ... and so on, and having to test 0.0000001, 0.0000002, 0.0000003 and so on. [We could still test 0.0000001, 0.0000002 and so on, but there would be no point because if our signal has a frequency of 4 cycles per second, those frequencies could not be in the signal.]

The interesting aspect of this is that even if the periodic signal were not literally created by adding waves, we can still make a list of frequencies that could be added to recreate it (or recreate it approximately).

Step 2: Centre the signal

If the original signal has a non-zero mean level, then we have to make a note of that mean level and centre the signal on y = 0. In other words, we have to remove the mean level from the signal. We subtract the mean level from the signal so that the signal has a mean level of zero units. The original mean level of the signal will be part of the addition when all the constituent waves have been discovered.

Step 3: Test each wave

In Chapter 16 on multiplying waves, we looked at the mean levels of signals created by multiplying two waves of the same or different frequencies. A summary of the important rules related to multiplication is as follows:

1. If we multiply two Sine waves with zero mean levels and *different* frequencies, the resulting signal will *always* have a mean level of zero.

   If the frequencies of the Multiplication Waves are different and they both have zero mean levels, then there will be two equivalent Addition Waves, each with zero mean levels and *non-zero* frequencies. Therefore, the resulting signal will have a zero mean level. This will be true regardless of any other characteristics. For example "y = 5.5 sin ((360 * 7t) + 99)" multiplied by "y = 2.7 sin ((360 * 8t) + 3)" will produce two Addition Waves with zero mean levels, and therefore a signal that has a zero mean level.

2. If we multiply two Sine waves with zero mean levels, zero phases and the *same* frequency, the resulting signal will *always* have a non-zero mean level that will be equal to half the square of the amplitudes.

   If the frequencies of the Multiplication Waves are the same and have zero mean levels and zero phases, then there will be two equivalent Addition Waves, each with zero mean levels, but one of them will have zero frequency, and thus end up acting as a mean level to the other one. Therefore, the resulting signal will have a non-zero mean level.

For example:
“y = 2.3 sin (360 * 11t)”
... multiplied by:
“y = 5.7 sin (360 * 11t)”
... will produce two Addition Waves with zero mean levels:
“y = 6.555 sin ((360 * 0t) + 90)”
... and:
“y = 6.555 sin ((360 * 22t) + 270)”.
The zero-frequency wave will end up as just its amplitude multiplied by the Sine of 90 degrees, which is the same as its amplitude multiplied by 1. Therefore, it ends up as “y = 6.555”, and acts as a mean level to the other wave. Therefore, the resulting signal will have a mean level of 6.555 units. The value of 6.555 is the first Addition Wave’s amplitude, which was:
0.5 * (a_1 * a_2).

3. If we multiply two Sine waves with zero mean levels, the *same* frequency, and phases that differ by exactly 90 degrees, then the resulting signal will *always* have a zero mean level.

   If the frequencies of the Multiplication Waves match, one of the Addition Waves will have zero frequency and end up as a mean level to the other one. However, as the phase of that wave will be calculated by $\phi_1 - \phi_2 + 90$, if the phases are 90 degrees apart, the phase will end up as 180 or 0 degrees. Therefore, the wave will be equal to its amplitude multiplied by the Sine of 180 or 0, which is zero, and so the wave will end up as “y = 0”. This means it acts as a mean level of zero to the other wave, which means that the resulting signal will have zero mean level.

4. If we multiply two Sine waves with zero mean levels, the same frequency, and phases that are anything but 90 degrees apart, then the resulting signal will have a non-zero mean level.

   If the frequencies of the Multiplication Waves match, one of the Addition Waves will have zero frequency and end up as a mean level to the other one. The phase of that wave will be calculated by $\phi_1 - \phi_2 + 90$. This means that that wave will end up as its amplitude multiplied by the Sine of $\phi_1 - \phi_2 + 90$. For two Multiplication Waves that have phases that are not 90 degrees apart, “$\phi_1 - \phi_2 + 90$” will be any number from the positive of the amplitude to the negative of the amplitude, except for zero.

5. If we multiply a Sine wave against a signal made from adding Sine waves, the result will be as if that wave had been multiplied against each individual wave in that sum, and then each of those results added together. This is true whether the multiplication is done by multiplying corresponding y-axis values from the wave and the signal, or if it is done by multiplying corresponding y-axis values from the wave and each of the signal's constituent waves, and then adding up the results.

These are all very important ideas. The fifth rule means that all the effects mentioned in rules 1 to 4 can happen to the constituent waves in a signal when it is multiplied by a wave. In other words, if a Sine wave with zero mean level is multiplied by a signal with zero mean level that is a sum of Sine waves, and if that Sine wave's frequency does not match any of the frequencies of the constituent waves, then the mean level of the resulting signal will be zero. If that Sine wave's frequency *does* match that of one of the constituent waves, and if the phases of it and the "matching-frequency" wave in the sum are anything but 90 degrees apart, then the resulting signal will end up with a non-zero mean level. If that Sine wave's frequency matches the frequency of one of the constituent waves, and if the phase of it and the wave in the sum are exactly 90 degrees apart, then the resulting signal will have a zero mean level. [This is also true if the phases are 270 degrees apart because if the phases are 270 degrees apart, then they are also 90 degrees apart.]

Rule 2 is important because, given rule 5, it means that if the phases of a multiplied Sine wave and the constituent wave of the same frequency in the signal have zero phases (or the same phase), the resulting mean level of the entire signal will be exactly half the two amplitudes multiplied together. This idea, among others, can be used to discern the amplitude of the constituent wave.

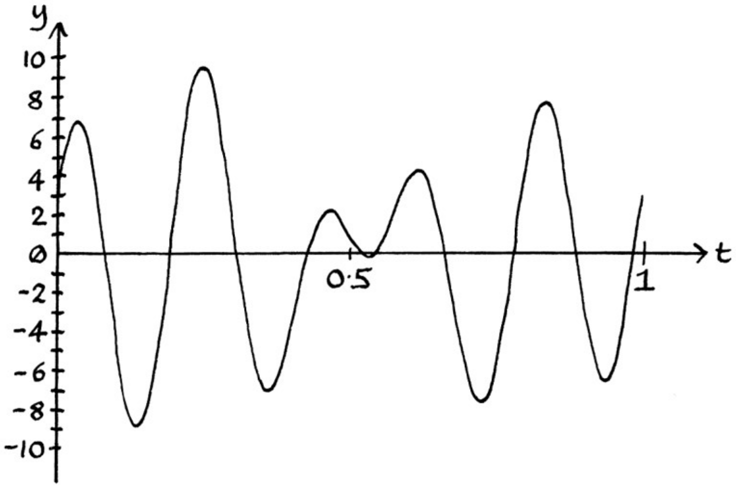
All of the above ideas mean that if we multiply a signal with zero mean level by a wave with zero mean level, the mean level of the result will indicate the characteristics of the waves that were added to make that signal. Depending on how the phases relate to each other, we will be able to determine if that frequency exists in the original signal (by the presence of a non-zero mean level), as well as the amplitude and phase of the wave with that frequency.

Example 1

To demonstrate all of the above, we will go through an example where we know what the constituent waves are. This will explain why and how everything works, which will enable us to use the process when we do not know what the constituent waves are. We will say we have a signal that is the sum of these three waves:
"y = 2.5 sin ((360 * 3t) + 200)"
"y = 3.4 sin ((360 * 4t) + 35)"
"y = 5.6 sin ((360 * 5t) + 20)"

Despite the signal being its own entity, it will always really remain a sum of these constituent waves because waves of different frequencies do not mix. When we multiply the signal by a particular wave, it will be as if we are multiplying each of these constituent waves by that wave, and then adding up the result.

The signal looks like this for one cycle:



Example 1: Non-matching test frequency

First, we will multiply the signal by the following wave, which we will call the "test wave":
"y = 1.4 sin (360 * 6t)"

As a frequency of 6 cycles per second is not one of the constituent frequencies of our signal, this should result in a signal that has a zero mean level.

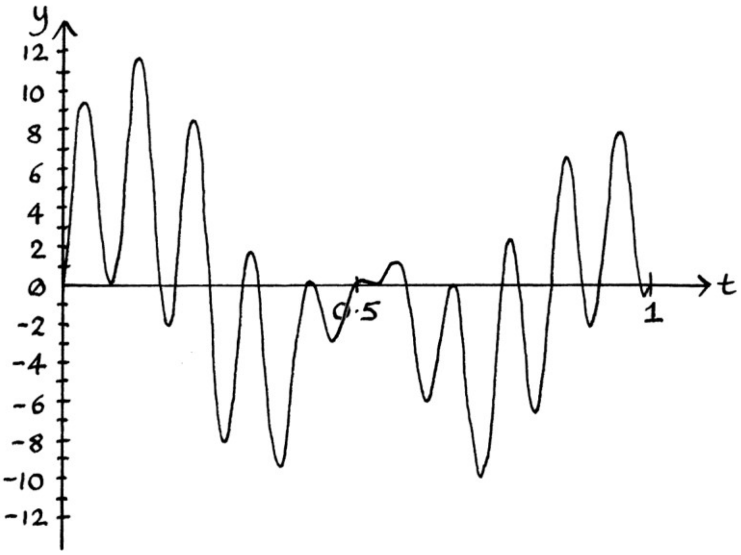
The result of multiplying the signal by the test wave can either be calculated by multiplying the corresponding (by time) y-axis values from the signal and the test wave, or it can be done by multiplying each constituent wave by the test wave and adding up the results. Each method will produce the same resulting signal. In practice, we would not know the constituent waves, so we would have to use the first method. [If we knew the constituent waves, there would be no point in analysing the signal.] As I am trying to explain how and why the multiplications work, we will use the second method, and the reason for the result will therefore be clearer. To do this we will be calculating the sum of these multiplications:

"y = 2.5 sin ((360 * 3t) + 200)" multiplied by "y = 1.4 sin (360 * 6t)"
"y = 3.4 sin ((360 * 4t) + 35)" multiplied by "y = 1.4 sin (360 * 6t)"
"y = 5.6 sin ((360 * 5t) + 20)" multiplied by "y = 1.4 sin (360 * 6t)"

These end up as the sum of:

"y = 1.75 sin ((360 * 3t) + 250)"
"y = 1.75 sin ((360 * 9t) + 110)"
"y = 2.38 sin ((360 * 2t) + 55)"
"y = 2.38 sin ((360 * 10t) + 305)"
"y = 3.92 sin ((360 * 1t) + 70)"
"y = 3.92 sin ((360 * 11t) + 290)"

The mean level of the signal made from adding these waves is zero. The resulting signal looks like this for one cycle:



The test wave's frequency was different from each of the constituent waves of the original signal, and therefore, the multiplication did not produce a result with a non-zero mean level. To clarify this, it does not matter what the amplitude or phase of the test wave is. As long as the frequency of the test wave does not match that of any of the constituent waves, the resulting signal will have a zero mean level.

Example 1: Matching test frequency

Now, we will multiply the original signal by this test wave:
"y = 1.4 sin (360 * 4t)"

This test wave has a frequency that matches one of the constituent waves in the signal – it is 4 cycles per second.

We can calculate the result of the multiplication by multiplying each of the constituent waves by the test wave, and then adding the results together:

"y = 2.5 sin ((360 * 3t) + 200)" multiplied by "y = 1.4 sin (360 * 4t)"
+
"y = 3.4 sin ((360 * 4t) + 35)" multiplied by "y = 1.4 sin (360 * 4t)"
+
"y = 5.6 sin ((360 * 5t) + 20)" multiplied by "y = 1.4 sin (360 * 4t)"

These end up as the sum of:

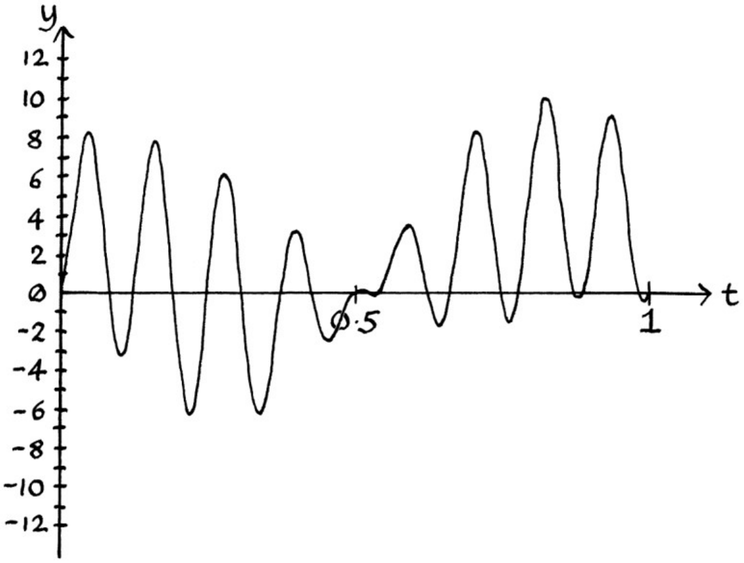
"y = 1.75 sin ((360 * 1t) + 250)"
"y = 1.75 sin ((360 * 7t) + 110)"
"y = 2.38 sin ((360 * 0t) + 55)"
"y = 2.38 sin ((360 * 8t) + 305)"
"y = 3.92 sin ((360 * 1t) + 110)"
"y = 3.92 sin ((360 * 9t) + 290)"

There is one wave with a zero frequency, and this ends up as:
"y = 2.38 sin (55)"
... which is:
"y = 1.94958187"

Therefore, we can say that the resulting signal consists of these waves:
"y = 1.75 sin ((360 * 1t) + 250)"
"y = 1.75 sin ((360 * 7t) + 110)"
"y = 2.38 sin ((360 * 8t) + 305)"
"y = 3.92 sin ((360 * 1t) + 110)"
"y = 3.92 sin ((360 * 9t) + 290)"
... and a mean level of 1.9496 units.

[I have not added together the two resulting waves with the same frequency, as for the purposes of this exercise, there is no point, and it is clearer if I do not.]

The resulting signal, therefore, has a mean level of 1.9496 units, and it looks like this for one cycle:



The test wave had a frequency that was one of the constituent waves of the original signal (and the phases were not 90 degrees apart), and therefore, the multiplication produced a result with a non-zero mean level.

Example 1: Matching test frequency and phase

Now, we will multiply the original signal by the following test wave, which not only matches one of the constituent waves' frequencies, but also has the same phase:
"y = 1.4 sin ((360 * 4t) + 35)"

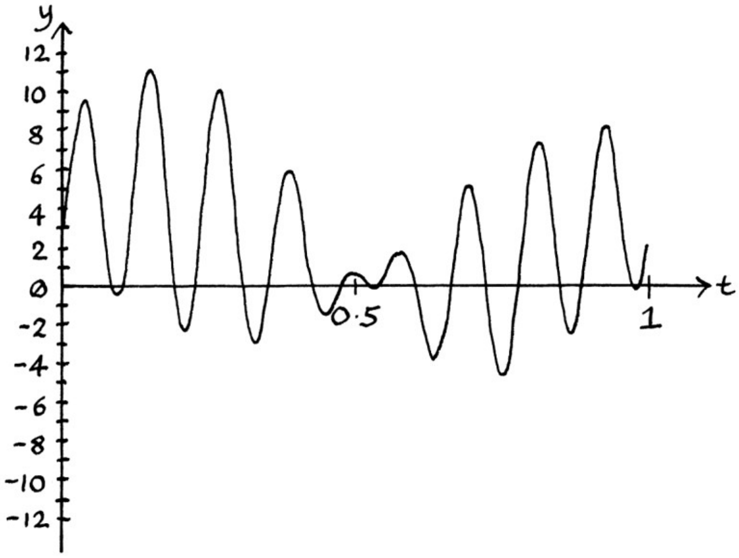
We calculate the result of the multiplication by multiplying each of the constituent waves by the test wave:
"y = 2.5 sin ((360 * 3t) + 200)" multiplied by "y = 1.4 sin ((360 * 4t) + 35)"
"y = 3.4 sin ((360 * 4t) + 35)" multiplied by "y = 1.4 sin ((360 * 4t) + 35)"
"y = 5.6 sin ((360 * 5t) + 20)" multiplied by "y = 1.4 sin ((360 * 4t) + 35)"
... and then adding the results.

We end up with the sum of:
"y = 1.75 sin ((360 * 1t) + 285)"
"y = 1.75 sin ((360 * 7t) + 145)"
"y = 2.38 sin ((360 * 0t) + 90)"
"y = 2.38 sin ((360 * 8t) + 340)"
"y = 3.92 sin ((360 * 1t) + 75)"
"y = 3.92 sin ((360 * 9t) + 325)"

The zero-frequency wave ends up as:
"y = 2.38 sin (90)"
... which is:
"y = 2.38 * 1"
... which is:
"y = 2.38"

Therefore, the sum of waves can be written as:
"y = 1.75 sin ((360 * 1t) + 285)"
"y = 1.75 sin ((360 * 7t) + 145)"
"y = 2.38 sin ((360 * 8t) + 340)"
"y = 3.92 sin ((360 * 1t) + 75)"
"y = 3.92 sin ((360 * 9t) + 325)"
... and a mean level of 2.38 units.

The resulting signal has a mean level of 2.38 units and looks like this for one cycle:



The test wave had a frequency and phase that matched those of one of the constituent waves in the original signal, and therefore, the multiplication produced a result with a non-zero mean level. What is more, the mean level of the whole signal is actually half the result of multiplying the amplitude of the test wave by the amplitude of the relevant constituent wave. It is:
0.5 * 3.4 * 1.4 = 2.38.

If we think of the multiplication formula from Chapter 16, the mean level of the resulting signal is actually:
0.5 * (a_1 * a_2)

Because the frequencies of the test wave and the constituent wave match, that particular result in the final sum has a zero frequency.

Because the frequencies match, the phases are relevant. [Note how I said "*relevant*", and not "irrelevant".] Because the phases are the same, that particular result has a phase of 90 degrees.

As the frequency of that particular result in the final sum is zero and the phase is 90 degrees, that particular result ends up as 0.5 * (a_1 * a_2), multiplied by the Sine of 90 degrees, which is 1. Therefore, that particular result ends up as just
0.5 * (a_1 * a_2). This acts as a mean level that affects the mean level for the whole resulting signal. The mean level of the whole resulting signal will end up as
0.5 * (a_1 * a_2), where a_1 and a_2 are the amplitudes of the test wave and the constituent wave of the same frequency.

Note that the mean level of the resulting signal will only be equal to 0.5 * (a_1 * a_2) if the Sine of the phase of the zero-frequency wave in the resulting sum is equal to 1. Therefore, the phases of the test wave and the constituent wave must match. If the phases were 180 degrees apart, we would end up with −0.5 * (a_1 * a_2). If the phases were 90 degrees apart, we would end up with 0, and the mean level of the resulting signal would be zero.

All of this means that we can work out the amplitude of the constituent wave from the mean level of the resulting signal. If the phases match (which admittedly, in practice, we would not be able to tell for sure without more work), we would know that:
mean level = 0.5 * (a_1 * a_2)
... where we will say a_1 is the amplitude of the test wave, and a_2 is the amplitude of the constituent wave. Therefore, the amplitude of the constituent wave would be:
(2 * mean level) ÷ a_1

In other words, we double the mean level and divide it by the amplitude of the test wave, and the result will be the amplitude of the constituent wave.

We can make the maths easier by only using test waves that have amplitudes of 1 unit. In that case, the amplitude of the constituent wave will be twice the resulting mean level. We can make the maths easier still – if we use test waves with amplitudes of 2 units, the amplitude of the constituent wave will be equal to the resulting mean level, and there will be no need for any more maths.

All of this is good to know, but if we are dealing with a periodic signal for which we do not know the constituent waves, we cannot know if the phases match or not. The solution to this is to try a range of phases and see how the resulting mean levels fluctuate. If we test a range of phases, then we can see what the maximum and minimum resulting mean levels are. We will know if we have a matching phase because the mean level will be at its maximum; we will know if we have a matching but negative phase because the mean level will be at its minimum. We will know if the phases are 90 degrees (or 270 degrees) apart because the mean level will be zero.

Example 1: Matching frequencies with a range of phases

Now we will multiply our original signal against a test wave with a range of phases. As a reminder, our original signal is made up of the sum of these waves:
"y = 2.5 sin ((360 * 3t) + 200)"
"y = 3.4 sin ((360 * 4t) + 35)"
"y = 5.6 sin ((360 * 5t) + 20)"

We will use the following test wave:
"y = 2 sin ((360 * 4t) + ϕ)"
... where ϕ will be a range of phases from 0 to 360 degrees, spaced at intervals of 5 degrees.

Note how this test wave has an amplitude of 2 units. This means that if the phase of the test wave matches the phase of the constituent wave with the same frequency, the mean level will be at its maximum, and it will also be exactly equal to the amplitude of the constituent wave. Using an amplitude of 2 units in the test wave will save us having to do any more maths.

Side note for this particular example

In this example, I am only trying to explain *why* the process works. For the purposes of doing the maths for this particular example, it is useful to know the following: we are only interested in the mean level of the resulting signal and nothing else. We also know that the mean level of the resulting signal will be unaffected when our test wave is multiplied against the 3-cycle-per-second and 5-cycle-per-second constituent waves. Therefore, for the purposes of calculating the maths for this particular example, we only need to multiply our test wave against the 4-cycle-per-second wave. Therefore, we only need to do this calculation:

"y = 3.4 sin ((360 * 4t) + 35)" multiplied by "y = 2 sin ((360 * 4t) + ϕ)"

... and even then we only need to pay attention to the first of the Addition Waves (which is the one with zero frequency) because the second one will not end up as a mean level. If we were testing for real, we would not have the luxury of being able to isolate one constituent wave for testing, and we would have more work to do. However, for explaining how this all works, we can take this shortcut, and the results will be the same.

The formula for the relevant part is this:
“y = 0.5 * (a_1 * a_2) * sin ((360 * (f_1 – f_2)) + (ϕ_1 – ϕ_2 + 90)”
... which, as we know that “f_1 – f_2” will be zero, means that we only need to calculate this:
“y = 0.5 * (a_1 * a_2) * sin (ϕ_1 – ϕ_2 + 90)”.

This means that we can find the results for this particular example with a lot less effort than we might otherwise need.

Continuing with the calculations

We will return to the calculations. First, we multiply the entire signal by our first test wave:
“y = 2 sin ((360 * 4t) + 0)”.
The mean level of the resulting signal will be 2.7851 units.

Next, we will use this test wave:
“y = 2 sin ((360 * 4t) + 5)”
The mean level of the resulting signal will be 2.9445 units.

Next, we will use this test wave:
“y = 2 sin ((360 * 4t) + 10)”
The mean level of the resulting signal will be 3.08145 units.

We continue in this way, increasing the phase of the test wave by 5 degrees each time. We end up with 72 different mean levels. These are the mean levels of the signals created by multiplying the original signal by each test wave.

Portrayed in a table, the first few mean levels are as so:

Phase of test wave	**Mean level of resulting signal**
0	2.7851
5	2.9445
10	3.08145
15	3.1950
20	3.2841
25	3.3483
30	3.3871
35	3.4
40	3.3871
45	3.3483
50	3.2841
55	3.1950

... and so on.

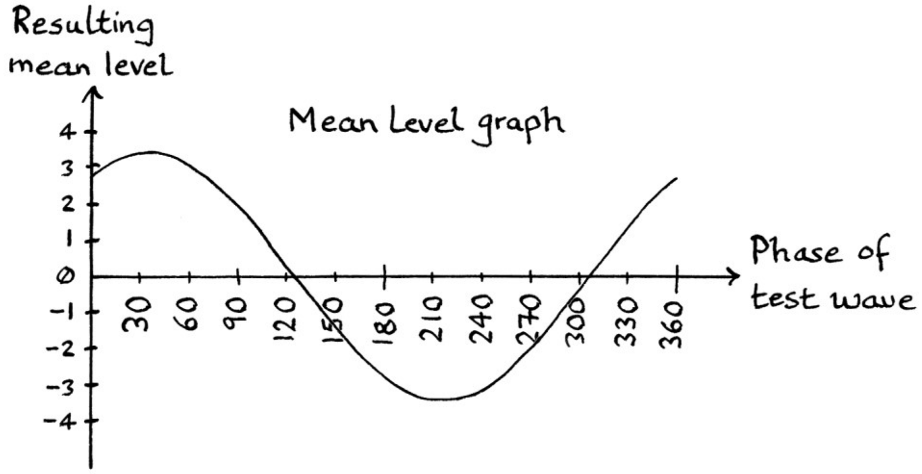
We can see from this part of the table that when the phase of the test wave is 35 degrees, the mean level of the resulting signal is at its highest – it is 3.4 units.

[If we had used the shortcut for calculating these values, as described in the side note above, then the values would have been calculated as so:

$0.5 * (2 * 3.4) * \sin(\phi_1 - 35 + 90)$

From this, it might be apparent that the values in the table are really the y-axis values of a Sine wave (one that relates to angles, and not time). This should also be apparent if you remember Chapter 16 on the multiplication of waves. Note that the shortcut is only relevant if we actually knew the constituent waves to start with. Normally, we would not know them, so the shortcut would not be of much use, except for demonstrating this point.]

If we plot the values from the whole table as a graph, we will see this wave:



We will call this the "Mean Level wave" or the "mean level graph". The graph shows the mean levels of the resulting signal after we have multiplied the original signal by test waves with phases from 0 to 360 degrees. The wave on the mean level graph has the formula "y = 3.4 sin (θ + 55)". Note that the Mean Level wave is not the same as the constituent wave, although it is very similar.

Note that if we had not known anything about the original signal, and had done the multiplications as the y-axis values from our test wave multiplied by the corresponding (by time) y-axis values from the original signal, then the mean level graph and its formula would be *identical* to those here. The process works on the signal itself – it does not matter whether that signal is treated as being made up of y-axis values or made up of constituent waves. This is because, as I have said many times before, different frequencies do not mix.

If we had multiplied a test wave with a range of phases against the signal, and the frequency of the test wave had not been in the signal, every resulting mean level would have been zero. The mean level graph would have just been a straight line and would have had the formula "y = 0" (or "y = 0t" if we want to emphasise that it still relates to time).

We know that the constituent wave has a frequency of 4 cycles per second, as we have been multiplying the signal by a wave of 4 cycles per second, and achieving a resulting signal with a non-zero mean level. As the original signal had a zero mean level, we know that the sum of the mean levels of all the constituent waves was zero. Given that it would be impossible to distinguish the mean levels once added together, we will have to assume that they all had zero mean levels. [Technically, they could have had mean levels that added up to zero, but there would be no way for us to detect this.]

The Mean Level wave contains all the other information about the constituent wave:

- The amplitude of the Mean Level wave is the same as the amplitude of the constituent wave – it is 3.4 units.
- The phase of the *constituent* wave is the angle on the θ-axis of the Mean Level wave where the y-axis value is highest. We know this is the case because if the frequencies of two multiplied waves match, then the first Addition Wave will end up as just a mean level, and if the phases match too, it will end up as the highest possible mean level. Therefore, the highest possible y-axis value on the mean level graph must be where the phases match. If the phases match, then the phase of the constituent wave will be the same as the phase of the test wave at that point. In this case, the phase of the constituent wave is 35 degrees.

Therefore, we can know that the constituent wave must be:
"y = 3.4 sin ((360 * 4t) + 35)"
... which is exactly correct.

We have successfully discovered every characteristic of one of the constituent waves that was added to make our original signal.

Example 1: Shortcut to the Mean Level wave

From looking at the above Mean Level wave graph, we can see that the curve is zero whenever the phases of the test wave and the constituent wave are 90 degrees apart. [This is the same as being 270 degrees apart.] If we had only done one multiplication instead of a range of multiplications, we might have accidentally tested against one of the phases that was exactly 90 degrees away from that of the constituent wave. If we had done that, we would not have been able to tell that the test wave's frequency was in the signal. Similarly, if we had only tested one wave, we would not have been able to know if the resulting mean level was the maximum possible, and therefore, we would not have been able to calculate the amplitude and phase of the constituent wave.

As it is, we multiplied by a range of phases, which gave us a good idea of the characteristics of the Mean Level wave. However, this is not foolproof either. Supposing our constituent wave had had a phase of 32.5 degrees instead of 35 degrees, then, because we were using phases at intervals of 5 degrees, we would never have calculated the exact point at 32.5 degrees. The Mean Level wave would have been slightly inaccurate, and our calculation of the amplitude and phase of the constituent wave would have been slightly off.

One solution to this is to multiply by many more test waves. However, without doing hundreds of test waves, this would still risk being inaccurate for situations when the phase of the constituent wave was something such as 32.005 degrees.

In Chapter 12, we looked at recreating a circle from just one of its derived waves. It turned out that, if a circle has zero mean levels, it is possible to recreate the circle from which a wave is derived, and have that circle represent the amplitude and phase, by reading the y-axis values from just two points from the Sine wave graph. These will be the y-axis value at 0 degrees, and the y-axis value at 90 degrees. These two values are used as a pair of coordinates, with the first value being the y-axis coordinate, and the second value being the x-axis coordinate. The coordinates indicate the phase point of the circle, which is enough information to know the radius of the circle, and the amplitude and phase of the derived waves.

To put this more concisely, knowing the y-axis values of a Sine wave at 0 degrees and 90 degrees is enough to recreate the circle from which the wave is, or could be said to be, derived (if the circle has zero mean levels). This also means that all the characteristics of a wave can be inferred from just those two y-axis values (if the wave has zero mean level).

This idea is useful for us here because it means we can recreate the entire *Mean Level* wave's circle with just the points at 0 degrees and 90 degrees. We do not need to bother calculating countless other points, and we do not need to worry about missing out any phases. If we know the Mean Level wave's circle, then we know everything about the Mean Level wave.

Therefore, for our original signal, and in fact for *any* signal, if we are testing for a frequency of 4 cycles per second, we only need to multiply it by the test waves:
"y = 2 sin ((360 * 4t) + 0)"
... and:
"y = 2 sin ((360 * 4t) + 90)"

For our original signal in this example, the mean level for the first multiplication is 2.7851 units. This is the y-axis coordinate of the Mean Level wave's circle's phase point. The mean level for the second multiplication is 1.9502 units. This is the x-axis coordinate of the Mean Level wave's circle's phase point.

The full coordinates are: (1.9502, 2.7851). This is the phase point of the Mean Level wave's circle.

From these two points, we can recreate the circle from which the *Mean Level* wave is, or could be said to be, derived. We use Pythagoras's theorem to calculate the radius: $\sqrt{2.7851^2 + 1.9502^2}$ = 3.4 units.

We use arctan to calculate the angle of the phase point:
arctan (2.7851 ÷ 1.9502) = 55 degrees.
[As we are using arctan, we check to see if this is in the correct quarter of the circle, which it is.]

This means that the Mean Level wave will have an amplitude of 3.4 units and a phase of 55 degrees. We have found the characteristics of the Mean Level wave with just two test waves.

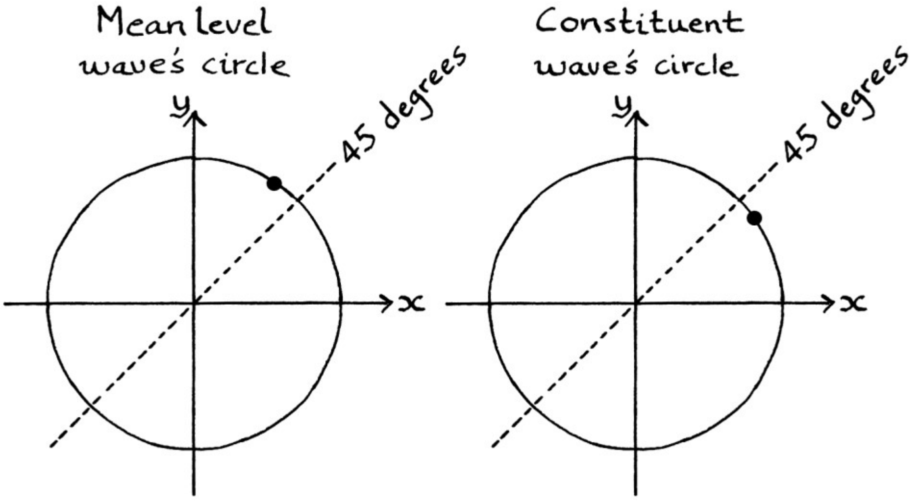
One beneficial consequence of using exactly two test waves that have phases 90 degrees apart is that if one of the phases should by chance be 90 degrees away from the phase of the constituent wave, then the other wave cannot possibly be 90 degrees away from the phase of the constituent wave. Therefore, if both resulting mean levels are zero, we can be completely sure that the tested frequency does not exist in the signal.

Example 1: Calculating the actual wave

At the moment, we end up with the Mean Level wave, which allows us to work out the amplitude and the phase of the constituent wave. As I said before, the Mean Level wave is *not* the constituent wave. For one thing, the Mean Level wave relates to angles and not time. Another thing is that the phase is wrong. The phase of our *Mean Level* wave is 55 degrees, while the phase of the *constituent* wave is 35 degrees. We can tell that the phase of the constituent wave is 35 degrees because on the Mean Level wave, the highest y-axis value occurs at 35 degrees – it indicates the place where the phase of a test wave matches the phase of the constituent wave.

Instead of reading off the angle for the highest y-axis value on the Mean Level wave, we can use another way to find the constituent wave's phase. We can turn the phase of the Mean Level wave into the phase of the constituent wave by taking the phase point of the Mean Level wave's circle, and flipping it along a 45-degree line. In other words, however above or below the phase point is from 45 degrees, we make it that number below or above 45 degrees.

The phase of 55 degrees is 10 degrees above 45 degrees, so it becomes 10 degrees below 45 degrees, which is 35 degrees.



We therefore end up with "y = 3.4 sin (θ + 35)". This has the characteristics of the constituent wave, but of course is still an angle-based wave and not a time-based wave, but we will sort this out in a moment.

A quick way to find the angle the other side of 45 degrees is to subtract the angle from 90 degrees (and possibly add on 360 afterwards to make it into a positive angle). For example, if we have 55 degrees, we calculate 90 – 55 = 35 degrees.

Another way to achieve exactly the same thing is to swap the "x" and "y" coordinates of the phase point on the circle for that wave. This works because it is equivalent to swapping the lengths of the adjacent and opposite sides of a right-angled triangle. If the phase point of the Mean Level wave's circle has the coordinates (A, B), then the phase point of the constituent wave's circle will have the coordinates (B, A). For our example, our Mean Level wave's circle's phase point was at:
(1.9502, 2.7851)
... so the constituent wave's circle's phase point will be at:
(2.7851, 1.9502).

Finding the angle of the phase point will still use arctan, but in this case it will be arctan (1.9502 ÷ 2.7851) = 35 degrees [after checking it is the correct result of the two possible ones].

The reason that finding the angle on the other side of 45 degrees produces the phase of the constituent wave becomes clearer with some thought. To find the phase of an angle-based wave (as described in Chapter 7), we take the first angle in the positive half of the θ-axis where the curve has a y-axis value of zero and is rising, and subtract that value from 360 degrees. For our Mean Level wave, this point is at 305 degrees. Therefore, the phase of the Mean Level wave is 360 – 305 = 55 degrees. We know that the angle where our Mean Level wave has the highest y-axis value will be the phase of the constituent wave – this is because it is the place where the phase of the test wave matches the phase of the constituent wave. The point where the highest y-axis value appears on our Mean Level wave will always be exactly 90 degrees after the curve has a y-axis value of zero and is rising. Therefore, to find the phase of the constituent wave, we take the angle where the Mean Level wave's y-axis value is zero and rising, add on 90 degrees, and subtract that from 360. In other words, we calculate this:
"360 – x + 90"
... where "x" is the angle on the mean level graph where y = 0 and the curve is rising. Given that 360 degrees is the same as 0 degrees, this calculation is the same as:
"– x + 90"
... or:
"90 – x"
... and subtracting a value from 90 degrees finds the value equidistant from, and on the other side of, 45 degrees. Therefore, the phase of the constituent wave will always be on the other side of 45 degrees from the phase of the Mean Level wave.

After finding the angle on the other side of 45 degrees, we still have an angle-based wave, but it is not difficult to make it into a time-based wave. Either we can put its characteristics into a 4-cycle-per-second Sine wave, or we can say it is a zero-frequency Sine wave with the formula, "y = 3.4 sin ((360 * 0t) + 35)", and change the frequency to 4 cycles per second.

Summary

A summary of the steps to find one constituent wave is as follows. We multiply the signal by two test Sine waves with the following characteristics:

- Both have the frequency that we want to test.
- Both have amplitudes of 2 units (we could use other amplitudes, but an amplitude of 2 makes the maths afterwards simpler).
- Both have zero mean levels.
- The first wave has zero phase; the second has a phase of 90 degrees.

If both the mean levels of the two signals resulting from the multiplications are zero, then the tested frequency does not exist in the original signal. Otherwise, the tested frequency does exist, and we continue with calculating its details.

We treat the mean levels as the coordinates of a "Mean Level" circle's phase point, with the first mean level as the "y" coordinate, and the second mean level as the "x" coordinate. We then swap the coordinates to turn them into the phase point of the constituent wave's circle. [The first mean level becomes the "x" coordinate; the second mean level becomes the "y" coordinate.] The phase of the constituent wave will be the angle of this phase point, found using arctan. The amplitude of the constituent wave will be the radius of this circle (which is the distance of the phase point from the origin of the axes).

To put this slightly more mathematically:

- First mean level is that of "signal * (2 sin (360 * test frequency * t))"
- Second mean level is that of "signal * (2 sin ((360 * test frequency * t)+90))"
- The Mean Level wave's circle's phase point coordinates are:
  (second mean level, first mean level)
- The Constituent wave's circle's phase point coordinates are:
  (first mean level, second mean level)
- Constituent wave's amplitude = $\sqrt{\text{first mean level}^2 + \text{second mean level}^2}$
- Constituent wave's phase = arctan (second mean level ÷ first mean level)

We can make this more concise as:

- First mean level is that of "signal * (2 sin (360 * test frequency * t))"
- Second mean level is that of "signal * (2 sin ((360 * test frequency * t)+90))"
- Constituent wave's amplitude = $\sqrt{\text{first mean level}^2 + \text{second mean level}^2}$
- Constituent wave's phase = arctan (second mean level ÷ first mean level)

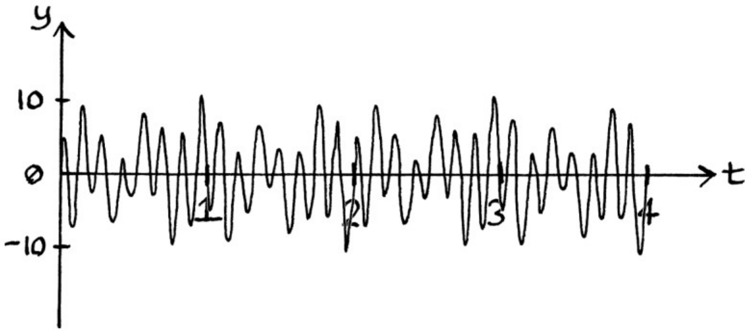
Although it took a long explanation to get here, the actual process is very simple.

All Steps Together

Now, we will combine the steps to calculate which waves were added to make up a signal for which we do not know the constituent waves.

Example 2

We will say that we have been given this signal:



We will pretend that we can read the graph with complete accuracy, and that we are able to perform mathematical processes on the signal. This signal has a zero mean level – this means we do not need to bother centring it on "y = 0", because it is already there.

The first step is to make a list of the frequencies that could possibly be in this signal. The signal repeats its pattern once every two seconds, which means that its frequency is 0.5 cycles per second. Therefore, every wave that was added to make it must have had a frequency that was an integer multiple of 0.5 cycles per second. The frequencies of the waves that were added to make up this signal must be in the following list: 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0, 6.5, 7.0, 7.5, 8.0, 8.5, and so on. This will be the list of frequencies that we will test for in the next stage.

Next, we will need to multiply the signal by our test Sine waves. We will have two test Sine waves for each frequency. The two waves in each pair will have amplitudes of 2 units. The first wave in each pair will have a phase of zero degrees; the second will have a phase of 90 degrees. We will multiply each y-axis value from the signal against the corresponding (by time) y-axis value from each test wave. [Obviously, in practice, we would take a series of evenly spaced y-axis values.] We then note down the mean levels of the signals created by each multiplication.

[To clarify how the multiplication works, either we can multiply infinitely long versions of a test wave and the signal, or we can multiply one cycle of the signal with the part of a test wave that lasts for that cycle. The first is obviously impossible in practice. The second is easy, especially because either the test wave will have the same frequency as the signal, or it will have a frequency that is an integer multiple of the frequency of the signal. For one cycle of the signal, there will always be an integer number of cycles of the test wave. Therefore, we only need to concentrate on one cycle of the original signal, and we can be sure that the later cycles of the multiplication would behave in the same way.]

The first frequency in the list is 0.5 cycles per second. This means that we will start with the test waves:
"y = 2 sin (360 * 0.5t)"
... and:
"y = 2 sin ((360 * 0.5t) + 90)"

The first multiplication produces a signal with a mean level of 0.4698 units; the second multiplication produces a signal with a mean level of −0.1710 units. The fact that at least one of these is non-zero means that one of the constituent waves has a frequency equal to that of the test wave. Instead of calculating the details of this constituent wave now, we will continue to test other frequencies.

Now, we will multiply the signal by test waves with the next frequency in the list. This is 1 cycle per second. The test waves will be:
"y = 2 sin (360 * 1t)"
... and:
"y = 2 sin ((360 * 1t) + 90)".

Both of these multiplications produce signals with a mean level of 0 units. This means that the original signal was not created by adding a wave with a frequency of 1 cycle per second.

The next frequency in the list is 1.5 cycles per second. The test waves will be:
"y = 2 sin (360 * 1.5t)"
... and:
"y = 2 sin ((360 * 1.5t) + 90)".

The results of these multiplications produce signals with zero mean levels, which means that there is no constituent wave with a frequency of 1.5 cycles per second.

The next frequency in the list is 2 cycles per second. The test waves will be:
“y = 2 sin (360 * 2t)”
... and:
“y = 2 sin ((360 * 2t) + 90)”.

The resulting mean levels are both zero. Therefore, there is no constituent wave with a frequency of 2 cycles per second in our original signal.

The next frequency in the list is 2.5 cycles per second. The test waves will be:
“y = 2 sin (360 * 2.5t)”
... and:
“y = 2 sin ((360 * 2.5t) + 90)”.

The mean level of the first result is 0.3430 units; the mean level of the second result is −3.0105 units. Therefore, one of the constituent waves definitely has a frequency of 2.5 cycles per second. Before we calculate the characteristics of this constituent wave, we will continue testing more waves.

The next frequency in the list is 3 cycles per second, so we test the relevant waves for that. We then continue through the rest of the list, testing each frequency and calculating the mean levels of the results. We will skip the details of these calculations as they work in the same way as the ones that we have already seen. For this example, we will stop (arbitrarily) when we reach 10 cycles per second. If we were calculating the characteristics of each constituent wave as we went along, we could add them together, and then we would know to stop when the total matched the original signal. [Although, in practice, this would presume that our calculations were perfectly accurate.]

Here is a table showing the resulting mean levels for each test wave:

Frequency being tested	**Mean level of the result of the multiplication with a 0-degree phase test wave**	**Mean level of the result of the multiplication with a 90-degree phase test wave**
0.5	0.469846	−0.171010
1	0	0
1.5	0	0
2	0	0
2.5	0.343006	−3.010523
3	0	0
3.5	0	0
4	0	0
4.5	0	0
5	0	0
5.5	0	0
6	0	0
6.5	0	0
7	0	0
7.5	6.06	0
8	0	0
8.5	1.845076	0.771812
9	0	0
9.5	0	0
10	0	0

The table shows that there are four waves in the signal, and they have frequencies of 0.5 cycles per second, 2.5 cycles per second, 7.5 cycles per second and 8.5 cycles per second. We can ignore all the test frequencies that produced two zero mean levels.

We will now find the details for each constituent wave.

The 0.5-cycle-per-second *Mean Level* wave, if thought of as a circle, has a phase point at the coordinates (−0.171010, 0.469846). Therefore, the 0.5 cycle-per-second *constituent* wave's circle will have a phase point at the coordinates (0.469846, −0.171010). It will have an amplitude of: $\sqrt{-0.469846^2 + -0.171010^2}$ = 0.5 units. It has a phase of arctan (−0.171010 ÷ 0.469846) = −20 degrees = 340 degrees. [To check if this is the correct arctan answer we want out of the two possible ones, we look at where the phase point is on the *constituent wave's* circle – it is in the bottom right-hand quarter of the circle. As 340 degrees is also in that quarter of the circle, 340 degrees is the correct answer.] The full formula for this constituent wave is:
"y = 0.5 sin ((360 * 0.5t) + 340)".

The 2.5-cycle-per-second Mean Level wave, if thought of as a circle, has a phase point at the coordinates (−3.010523, 0.343006). Therefore, the 2.5 cycle-per-second *constituent wave*'s circle has a phase point at the coordinates (0.343006, −3.010523). The constituent wave has an amplitude of:
$\sqrt{0.343006^2 + -3.010523^2}$ = 3.03 units. It has a phase of:
arctan (−3.010523 ÷ 0.343006) = −83.5 = 276.5 degrees. This phase point would be in the bottom right-hand corner of the constituent wave's circle, so this is the correct arctan answer. The full formula for this constituent wave is:
"y = 3.03 sin ((360 * 2.5t) + 276.5)"

The 7.5-cycle-per-second *Mean Level* wave's circle has a phase point at the coordinates (0, 6.06). Therefore, the *constituent* wave's circle has its phase point at the coordinates (6.06, 0). The constituent wave has an amplitude of: $\sqrt{6.06^2 + 0^2}$ = 6.06 units. It has a phase of arctan (0 ÷ 6.06) = 0 degrees. We do not really need to use arctan in this situation. The phase point on the circle has the coordinates (6.06, 0), so the angle is clearly 0 degrees. The full formula for this constituent wave is:
"y = 6.06 sin (360 * 7.5t)".

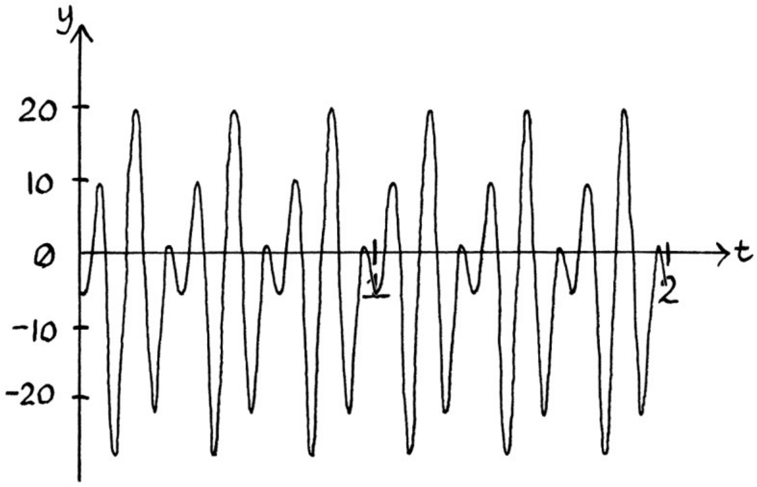
The 8.5-cycle-per-second Mean Level wave's circle has a phase point at the coordinates (0.771812, 1.845076). Therefore, the *constituent* wave's circle has a phase point at the coordinates (1.845076, 0.771812). The constituent wave has an amplitude of: $\sqrt{1.845076^2 + 0.771812^2}$ = 2 units. It has a phase of arctan (0.771812 ÷ 1.845076) = 22.7 degrees. The phase point on the circle is in the top right-hand quarter of the circle, so this is the correct arctan result. The full formula for this constituent wave is:
"y = 2 sin ((360 * 8.5t) + 22.7)"

After all that, we can say that the original signal is the sum of these waves:
“y = 0.5 sin ((360 * 0.5t) + 340)”
“y = 3.03 sin ((360 * 2.5t) + 276.5)”
“y = 6.06 sin (360 * 7.5t)”
“y = 2 sin ((360 * 8.5t) + 22.7)”

In fact, these waves are *exactly* the ones that were added to make the original signal, which shows that the process works.

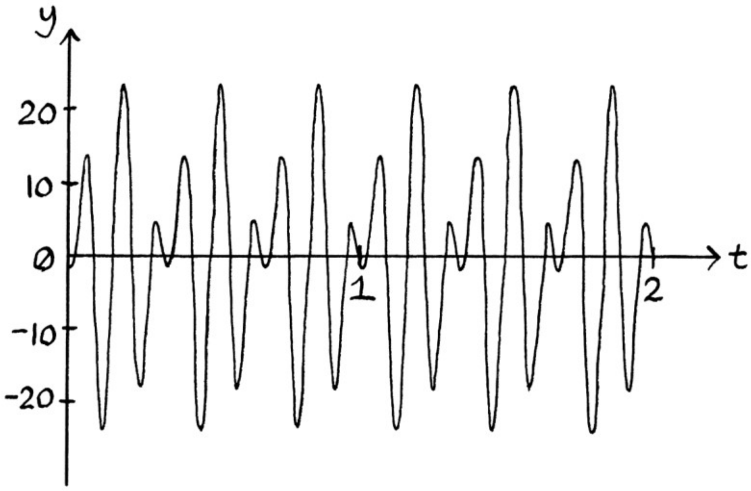
Example 3: a signal with a mean level

In this example, we will analyse a signal that has a non-zero mean level. We will analyse this signal:



Although it is hard to tell, the signal has a mean level of −4 units. The process of analysing a signal will not be as straightforward if the signal is not centred on the y-axis. Therefore, the first step is to remove the mean level from the signal (while making a note of it for later), and re-centre the signal on y = 0.

The centred signal looks like this:



Looking at the graph closely, we can see that it takes a third of a second to repeat its shape. Therefore, it has a frequency of 3 cycles per second. This means that the constituent waves that were added to create it would have had frequencies that were integer multiples of 3 cycles per second. They would have had frequencies from the following set: 3, 6, 9, 12, 15, 18, 21, 24, 27, 30, 33, 36, 39, 42, 45 and so on. We will use two test waves for each of these frequencies until we have found all the constituent waves.

The first pair of test waves will be:
"y = 2 sin (360 * 3t)"
... and:
"y = 2 sin ((360 * 3t) + 90)".

Multiplying the signal by the first of these test waves produces a resulting signal with a mean level of 0 units. Multiplying the signal by the second of these test waves also produces a resulting signal with a mean level of 0 units. This means that there is no wave of 3 cycles per second in the signal.

The second pair of test waves will be:
"y = 2 sin (360 * 6t)"
... and:
"y = 2 sin ((360 * 6t) + 90)".

Multiplying the signal by the first of these test waves produces a resulting signal with a mean level of 10.676306 units; multiplying the signal by the second of these test waves produces a resulting signal with a mean level of 4.978443 units. This means that there is definitely a wave with a frequency of 6 cycles per second in the signal. We will work out its characteristics later.

The third pair of test waves will be:
“y = 2 sin (360 * 9t)”
... and:
“y = 2 sin ((360 * 9t) + 90)”.

Multiplying the signal by the first of these test waves produces a resulting signal with a mean level of −12.163667 units; multiplying the signal by the second of these test waves produces a resulting signal with a mean level of −5.415613 units.

We carry on multiplying the original signal by pairs of waves with frequencies from our list. If we were calculating the characteristics of the constituent waves as we did this, we could add them up as we went along, and, if we were accurate enough, we would know when to stop when our sum matched the original signal. However, to make this explanation clearer, we will do the sum afterwards, and perform ten pairs of multiplications with the hope that that is enough to cover all the frequencies in the signal. A table showing the resulting mean levels for each pair of multiplications looks like this:

Frequency being tested	**Mean level of the result of the multiplication with a 0-degree phase test wave**	**Mean level of the result of the multiplication with a 90-degree phase test wave**
3	0	0
6	10.676306	4.978443
9	−12.163667	−5.415613
12	0	0
15	1.000000	0
18	0	0
21	0	0
24	0	0
27	0	0
30	0	0

From the table, we can see that there are non-zero mean levels for the frequencies of 6 cycles per second, 9 cycles per second and 15 cycles per second. This means that there are 3 constituent waves and these are their frequencies.

For the 6-cycle-per-second *Mean Level* wave, we know that the coordinates of the phase point of its circle will be at (4.978443, 10.676306). Therefore, the coordinates of the phase point of the corresponding *constituent wave* circle will be (10.676306, 4.978443). The radius of the circle, and therefore, the amplitude of the constituent wave will be $\sqrt{10.676306^2 + 4.978443^2}$ = 11.78 units. The angle of the phase point, and therefore, the phase of the constituent wave will be arctan (4.978443 ÷ 10.676306) = 25 degrees. As the phase point is in the top right-hand quarter of the circle, this is the arctan result we want. The formula for this constituent wave is:
"y = 11.78 sin ((360 * 6t) + 25)"

The coordinates of the phase point of the 9-cycle-per-second *Mean Level wave's* circle are (−5.415613, −12.163667). Therefore, the coordinates of the phase point of the corresponding *constituent wave's* circle will be (−12.163667, −5.415613). The radius of the circle, and therefore, the amplitude of the constituent wave will be $\sqrt{-12.163667^2 + -5.415613^2}$ = 13.3148 units. The angle of the phase point, and therefore, the phase of the constituent wave will be arctan (−5.415613 ÷ −12.163667) = 24 degrees. The phase point is in the bottom left-hand quarter of the circle, yet 24 degrees is in the top right-hand quarter of the circle. Therefore, this is *not* the arctan result we want, and we actually want 24 + 180 = 204 degrees. The formula for this constituent wave is:
"y = 13.3148 sin ((360 * 9t) + 204)"

The coordinates of the phase point of the 15-cycle-per-second *Mean Level wave's* circle are (0, 1). Therefore, the coordinates of the phase point of the corresponding *constituent wave* circle will be (1, 0). The radius of the circle, and therefore, the amplitude of the constituent wave will be $\sqrt{1^2 + 0^2}$ = 1 unit. The angle of the phase point, and therefore, the phase of the constituent wave will be arctan (0 ÷ 1) = 0 degrees. [We did not really need to use Pythagoras's theorem or arctan to calculate these, as we can know what the results will be by thinking about the circle.] The formula for this constituent wave is:
"y = sin (360 * 15t)"

[Note that we do not really need to think about the Mean Level waves when doing these calculations, but I am leaving that step in to emphasise how and why the method works.]

We can now say that the original signal is made up of the sum of these three constituent waves and a mean level: [Do not forget the mean level that we removed from the original signal earlier]

“y = 11.78 sin ((360 * 6t) + 25)”
“y = 13.3148 sin ((360 * 9t) + 204)”
“y = sin (360 * 15t)”
... and a mean level of −4 units.

In reality, the original signal was created by adding these waves:
“y = 2 + 11.78 sin ((360 * 6t) + 25)”
“y = −6 + 7 sin ((360 * 9t) + 186)”
“y = 7 sin ((360 * 9t) + 222)”
“y = sin (360 * 15t)”

In this addition, the two waves with a frequency of 9 cycles per second became added together to produce a wave with the formula:
“y = −6 + 13.3148 sin ((360 * 9t) + 204)”.
Once added together, it is impossible to tell that there were originally two separate waves with a frequency of 9 cycles per second.

The mean levels in the sum of waves also became combined, meaning that it is impossible to tell that there was anything other than one mean level.

Although we have not discovered the *actual* waves that were added together to make up the signal, we have discovered the waves after any additions were performed. It would be impossible to find the actual waves in this case. This situation is analogous to the buckets of unmixable liquids – we cannot tell how many containers of each liquid were used to fill a bucket, but we can see how much of each liquid is in the bucket. In this example, if we added together the waves we discovered through analysis, it would be an exact match for the original signal. However, those waves are not the exact set of waves that were added to make up the original signal.

The process has found as much information as it is possible to find, and we have a list of waves that when added would create the signal exactly.

Example 4: working from a picture

For interest's sake, we will find the constituent waves from a drawing of a signal. The accuracy of the results will be much less, but we will still be able to find an approximation of the constituent waves. [As this is a PDF, the dimensions of any image are dependent on how the PDF is viewed. Therefore, we will have to pretend that we are reading from a picture.]

To do this, we will need to be able to measure values from the signal, and we will need to measure values from some test waves. Instead of drawing test waves, it is easier and quicker to draw a circle and measure values from that. In this way, we can do the whole process without requiring a calculator (as long as we can do basic maths in our heads or on paper).

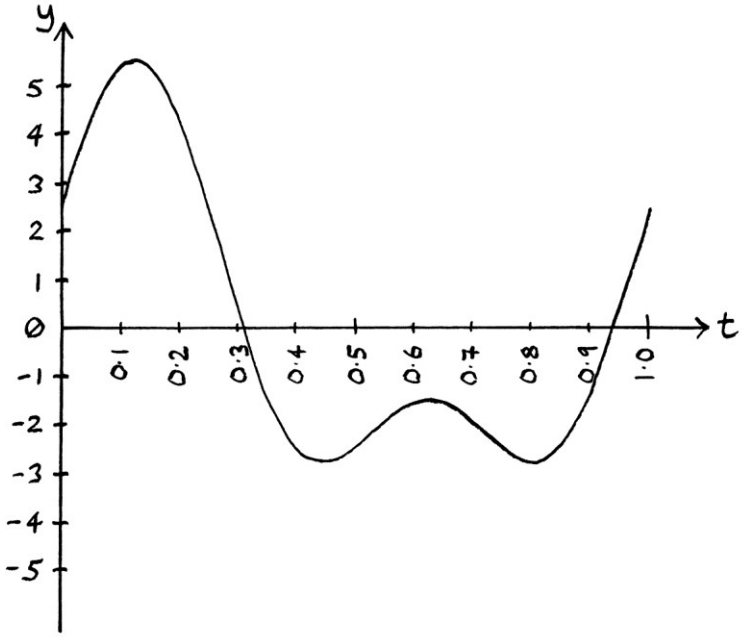
To calculate the constituent waves, we need a drawing of the signal for one cycle and a drawing of a 2-unit-radius circle.

If the signal is drawn with one cycle taking up 10 centimetres along the t-axis, then we will be able to read the scale of the t-axis fairly well. The length of 10 centimetres will represent 0.5 seconds; 1 centimetre will represent 0.05 seconds; 1 millimetre will represent 0.005 seconds. This means that we will be able to distinguish time values down to 0.005 seconds. [We will be able to distinguish between say, 0.055 and 0.060 seconds, but not between 0.059 and 0.060 seconds].

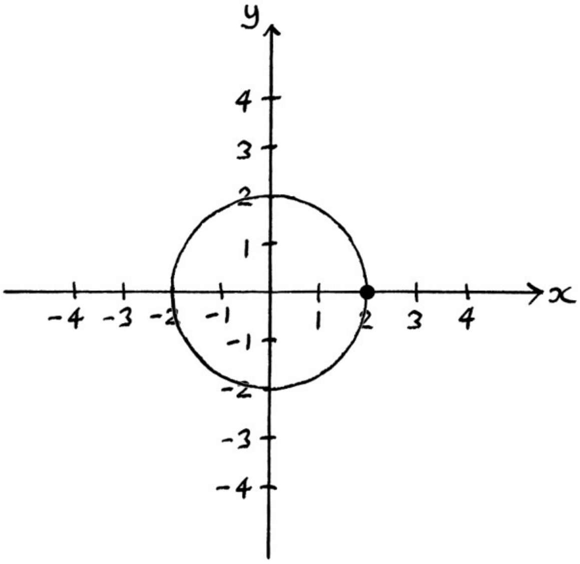
If the y-axis of the signal's graph extends ten centimetres above and ten centimetres below the line of the t-axis, and the signal is drawn so that its highest point reaches the top, or its lowest point reaches the bottom, then one millimetre will represent one hundredth of the maximum y-axis value. Half a millimetre will represent one two-hundredths of the maximum y-axis value. However, we will ignore such possible accuracy and instead just read values to one decimal place.

If the circle is drawn with a five-centimetre radius, we will be able to read Sine results off it to an accuracy of one decimal place.

We will analyse this simple signal (which, in this drawing, is not to the exact size mentioned above):



We will use this circle (which, again, is not drawn to its exact size):



The signal repeats once every second. We want the details of 1 cycle, so we will read points from the drawing of the signal at intervals of 0.05 seconds for 1 second. Then, we can use the readings in all our future calculations. Note that we do not read the value at 1 second as that is the start of the second cycle, and including it in calculations will make the results incorrect.

Time (seconds)	**y-axis value read from the drawing of the signal**
0.00	2.5
0.05	4.3
0.10	5.4
0.15	5.4
0.20	4.3
0.25	2.5
0.30	0.4
0.35	−1.4
0.40	−2.4
0.45	−2.8
0.50	−2.5
0.55	−1.9
0.60	−1.6
0.65	−1.6
0.70	−1.9
0.75	−2.5
0.80	−2.8
0.85	−2.4
0.90	−1.4
0.95	0.4

The signal repeats once a second, so it has a frequency of 1 cycle per second. This means that the list of test frequencies that we will use will be as so:
1, 2, 3, 4, 5, 6, 7, 8, 9, 10... and so on.

Our first pair of test waves will be “y = 2 sin (360t)” and “y = 2 sin ((360t) + 90)”. We can read the y-axis values for these waves from our picture of a circle with a radius of 2 units. The first of this pair of test waves will be the y-axis values from the circle. As “y = 2 sin ((360t) + 90)” is the same as “y = 2 cos (360t)”, we can read the values for the second of this pair from the x-axis values of our circle.
As our signal repeats its cycles in one second, and so do our test waves, we only need to read 1 second’s worth of values for our test waves. These will need to be spaced in the same way as the signal’s values, so we can multiply them together later. To be spaced at intervals of a twentieth of a second, the y-axis and x-axis values will need to read off the circle at intervals of 360 ÷ 20 = 18 degrees.

The table of values for each test wave is as follows:

Time (seconds)	**Values for "y = 2 sin 360t" (y-axis values from the circle)**	**Values for "y = 2 sin ((360t) + 90)" (x-axis values from the circle)**
0.00	0	2
0.05	0.6	1.9
0.10	1.2	1.6
0.15	1.6	1.2
0.20	1.9	0.6
0.25	2	0
0.30	1.9	−0.6
0.35	1.6	−1.2
0.40	1.2	−1.6
0.45	0.6	−1.9
0.50	0	−2
0.55	−0.6	−1.9
0.60	−1.2	−1.6
0.65	−1.6	−1.2
0.70	−1.9	−0.6
0.75	−2	0
0.80	−1.9	0.6
0.85	−1.6	1.2
0.90	−1.2	1.6
0.95	−0.6	1.9

If we multiply each point we have read from the signal against each point we have read from the first of this pair of test waves, and add up the results, we end up with 49.66. The average of this is 2.483. This is the mean level of the result of multiplying the original signal by the first of this pair of test waves. We have calculated this by multiplying equally spaced corresponding points of the signal and the wave together. We can be sure that this result will be slightly inaccurate, but by how much, we will not know until later.

If we multiply each point we have read from the signal against each point we have read from the second of this pair of test waves, and add up the results, we end up with 49.66 again. The average of this is 2.483, as before. This is the mean level of the result of multiplying the original signal by the second of this pair of test waves.

The two results are the coordinates of our Mean Level wave's circle's phase point. The coordinates are (2.483, 2.483). Without doing any maths, we can tell that this point is at 45 degrees because the x-axis and y-axis coordinates are the same. The phase point of the constituent wave will have these coordinates the other way around, which, because the coordinates are the same, are also (2.483, 2.483). The radius of the circle, and the amplitude of the constituent wave will be $\sqrt{2.483^2 + 2.483^2}$ = 3.5115 units. Because all of our readings were to 1 decimal place, we will round this up to 1 decimal place, and we end up with an amplitude of 3.5 units. The angle of the phase point is clearly 45 degrees, and arctan would confirm this.

From all of this, we can say that our first constituent wave is:
"y = 3.5 sin ((360t + 45)"

We should be aware that problems with accuracy might mean this is slightly (or completely) wrong. The way to tell is to add up the found constituent waves afterwards and see if they match the original signal.

The second pair of test waves will be:
"y = 2 sin (360 * 2t)"
... and:
"y = 2 sin ((360 * 2t) + 90)"

We can read the values for these waves from our picture of a circle with a radius of 2 units. We still only need to read one second's worth of values for our test waves, but as these waves have twice the frequency as before, we will now be reading two cycles and not one cycle in that second. An object rotating around a circle at 2 cycles per second will complete 360 degrees every half second, which is 720 degrees every one second. To be spaced at intervals of a twentieth of a second, the y-axis and x-axis values will need to be read off the circle at intervals of 720 ÷ 20 = 36 degrees. [Strictly speaking, we only need to read off one cycle's worth, and then copy the values for that cycle for the second cycle because we know that they will be the same].

The table of values for each of these test waves is as follows:

Time (seconds)	**Values for "y = 2 sin (360 *2t)" (y-axis values from the circle)**	**Values for "y = 2 sin ((360 * 2t) + 90)" (x-axis values from the circle)**
0.00	0	2
0.05	1.2	1.6
0.10	1.9	0.6
0.15	1.9	−0.6
0.20	1.2	−1.6
0.25	0	−2
0.30	−1.2	−1.6
0.35	−1.9	−0.6
0.40	−1.9	0.6
0.45	−1.2	1.6
0.50	0	2
0.55	1.2	1.6
0.60	1.9	0.6
0.65	1.9	−0.6
0.70	1.2	−1.6
0.75	0	−2
0.80	−1.2	−1.6
0.85	−1.9	−0.6
0.90	−1.9	0.6
0.95	−1.2	1.6

If we multiply each value for the first of this pair of test waves in this table against the corresponding value from the signal in the first table, and take the average of the results, we end up with 2.02. If we do the same for each value for the second of this pair of test waves, we end up with 0.

These two values are the coordinates of the phase point of our Mean Level wave's circle: (0, 2.02). We swap them around to find the coordinates of our constituent wave's circle: (2.02, 0). As the phase point is on the x-axis, we do not need to use Pythagoras's theorem or arctan – we know that the radius of the circle, which is the amplitude of the constituent wave, is 2.02 units. The angle of the phase point, which is the phase of the constituent wave, is 0 degrees. To allow for the accuracy we have been using, we will round up the amplitude to 1 decimal place, and say it is 2.0 units, which we will call 2 units.

The constituent wave is therefore:
“y = 2 sin (360 * 2t)”

If we added up the constituent waves we have calculated so far [which are the waves “y = 3.5 sin ((360t + 45)” and “y = 2 sin (360 *2t)”], we would see that the sum is equal to the original signal. Therefore, we have found the constituent waves, and we do not need to do any more testing. The original signal is made up of the sum of:
“y = 3.5 sin (360t + 45)”
... and:
“y = 2 sin (360 * 2t)”

In fact, these are exactly the waves that were added to make the original signal. This shows that the process can work even if we are reading values off a picture. However, it is very important to note that we only found the exact waves by rounding up the answers, and it just so happened that the actual constituent waves did not have values with more than 1 decimal place in them. If the first constituent wave had had an amplitude of 3.51 units, we would never have been able to know that for sure, given the accuracy of our readings. Likewise, if its phase had been 44.95, we would not have been able to tell if we had found the right answer or not.

To find answers with more accuracy, we would need to take more readings and to measure them more accurately. When reading values from pictures, this is difficult to do. When analysing waves that have been gathered in other ways, it is still something to bear in mind.

Because our constituent waves had the low frequencies of 1 cycle per second and 2 cycles per second, we avoided a potential problem with the process. If we had needed to test much faster waves, we would have found that the spacing of the values we were using would become less adequate. Reading values every 0.05 seconds is fine for frequencies of 1 cycle per second or 2 cycles per second, but would end up giving *wrong* results for higher frequencies. We can see how this is so by thinking about what would happen if we had wanted to test a frequency of 10 cycles per second. An object rotating around a circle at 10 cycles per second would complete 360 degrees once every 0.1 seconds. Therefore, if we only took readings from the circle every 0.05 seconds, either every y-axis reading would be from when the object was at 0 degrees, or it would be from when the object was at 180 degrees – both of these would be values of 0 units. Therefore, every value taken for the first of the pair of test waves would be zero. The values for the second of the pair of test waves would all be either +2 units or −2 units. The table would look like this:

Time (seconds)	**Values for "y = 2 sin (360 * 10t)" (y-axis values from the circle)**	**Values for "y = 2 sin ((360 * 10t) + 90)" (x-axis values from the circle)**
0.00	0	2
0.05	0	−2
0.10	0	2
0.15	0	−2
0.20	0	2
0.25	0	−2
0.30	0	2

... and so on.

This means that our record of the first test wave is not that of a Sine wave with a 10 cycle per second frequency, but that of a Sine wave with zero frequency and zero phase. Our record of the second test wave only shows two points from each cycle of the wave. The limited number of values we have do not reflect the characteristics of the two waves, and therefore, any tests done using these values will produce incorrect results. If our original signal had contained a wave of 10 cycles per second, we would not have been able to find it using the time intervals we were using.

The problem also exists in our reading of the signal, although this is much less obvious. Supposing our signal had contained a zero-phase constituent wave of 10 cycles per second, there would be no trace of it if we had used readings at intervals of 0.05 seconds.

As it is, we could have had read more time values from the signal and the circle for a cycle. If the time axis of the drawing of the signal were 10 centimetres long, then it would be possible to read values along it spaced at 0.005 seconds [They would be one millimetre apart]. There would be 200 readings for 1 second. We can just about read y-axis and x-axis values from the circle at these intervals, but we lose some accuracy – at a frequency of one cycle per second, we would need to distinguish between angles of 360 ÷ 200 = 1.8 degrees. At higher frequencies, the angles would be larger, so it would be less of a problem. Using 200 readings for one second would require a lot more work, but we would be able to deal with higher frequencies *up to a certain level*. There will always become a point when the frequency of the test wave we want to use requires a higher interval in our test wave readings than we can achieve. At that point, our record of the test wave does not reflect the actual test wave, so the process will fail.

Obviously, it is rare that will ever need to analyse a drawing of a signal, but the problems explained here play a huge part in the storage and analysis of discrete waves as we will see in Chapter 42.

Thoughts so far

Mean levels

In practice, if we were doing the process with real-world waves, it might be difficult to calculate the mean level. It all depends on the form in which the waves are being presented to us. If we were dealing with "discrete waves", where the waves are treated as a series of y-axis values at equally spaced moments in time, and we had the waves in a file on a computer, then the mean level would be easy to find. We would just average the given y-axis values for one cycle of the signal. Supposing we had a picture of a wave, we could read y-axis values off it at evenly spaced moments of time, and work out the average in that way. The accuracy would be inferior, but it would still be adequate for some purposes.

Similarly, it is harder to do a multiplication on a signal if the signal is not in the form of a discrete wave.

Number of test frequencies

In Example 2 of this chapter, we used 20 test frequencies (40 test waves), and in Example 3, we used 10 test frequencies (20 test waves). This was mainly to make the explanation easier to follow. We could have added up the calculated constituent waves as we went along, and when the sum matched the original signal, we would know when to stop. For Example 2, we would have needed to test 17 frequencies; in Example 3, we would have needed to test 5 frequencies. In the real world, a lack of accuracy in reading the original signal will mean that we are unlikely to find a sum of calculated constituent waves that exactly matches the original signal. Therefore, knowing when to stop testing for ever-higher frequencies becomes an arbitrary judgement. It becomes a compromise between wanting accuracy and wanting to move on to the next stage in whatever it is we are doing. If, for example, we always tested a million frequencies every time we analysed a signal, we would be more likely to find all the constituent frequencies, but with the downside that most of our testing would be a waste of effort. On the other hand, if we only ever tested 10 frequencies, we would be more likely to miss out constituent frequencies. What is best to do depends on the situation and what it is we want to achieve.

Accuracy

If we were dealing with real-world waves, it is unlikely that the accuracy of the results would ever be perfect. The slightest inaccuracy at any point in observing the wave, storing the wave, and performing all the stages of maths on the wave, will cause the results to stray slightly from what they should be.

Periodic signals

It pays to remember that this method only works with periodic signals. It will not work if the signals do not repeat. One obvious reason for this is that we cannot make a list of test frequencies if we have no frequency on which to base them.

Mean Level wave and constituent wave

At the moment, we are calculating the Mean Level wave's circle's phase point, and then switching the coordinates to find the constituent wave's circle's phase point. Doing it this way helps reinforce how and why this all works. Of course, if we were doing this efficiently, we would skip straight to using Pythagoras's theorem and arctan, and we would not bother thinking about the Mean Level wave or its circle's phase point at all.

Fourier

The process of finding the constituent waves of a periodic signal in this way is usually called "Fourier series analysis", after Jean Baptiste Joseph Fourier, who made progress in this field of maths. The essence of Fourier's ideas is that any periodic signal can be treated as if it were the sum of pure waves.

[Fourier's actual argument was that any periodic signal can be treated as the sum of *pairs* of Sine waves and Cosine waves with zero phase. In other words, instead of ending up with constituent waves, we would end up with pairs of Sine and Cosine waves with zero phase that when added together would make up those constituent waves. This amounts to exactly the same thing as ending up with constituent waves, but it is a slightly less useful result. Such an idea was discussed in Chapter 15 on the frequency domain.]

Nowadays, some people argue that Fourier's ideas about periodic signals was completely correct, while most people argue that he was correct with certain exceptions. In my view, a square wave is obviously one of those exceptions.

The general term "Fourier analysis" is used to refer to the analysis of both periodic and aperiodic signals. Fourier *series* analysis is a type of Fourier analysis. Sometimes, people use the term "Fourier analysis" to refer to *any* method of finding the constituent waves, even if it has no basis in Fourier's work.

The term "Fourier series" is used to refer to the waves that if added together would recreate a given periodic signal. It is just another term for the list of constituent waves that make up a signal. However, the constituent waves in a Fourier series are usually given in terms of the zero-phase Sine waves and zero-phase Cosine waves that, if added together, would produce each constituent wave. The term "Fourier series" is also used as shorthand for "Fourier series analysis" (so in other words it is used to refer to the actual process that calculates the constituent waves of a periodic signal).

[The "Fourier transform" is the name of a process that finds the constituent waves of an *aperiodic* signal. It is more complicated than the Fourier series. Some people think of the Fourier series as being a type of Fourier transform, while other people think of the Fourier series as being different from the Fourier transform. Some people do not distinguish between them. This confusion means that some seemingly promising explanations of the Fourier transform unhelpfully turn out just to be explanations of Fourier series analysis. Fourier series analysis is straightforward to understand and execute, but it takes a while to explain. The Fourier transform and all its variations are straightforward to execute, but take much longer to explain.]

Most explanations of Fourier series analysis involve giving complicated mathematical formulas that explain what it is, while spending very little time explaining the process itself. The excessive formulas are what make it seem complicated, and are the reason why so many people, even many who use it, do not understand it. If you can understand the multiplication of waves and the idea that different frequencies do not mix, then you have enough information to understand Fourier analysis of periodic signals.

Analysis and synthesis

If we analyse a signal to find the constituent waves, the process is called “analysis” (obviously). If we recreate a signal by adding up its constituent waves, the process is called “synthesis”.

Infinitely long signals

Often in explanations of Fourier series analysis, the signals and constituent waves are treated as if they repeat forever. The signals and waves are considered to exist for an infinitely long time. It is really a personal choice if you want to think in this way, and in practice, we only need one cycle of the original signal to analyse it.

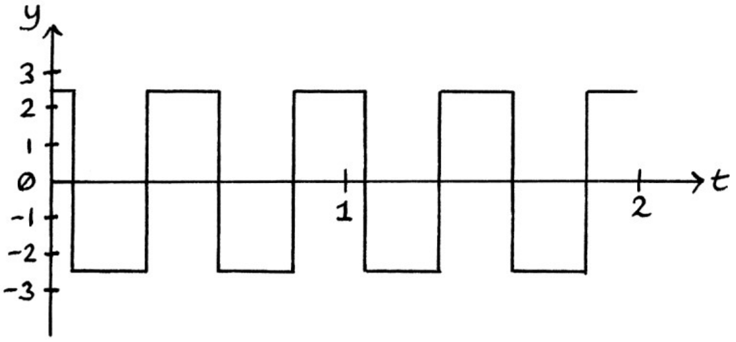
Related to this, some people say that a Sine wave or a Cosine wave *always* exists for an infinite amount of time. This is not true, as a wave exists for the length of time of the entity it is describing. An object rotating around a circle for one second will produce a Sine wave and a Cosine wave that last for one second. The movement of the end of a piston for half an hour can be described using a Sine wave that is half an hour long. Having said all that, when dealing with theoretical formulas and ideas, sometimes it can be useful to think of waves as repeating forever.

Signals that are not literally the sum of waves

So far in this chapter, we have looked at signals that were created by adding pure waves together. The signals were literally the sums of pure waves. The analysis process also works on periodic signals that were not literally created by adding pure waves together. It can even be used on periodic signals that could not have been created by adding pure waves together, but in such cases, at best it will find an approximate sum of constituent waves.

Square waves

We will use the process on this square wave to see what happens:



The above square wave is not a pure wave. [Square waves might best be called “square signals”, but they are too well known for their name to be changed.] If we were to categorise square waves in the same way that we would pure waves, we would say that this has an amplitude of 2.5 units, a frequency of 2 cycles per second, a phase of 125 degrees and a mean level of zero units.

To find the waves that could have been added to create this signal, we will use the same process as before. First, we make a list of the frequencies that could possibly be in the signal. As the square wave has a frequency of 2 cycles per second, we know that the list will consist of frequencies that are multiples of 2. Therefore, the waves that make up this signal must have frequencies in the following list:
2, 4, 6, 8, 10, 12, 14, 16, 18, 20... and so on.

Next, we multiply the signal by pairs of test waves. One thing to note in this example is that we will start to see variations in the accuracy of our results. The later digits after the decimal points of the resulting mean levels might be completely different depending on how accurately we measure the waves, or how accurately the maths is done. For this reason, giving the final results to many decimal places is not really appropriate, as the later digits might be wrong.

The first pair of test waves will be:
"y = 2 sin (360 * 2t)"
... and:
"y = 2 sin ((360 * 2t) + 90)"

Multiplying the signal by the first of this pair produces a resulting signal with a mean level of −1.82569362 units. To be more pedantic, I should say, *when I do the calculation using the readings of the waves that I have and the tools I use to do the calculation, I end up with −1.82569362 units*. [If you create the same square wave in a computer program or Excel, and then try to analyse it yourself, your results might be slightly different, but you should find something between −1.82 and −1.83.] When we have calculated the final constituent waves, it is worth rounding up the values so that they reflect the achievable accuracy. The concept of accuracy and how to deal with it becomes much more evident in this example.

When I multiply the signal by the second of these test waves, I end up with a resulting signal with a mean level of 2.60748177 units. These two mean levels mean that there is definitely a constituent wave with a frequency of 2 cycles per second in the sum. We will calculate its details later.

The next pair of test waves will be:
"y = 2 sin (360 * 4t)"
... and:
"y = 2 sin ((360 * 4t) + 90)"

When I multiply the signal by the first of these test waves, I produce a resulting signal with a mean level of 0.00009788 units. When I multiply the signal by the second of these two test waves, I produce a resulting signal with a mean level of −0.00003563 units. Although, these results are not literally zero, they are very small in comparison to the previous mean level results. Therefore, we will consider them both as being zero. If we had perfect accuracy, they would be zero, but perfect accuracy is not possible to achieve in this situation. All this means that my calculations in this example are only valid up to about 3 decimal places. As these two mean levels are zero (or are going to be treated as being zero), we know that there is no constituent wave with a frequency of 4 cycles per second in the original signal.

The next pair of test waves will be:
“y = 2 sin (360 * 6t)”
... and:
“y = 2 sin ((360 * 6t) + 90)”

When I multiply the signal by the first of this pair, I end up with a resulting signal with a mean level of 1.02489710 units. When I multiply the signal by the second of this pair, I have a resulting signal with a mean level of 0.27454846 units. Therefore, 6 cycles per second is the frequency of one of the constituent waves.

The next pair of test waves will be:
“y = 2 sin (360 * 8t)”
... and:
“y = 2 sin ((360 * 8t) + 90)”

Multiplication by the first of these test waves produces a signal with a mean level of −0.00006696 units. Multiplication by the second test wave produces a signal with a mean level of −0.00007979 units. These are small enough mean levels to be treated as zero. Therefore, there is no wave with a frequency of 8 cycles per second in the constituent waves.

We continue in this way until we have done the arbitrary number of 10 pairs of test waves. Unlike with the previous examples, if we had continued testing more frequencies, we would have found more constituent waves. For a square wave such as this, there are really an infinite number of constituent waves, with each one having a smaller amplitude than the one before it. Where to stop is an arbitrary decision.

The table of results looks like this:

Frequency being tested	**Mean level of the result of the multiplication with a 0-degree phase test wave**	**Mean level of the result of the multiplication with a 90-degree phase test wave**
2	−1.82569362	2.60748177
4	0.00009788	−0.00003563
6	1.02489710	0.27454846
8	−0.00006696	−0.00007979
10	−0.05555425	−0.63419118
12	−0.00005207	0.00009022
14	−0.41209454	0.19223946
16	0.00010259	0.00001807
18	0.25013696	0.25003875
20	−0.00001811	−0.00010258

If we give the values in the table to 3 decimal places, it is more obvious as to which frequencies exist in the signal:

Frequency being tested	**Mean level of the result of the multiplication with a 0-degree phase test wave**	**Mean level of the result of the multiplication with a 90-degree phase test wave**
2	−1.826	2.607
4	0	0
6	1.025	0.275
8	0	0
10	−0.056	−0.634
12	0	0
14	−0.412	0.192
16	0	0
18	0.250	0.250
20	0	0

[Despite having just given the values to 3 decimal places, we will perform the following calculations with 8 decimal places.]

In the following calculations, we will skip the idea of the Mean Level wave's circle, and go straight to the constituent wave's circle.

The first constituent wave's circle's phase point is at (−1.82569362, 2.60748177). The radius of the circle, which is the amplitude of the wave, is:
$\sqrt{-1.82569362^2 + 2.60748177^2}$ = 3.18309886 units. The angle of the phase point is arctan (2.60748177 ÷ −1.82569362) = −55.00124993 = 304.99875007. The phase point is in the top left quarter of the circle, so the arctan result we want is the other possible one, which is 304.99875007 – 180 = 124.99875007 degrees. Giving the result to 2 decimal places, we will say that the full formula of this constituent wave is:
"y = 3.18 sin ((360 * 2t) + 125)"

Note how this has the same phase and frequency as the square wave we are analysing. Also, notice how it has a higher amplitude than the square wave. When it comes to adding all the constituent waves together, the later waves will reduce the shape of this Sine wave and flatten its top to make it more resemble the square wave.

The next wave to calculate is the one at 6 cycles per second. The constituent wave's circle's phase point will be at (1.02489710, 0.27454846). This circle has a radius of $\sqrt{1.02489710^2 + 0.27454846^2}$ = 1.06103295 units, so the constituent wave has an amplitude of 1.0613295 units. The phase point is at an angle of:
arctan (0.27454846 ÷ 1.02489710) = 14.99625020 degrees. The phase point is in the top right-hand quarter of the circle, so this is the angle that we want. Giving the result to 2 decimal places, the full formula of this constituent wave is:
"y = 1.06 sin ((360 * 6t) + 15)"

The next wave to calculate is the one at 10 cycles per second. The constituent wave circle's phase point will be at (−0.05555425, −0.63419118). The amplitude of the constituent wave is 0.63661977 units; its phase is 264.99374986 degrees. The formula for the wave is:
"y = 0.64 sin ((360 * 10t) + 264.99)"

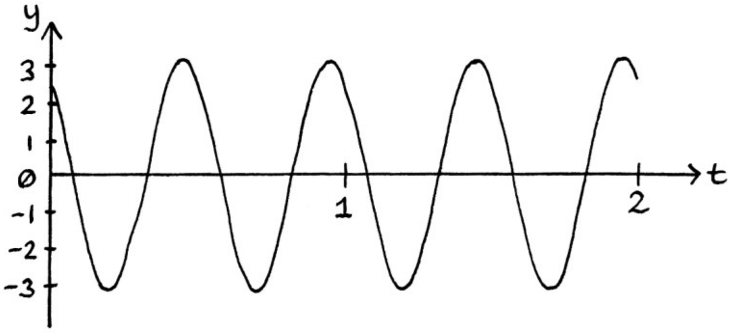
We will skip the details of the later calculations. The full list of constituent waves, discovered by testing for 10 frequencies, is as follows:

"y = 3.18 sin ((360 * 2t) + 125)"
"y = 1.06 sin ((360 * 6t) + 15)"
"y = 0.64 sin ((360 * 10t) + 264.99)"
"y = 0.45 sin ((360 * 14t) + 154.99)"
"y = 0.35 sin ((360 * 18t) + 44.99)"

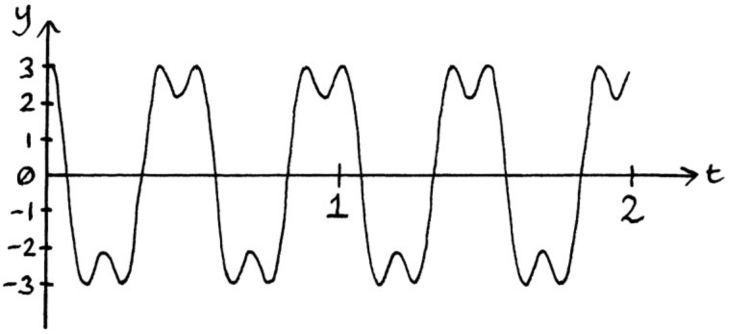
Notice how the amplitude of each successive wave is less than the one before. This is caused by the nature of a square wave, and would not necessarily happen when doing this process with other signals.

We will add up each of these discovered constituent waves one by one. To start with, we have:

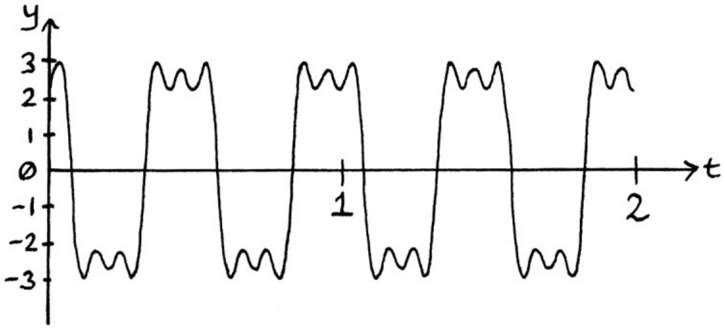
"y = 3.18 sin ((360 * 2t) + 125)":



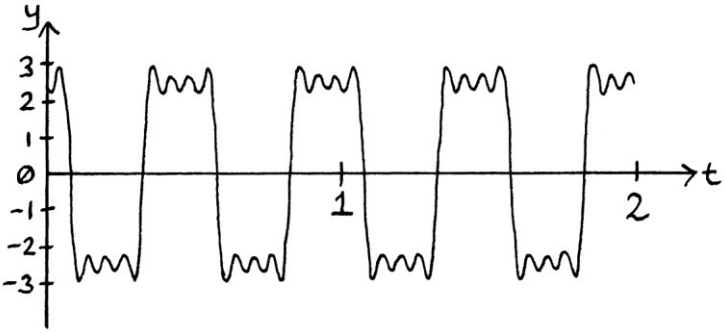
The result of adding on "y = 1.06 sin ((360 * 6t) + 15)" is this:



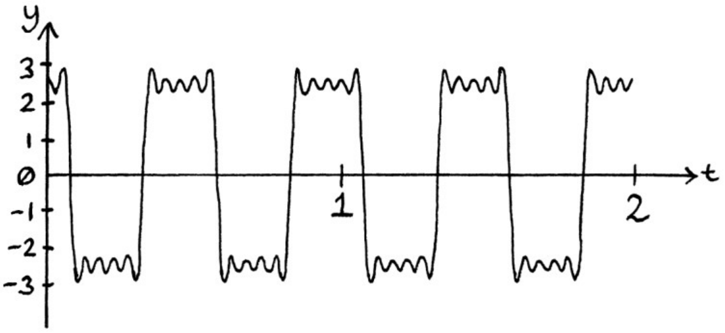
After we have added on “y = 0.64 sin ((360 * 10t) + 264.99)”, we have this:



After we add “y = 0.45 sin ((360 * 14t) + 154.99)”, we have this:

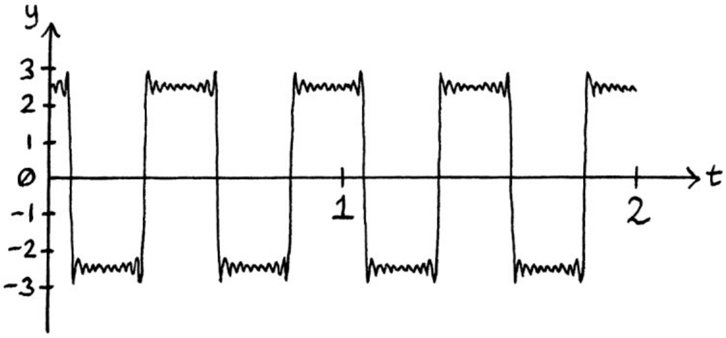


After we add “y = 0.35 sin ((360 * 18t) + 44.99)”, we end up with this:



We started with a Sine wave with the same frequency and phase of the square wave, and then with each extra added wave, the resulting signal becomes slightly more like our square wave. However, the result is only an approximation to the square wave. It will become more accurate as each new wave is added, but no matter how many constituent waves we find and add to the sum, the sum will never be exactly correct. The most obvious sign of this is in the ripples around the edges of the peaks. [The ripples are called the “Gibb’s phenomenon” after the mathematician Josiah Willard Gibbs].

Supposing we had not stopped analysing the signal after 10 pairs of test waves, but instead had done 20 pairs of test waves, the resulting sum would have looked like this:



It more closely resembles a square wave, but it is still not perfect. We could try a million pairs of test waves, but the result would still not be perfect.

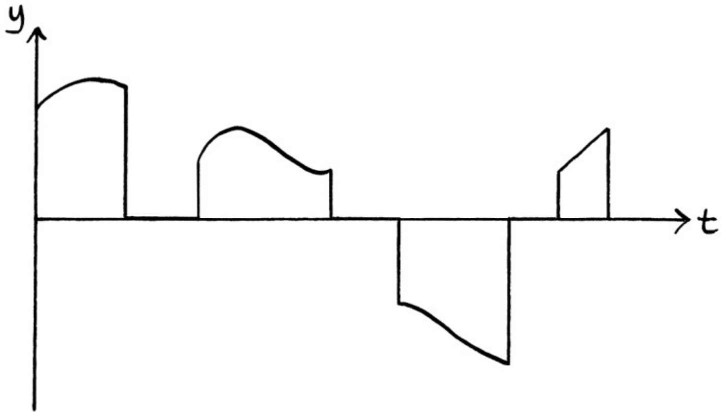
A square wave is a good example of the abilities and limitations of the process. We cannot analyse a square wave to find the list of waves that could be added to create it because it is not possible to portray a square wave as a sum of waves. A square wave is not a sum of waves. To treat it as being a sum of waves is really an attempt to draw straight lines using curves. However, we can analyse a square wave to find an approximation, and for much of the world of waves, this is sufficient.

Square waves are a good example of how not all periodic signals are the sum of two or more pure waves. Strangely, the imperfections in recreating a square wave are frequently overlooked, and square waves are often given as an example of the opposite being true.

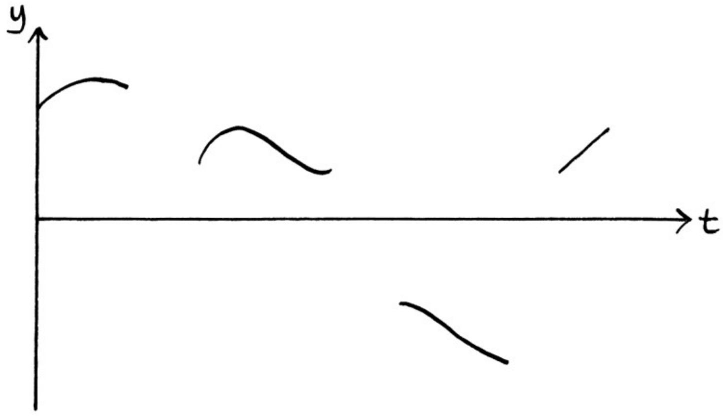
Sudden jumps

Any signal with instant drops or rises will be impossible to portray accurately as the sum of pure waves. When recreating signals by adding pure waves, only gradual rises and drops can be reasonably well portrayed as sums. Instant vertical jumps in a signal's curve are often called "discontinuities".

Another example of a curve containing instant vertical jumps is as so:



Occasionally, such a graph might appear as follows, where the vertical lines are not drawn, and it is not obvious that the gaps between the curved parts are horizontal lines at y = 0. [Such graphs are commonly drawn by computer programs.]



Other ways to think about the process

In this chapter, I have explained how to analyse a periodic signal to find the constituent waves in a way that is consistent with what we have learnt so far in this book.

Usually, Fourier series analysis is explained in slightly different ways. Here, I will explain how to understand other common ways.

Test waves: Sine and Cosine

I gave each pair of test waves in the form of a Sine wave with no phase, and a Sine wave with a 90-degree phase. Often, you will see these given in terms of a Sine wave and a Cosine wave. Therefore, where I would give these formulas for 2-cycle-per-second test waves:
“y = 2 sin (360 * 2t)”
“y = 2 sin ((360 * 2t) + 90)”
... other people might give these formulas instead:
“y = 2 sin (360 * 2t)”
“y = 2 cos (360 * 2t)”

More usually, you will see the test waves given with amplitudes of 1 unit, and the results will be doubled afterwards. Therefore, the test waves for 2 cycles per second will be:
“y = sin (360 * 2t)”
“y = cos (360 * 2t)”

For the purposes of teaching people who are trying to understand waves, I think it is better to give both test waves in terms of Sine. Particularly when it comes to Fourier series analysis, there is a risk that people start to think of Sine and Cosine as being completely different entities or having different properties. There is nothing special about a Cosine wave with zero phase – it is just a Sine wave with a 90-degree phase.

Another reason for using Sine is that it fits in with the ideas behind recreating the circle in Chapter 12. I also think it pays to get used to non-zero phases. Some people seem to have a fear of non-zero phases.

Giving the results in terms of two separate waves

Frequently in lessons on waves, you will see the results for each constituent wave given in terms of a Sine wave with zero phase and a Cosine wave with zero phase. This is instead of taking the final step and giving the actual amplitude and phase of the constituent wave.

In Chapter 15 on the frequency domain, in the section on "Phase point coordinate pairs", we saw how it was possible to describe a single wave in terms of graphs with two entries for each wave. Each wave was represented by two lines either side of the frequency of that wave. This enabled the graph to show waves with non-zero phases on a simple wave frequency domain graph. There are two ways of thinking about those lines:

- We could think of them as showing the coordinates of the phase points of the circle from which that wave was, or could have been, derived – one line represented the x-axis coordinate, and the other line represented the y-axis coordinate.

- We could think of them as showing the amplitudes of a Sine wave with zero phase and a Cosine wave with zero phase – one line represented the Sine wave's amplitude, and the other line represented the Cosine wave's amplitude. These waves are added together to make the wave indicated on the graph.

Both ways of thinking amount to exactly the same thing. If the first line represents the phase point's x-axis coordinate, and the second line, the phase point's y-axis coordinate, then it would be the same as if the first line represented a zero-phase Sine wave's amplitude, and the second line represented a zero-phase Cosine wave's amplitude. If we treat the lines as phase point coordinates, then we use Pythagoras's theorem and arctan to calculate the amplitude and phase of the wave. If we treat the lines as a Sine wave with zero phase and a Cosine wave with zero phase, then we add them together, and doing that involves Pythagoras's theorem and arctan to calculate the resulting amplitude and phase. Both ways of thinking use exactly the same maths and produce exactly the same result.

The use of such a frequency domain graph demonstrates that it is possible to portray a wave with any phase in terms of one Sine wave with zero phase, and one Cosine wave with zero phase.

Often when you see Fourier series analysis, you will see the results given, not as a single wave with any phase, but as a Sine wave with zero phase and a Cosine wave with zero phase. This is equivalent to leaving the result as the phase point of the constituent wave's circle and stopping there.

As described earlier in this chapter, if we do Fourier series analysis the long way to find one frequency, we multiply the original signal by countless waves with the same attributes but a range of phases. The results produce a Mean Level wave that indicates the characteristics of the constituent wave. The phase of the constituent wave is the place on the θ-axis of the Mean Level wave where the y-axis value is highest. The amplitude is that y-axis value. A shortcut is to just multiply the signal by a Sine wave with zero phase and a Sine wave with a 90-degree phase, and use these to describe the coordinates of the phase point of the Mean Level wave's circle (from which we could reconstruct the entire Mean Level wave). The phase point of the constituent wave's circle will be the same but with the coordinates swapped.

The more common explanation of Fourier series analysis completely skips over the fact that we are taking a shortcut for multiplying a range of phases. It does not mention that there is a Mean Level wave at all. It just says that we are multiplying the signal by a Sine wave and a Cosine wave, without any of the background as to why we would want to do that. The result of the multiplication by the Sine wave, instead of being treated as the y-axis coordinate of a Mean Level wave's circle's phase point, or even as the x-axis coordinate of a constituent wave's circle's phase point, is treated as the amplitude of a Sine wave with zero phase. The result of the multiplication by the Cosine wave, instead of being treated as the x-axis coordinate of a Mean Level wave's circle's phase point, or even as the y-axis coordinate of a constituent wave's circle's phase point, is treated as the amplitude of a Cosine wave with zero phase. Therefore, instead of ending up with a phase point, we end up with a Sine wave with zero phase and a Cosine wave with zero phase. These two waves, when added together will result in the actual constituent wave. The process to add these together is the same one as that to find the constituent wave from its phase point – in other words, we use Pythagoras's theorem and arctan, and on exactly the same values, and in exactly the same way.

To summarise all of this, the more common explanation for Fourier series analysis neglects to account for why we are doing multiplications or what the results mean. It also gives the result in a way that, although it has the same ultimate meaning, does not relate to the actual purpose behind the method. I am sure this is one of the main reasons that Fourier series analysis is considered complicated, when really it should be considered simple but requiring a lengthy explanation.

A description instead of an explanation

Usually, the *concept* of the Fourier analysis of periodic signals is given in terms of a formula such as this one:

$$f(t) = h + \sum_{n=1}^{\infty} a_n \sin\left((360 * f_n * t) + \phi_n\right)$$

This looks complicated at first glance, but if we go through it step by step, it will become much easier to understand.

The first thing to realise is that this is a *description* of the idea that any periodic signal can be said to be made up of the sum of pure waves with various amplitudes, frequencies, and phases. This formula does not help us in analysing a signal to find its constituent waves, or explain how the process for doing so works. To use the analogy of different liquids in a bucket, this is a bit like announcing, “Any bucket of liquids contains a number of different unmixable liquids.” [Being more specific, it is like saying, “Any bucket of liquids can be said to contain up to an infinite number of different unmixable liquids”.] Usually, this formula will be given alongside other formulas that indicate how the process works – therefore, although the formula does not tell us how to analyse signals, it is usually part of a set of formulas that does. [I will introduce the other formulas in Chapter 30 on calculus.]

The second thing to realise is that this is a very vague description – what the formula is *intended* to mean cannot be deduced from what the formula actually says. Several aspects of the formula require an explanation, even if you understand what similar mathematical formulas mean.

The “f(t)” in the formula essentially means “any function that is working on increasing values of time”. Although it is not explicitly stated, here it is referring to any time-based *periodic* signal. It is saying that any time-based periodic signal is equivalent to the following part of the equation.

The “h” stands for mean level. This is the abbreviation for mean level that I have been using in this book. [Other people might use other symbols such as “DC”, which is short for “direct current”, even if the signal has no connection to electricity.] Given that we remove the mean level from a signal before we start analysing it, this refers to that original overall mean level. This mean level is added to the rest of the equation.

The “Σ” is the upper-case Greek letter sigma, which is the Greek equivalent to the Latin letter “S” for “Sierra”. In this context, it is being used as shorthand to indicate a series of sums – it means that we will start with the variable “n” as the number 1, and fill in each “n” in the rest of the formula with the number 1. Once we have worked out what the formula would be with “n” as 1, we add that to whatever the result of the formula would be if “n” were 2, and then we add all that to whatever the formula would be if “n” were 3. We continue in this way until “n” has risen to infinity (which of course can never happen because it would take forever). The “n” will always represent an integer, and it will always increase in steps of 1.

This simpler example of a sum:

$$\sum_{n=1}^{10} n$$

... results in the numbers from 1 to 10 being added together. It is equivalent to the following sum:
1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10

In a C program, such an idea might be given by something along these lines:

```
result = 0;
for(n = 1; n <= 10; n++)
{
  result = result + n;
}
```

The Fourier formula might appear as so in a C program:

```
result = 0;
for(n = 1; n < 99999999; n++)
{
  result = result + ResultOfProcedureInvolvingn(n);
}
```

... where 99999999 is just a very high number as a replacement for infinity.

The “Σ” sum is called a “sigma sum” on account of it using the Greek letter sigma. If you are ever confused by a complicated “Σ” sum, write out the first few parts of the addition in full to see what it means. Every line will have a similar layout, but differ depending on the value of “n”. Although such sums can seem complicated at first sight, they become easier to understand as you see them more often.

The rest of the formula is a straightforward pure wave, where the amplitude, frequency and phase have subscripts of "n". In this case, the subscript "n" is just to distinguish between the amplitudes, frequencies and phases of each wave. In this formula, the value of "n" is not being used in any calculations. It is just an identifier.

The whole equation and the "Σ" sum (sigma sum) formula can be written out in full as so:

$f(t) = h +$
$a_1 \sin((360 * f_1 * t) + \phi_1) +$
$a_2 \sin((360 * f_2 * t) + \phi_2) +$
$a_3 \sin((360 * f_3 * t) + \phi_3) +$
$a_4 \sin((360 * f_4 * t) + \phi_4) +$
$a_5 \sin((360 * f_5 * t) + \phi_5) +$
$a_6 \sin((360 * f_6 * t) + \phi_6) +$
$a_7 \sin((360 * f_7 * t) + \phi_7) +$
$a_8 \sin((360 * f_8 * t) + \phi_8) +$
... and so on.

In this case, numbering the amplitudes, frequencies and phases is just to say that these are different entities (although, with the exception of the frequencies, they might or might not have the same values). The subscripted numbers tell us nothing about the actual values of the wave attributes. They essentially mean "the first wave's amplitude, whatever it might be", "the second wave's amplitude, whatever it might be", and so on. The subscripts are acting as identification tags. The same is true for each subscripted "n" in the original formula.

To summarise of all the above, the formula says that a periodic signal is equivalent to the sum of a mean level added to a pure wave with certain characteristics, added to a second pure wave with certain characteristics, added to a third pure wave with certain characteristics, ... and so on forever. As I have said before, this idea is not strictly true – there are exceptions that disprove it. However, it is true enough to be useful in the analysis of waves. Another point is that the formula gives the sum as an infinite number of waves, when an infinite number of waves would not be needed for many signals (especially those that were literally created by adding a particular number of waves).

The formula does not help with actually performing Fourier series analysis – it only describes a situation. At the level of maths reached so far in this book, it slightly hinders doing Fourier series analysis as it makes a straightforward process appear far more complicated than it really is, thus preventing people from thinking that they can learn about it. However, the formula is still useful because it forms the basis of more complicated techniques and ideas. By being able to describe Fourier series analysis with such a formula, future methods that are based on it can be more succinctly summarised.

We can write the formula with notes to make it slightly clearer, and to indicate facts about its parts that are important but not specifically stated:

$$f(t) = h + \sum_{n=1}^{\infty} a_n \sin((360 * f_n * t) + \phi_n)$$

... where:

- "f(t)" refers to the original *periodic* signal that we are discussing.
- "=" means "is equal to", but in this situation should be "is equal to, or is approximately equal to, depending on the nature of the periodic signal that we are discussing".
- "h" is the mean level of the original signal.
- "Σ" is shorthand for a series of sums, where we add up variations of the part to the right of the "Σ", and each variation will differ by having a different value of "n". The first variation will have each "n" replaced by "1", the second variation will have each "n" replaced by "2", and so on. We start with "n" as 1 and continue forever, although in reality, we probably would not *need* to continue forever, and in practice, we would not be *able* to continue forever.
- "a_n" is the amplitude of each wave in each part of the sum. In the first part of the sum, "a_n" will become "a_1". In the second part, it will become "a_2". In the third part, it will become "a_3", and so on. Each "a" will refer to the amplitude of the wave in that part of the addition. Some or all of the amplitudes might have the same value as each other, or they might all be different. It depends on what is needed to make up the original signal.
- "f_n" refers to the frequency of each wave in the sum. In the first part of the sum, "f_n" will become "f_1". In the second part, it will become "f_2", and so on. Each "f" will refer to a *different* value. [If we had two or more waves with the same frequency, those waves would become added together to make another pure wave with that frequency. Therefore, there is no point in

having duplicate frequencies in the sum.] Each "f" will be an integer multiple of the frequency of the original signal "f(t)".

- "ϕ_n" refers to the phase of each wave in the sum. In the first part of the sum, "ϕ_n" will become "ϕ_1". In the second part, it will become "ϕ_2" , and so on. The phases might all be different, or some or all might be the same as each other. It all depends on what is necessary to make up the original signal.

Another description

Another way of phrasing the formula from above is as so:

$$f(t) = h + \sum_{n=1}^{\infty} a_n \, sin(360 * f_n * t) + b_n \, cos(360 * f_n * t)$$

This formula says exactly the same thing as before, but instead of showing a single pure wave *with* a phase, it is treating that wave as being the sum of a Sine wave with *zero* phase and a particular amplitude, and a Cosine wave with *zero* phase and a particular amplitude.

As the Sine wave and Cosine wave may or may not have different amplitudes, the amplitudes are portrayed with the letters "a" and "b" to indicate that they could be different. The frequency in each pair of waves will be the same, so we have the same "f_n" in the Sine wave and the Cosine wave.

Note that the amplitudes of each wave are *not* the amplitude of the constituent wave, but the amplitudes of the zero-phase Sine and Cosine waves that would need to be added together to make up that constituent wave.

The formula could be rewritten to be:

f(t) = h +
a_1 sin (360 * f_1 * t) + b_1 cos (360 * f_1 * t) +
a_2 sin (360 * f_2 * t) + b_2 cos (360 * f_2 * t) +
a_3 sin (360 * f_3 * t) + b_3 cos (360 * f_3 * t) +
a_4 sin (360 * f_4 * t) + b_4 cos (360 * f_4 * t) +
a_5 sin (360 * f_5 * t) + b_5 cos (360 * f_5 * t) +
a_6 sin (360 * f_6 * t) + b_6 cos (360 * f_6 * t) +
... and so on.

... where:

- The different amplitudes (a_1, b_1, a_2, b_2, a_3, b_3 etc) might or might not be the same as each other.
- The frequencies within one pair of Sine and Cosine waves will be the same as each other.
- The frequencies on different lines of the sum will be different from each other.

The formula in this case is saying that any periodic signal is equivalent to a mean level added to a Sine wave (with zero phase, a particular amplitude, and a particular frequency), added to a Cosine wave (with zero phase, a particular amplitude, and the same frequency as the Sine wave), added to a second Sine wave (with zero phase, a particular amplitude, and a particular frequency), added to a second Cosine wave (with zero phase, a particular amplitude, and the same frequency as the Sine wave), and so on.

This way of expressing the formula is closer to the more commonly used way of interpreting the two points from the Mean Level wave [which is a way that skips the idea of a Mean Level wave altogether, and treats the values as amplitudes of a Sine wave and a Cosine wave about to be added together]. However, this formula adds another level of complexity to what, as always, can be stated non-mathematically much more succinctly.

Fourier's original idea about periodic signals was that they all could be reduced to a sum of Sine and Cosine waves with zero phases in this way. Therefore, thinking about the results in this way is consistent with Fourier's thinking. However, in my view, it defeats the point of the process – the whole point of the process is to find the constituent waves, not to find pairs of waves that when added together result in the constituent waves. As before, this formula is still useful for being the basis of more complicated ideas, so despite it not being as useful for actually doing the procedure, it is useful in explanations of more advanced maths. [Being able to think of a wave with a non-zero phase as being the sum of a Sine wave and a Cosine wave with zero phases also helps in developing a more thorough understanding of waves.]

More descriptions

With these types of formulas, when there is only one wave, you will often see them with Cosine instead of Sine:

$$f(t) = h + \sum_{n=1}^{\infty} a_n \cos\left((360 * f_n * t) + \phi_n\right)$$

The above formula means exactly the same thing as if it mentioned Sine. The only difference would be that, for any given signal, the phases would be different if we said it consisted of a sum of Cosine waves, instead of a sum of Sine waves.

In all of the above formulas, the mean level ("h") is kept separate to the rest of the formula. If the mean level is treated as a wave with zero frequency, then it can be included in the right-hand part of the formula. This idea is usually suggested by starting the counting from zero – the "n" part will start at 0, and not 1.

$$f(t) = \sum_{n=0}^{\infty} a_n \sin\left((360 * f_n * t) + \phi_n\right)$$

Strictly speaking, as the actual values of the frequencies are not specified in the formula, the change from "n = 1" to "n = 0" is an unnecessary one – we could remove the "h" and still start counting from 1. There is nothing in the previous formulas that states that any frequency is not zero. The values of "f_1", "f_2", f_3" can refer to any frequency, and there is no reason why one of them should not be zero cycles per second. Similarly, if we start with "f_0", there is no reason why *it* should be zero cycles per second. The subscripted letters refer to different unspecified values, and their meaning is not stated in the formula. Having said all of that, the convention is to use this formula starting at n = 0.

As the actual values of the characteristics of the wave are not stated in the formulas, and as there is a potential for any phase, it does not matter whether the wave is a Sine wave or a Cosine wave, and it will still be valid.

Usually, the waves will be given in terms of radians and not degrees. In that case, the 360 is replaced by 2π, Sine would be working in radians, and the phase would be an angle given in radians. The only visible difference in the actual formula would be that "360" would be replaced by "2π".

In radians, the above formula would look like this:

$$f(t) = \sum_{n=0}^{\infty} a_n \sin\left((2\pi * f_n * t) + \phi_n\right)$$

[For the purposes of learning about waves, and being able to see patterns in results, it is easier to start with degrees. Radians have their uses, but they unnecessarily complicate explanations such as this one. I will explain radians in Chapter 22.]

Sometimes, you will see letters other than “n” being used. A common alternative is “k”, in which case, the formula would look like this:

$$f(t) = \sum_{k=0}^{\infty} a_k \sin\left((360 * f_k * t) + \phi_k\right)$$

It does not matter which letter is used as long as it does not conflict with the other symbols.

Sometimes, you will see the formula given in terms of Imaginary powers of the number “e”, but I will explain what this means in later chapters.

It does not matter if you do not understand any of these formulas at the moment. You do not need to understand any of them to analyse a periodic signal. Understanding how and why we can analyse a periodic signal is completely independent of understanding these formulas. Having said that, it can be useful to understand the formulas (even if only vaguely) if you want to read typical textbooks on waves, and the formulas are useful if you eventually want to understand the Fourier transform. The more you see formulas similar to those described in this section, the more easily you will understand them. Similarly, the more you see complicated formulas *of any nature*, the more easily you will understand the formulas in this section. You will develop a mind for formulas as you learn more about waves. Formulas that might have seemed totally incomprehensible before this chapter will end up seeming straightforward.

Another formula

Here is a variation of the formulas that we have seen so far:

$$f(t) = \sum_{n=0}^{\infty} a_n \sin((360 * f * n * t) + \phi_n)$$

The difference between this formula and the previous ones is that this formula has "n" as part of the multiplication being Sined, as opposed to being just a subscript to distinguish between the different occurrences of "a", "f" and "ϕ".

In the above formula, "f" is intended to be specifically the frequency of the signal that we are describing, or in other words, the "fundamental frequency". Given that every constituent wave will be an integer multiple of this frequency, this formula will run through every possible constituent frequency.

The "Σ" sum (sigma sum) written out in full is as so:

a_0 sin ((360 * f * 0 * t) + ϕ_0) [This will act as a mean level]
+
a_1 sin ((360 * f * 1 * t) + ϕ_1)
+
a_2 sin ((360 * f * 2 * t) + ϕ_2)
+
a_3 sin ((360 * f * 3 * t) + ϕ_3)
... and so on.

This sum is more specific than the earlier formulas because every wave has a frequency that is an integer multiple of that of the signal we are describing. The earlier formulas never specified the characteristics of the frequencies – they just stated that "f_0", "f_1", "f_2", "f_3" and so on were constituent frequencies with unspecified values that were all different. [Actually, the fact that they were all different was never stated in the formulas, but it was implied, or else the formulas would not be as useful.]

In the previous formulas, the actual frequencies were never stated, and could have had any value. There was the implication that one of the frequencies would have been zero to provide for the mean level, if any. With this new formula, the frequencies are specified – they are all sequential integer multiples of the frequency of the original signal. Given that, we must assume that many of the amplitudes are zero, or else we would be saying that every periodic signal literally contained an infinite number of waves.

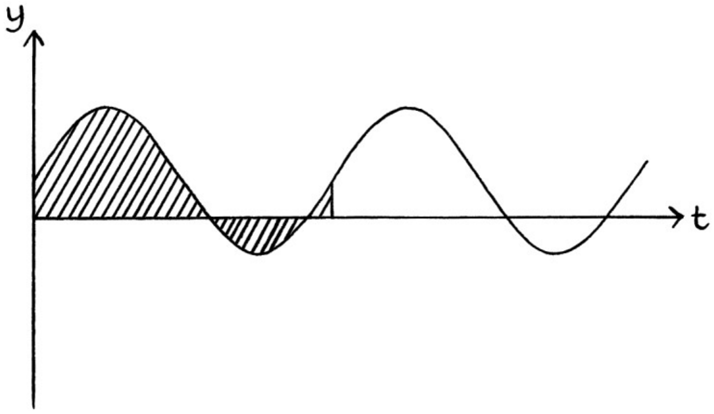
In many ways, this formula with a multiplication by "n" is superior to the previous ones that relied on "f" with a subscript. All of the previous formulas can be adapted to this idea.

One minor improvement we could make is to indicate that the "f" in the formula was the frequency of the original signal. We could do this by giving it a subscript such as "f_{signal}" or "f_t". [We cannot use "f_s" as that is used for another purpose as we will see later when we look at discrete signals.] A subscript has the benefit of suggesting that "f" is the frequency of the original signal, but with the drawback that it makes the formula more cluttered.

Areas instead of mean levels

In traditional explanations of Fourier series analysis, people do not pay attention to the mean levels of the signals, but to the total area between the curve of the signal and the t-axis over one cycle, all divided by the period of the signal. This ends up as being identical to the mean level, but it is another way of thinking about it.

The area between the curve of a signal and the t-axis is treated as the area above the t-axis minus the area beneath the t-axis.

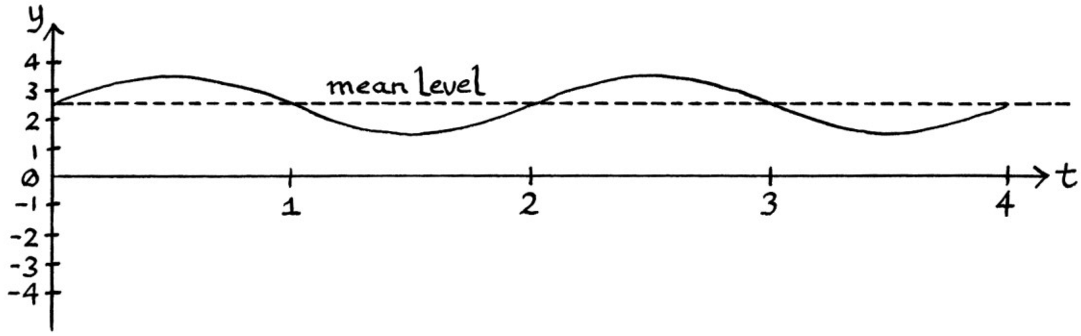


Using the language of calculus, this area is the definite integral of the signal over one cycle. [It does not matter if you do not know anything about calculus. I will explain basic calculus in Chapter 30.] The area can be calculated exactly with calculus *if we know the formula for the signal*, or it can be approximated by careful measuring.

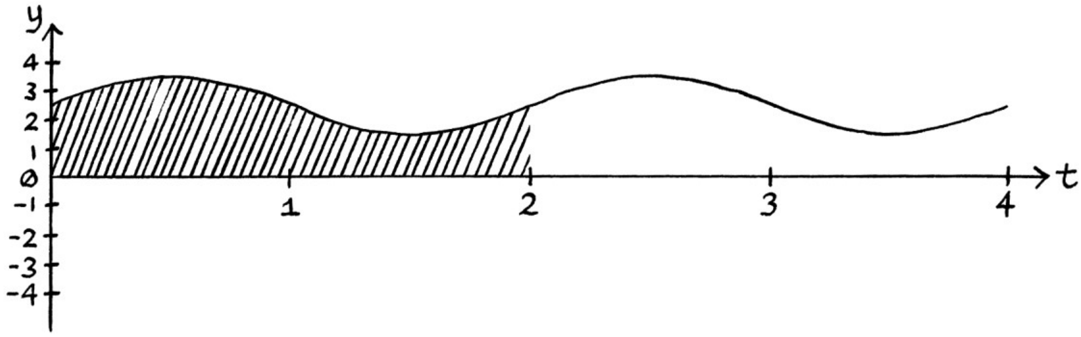
When calculating a mean level, we can calculate it over one cycle, or we can calculate it over all time. As the mean level is the average y-axis value, it will be the same whether we measure it for one cycle or for an integer number of cycles. When it comes to measuring the area, the total area for all time might be infinitely high, so it is better to calculate it over just one cycle.

The mean level over one cycle, and the area between the curve and the t-axis for one cycle, are directly connected. If we have the area, we can divide it by the amount of time along the t-axis, and we will have the mean level. If we have the mean level, we can multiply it by the amount of time along the t-axis for one cycle, and we will have the area. Such an idea is intuitive if we were thinking about square waves, where the area of a section would be the height multiplied by the width. When it comes to curved areas, the maths still works in the same way, but the area is the average height multiplied by the width.

As an example, we will look at the wave "y = 2.5 + sin (360 * 0.5t)". One cycle lasts for 2 seconds, so the period is 2 seconds. The mean level over one cycle (and over all cycles) is 2.5 units, which we can know from the formula:



The area over one cycle is 5 square units, which we can know either by using calculus or by multiplying the mean level by the length of one cycle. The length of one cycle is 2 seconds (or we can say that the period is 2 seconds), so the area is 2 times the mean level.



If the area between the curve of a signal and the t-axis is zero, then the mean level will also be zero. Conversely, if the mean level is zero, the area will be zero.

When it comes to testing waves, if the result of multiplying an original signal by a test wave is a signal with zero *area* between its curve and the time axis over one cycle, then that test frequency does not exist in the original signal. This is the same as if the mean level were zero. If the area is not zero, then it must be divided by the length of time of one cycle (the period) to find the mean level, and then that can be used to find the amplitude and phase of the constituent wave.

In traditional wave explanations, the mean levels of the results of multiplications with test waves are still used, but people do not think of them as "mean levels", but as the area under the curve divided by the period, which is ultimately exactly the same thing.

The generally accepted symbol for period is the letter "T". Therefore, in formulas for Fourier series calculations, you can expect to see the calculated area divided by "T" or multiplied by $\frac{1}{T}$ both of which mean the same thing. In this way, "T" is the period, but it also the "width" of the area for one cycle.

Using the area instead of the mean level complicates things because it introduces calculus into the process. It is also easier to think about mean levels than it is to think about some long-winded calculation that ends up with the mean level anyway. However, in some more complicated signals, *where we know the formula of the signal*, it might be easier to use calculus to work out the area, and then use that to work out the mean level, than to work out the mean level using other ways. When it comes to discrete signals (where a signal is given as a sequence of y-axis values taken from equally spaced moments in time), calculating the mean level is as simple as averaging the given y-axis values over one cycle. In such cases, there is no need to calculate the area, and thinking about calculus is an unnecessary complication. Generally, mean levels are much easier to understand.

Other people’s terminology

These are some commonly used terms in analysing signals:

Coefficient
A coefficient is a value that is multiplied by a variable in a formula. The amplitude in a wave formula can be thought of as a coefficient – it is multiplied against the results of the Sine or Cosine functions. When people leave the analysis process unfinished in the form of a Sine wave and a Cosine wave with zero phases, it is common for the amplitudes of the Sine wave and Cosine wave to be called the “coefficients”. Although having another name for “amplitude” seems unnecessary now, it can be a useful term when things become more complicated.

Harmonic
When we make a list of test frequencies for analysing a signal, they are part of what is called “a harmonic series”. By this, I mean that they are all sequential integer multiples of a common value. For example, the series of numbers: 5, 10, 15, 20, 25, 30, 35, 40, 45 and so on, are all integer multiples of the number 5. They are a harmonic series. The number 5 in this harmonic series is called the “fundamental” value. It is the base value from which all the other values are calculated. When making our list of test wave frequencies, the frequency of the signal can be thought of as the “fundamental frequency”. The test frequencies are integer multiples of this.

Frequency filters

Now that we can find the waves that were added to make up a periodic signal, we can perform very accurate frequency filtering on signals. First, we reduce the signal to its constituent waves, and then we add up only those constituent waves that have the frequencies we desire.

As an example, we will say that we want to perform a high pass filter on a signal to remove any waves with frequencies less than or equal to 3 cycles per second. To do this, we find the constituent waves of the signal, and then add them together while ignoring any that have frequencies less than or equal to 3 cycles per second. The resulting signal will be the same as the original signal, but filtered to have only frequencies above 3 cycles per second.

We can perform any type of frequency filtering in this way. We could filter out even frequencies or odd frequencies, or even frequencies that were prime numbers if, for some reason, that is what we wanted.

The downside to filtering in this way is that, although we can have very accurate filters, it can take a lot of work depending on the nature of the signals. There are quicker, but less accurate, methods of filtering.

Chapter conclusion

Fourier series analysis is a very important part of the study of waves. It is usually explained using complicated mathematical formulas and in a way that makes it seem far more difficult than it actually is. Here is a simple summary for performing the whole process:

- Remove the mean level from the signal, if any, and make a note of it for later.

- Find the frequency of the signal, and make a list of test frequencies that are sequential integer multiples of this number. For example, if the signal has a frequency of 3 cycles per second, then the list will contain frequencies of 3, 6, 9, 12, 15, 18... and so on.

- For every frequency in the list, multiply the signal by each of these test waves in turn, swapping "f" for the frequency being tested:
  "y = 2 sin (360 * f * t)"
  ... and:
  "y = 2 sin ((360 * f * t) + 90)"

- Calculate the mean levels of the signals resulting from these multiplications. If both mean levels are zero, that frequency does not exist in the signal. Otherwise, the amplitude of the wave for that frequency will be the square root of the sum of the square of these mean levels. The phase of the wave for that frequency will be the arctan of the first mean level divided by the second mean level. [Check that you have the correct arctan result.]

- Do not forget the mean level, if any, that was removed from the signal at the start.

If you can program arrays in any programming language, then it is not particularly difficult to write a computer program that can add and multiply waves. If you can do that, then it will be within your ability to write a program to perform the above steps to analyse the signals that your program makes. The trickiest part will be calculating the fundamental frequency of the original signal – if you struggle with this, you could type the frequency in by hand instead.

It is a harder task to write a program to analyse the signals created by a *different* program, as there will be more factors to take into account.

You could also perform the tasks in a spreadsheet program such as Microsoft Excel, although doing that can be more cumbersome. We will look at how to do that in Chapter 43, which describes discrete Fourier series analysis.

Chapter 19: Corresponding signals

In this chapter, we will briefly look at the signals derived from circles and shapes. You might not need to know much from this chapter, but it will still increase your knowledge and understanding of waves.

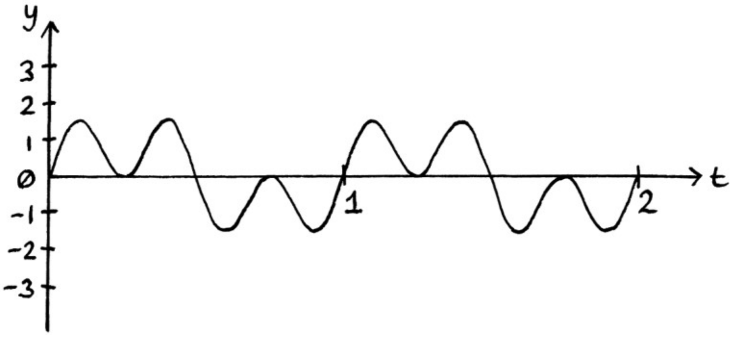
Corresponding signals

As we know, if we have a circle centred on the origin of the axes, the derived Sine wave has a corresponding derived Cosine wave and vice versa. Sine waves with zero mean levels and Cosine waves with zero mean levels have corresponding "twins" that have the same characteristics. For example:
"y = 4.5 sin ((360 * 11t) + 67)"
... comes from the same circle as:
"y = 4.5 cos ((360 * 11t) + 67)".
The two waves are corresponding waves.

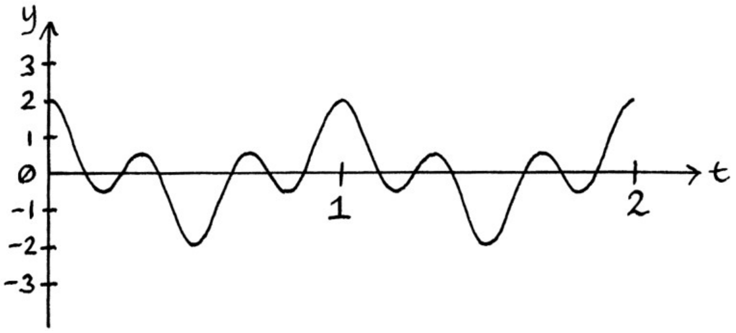
If we have just a Sine wave with zero mean level, we can recreate the corresponding Cosine wave (presuming it, too, has a zero mean level) by sliding it to the left by 90 degrees along the time or θ-axis. If we have just the Cosine wave, we can recreate the corresponding Sine wave by sliding it to the right by 90 degrees along the time or θ-axis.

Things become more complicated if we want to find corresponding *signals* that are not pure waves – in other words if we want to find the corresponding *sum* of two or more pure waves.

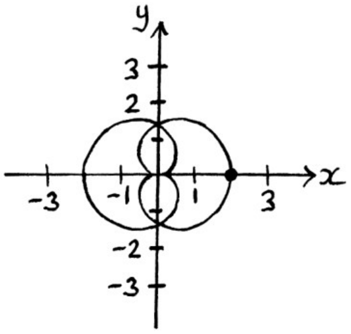
This is the graph of the sum of "y = sin 360t" and "y = sin (360 * 3t)":



... while this is the graph of the sum of “y = cos 360t” and “y = cos (360 * 3t)”:



The two signals are made up of corresponding waves, but the actual signals do not resemble each other. We can see that we could not recreate the graph of the sum of Cosines from the graph of the sum of Sines by sliding the signal left or right. This is because when we are adding Cosine waves, we are really adding Sine waves with an extra 90-degree phase, and the difference in frequency of the two added waves means that the 90-degree phases distort the addition. The difference in the two signals is also apparent in the shape on the circle chart:

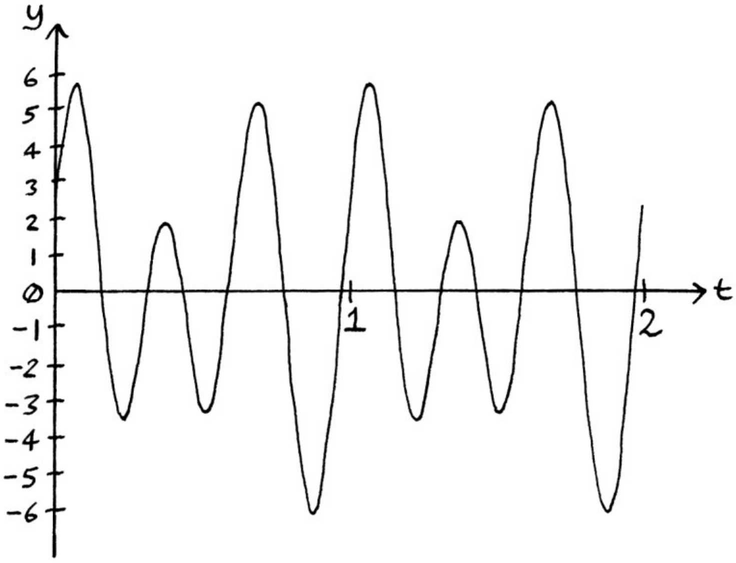


The shape is wider than it is high – the vertically derived signal will have different y-axis values to those of the horizontally derived signal.

A general rule is that we can create a corresponding *pure wave* by sliding it by 90 degrees left or right, but we cannot create a corresponding *impure signal* by sliding it 90 degrees left or right. However, as one might guess, we can create a corresponding signal by reducing the given signal to its constituent waves, shifting each of those waves by 90 degrees left or right, and then adding those shifted waves up to create a signal.

Example

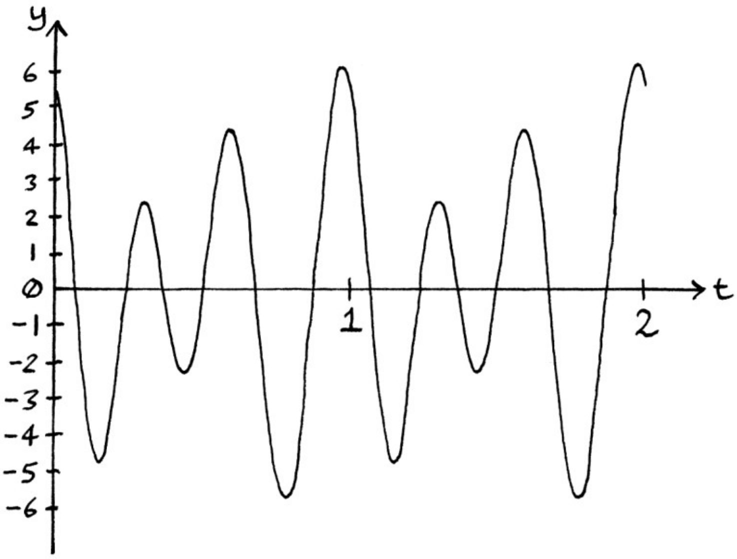
As an example, we will say we have the following signal, which we have been told is created by adding Sine waves of various amplitudes, frequencies and phases: [The signal can be thought of as a vertically derived signal.]



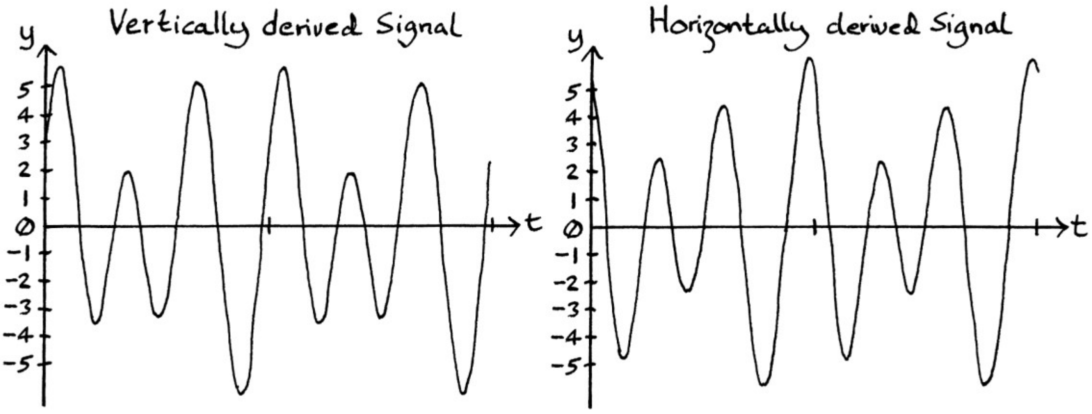
To find the corresponding signal created by adding Cosine waves [which is the corresponding horizontally derived signal], it is first necessary to reduce this signal to its constituent waves. After some calculations, we would discover that the constituent waves are:
"y = 2.5 sin ((360 * 2.0t)"
"y = 4.0 sin ((360 * 3.0t) + 45)"
"y = 0.5 sin ((360 * 4.0t) + 300)"

We then shift each of these Sine waves by 90 degrees to the left to make them into the corresponding Cosine waves. [If we kept them as Sine waves, we would add 90 degrees to each of their formulas, but here we will turn them into Cosine waves, so we can just change the word "sin" into "cos"]. We end up with:
"y = 2.5 cos ((360 * 2.0t)"
"y = 4.0 cos ((360 * 3.0t) + 45)"
"y = 0.5 cos ((360 * 4.0t) + 300)"

We then add these waves together, and we end up with the corresponding signal, which is this:

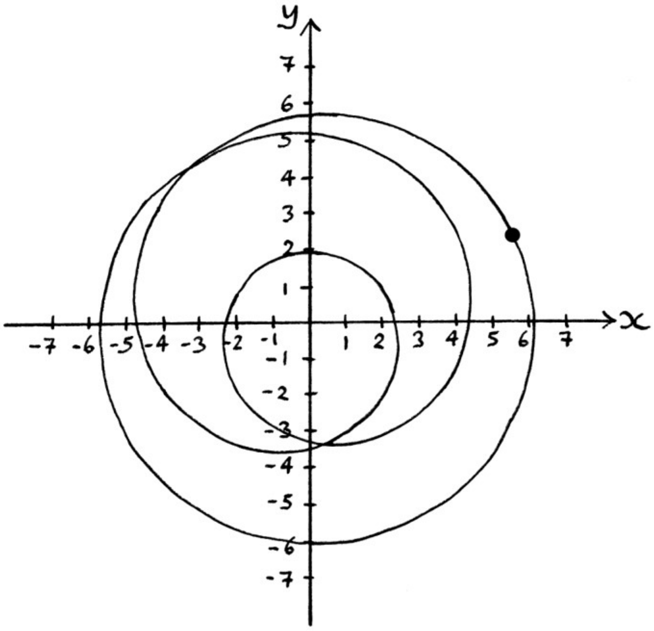


Therefore, we can say that these two signals are corresponding signals:



The first is the vertically derived signal from a shape, and is the sum of Sine waves; the second is the horizontally derived signal from the same shape, and is the sum of Cosine waves.

The shape, from which these two signals derive, looks like this:



The two types of corresponding signals

When it comes to corresponding signals, there are really two types:

- We can have corresponding signals in the sense that one signal is made up of, or is approximately made up of, the sum of Sine waves, and the other is made up of, or is approximately made up of, the sum of the corresponding Cosine waves.

- We can have corresponding signals in the sense that they both derive from the same shape.

When dealing with signals as in the previous example, it might seem that these will always be the same thing, but in practice, these are not necessarily connected ideas.

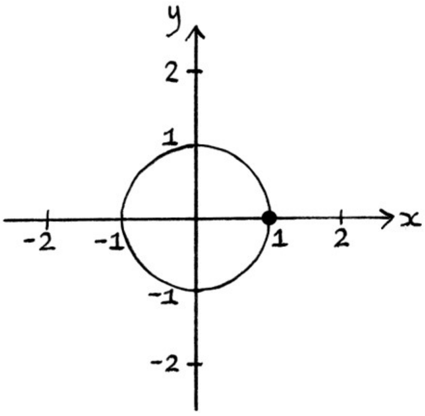
All corresponding signals of the first type are also corresponding signals of the second type, however not all corresponding signals of the second type are corresponding signals of the first type. There are shapes from which two signals are derived, where one signal can be said to be made up of the sum of particular Sine waves, but the other signal is *not* made up of the sum of the corresponding Cosine waves. Similarly, there are shapes where one derived signal can be said to be made up of the sum of particular Cosine waves, but the other signal is *not* made up of the sum of corresponding Sine waves.

Another way of expressing this is to say that there are multiple shapes that can produce the same vertically derived signal but a different horizontally derived signal, and there are multiple shapes that can produce the same horizontally derived signal, but a different vertically derived signal. For any vertically derived signal, there will always be one shape where the horizontally derived signal is made up of the Cosine wave sum of its Sine wave sum, but there will also be shapes where it is not.

I will call the two types of corresponding signals "corresponding-by-sum" and "corresponding-by-shape".

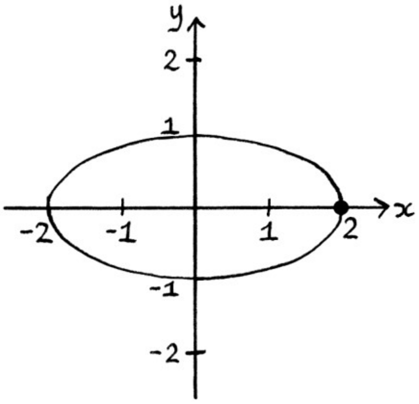
In the previous example, the "corresponding-by-sum" and "corresponding-by-shape" signals happened to be identical.

This circle...



... has its y-axis coordinates based on the wave "y = sin 360t" and its x-axis coordinates based on the wave "y = cos 360t". We can also say that the derived waves from the circle are "y = sin 360t" and "y = cos 360t", as that means the same thing. The two waves are corresponding-by-shape signals, in that they come from the same shape, and they are also corresponding waves in the sense that they are corresponding Sine and Cosine waves.

This shape, on the other hand...



... has its y-axis coordinates based on the wave “y = sin 360t”, but its x-axis coordinates based on the wave “y = 2 cos 360t”. In the sense that both waves derive from the same shape, the waves are corresponding-by-shape signals. However, they are clearly not corresponding waves as the amplitudes are different.

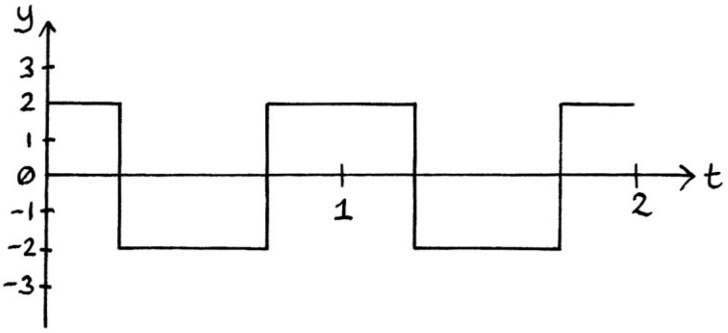
Both of the above shapes have the same vertically derived signal (“y = sin 360t”). However, they have different horizontally derived signals. There are an infinite number of shapes that will produce the same vertically derived signal, but only one of those shapes will produce a horizontally derived signal that is the corresponding Cosine wave to the Sine wave.

This was a simple example to demonstrate the point. When it comes to analysing signals that cannot be exactly recreated by adding pure waves, it is very easy to find situations where “corresponding-by-sum” signals and “corresponding-by-shape” signals are different.

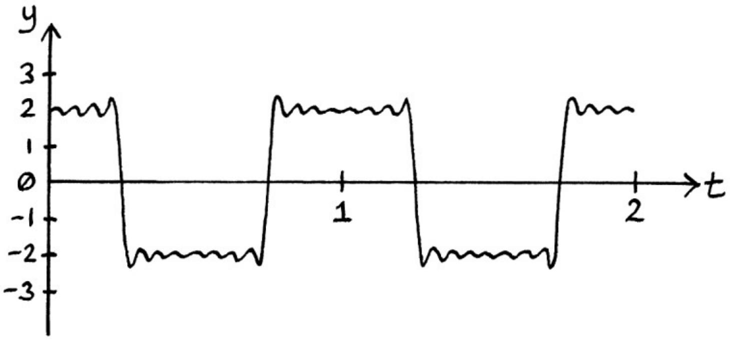
A square wave

A square wave is a good example of the difference between corresponding-by-sum signals and corresponding-by-shape signals. It is also a good example of how corresponding signals can look nothing like each other.

The following picture is of a square wave, which we will say is approximately the sum of some Sine waves:

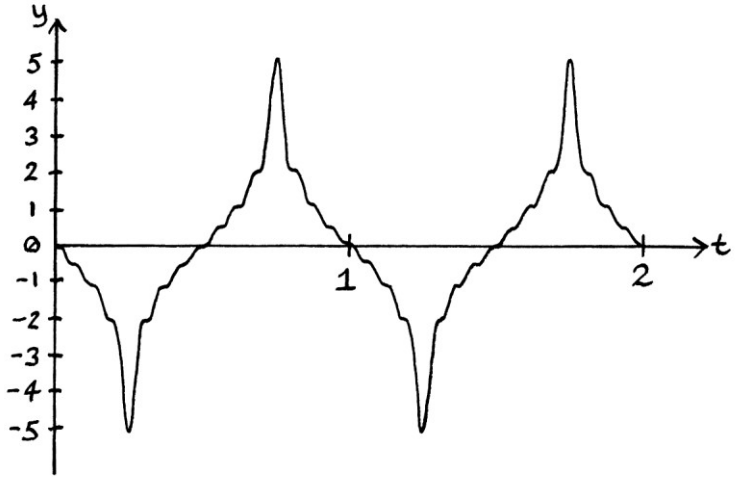


Using the same characteristics to describe this square wave as we would a pure wave, we will say that the square wave has an amplitude of 2 units, a frequency of 1 cycle per second, a phase of 90 degrees, and a mean level of zero units. After analysing the square wave, it turns out that it is approximately made up of these constituent Sine waves:
"y = 2.55 sin (360t + 90)"
"y = 0.85 sin ((360 * 3t) + 270)"
"y = 0.51 sin ((360 * 5t) + 90)"
"y = 0.36 sin ((360 * 7t) + 270)"
"y = 0.28 sin ((360 * 9t) + 90)"
"y = 0.23 sin ((360 * 11t) + 270)"
"y = 0.20 sin ((360 * 13t) + 90)"
"y = 0.17 sin ((360 * 15t) + 270)"
... although these waves when added together actually produce the following signal, which is just an approximation of a square wave:



To find the *corresponding-by-sum* signal (which will be made up of a sum of Cosine waves), we will add up these waves:
"y = 2.55 cos (360t + 90)"
"y = 0.85 cos ((360 * 3t) + 270)"
"y = 0.51 cos ((360 * 5t) + 90)"
"y = 0.36 cos ((360 * 7t) + 270)"
"y = 0.28 cos ((360 * 9t) + 90)"
"y = 0.23 cos ((360 * 11t) + 270)"
"y = 0.20 cos ((360 * 13t) + 90)"
"y = 0.17 cos ((360 * 15t) + 270)"

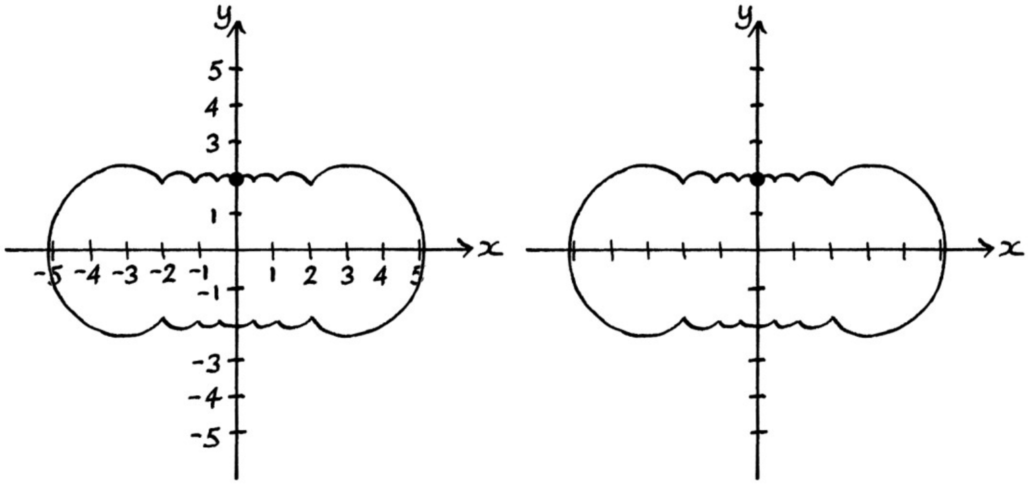
The sum of these Cosine waves looks like this:



An important thing to realise is that this is not the corresponding-by-sum signal to the original square wave, but the corresponding-by-sum signal to our *approximation* of the original square wave. Given how our approximation to the square wave looked similar to the real square wave, but not exactly the same, then it would follow that this will be similar to, but not exactly the same as, the actual corresponding-by-sum signal.

This corresponding-by-sum signal looks nothing like a square wave at all. It is unlikely that one would guess they came from the same shape.

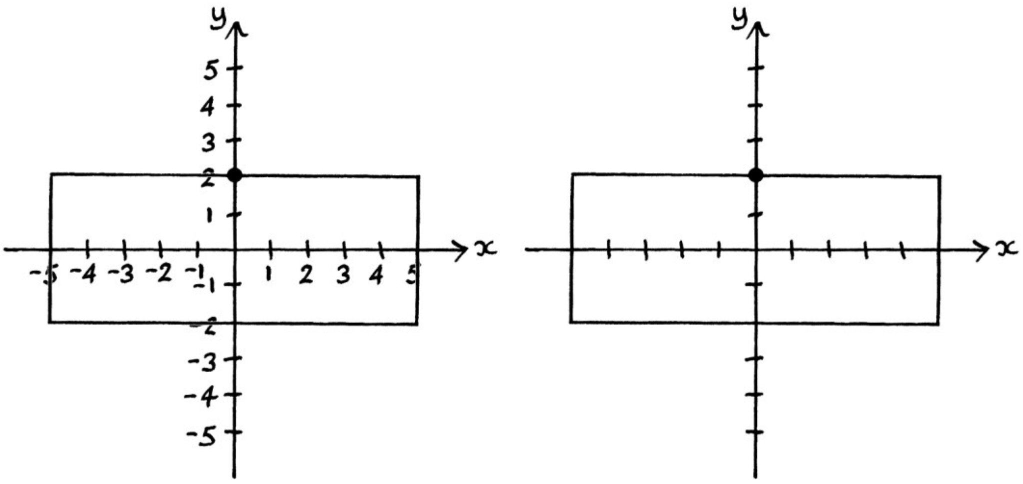
The shape for which the vertically derived signal is our *approximation* of a square wave, and for which the horizontally derived signal is our corresponding-by-sum signal is this (drawn with and without axis numbers to make it clearer):



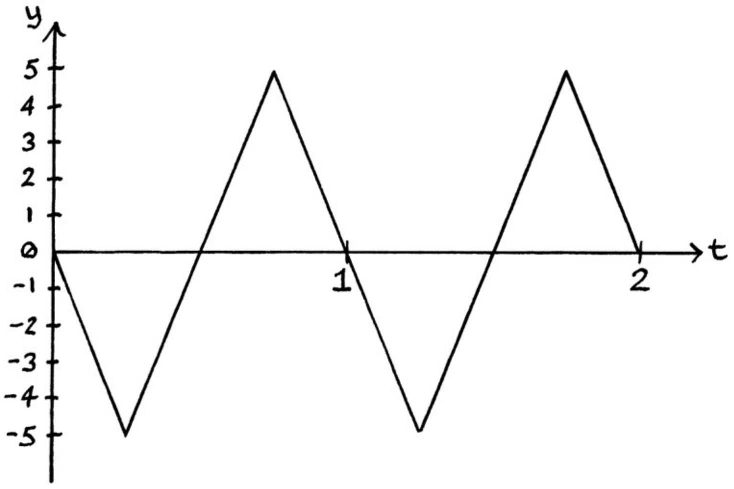
From thinking about the object moving around this shape, we can see why this shape produces the two signals that it does. The object starts at the coordinates (0, 2). It moves at a constant rate to the left while fluctuating up and down ever so slightly until it reaches near to the left-hand side when it begins to move gradually downwards. During the time that it was moving to the left, its y-axis value was fluctuating slightly, but staying at around y = 2. Its x-axis value was decreasing at a nearly constant rate. When the object moves downwards, its y-axis value decreases to y = 0 at the same time as its x-axis value reaches its minimum at an ever-decreasing rate. The object continues to move downwards, which means its y-axis value falls to its lowest value. At the same time, its x-axis value starts to increase at an ever-increasing rate. When the object is moving to the right along the bottom of the shape, its y-axis value is fluctuating slightly around the minimum, while its x-axis value is increasing at a steady rate. It continues to move to the right while fluctuating up and down slightly, until it gets near to the right-hand side when it moves upwards, and so on.

The object's y-axis values fluctuate around y = 2 when the object is on top of the shape, and around y = −2 when the object is on the bottom of the shape. This behaviour is shown in the vertically derived signal, which is our approximation of the square wave. The object's x-axis values are consistently going from just over x = 5 to just under x = −5 and back again. This is all reflected in the horizontally derived signal.

The exact shape that is being estimated by the reproduced shape is actually a rectangle (drawn here with and without axis numbers to make it clearer):



From this rectangle, our original square wave is the vertically derived signal. The rectangle is the exact shape from which the original square wave is derived. The horizontally derived signal from this rectangle looks like this:



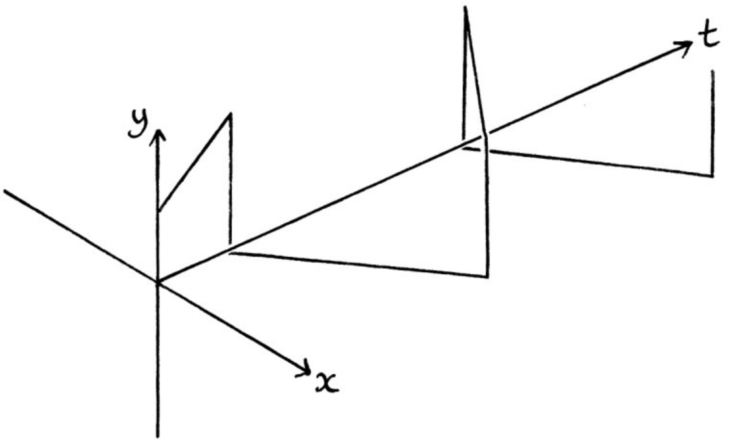
The rectangle is the shape that the two derived signals in this case would create if used as coordinates. The sawtooth signal above and the original square wave are corresponding-by-shape signals and corresponding-by-sum signals. [If I were being pedantic, I would say that they would be the corresponding-by-sum signals if it were possible to have a sum of waves that could create a square wave.]

We will think about the object's journey around the rectangle shape. The object starts at the coordinates (0, 2) on the rectangle, which is in the middle of the top. It moves to the left corner of the rectangle, staying at a constant height. While it does this, its y-axis position stays at 2, and its x-axis position falls at a constant rate to its minimum. Then the object's height drops *instantly* to y = −2. Because it drops instantly, the x-axis position of the object is the same at the start and end of the drop.

[If it dropped slowly, the x-axis would remain the same for the duration of the drop. This is a good example of how one cannot tell an object's frequency by looking at a completed shape, without knowing other information. If the object moved at a constant speed around the shape, the derived signals would look different – the square wave would no longer have steep edges, but instead sloped edges; the other signal would spend more time at its maximums. The two signals would look similar. The way in which the object drops instantly means that the drawing of the rectangle would probably be more instructive if it were drawn without the vertical lines connecting the top and bottom lines.]

After the drop, the object moves to the right-hand side of the bottom of the rectangle. During this time, the x-axis position of the object is increasing from its minimum to its maximum at a constant speed. When the object reaches the right-hand lower corner of the rectangle, it instantly jumps up to y = 2.

The "helix" for the square wave's shape looks like this:

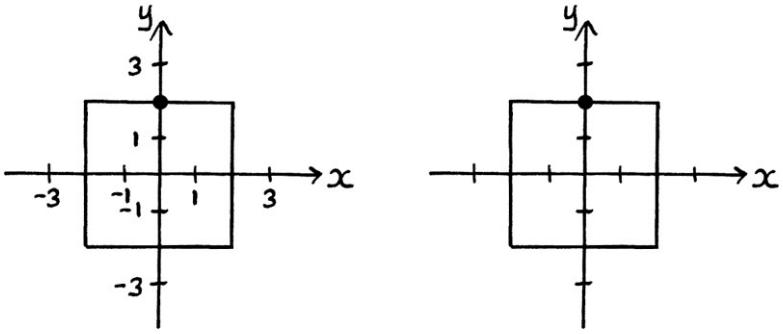


Seeing the square wave as a helix helps visualise exactly what is happening. If we could view the "helix" from the side with the t-axis to the right, we would see the square wave; if we could view the "helix" from below with the t-axis to the right, we would see the sawtooth signal (the corresponding-by-shape signal to the square wave.) If we could view the "helix" end on, with the t-axis pointing away from us, we would see the rectangle shape.

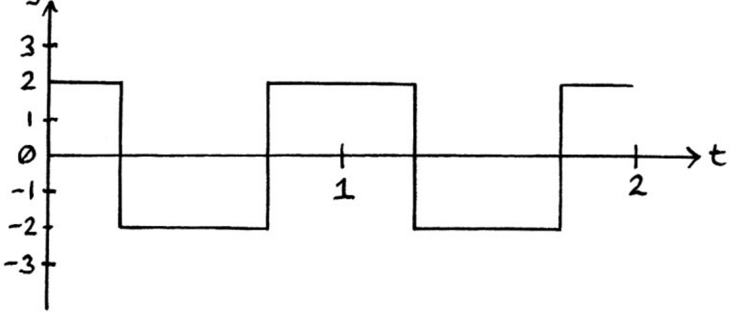
Our approximation of the square wave and the approximation's corresponding-by-sum signal are, in this case, reasonably good approximations of the actual square wave and its corresponding-by-sum signal. In this case, the corresponding-by-sum signals are actually the same as the corresponding-by-shape signals. One of the signals has been created by adding Sine waves; the other by adding the corresponding Cosine waves. Both signals are derived from the same shape.

There are other signals that could be used with the square wave to build a shape, or to put this the other way around, there are other shapes that would have a vertically derived signal as the same square wave, but would have a different horizontally derived signal. In such cases, the horizontally derived signal and the vertically derived signal would be corresponding-by-shape signals but not corresponding-by-sum signals.

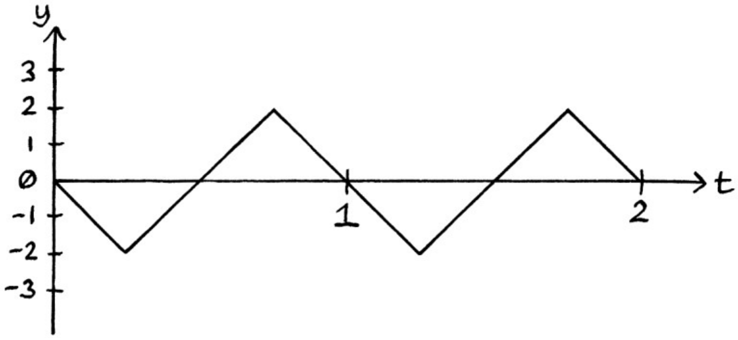
Our square wave could just as easily have come from this shape, which is a square (drawn with and without axis numbers):



This square shape's vertically derived signal is our square wave from before:



... but the square shape's horizontally derived signal is much shallower than the corresponding sawtooth signal we calculated before: [One might guess it would be shallower because the square shape is narrower than the rectangle.]



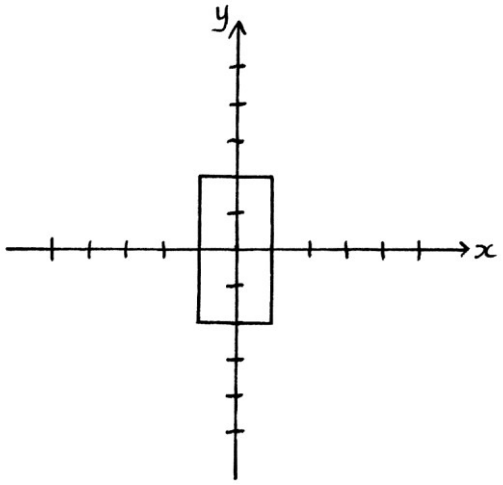
These two signals are corresponding-by-shape signals, but not corresponding-by-sum signals.

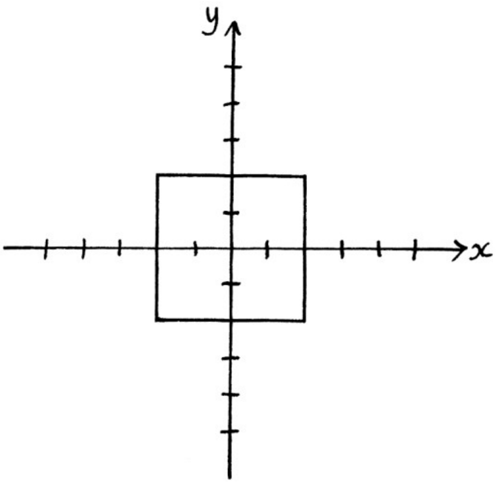
This particular "shallow sawtooth" horizontally derived signal is approximately the sum of:
"y = 1.62 cos (360t + 90)"
"y = 0.18 cos ((360 * 3t) + 270)"
"y = 0.06 cos ((360 * 5t) + 90)"
"y = 0.03 cos ((360 * 7t) + 270)"
"y = 0.02 cos ((360 * 9t) + 90)"

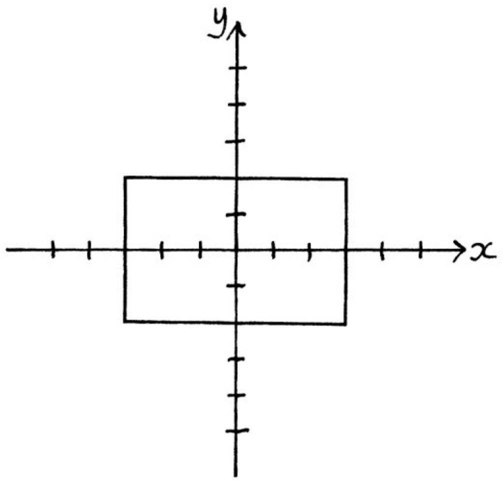
The "corresponding-by-sum" sawtooth signal to the original square wave from earlier in this chapter was approximately the sum of:
"y = 2.55 cos (360t + 90)"
"y = 0.85 cos ((360 * 3t) + 270)"
"y = 0.51 cos ((360 * 5t) + 90)"
"y = 0.36 cos ((360 * 7t) + 270)"
"y = 0.28 cos ((360 * 9t) + 90)"
... and so on. These are the same frequencies and phases, but larger amplitudes.

There are an infinite number of shapes that will produce any given vertically derived signal, but different horizontally derived signals. The same is true for our square wave.

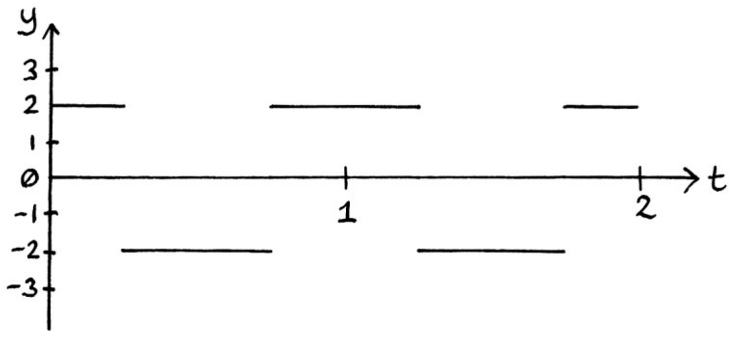
As an example, the following shapes have our square wave as the vertically derived signal, but have a different horizontally derived signal:



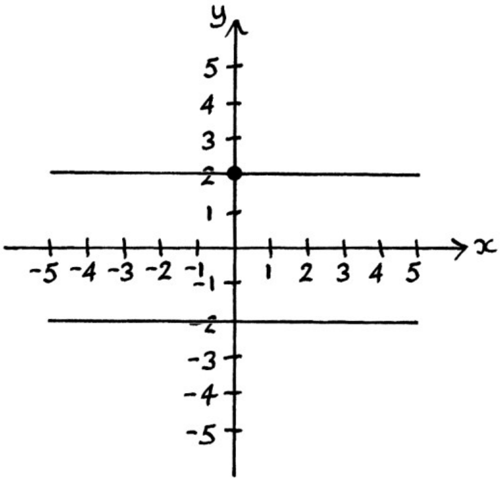




One interesting aspect of square waves is that the "curve" drops and rises instantly. This is not always clear in drawings of square waves, where the top lines are usually drawn connected to the bottom lines. For many situations, it would be clearer if a square wave were drawn without the vertical lines. Then, the drawings would emphasise the instant jumps:

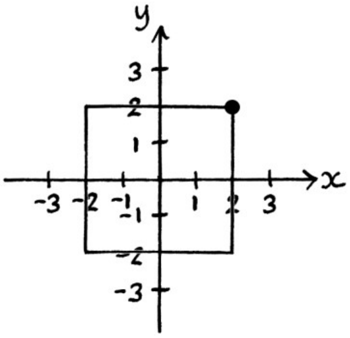


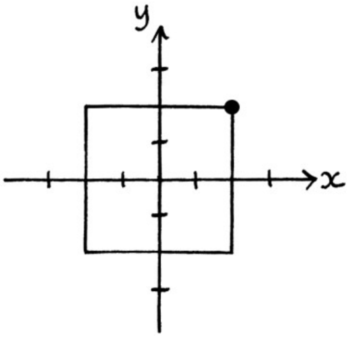
The rectangle or square from which a square wave is derived could be drawn as two unconnected horizontal lines to emphasise the sudden jumps:



An object moving around a square

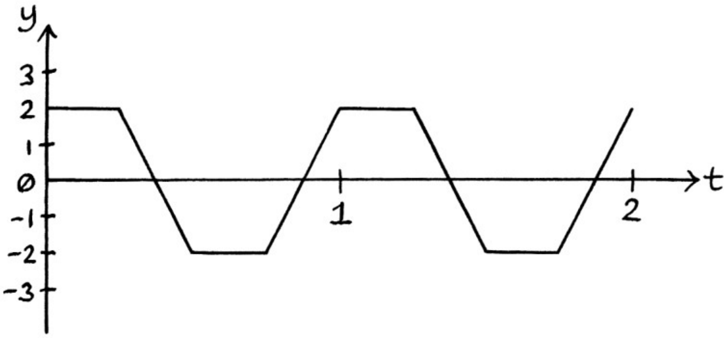
Supposing an object moved around a square at an even rate, starting at the top right-hand corner, and did not suddenly jump upwards or downwards, then the square shape would look like this (drawn with and without axis numbers):



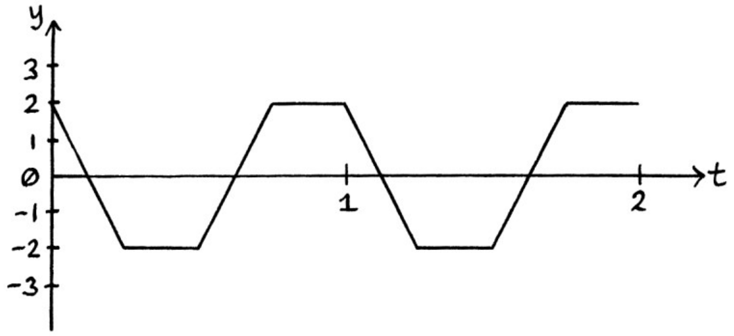


... and the two derived signals would be as so:

The vertically derived signal:



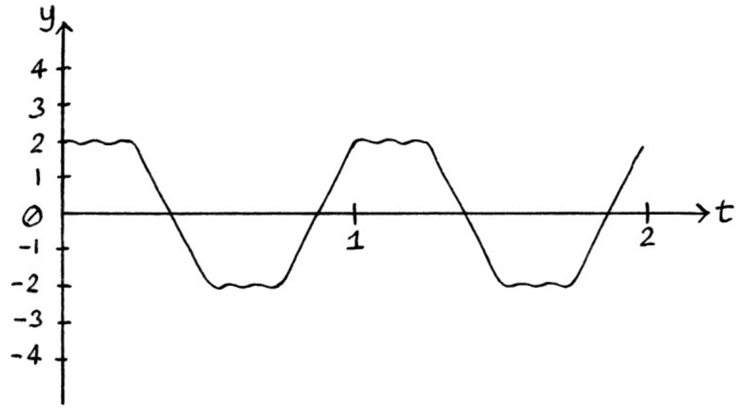
The horizontally derived signal:



If we analysed the vertically derived signal, we would find that an approximation of it was the sum of these Sine waves (ignoring any waves with smaller amplitudes):

“y = 2.29264 sin (360t + 45)”
“y = 0.25474 sin ((360 * 3t) + 135)”
“y = 0.09171 sin ((360 * 5t) + 45)”
“y = 0.04679 sin ((360 * 7t) + 135)”
“y = 0.02830 sin ((360 * 9t) + 45)”
... and a mean level of zero units.

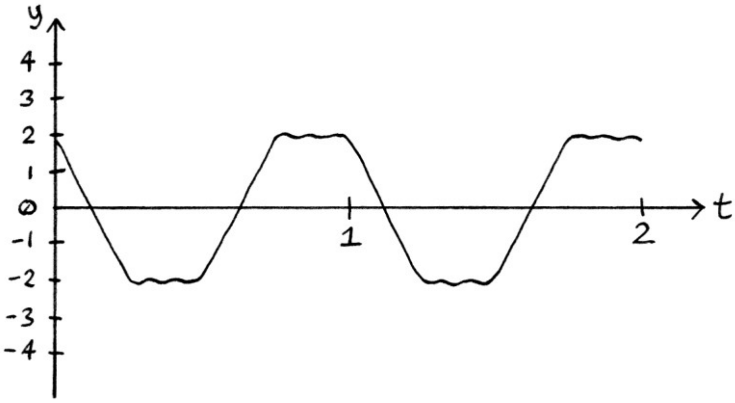
When those waves are added together, we end up with this signal:



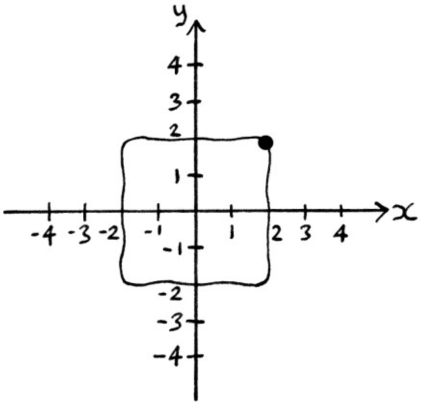
Note how this approximation is reasonably accurate. This is because there are no moments when the y-axis values instantly jump upwards or downwards.

If we analysed the horizontally derived signal, we would find that it is made up of the sum of these Cosine waves:
“y = 2.29260 cos (360t + 45)”
“y = 0.25475 cos ((360 * 3t) + 315)”
“y = 0.09170 cos ((360 * 5t) + 45)”
“y = 0.04679 cos ((360 * 7t) + 315)”
“y = 0.02830 cos ((360 * 9t) + 45)”
... and a mean level of zero units.

The signal created by adding these waves looks like this:



The signal created by using our two *approximated* signals as coordinates looks like this:



This is a reasonably good approximation to our original square shape, and shows that our analysis of the signals was correct.

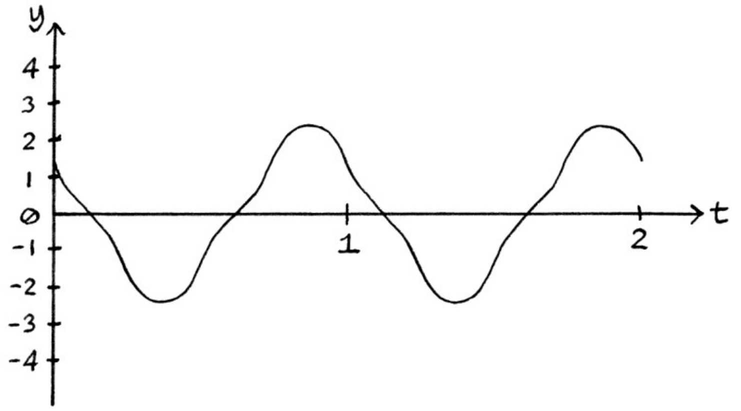
To make it easier to compare them, the two sums for the derived signals put next to each other are as so:

"y = 2.29264 sin (360t + 45)"	"y = 2.29260 cos (360t + 45)"
"y = 0.25474 sin ((360 * 3t) + 135)"	"y = 0.25475 cos ((360 * 3t) + 315)"
"y = 0.09171 sin ((360 * 5t) + 45)"	"y = 0.09170 cos ((360 * 5t) + 45)"
"y = 0.04679 sin ((360 * 7t) + 135)"	"y = 0.04679 cos ((360 * 7t) + 315)"
"y = 0.02830 sin ((360 * 9t) + 45)"	"y = 0.02830 cos ((360 * 9t) + 45)"

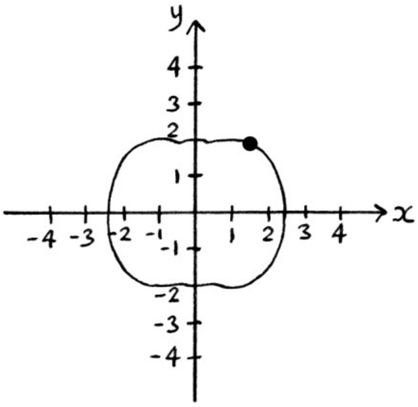
Ignoring minor differences in the accuracy of the amplitudes, we can see that these are not corresponding-by-sum signals – they are only corresponding-by-shape signals. The list of Sine waves is not the same as the list of Cosine waves – the amplitudes are the same and the frequencies are the same, but the phases of the 3 cycle-per-second and 7 cycle-per-second waves are different by 180 degrees. The corresponding-by-sum Cosine waves would be:

"y = 2.29264 cos (360t + 45)"
"y = 0.25474 cos ((360 * 3t) + 135)"
"y = 0.09171 cos ((360 * 5t) + 45)"
"y = 0.04679 cos ((360 * 7t) + 135)"
"y = 0.02830 cos ((360 * 9t) + 45)"

The sum of these waves produces this signal:



The shape that the two corresponding-by-sum signals would make has a vague similarity to the corresponding-by-shape signal, but it is clearly different:



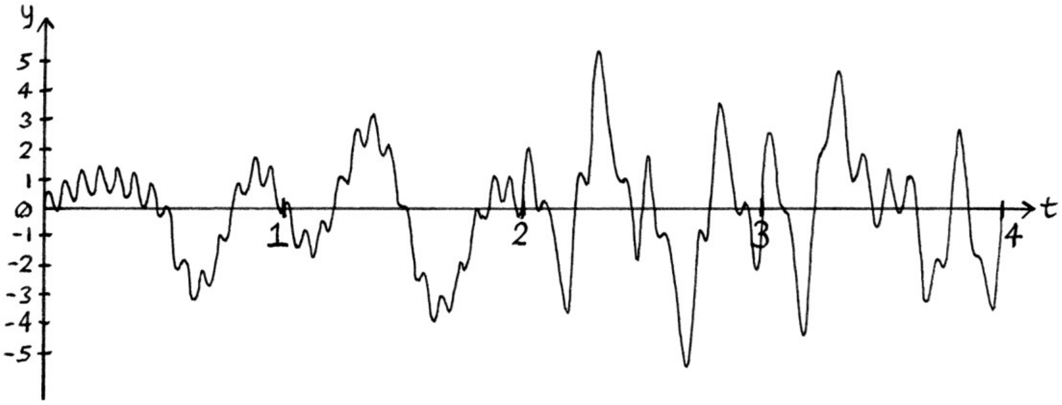
Thoughts on corresponding signals

There are an infinite number of shapes that will produce a given vertically derived signal, and there are an infinite number of shapes that will produce a given horizontally derived signal. However, for a given vertically derived signal, there is only one shape for which the horizontally derived signal is made up of the sum of the corresponding Cosine waves. Also, for a given horizontally derived signal, there is only one shape for which the vertically derived signal is made up of the sum of the corresponding Sine waves.

It is possible to use Fourier series analysis to find the constituent waves of a signal, and from that, it is possible to make up the sum of the corresponding waves to produce the corresponding-by-sum signal.

Chapter 20: Aperiodic signals

In Chapter 18, we looked at finding the constituent waves for periodic signals that were, or could have been, the sums of pure waves, and we also looked at finding the constituent waves of periodic signals that could not have been the sums of pure waves. In this chapter, we will take a brief look at finding the constituent waves for aperiodic signals, or in other words, signals that are not periodic – signals that do not repeat, such as this one:



This chapter will not teach you how to analyse aperiodic signals. Instead, it is a very basic introduction to the *idea* of analysing aperiodic signals, and the pitfalls in doing so. There is much more that could be said about aperiodic signals, and there are better ways of analysing them than those described here. [I do not describe the Fourier Transform in this chapter.] After reading this chapter, you will probably be no nearer to being able to analyse aperiodic signals than before.

If a signal does not repeat, then either:

- it is the sum of pure waves including two or more that have a frequency ratio that is an irrational number to one. [As described in Chapter 13 in the section on the addition of different frequencies.]

... or:

- it is not, and cannot be, the sum of pure waves.

If the signal is the first of these, then we would be able to recreate the signal exactly using a sum of waves *if we could find the frequencies of those waves*. [We would not be able to use Fourier series analysis as there would be no frequency on which to base our test frequencies.] If the signal is the second of these, then there is no sum of pure waves that can recreate it. If there were a sum of pure waves that could recreate it, then the signal would not be aperiodic (unless there were an irrational frequency ratio in the sum). Therefore, at best we would be approximating it *if we could find the frequencies*.

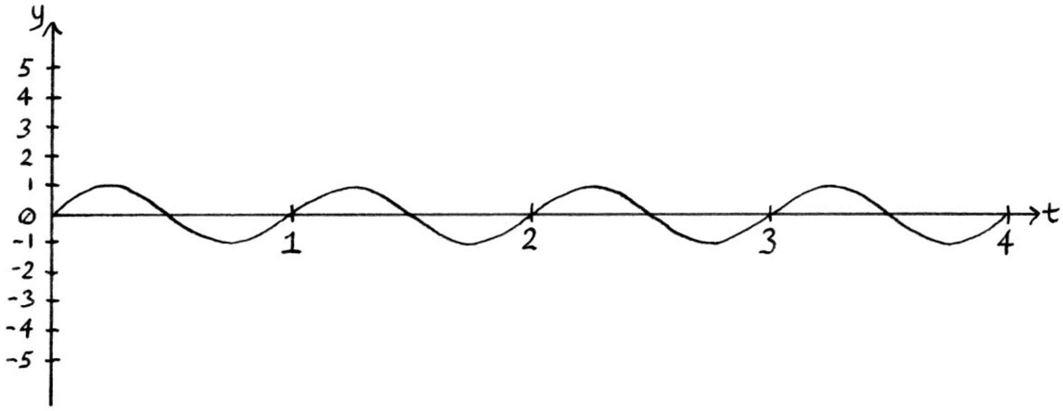
For now, we will ignore aperiodic signals created by adding waves with irrational frequency ratios.

One way of thinking about an aperiodic signal is to consider it the sum of sections of pure waves that are added at certain moments in time. For example, we could have a signal made up of the sum of:

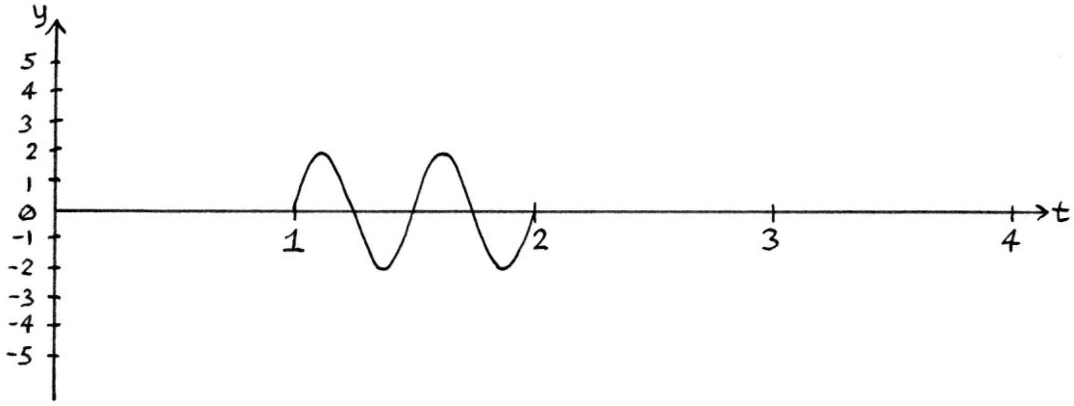
"y = sin 360t" for the length of the whole signal.
"y = 2 sin (360 * 2t)" for one second, starting at t = 1 second.
"y = 3 sin (360 * 3t)" for just 1.5 seconds, starting at t = 1.5 seconds.

These waves look like this:

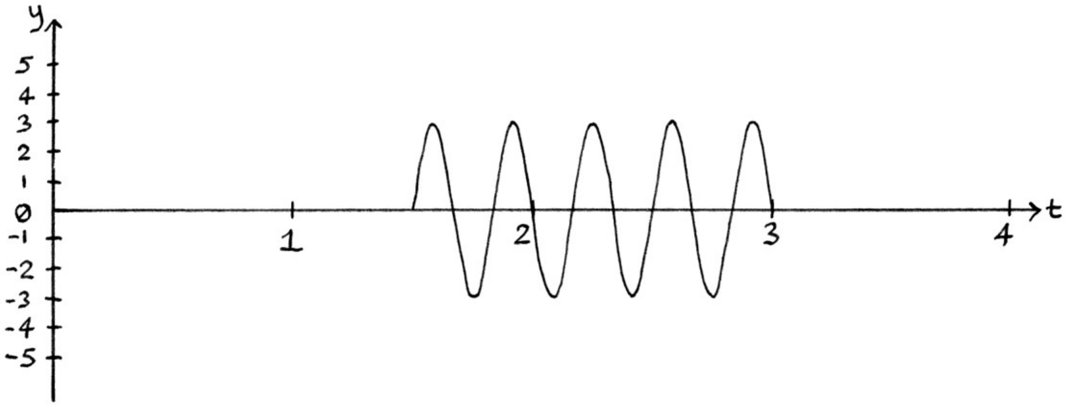
"y = sin 360t" (exists for the duration of the whole signal):



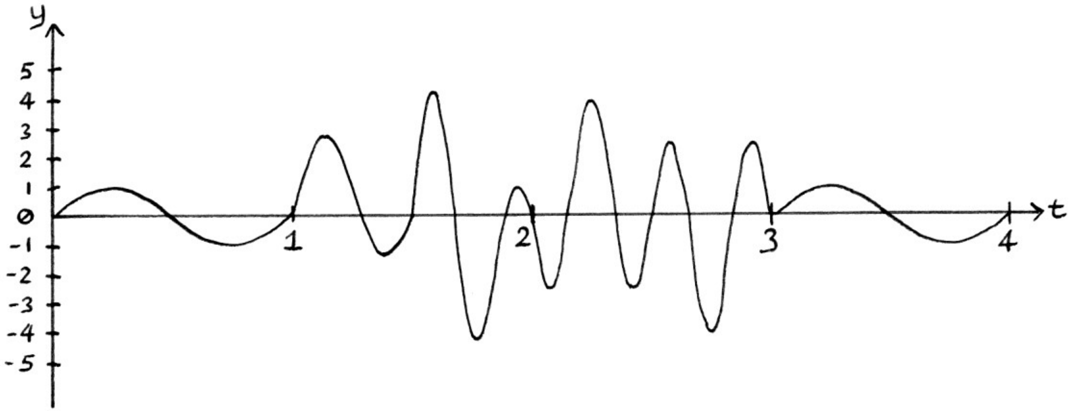
"y = 2 sin (360 * 2t)" (exists for one second, starting at t = 1 second):



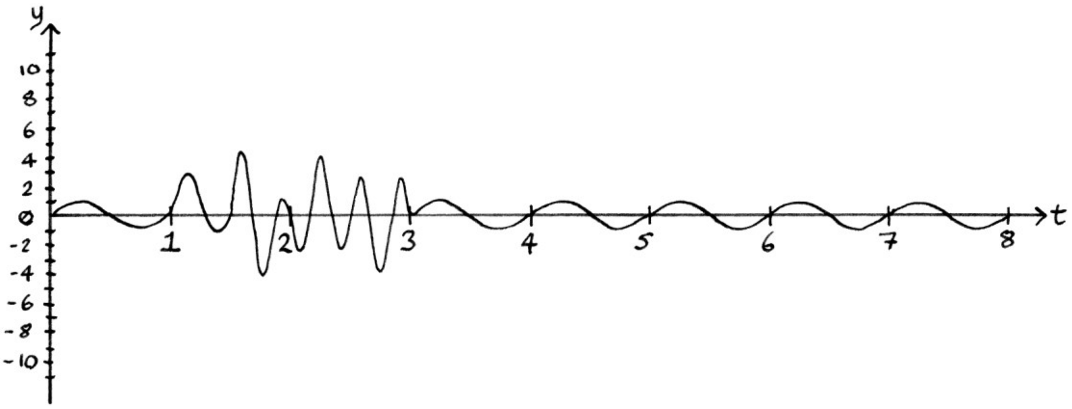
“y = 3 sin (360 * 3t)” (exists for 1.5 seconds starting at t = 1.5 seconds):



Added together, they produce this signal:

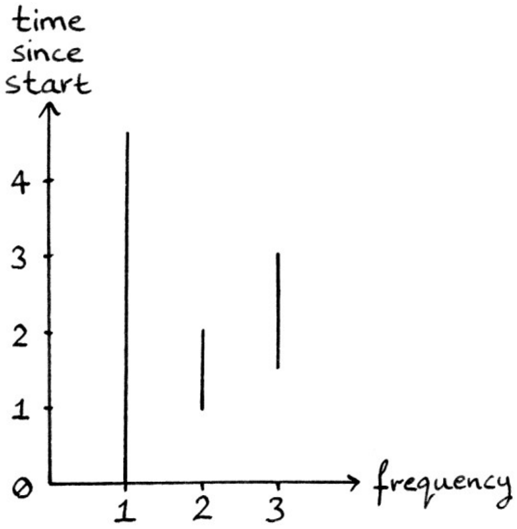


This signal is definitely not periodic. After 3 seconds, the signal continues in an identical way to “y = sin 360t”, but the overall signal never repeats its shape. Viewed over a longer time, the signal looks like this:

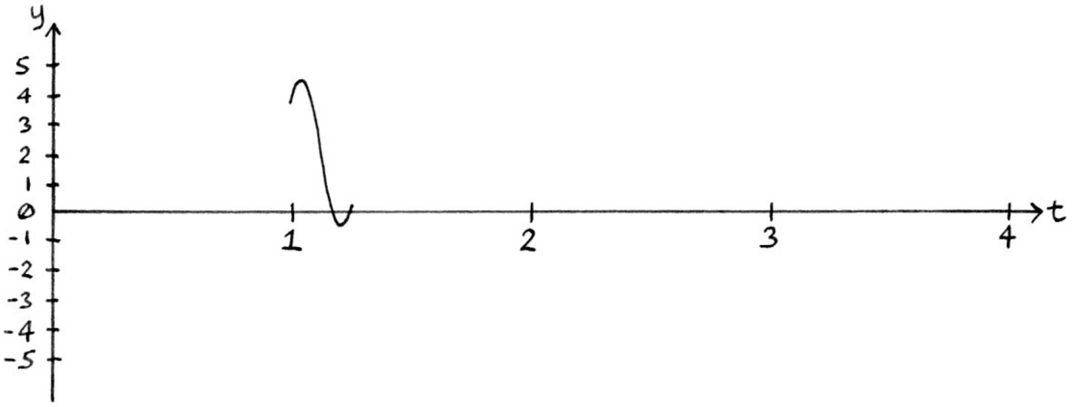


An important point to realise is that previously when we looked at *periodic* signals, all constituent waves were treated as if they existed for the entire duration of the signal. Now that we are dealing with (non-irrational ratio) aperiodic signals, at least one of the constituent waves will only exist for a fraction of the time that the whole signal exists.

Another significant fact about (non-irrational ratio) aperiodic signals is that they cannot (usefully) be portrayed on a frequency domain graph if that frequency domain graph does not relate to time in some way. In other words, either the frequency domain graph shows one distinct moment in time or else one of the axes must be time. [Technically, one could have a non-time-based frequency domain graph that covered all of the time from t = 0 up to t = infinity, but whether that would be useful or not is another matter]. The time-based frequency domain graph for the above signal looks like this:

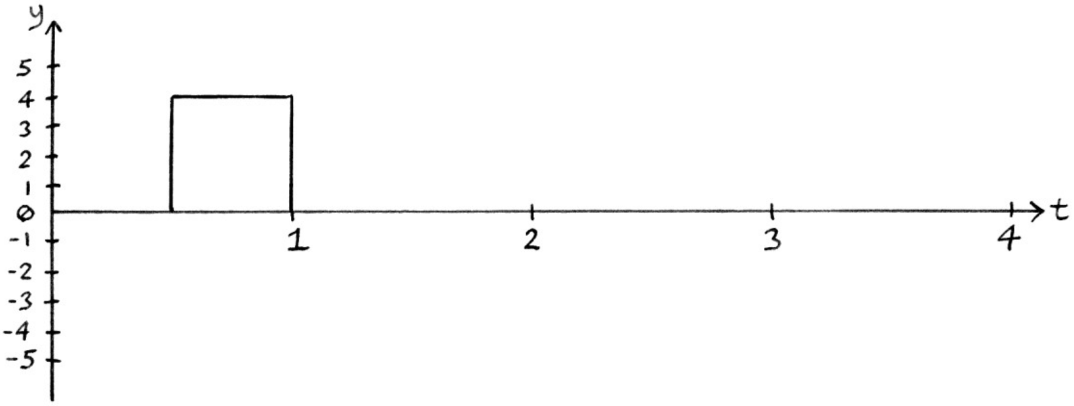


As another example, we will have a signal created solely of 0.25 seconds of:
“y = 2 + 2.5 sin ((360 * 3t) + 45)”
... at t = 1 second. The signal looks like this:



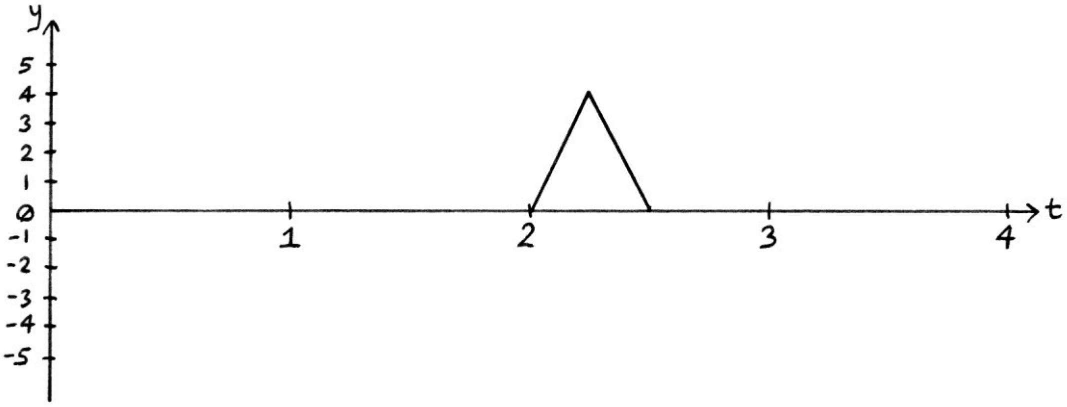
The above signal is aperiodic. The signal never repeats its shape. The y-axis values are zero for all time, except between 1 second and 1.25 seconds, where they follow the curve of "y = 2 + 2.5 sin ((360 * 3t) + 45)".

Here is another aperiodic signal:



The y-axis values for this signal are zero for all time, except at the small section between 0.5 and 1 seconds. If we had a periodic square wave, we would have to say that it could not be created exactly by adding waves. In the above graph, however, we can say that the "square section" of the signal can be created by adding *sections* of waves. In this particular case it can be created by adding 0.5 seconds' worth of "y = 4 sin ((360 * 0t) + 90)" placed at 0.5 seconds. However, if we were only considering non-zero frequencies, then we would only be able to approximate it.

Here is yet another aperiodic signal:



The above aperiodic signal has not literally been created by adding sections of waves at different times, however it would be possible to recreate an approximation of it by doing so – *if we could figure out the wave sections that, if added, would approximate it.*

In a similar way to how we can treat *periodic* signals that are not pure waves as being the sum of, or approximately the sum of, two or more pure waves of various amplitudes, frequencies, phases and mean levels, so can we treat *aperiodic* signals as being the sum of, or approximately the sum of, one or more *sections* of pure waves of various amplitudes, frequencies, phases and mean levels, placed at specific times.

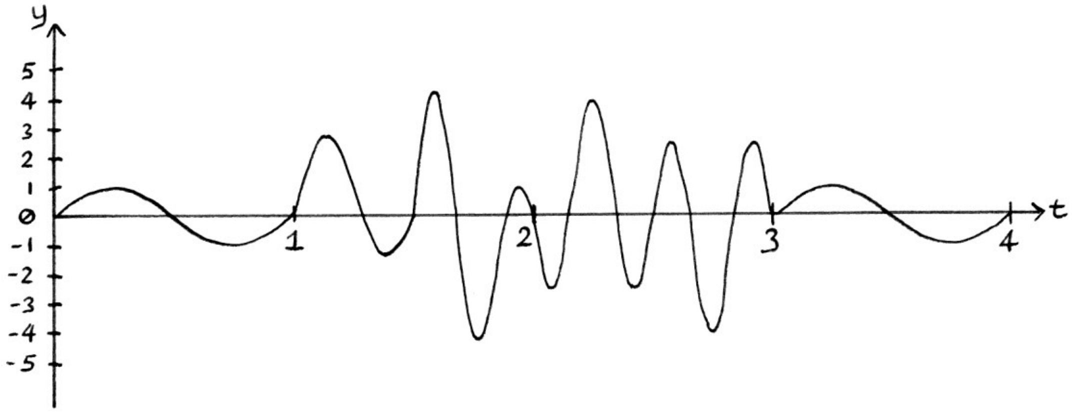
If a *periodic* signal is analogous to a bucket of unmixable liquids, then an *aperiodic* signal is analogous to a steadily moving river of unmixable liquids, where liquids are constantly added and removed along its length.

Circles and helices

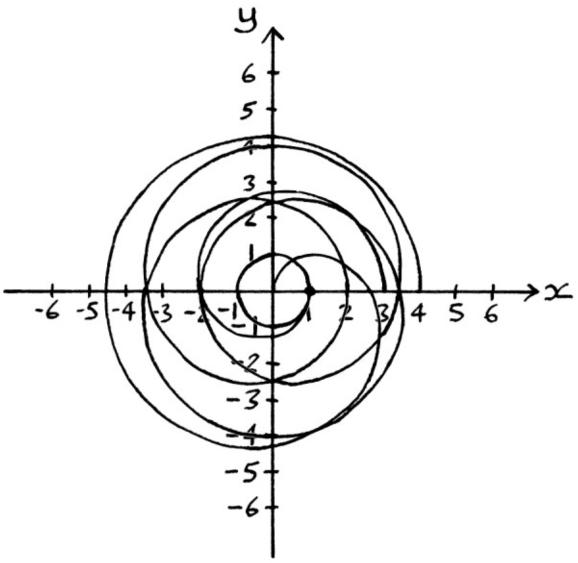
It is easiest to think about aperiodic signals as wave graphs on the time axis. However, it is also possible to think of them on the circle chart or helix chart.

Looking at an aperiodic signal on the circle chart does not particularly help in understanding it, as there is no way to distinguish the changes over time.

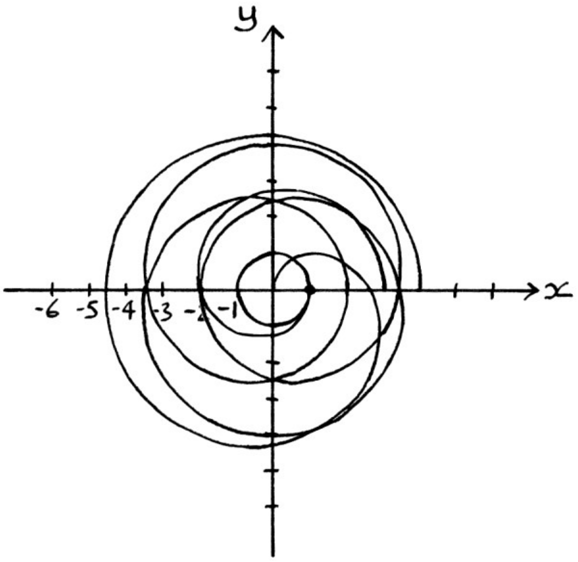
Our graph from earlier...



... looks like this on the circle chart:

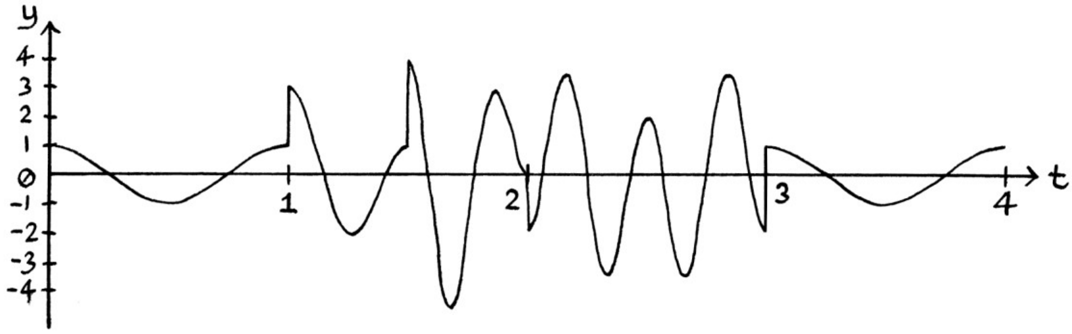


... or drawn without axis numbering to make it slightly clearer:

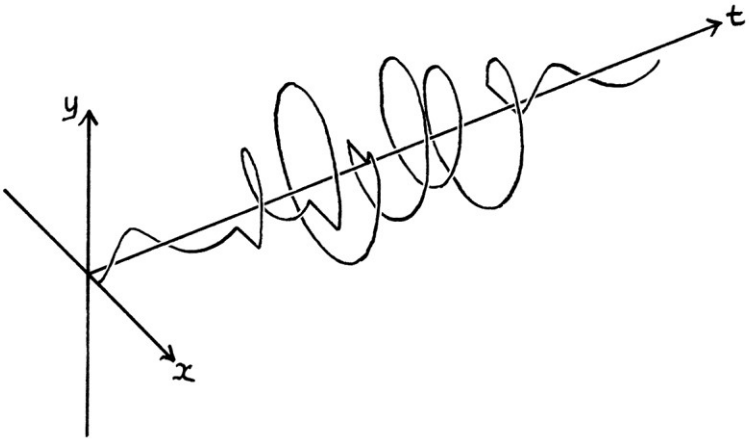


The reason the shape is not continuous, yet has breaks in it, becomes clearer when we look at the corresponding-by-sum (which is also the corresponding-by-shape) signal.

The corresponding-by-sum signal is made up of these waves:
“y = cos 360t” for the length of the whole signal.
“y = 2 cos (360 * 2t)” for one second, starting at t = 1 second.
“y = 3 cos (360 * 3t)” for just 1.5 seconds, starting at t = 1.5 seconds.



The signal based on Cosine waves has jumps in it, and therefore, the shape on the circle chart has jumps in it. The circle chart is not particularly useful in this situation, however, the “helix” on the helix chart is slightly easier to understand.



Analysing aperiodic signals

Aperiodic signals can be analysed to find the sections of constituent waves that were added, or could have been added, to make them (and when those sections existed).

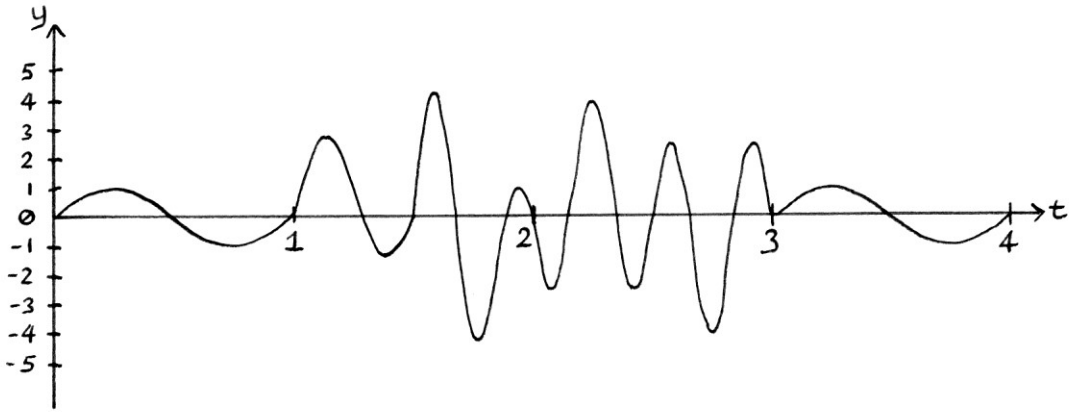
In this chapter, I will give a brief explanation of some simple, but inaccurate, ways of doing this. I am explaining these methods to demonstrate how the analysis of aperiodic signals is much more complicated than the analysis of periodic signals.

We cannot use the Fourier series method on its own to analyse aperiodic signals because if a signal does not repeat, then we cannot make a list of possible frequencies against which to test it. Fortunately, we can solve this problem by splitting the wave up into sections, and then using the Fourier series method to analyse each section in turn.

Splitting up the signal

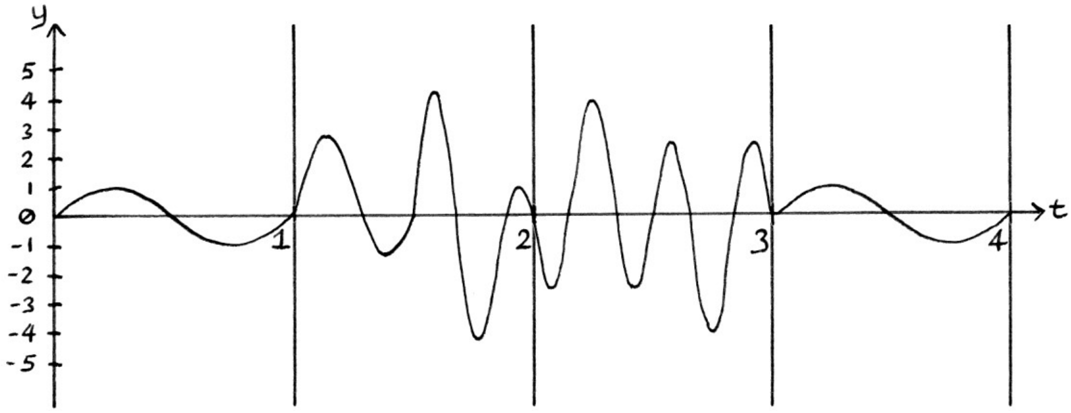
One way to analyse an aperiodic signal is to split the signal up into sections and treat each section as a cycle of a new *periodic* signal. Then we can analyse those periodic signals to find their constituent waves. In this way, we are finding the constituent waves that, if added together, would produce an approximation of one section of time of the original signal.

As an example, we will use this aperiodic signal from earlier:



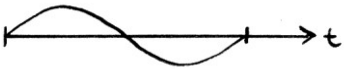
It consists of "y = sin 360t" for the length of the whole signal, added to "y = 2 sin (360 * 2t)" for one second, starting at t = 1 second, added to "y = 3 sin (360 * 3t)" for 1.5 seconds, starting at t = 1.5 seconds. For the purposes of this example, we will pretend we do not know any of this.

We will split the signal up into one-second sections, where one second is an arbitrarily chosen amount to demonstrate the method.

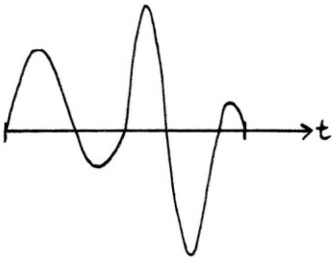


For this example, we will look at four sections of the signal:

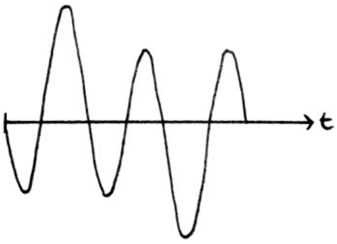
The first section, from 0 seconds to 1 second, looks like this:



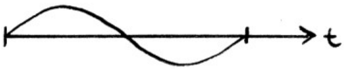
The second section, from 1 second to 2 seconds, looks like this:



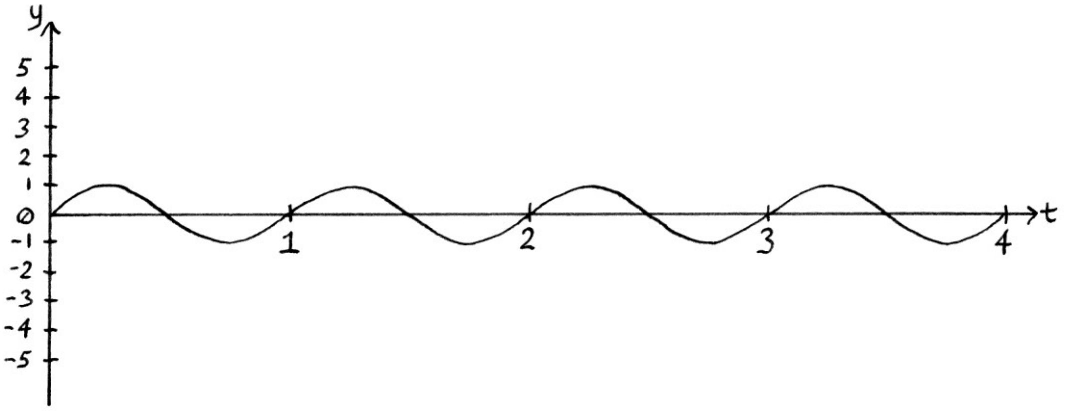
The third section, from 2 seconds to 3 seconds, looks like this:

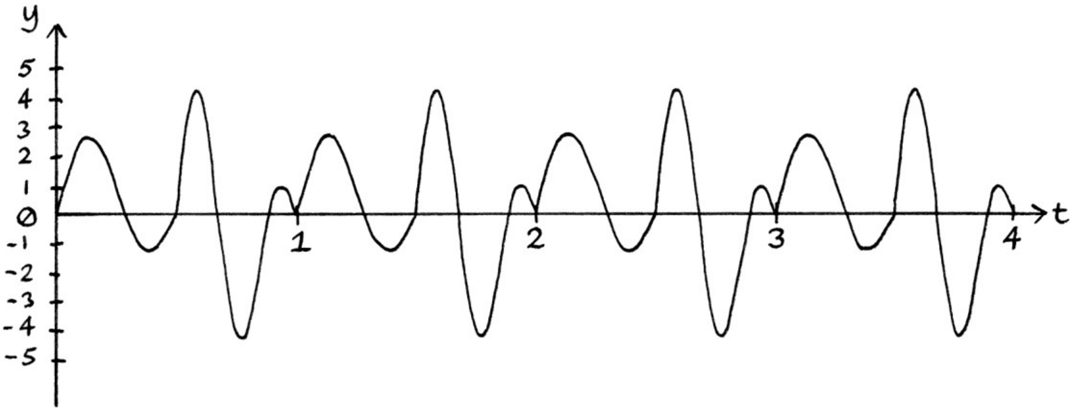


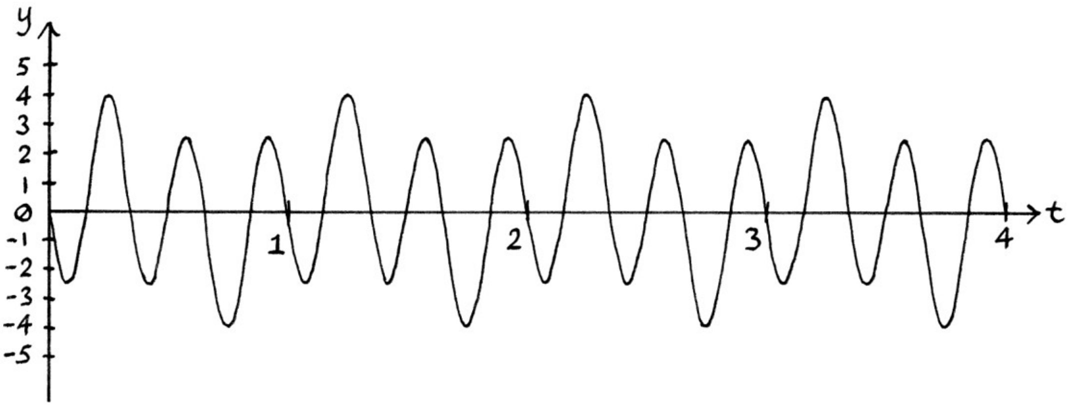
The fourth section, from 3 seconds to 4 seconds, looks like this (which is the same as the first section):

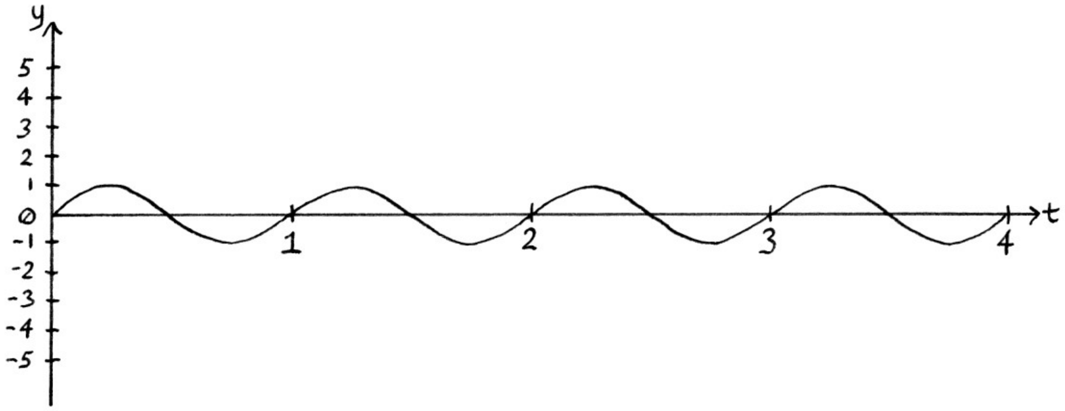


We then treat each section as a cycle of a new signal, in the sense that we repeat each section over and over again. In this way, we end up with four new periodic signals as follows:









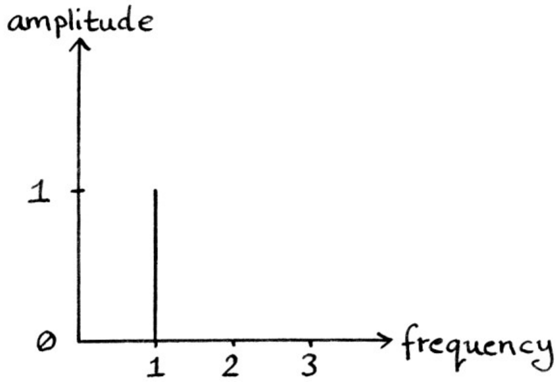
We will call these signals “child signals” to distinguish them from the original signal [Other people will probably call them something else]. Each child signal consists of a repeated section taken from the original signal. Each of our new child signals has a frequency of one cycle per second. This is because each cycle is the length of one section, and the sections last 1 second (they have a period of 1 second). Therefore, the frequency will be the reciprocal of 1 second: 1 ÷ 1 = 1 cycle per second.

Now that we have signals that repeat, we can use Fourier series analysis to analyse them. First, we create a list of test waves for each child signal. As all the child signals have the same frequency, we can use the same list of test waves for each one. The list will consist of frequencies that are integer multiples of one cycle per second: 1, 2, 3, 4, 5, 6, 7, 8, 9... and so on.

We test each of the pairs of test waves against each child signal, and we will find out which waves could be added to make a close approximation to each of our child signals. For this example, we will use 20 test frequencies for each signal. One thing to realise about the process is that the child signals were not literally created by adding waves, and therefore, it might not be possible to recreate them exactly by adding waves. Therefore, any analysis might produce an approximation.

After testing the first child signal, we would discover that it is made up of only one constituent wave: “y = sin 360t”. We can tell this is true just by looking at the signal. In this case, our calculations have discovered the exact wave that is equivalent to the child signal.

The *non-time* (i.e. amplitude) frequency domain graph for this child signal looks like this:

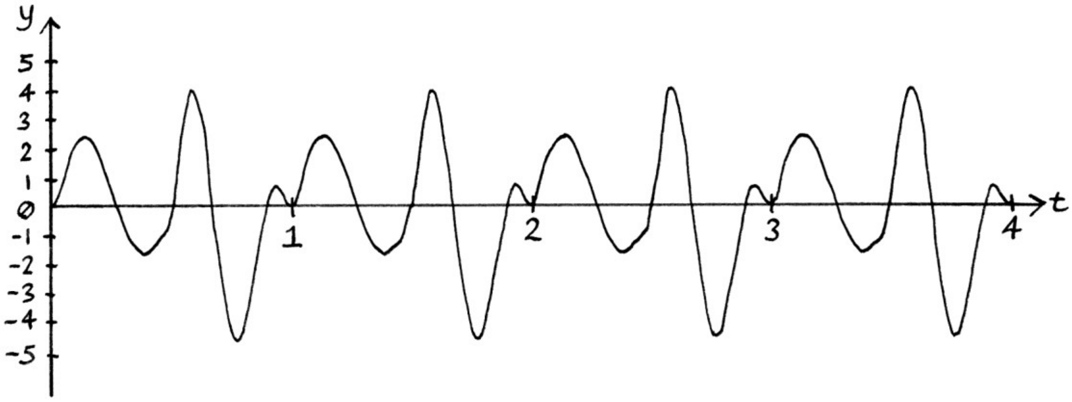


Analysing the second child signal is not as straightforward because we will end up with more constituent waves. If we use 20 test waves and ignore the lowest amplitudes, we end up with this list of constituent waves:
“y = 1 sin 360t”
“y = 2.31 sin ((360 * 2t) + 30)”
“y = 1.5 sin ((360 * 3t) + 180)”
“y = 0.82 sin ((360 * 4t) + 270)”
“y = 0.21 sin ((360 * 6t) + 270)”
“y = 0.10 sin ((360 * 8t) + 270)”
“y = 0.06 sin ((360 * 10t) + 270)”
... and a mean level of 0.32 units.

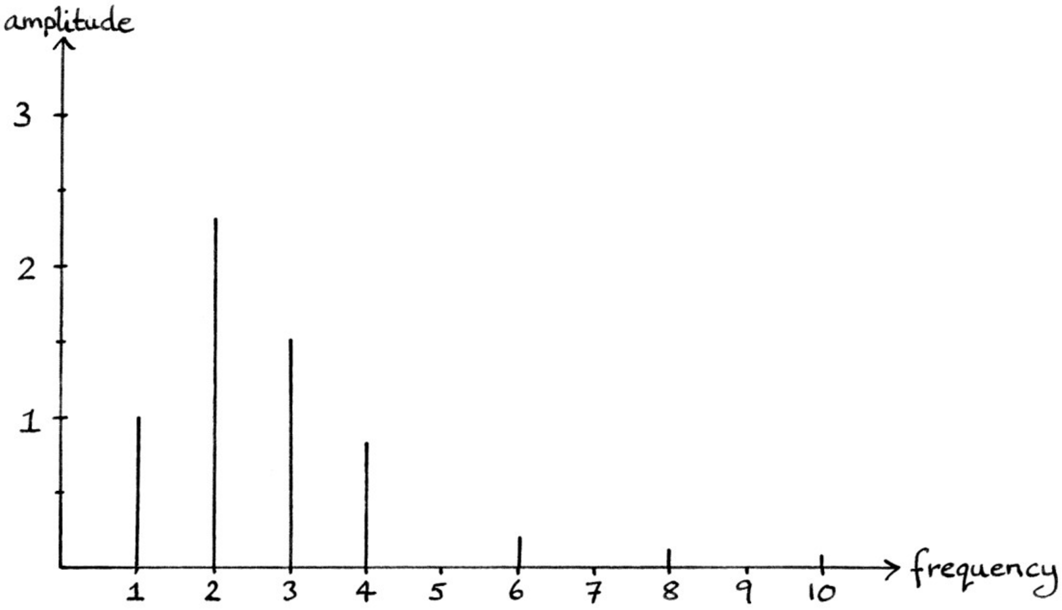
Analysing this child signal has produced 7 different constituent frequencies, despite how there were only 3 frequencies in the original signal. This is due to a new wave being introduced halfway through the section we are analysing. This has distorted our results.

If we add the constituent waves for this child signal, we end up with a reasonably good approximation of the child signal, which shows that our analysis has done a good job *on this section*. [Whether this means our analysis of the *whole original signal* is good or not is a different matter].

The reconstructed section looks like this, which is very close to the original child signal:



The non-time (i.e. amplitude) frequency domain graph for this child signal looks like this:



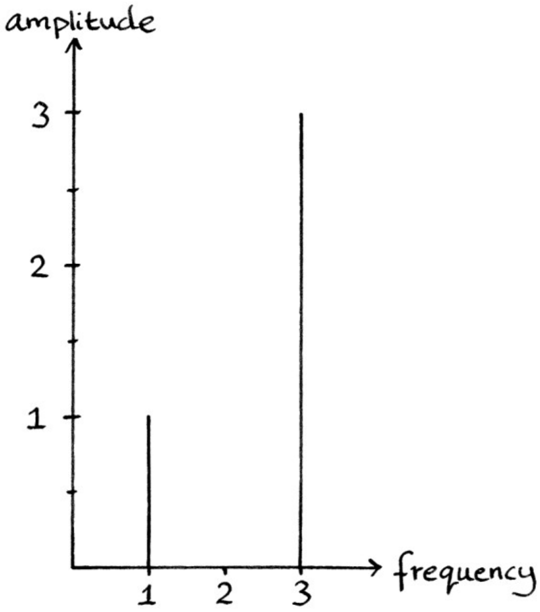
If we think back to the original signal, the waves that we knew existed in the second second were:
“y = sin 360t” for the whole of this second.
“y = 2 sin (360 * 2t)” for the whole of this second.
“y = 3 sin (360 * 3t)” for the second half of this second.

Our analysis of this section has found the 1 cycle-per-second wave perfectly, it has found the 2 cycle-per-second wave with a slightly reduced amplitude and an incorrect phase, and it has found the 3 cycle-per-second wave with half its amplitude and an incorrect phase. It has also found waves with frequencies of 4, 6, 8 and 10 cycles per second, but, with the exception of the 4-cycle-per-second wave, all of these have smaller amplitudes than the waves that actually existed in that second. It also discovered a mean level of 0.32 units that is not in the original signal. Therefore, we can say that this method has discovered the waves that existed for the second second, but with a lot of inaccuracies.

For the third child signal, the calculated constituent waves are:
“y = 1 sin 360t”
“y = 3 sin ((360 * 3t) + 180)”
... and a mean level of zero.

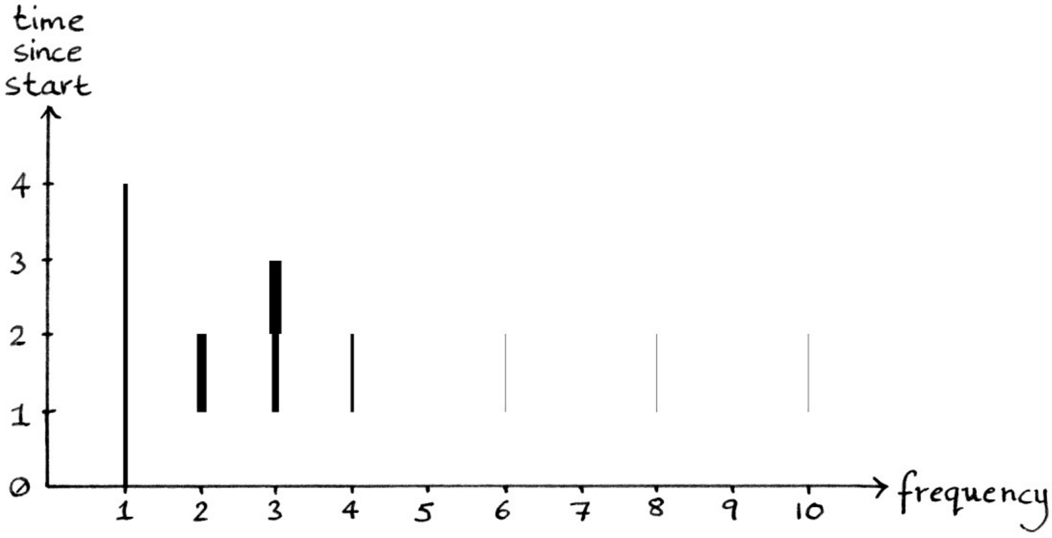
These happen to be exactly the waves that existed during the third second. The 3 cycle-per-second wave has a phase of 180 degrees because in the original signal, it started at 1.5 seconds. Therefore, at the moment it reached the third second, it was 180 degrees through a cycle. Observing it from that moment is the same as it having a phase of 180 degrees. The fact that the calculated constituent wave has a phase dependent on when we observed it, and not based on when it started, means that the resulting phases for all calculated waves are not likely to be particularly accurate. The phases would be required if we were going to build a copy of that particular section of the original signal, but they do not reflect the wave as it appears in the original signal.

The non-time frequency domain graph for this child signal looks like this:



For the fourth child signal, the results are identical to the first child signal. There is one constituent wave: “y = sin 360t”.

Before we think about the advantages and disadvantages of the method we have just used, we will plot all the results on to a single frequency domain graph with time as the y-axis. This frequency domain graph shows the frequencies that appear in the original signal as calculated second by second for the first 4 seconds. [I have marked the frequencies with higher amplitudes with thicker lines. This slightly goes against how graphs are meant to work. In this graph, the thicker lines do not mean that they refer to more than one frequency – they are just there to suggest that the amplitude is higher than a wave with a thinner line. If this book were in colour, I would have used different colours instead.]



For the duration of each second on the graph, the frequencies remain the same. This reflects how we split the original signal up into one-second sections.

Flaws with this method

There are several flaws with this method of analysing an aperiodic signal.

We are not finding the actual waves

The first flaw in the method is that we are not finding the pure waves that were added to create the original signal – instead we are finding the pure waves that could be added to create an approximation of *each second* of the original signal. Our analysis of the child signals gave us enough information to be able to recreate each second of the original signal fairly well. However, we do not necessarily have the exact waves that were added to create the *original* signal, and we definitely do not have the exact moments that all the added waves existed in the signal (unless by chance they existed for an exact multiple of one second).

Unknown starts and finishes

Another flaw is that we only see the frequencies over the duration of one second. If a wave with a particular frequency exists from, say, 0.5 to 1.5 seconds, we would never be able to detect exactly when it started or finished.

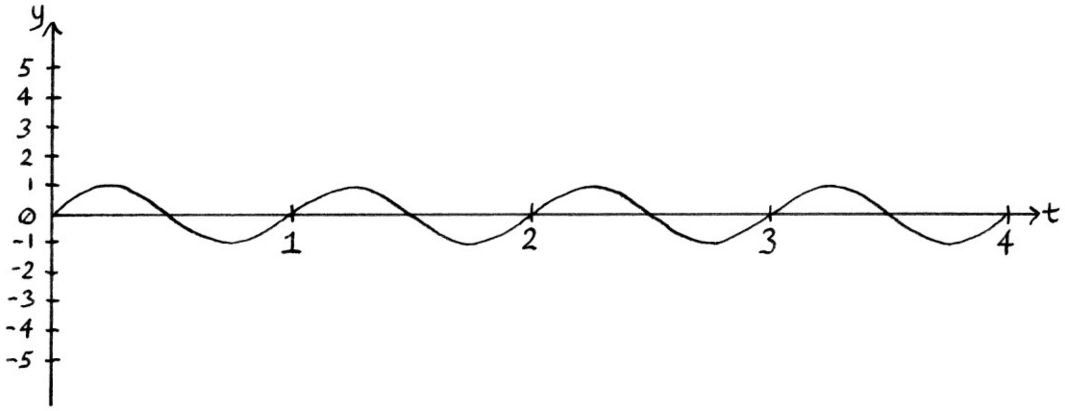
A related problem is that if a wave existed for less than a whole second (while still having one or more cycles within that time), and was within the boundaries of the one second sections, then we might not be able to detect its existence – it would produce spurious results.

A possible solution to this is to use sections of a shorter length than one second.

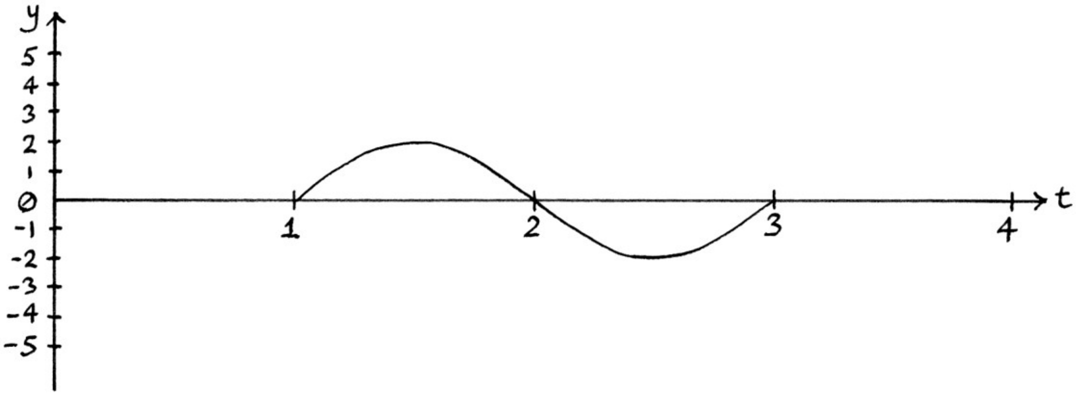
Slower frequencies

If there were a constituent wave with a frequency of less than 1 cycle per second, it would not appear in our results, as we are only testing for frequencies of integer multiples of 1. As an example, we will look at a signal made up of “y = sin 360t” for the whole duration of the signal, and “y = 2 sin (360 * 0.5t)” for 2 seconds, starting at 1 second:

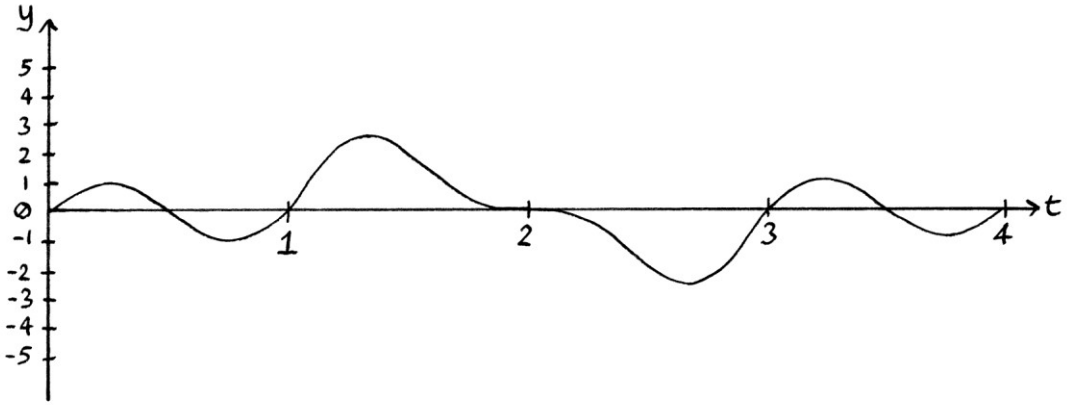
First wave:



Second wave:



The whole signal:



Analysing the child signals from this signal, we would have these results:
For the child signal created from the first second:
"y = 1 sin 360t"

For the child signal created from the second second:
"y = 1.31 sin (360t + 320)"
"y = 0.17 sin ((360 * 2t) + 270)"
"y = 0.07 sin ((360 * 3t) + 270)"
... and a mean level of 1.27 units.

For the child signal created from the third second:
"y = 1.31 sin (360t + 40)"
"y = 0.17 sin ((360 * 2t) + 90)"
"y = 0.07 sin ((360 * 3t) + 90)"
... and a mean level of −1.27 units.

For the child signal created from the fourth second:
"y = 1 sin 360t"

The sum of the waves found for the second and third seconds will be very close to *reproducing* the original curve for those seconds, however, with the exception of the 1 cycle-per-second wave, they are not the frequencies that existed during that second. This is an important distinction – we can recreate a second by second duplicate of the original signal, but the waves that are in that duplicate may or may not be those that were in the original signal.

One interesting observation is that if we have a signal with a zero mean level, that is made up of waves with zero mean levels, and one isolated section of it has a *non-zero* mean level, then it means that that section contains one of the following:

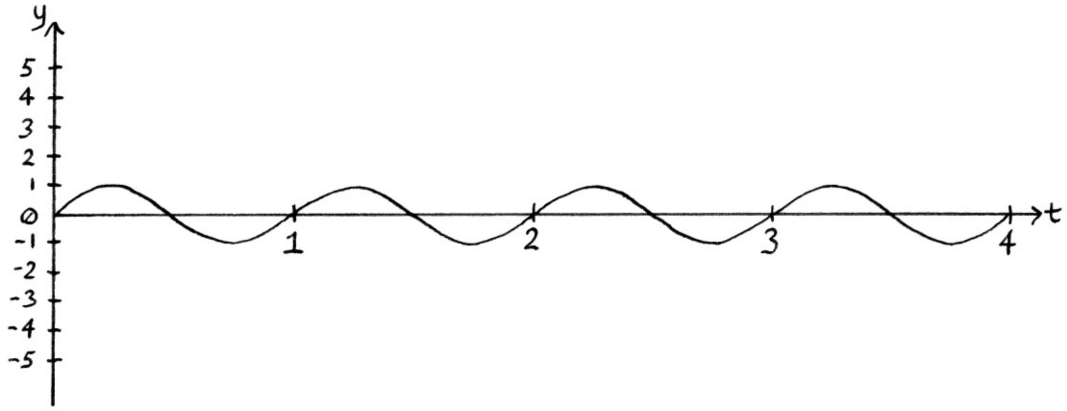
- A new wave starting some time after the start of the section.
- An existing wave ending some time before the end of the section.
- A wave with a period that is longer than the length of the section.
- A wave that does not complete a full cycle. (In some situations, this is the same as a wave with a period that is longer than the length of the section).

Given that, we have some clues as to the true nature of the original signal. With more thought, there might be ways to use this knowledge to improve our analysis. In this particular example, a 0.5-cycle-per-second wave is suggested in the second and third seconds by the nature of the resulting mean levels and phases.

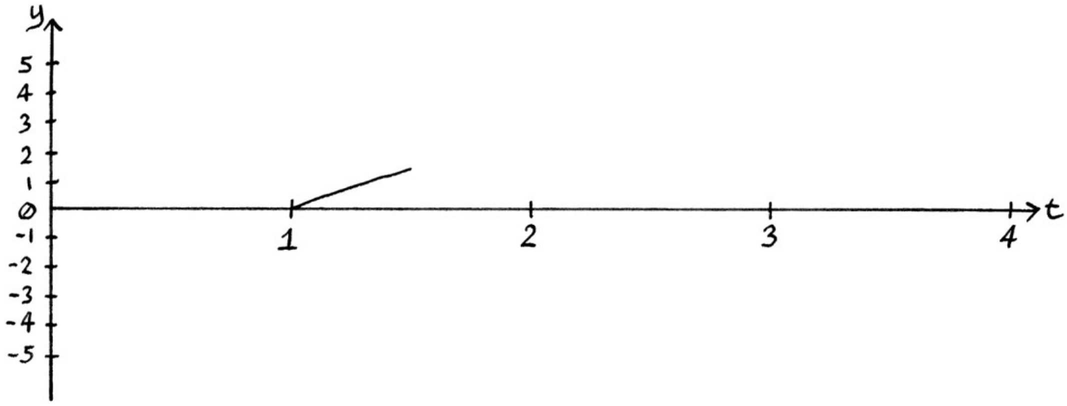
Incomplete cycles

If a wave exists for less time than it can complete one cycle, we would not be able to detect it correctly. As an example, we will look at a signal made up of "y = sin 360t" for the whole duration of the signal, and "y = 2 sin (360 * 0.25t)" for 0.5 seconds, starting at 1 second.

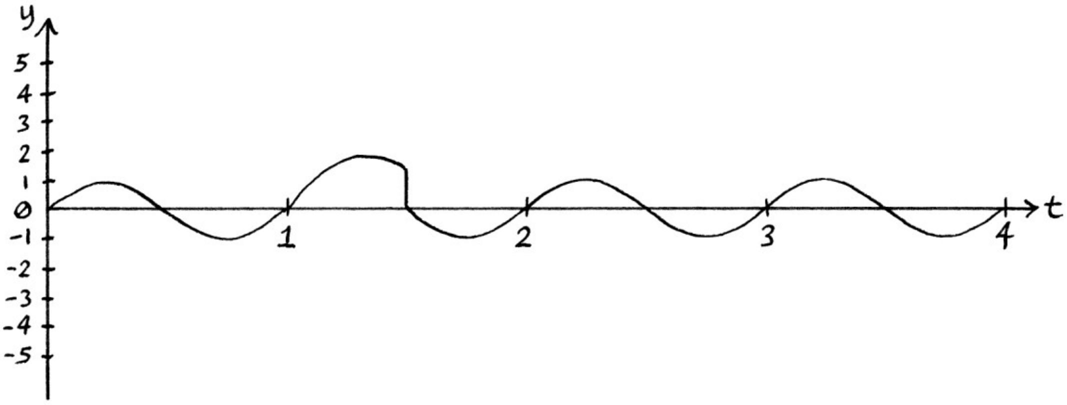
The first wave:



The second wave:



The full signal:



Analysing the child signals, we would end up with these results:
For the child signal created from the first second:
"y = 1 sin 360t"

For the child signal created from the second second:
"y = 1.51 sin (360t + 349)"
"y = 0.23 sin ((360 * 2t) + 183)"
"y = 0.15 sin ((360 * 3t) + 349)"
"y = 0.11 sin ((360 * 4t) + 181)"
"y = 0.09 sin ((360 * 5t) + 353)"
"y = 0.08 sin ((360 * 6t) + 181)"
"y = 0.06 sin ((360 * 7t) + 355)"
"y = 0.06 sin ((360 * 8t) + 181)"
"y = 0.05 sin ((360 * 9t) + 356)"
... and a mean level of 0.37 units.

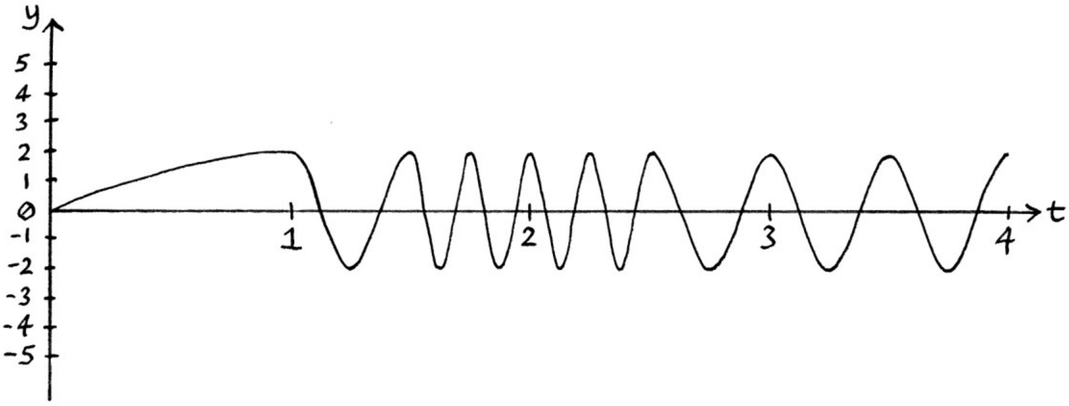
For the child signal created from the third second:
"y = 1 sin 360t"

As before, the results for the second second are enough to create a copy of that second in the original signal, but they bear little resemblance to the actual waves that existed then. Also as before, the non-zero mean level gives a clue as to what might be happening within the signal.

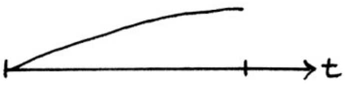
Although we cannot detect the partial wave, there are clues that there might be a partial wave. If we analysed more signals such as this, then it might be the case that we end up seeing patterns in the results that would lead us to being able to make a rule for detecting and identifying partial waves and waves with slow frequencies.

Instant vertical jumps

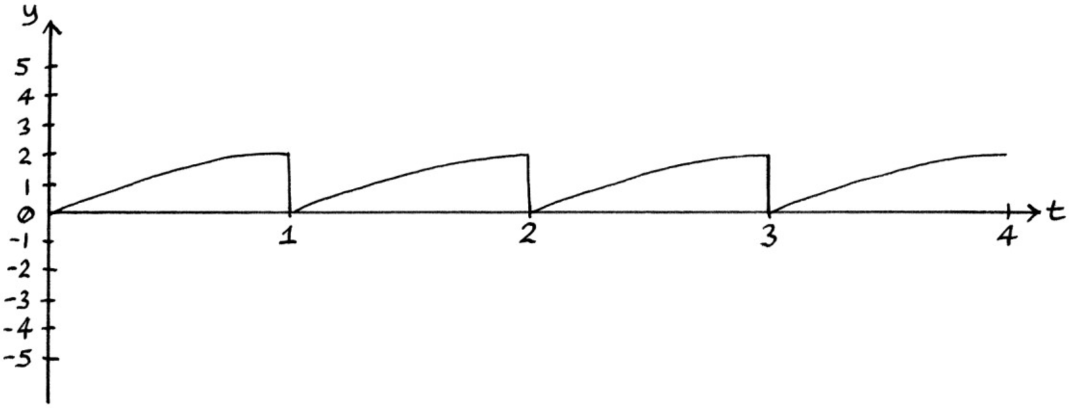
We will look at this signal:



The first second of the signal is as so:

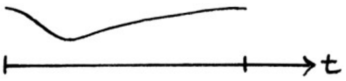


The child signal based on that section looks like this:

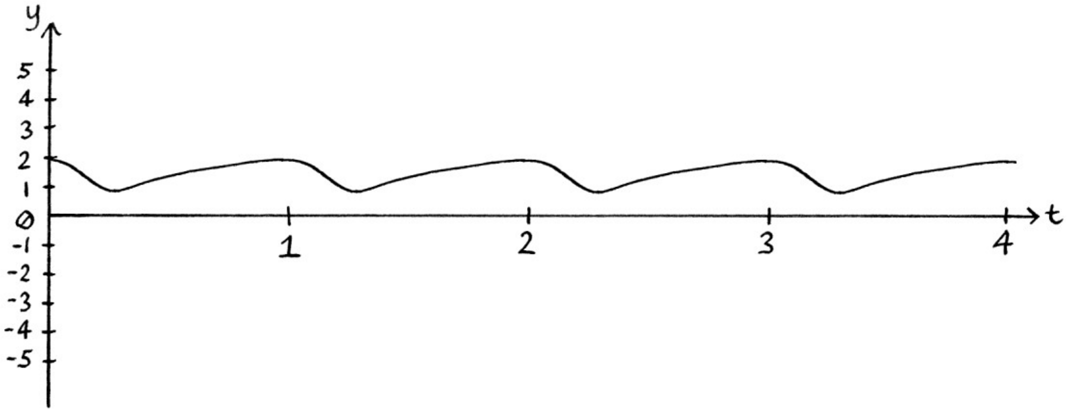


At the end of every cycle, the signal jumps vertically downwards to the start of the next cycle. Whenever there are instant vertical jumps in a signal, it means that our approximation of the constituent waves becomes much less accurate. As we saw in Chapter 18, it is impossible to portray an instant vertical jump accurately using the sum of pure waves. We need to reduce inaccuracies as much as possible, and these jumps do not help us.

One solution to these vertical jumps is to alter the start and ends of each cycle of the child signal, so that they join up to each other without jumps. For the above child signal, we could alter the start and the end of each cycle so that they start and end at the same y-axis value:

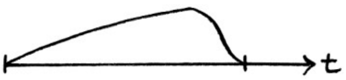


The child signal in full will be free of instant vertical jumps:

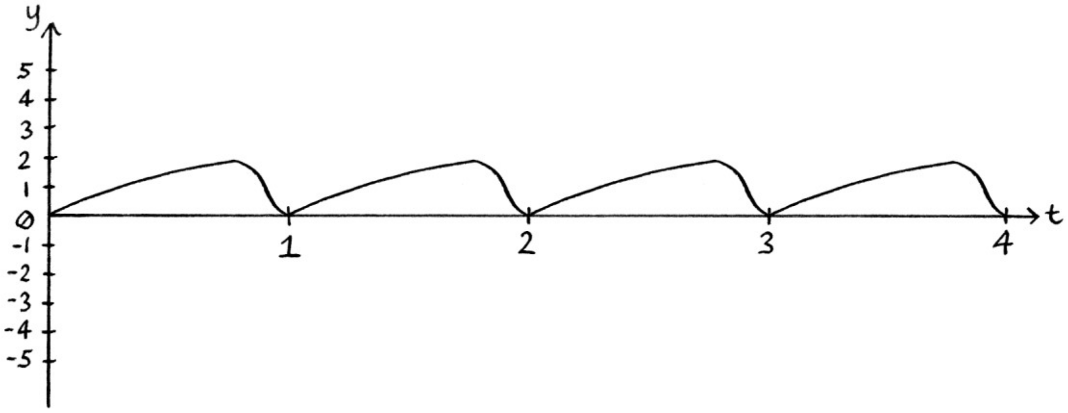


Altering our signal will mean that the child signals will be less faithful to the original signal. On the other hand, it means that we will be able to recreate the sections of the original signal more accurately. Whatever we do, there will be a compromise.

Another way to avoid the jumps is to make the starts of each cycle rise from zero, and to make the ends of each cycle fall to zero. In this way, the cycles are joined up at y = 0. For the first section of our signal, we could alter it to look like this:



The full child signal will look like this:



There are generally used methods of tapering the starts and ends of sections to reduce them to zero, and these are called "window functions". A window function is really a predefined way of scaling a section to fit certain characteristics. Different window functions taper a section in a different way – some are more aggressive than others. Although we used a window function here to remove vertical jumps, they are also used to improve the results of analysing signals whether there are vertical jumps or not. Using window functions in this example will complicate things more than is needed, so we will ignore the idea for now.

Phases

When we analyse a section of the original signal, we are finding the phases of the waves *as if they started at the beginning of that section*. This, of course, will not be the true phase of the actual constituent wave, unless, by chance, it happened to coincide with the beginning of that section. This means that knowing the phases of the calculated constituent waves may or may not be useful. Knowing the calculated phases would be useful if we wanted to reconstruct a copy of one particular section of the original wave, however it pays to remember that they are probably not the phases of the actual added wave sections. [Saying all that, for simple signals, we might be able to work out when a constituent wave started by seeing what its perceived phase is at the start of a section].

Thoughts

All constituent wave pieces will affect the y-axis of the curve. Therefore, if we cannot detect each one exactly, our calculations will be distorted and our results will be incorrect. Undetected wave pieces can be thought of as “noise”. In such cases, despite the analysis missing out pieces of waves, we will still be creating a good approximation of the original signal. We could recreate the signal very closely using our approximation. The trouble is that our approximation is not necessarily representative of the exact original constituent wave pieces and when they actually started or finished. Whether this is a problem depends on what it is we are trying to achieve. If we want to know some pieces of waves that, if added, would make a copy of the signal, then this is not a problem; if we want to know the exact original constituent waves for every moment in time, then it is a problem.

We can have a more accurate idea of when pieces of waves start and end by splitting the original signal up into overlapping sections, instead of independent sections, which is what we will do next.

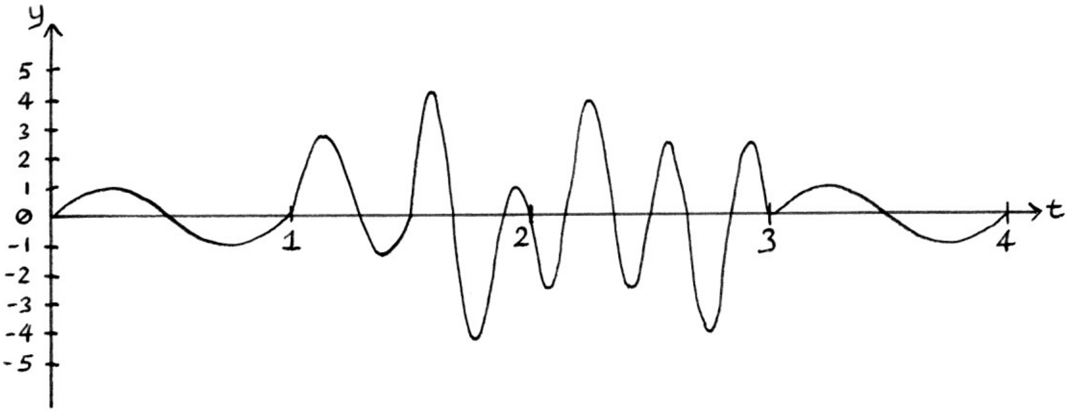
Something that may or may not be obvious is that we do not actually need to create entire child signals from each piece of the original signal. We only ever deal with the very first cycle of each child signal, so whether or not that section repeats is irrelevant to any calculations. However, having the child signals repeat makes what we are doing clearer. It helps in learning.

Overlapping split signals

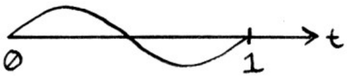
So far, we have been splitting the original signal up into equal sections of time, and analysing each section. Now we will analyse overlapping sections of the original signal. To do this we take sections of the same size as before, but each section will only be slightly further into the signal. For example, we can take 1 second’s worth of the signal starting at t = 0, then take another second’s worth of the signal at t = 0.1 seconds, then another second’s worth at t = 0.2 and so on. We then treat all of these pieces as cycles from which we can create our child signals.

We will have a lot more work to do, as we will be analysing many more child signals. However, the process will mean that we can see the variations in frequency for more moments in time.

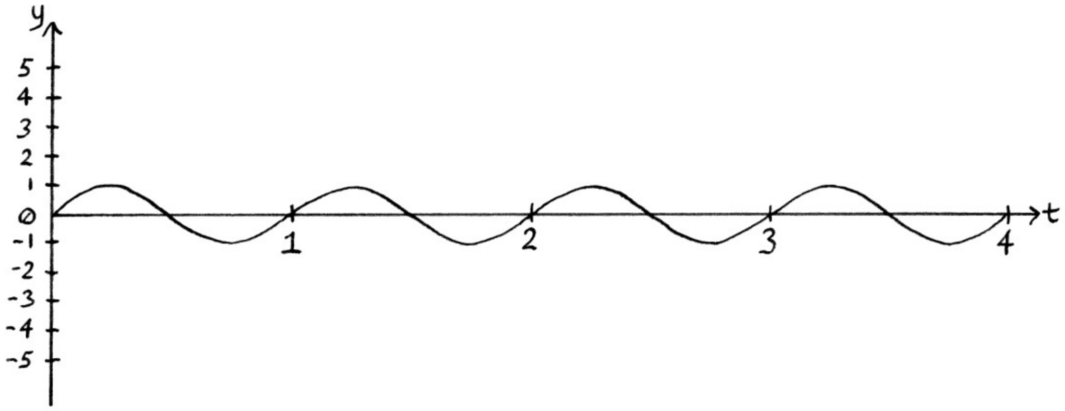
We will use the signal from earlier:



We will take one second from it, starting at t = 0:

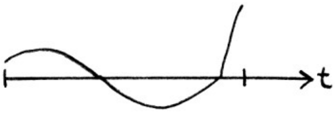


We then repeat this second to create our first child signal:

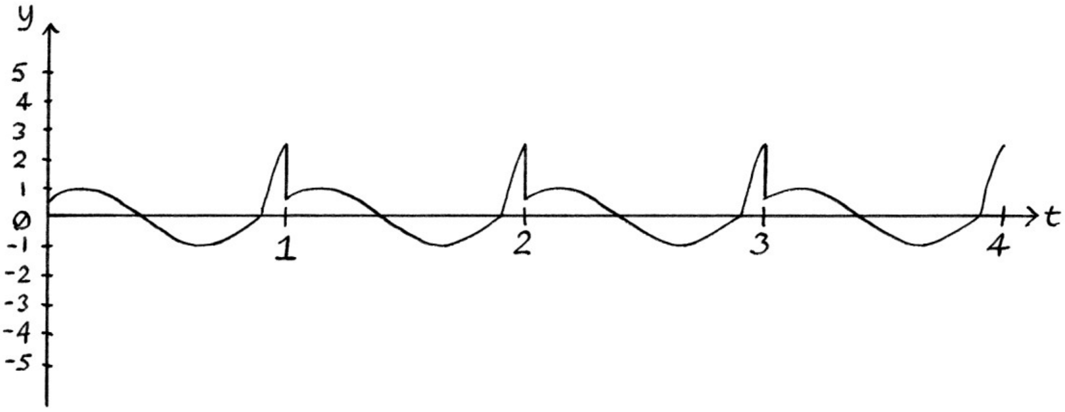


We analyse this child signal, and as before, we will find out that it consists entirely of "y = sin 360t".

We then take another one second from the original signal, starting at t = 0.1 seconds.

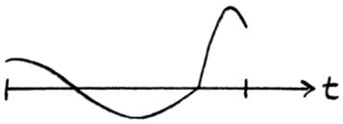


We turn this section into our second child signal:

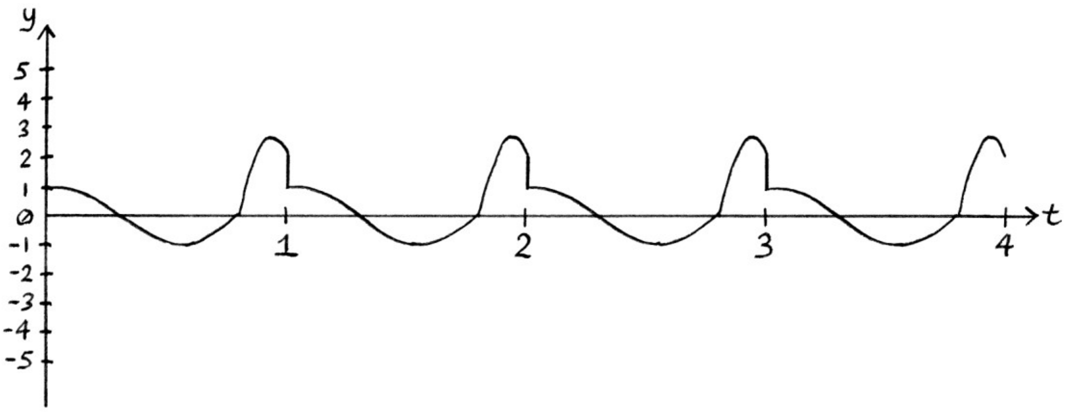


We will analyse this signal to find its constituent waves. The first few waves are as follows:
“y = 1.10 sin (360t + 46)”
“y = 0.21 sin ((360 * 2t) + 115)”
“y = 0.20 sin ((360 * 3t) + 128)”
“y = 0.18 sin ((360 * 4t) + 140)”
“y = 0.16 sin ((360 * 5t) + 151)”
“y = 0.14 sin ((360 * 6t) + 162)”
“y = 0.12 sin ((360 * 7t) + 172)”
“y = 0.10 sin ((360 * 8t) + 180)”
... and a mean level of 0.11 units.

We then take one second from the original signal, starting at t = 0.2 seconds.



We turn this into a child signal:



We analyse this signal. The first few constituent waves with amplitudes over 0.10 units are:
“y = 1.43 sin (360t + 89)”
“y = 0.48 sin ((360 * 2t) + 151)”
“y = 0.37 sin ((360 * 3t) + 180)”
“y = 0.25 sin ((360 * 4t) + 208)”
“y = 0.14 sin ((360 * 5t) + 231)”
... and a mean level of 0.28 units.

We continue in this way, taking one-second sections from the original signal, and turning them into child signals. This might seem straightforward, but we have to know when to stop. If the original signal continues forever, then we would continue forever too. If the original signal has a finite length, then we need to extend the signal for an extra second in which all the y-axis values are zero. We then continue collecting sections until we end up with a section with nothing in it. If we stopped earlier – at the last whole second of the original signal – we would not find out if any new waves appeared within the last second.

For our original signal example, we will go through it for 3 seconds at steps of 0.1 seconds. This means that we will have 30 child signals. For each of these child signals there will be a range of particular frequencies.

To make the results easier to digest, we will ignore constituent waves with amplitudes under 0.10 units. The constituent waves are as follows:

For the second starting at t = 0:
“y = sin 360t”

For the second starting at t = 0.1:
“y = 1.10 sin (360t + 46)”
“y = 0.21 sin ((360 * 2t) + 115)”
“y = 0.20 sin ((360 * 3t) + 128)”
“y = 0.18 sin ((360 *4t) + 140)”
“y = 0.16 sin ((360 * 5t) + 151)”
“y = 0.14 sin ((360 * 6t) + 162)”
“y = 0.12 sin ((360 * 7t) + 172)”
“y = 0.10 sin ((360 * 8t) + 180)”
... and a mean level of 0.11 units.

For the second starting at t = 0.2:
“y = 1.43 sin (360t + 89)”
“y = 0.48 sin ((360 * 2t) + 151)”
“y = 0.37 sin ((360 * 3t) + 180)”
“y = 0.25 sin ((360 * 4t) + 208)”
“y = 0.14 sin ((360 * 5t) + 231)”
... and a mean level of 0.29 units.

For the second starting at t = 0.3:
“y = 1.43 sin (360t + 126)”
“y = 0.53 sin ((360 * 2t) + 222)”
“y = 0.42 sin ((360 * 3t) + 280)”
“y = 0.25 sin ((360 * 4t) + 332)”
... and a mean level of 0.28 units.

For the second starting at t = 0.4:
“y = 1.27 sin (360t + 175)”
“y = 0.86 sin ((360 * 2t) + 298)”
“y = 0.61 sin (360 * 3t)”
“y = 0.18 sin ((360 * 4t) + 40)”
“y = 0.15 sin ((360 * 5t) + 344)”
“y = 0.14 sin ((360 * 6t) + 18)”
... and a mean level of 0.11 units.

For the second starting at t = 0.5:
“y = 1.31 sin (360t + 220)”
“y = 1 sin (360 * 2t)”
“y = 0.51 sin ((360 * 3t) + 90)”
“y = 0.12 sin ((360 * 5t) + 90)”
... and a mean level of zero units.

For the second starting at t = 0.6:
“y = 0.80 sin (360t + 236)”
“y = 1.50 sin ((360 * 2t) + 89)”
“y = 0.90 sin ((360 * 3t) + 161)”
“y = 0.53 sin ((360 * 4t) + 142)”
“y = 0.44 sin ((360 * 5t) + 169)”
“y = 0.41 sin ((360 * 6t) + 166)”
“y = 0.29 sin ((360 * 7t) + 180)”
“y = 0.28 sin ((360 * 8t) + 187)”
“y = 0.20 sin ((360 * 9t) + 187)”

“y = 0.17 sin ((360 * 10t) + 197)”
“y = 0.14 sin ((360 * 11t) + 186)”
“y = 0.10 sin ((360 * 12t) + 187)”
“y = 0.11 sin ((360 * 13t) + 180)”
“y = 0.10 sin ((360 * 14t) + 171)”
“y = 0.10 sin ((360 * 15t) + 177)”
“y = 0.11 sin ((360 * 16t) + 174)”
... and a mean level of 0.32 units.

There are clearly a lot of incorrect frequencies that were not in the original signal.

To save time, we will skip the rest of the results.

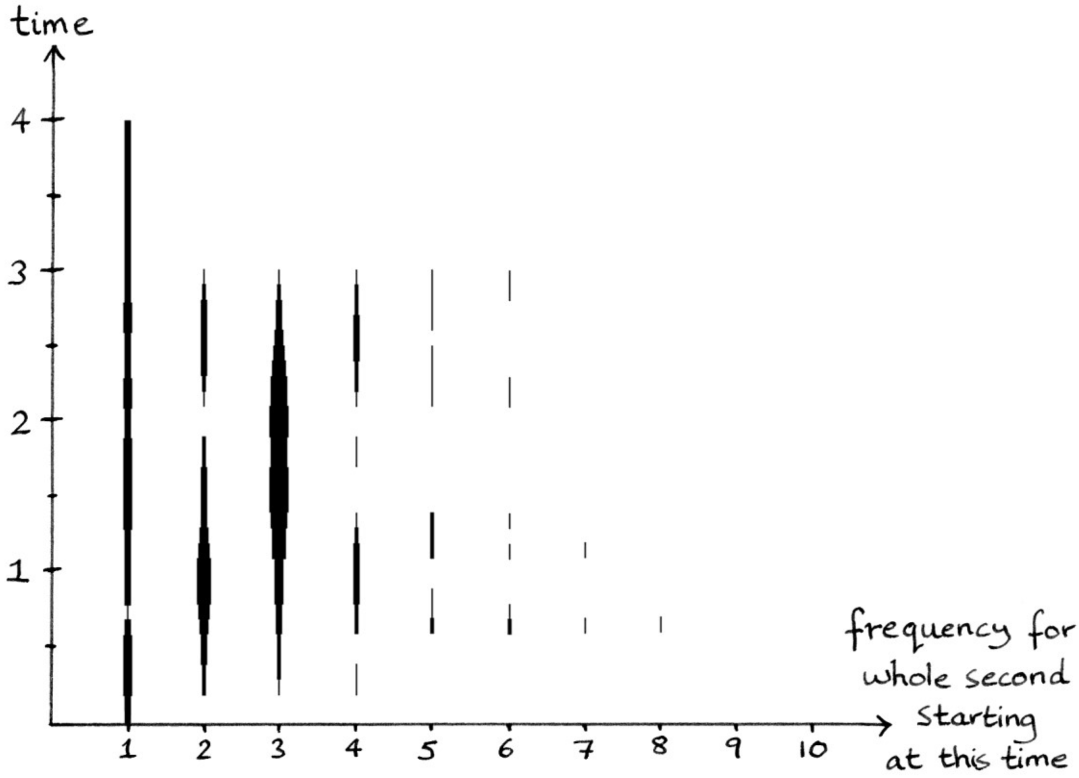
What the results mean

Earlier in this chapter, when we were taking separate sections of the original signal, we were really finding a list of constituent waves that if added would recreate a one second section of the signal. They were not literally the waves that were added, but they were the waves that would approximately recreate that second. Now, because we are overlapping each second, our results will not be able to recreate a copy of the original signal – at best, we could recreate a copy of any one second of the signal as long as it started on a 0.1 second boundary.

The meaning of what we have is apparent if we try to plot the results on a frequency domain graph. The time axis will be divided into 0.1 second sections. The plotted waves existing in each 0.1 second section are not the list of frequencies for that 0.1 second, but the list of frequencies for one second, *starting at that* 0.1 second time.

The frequency domain graph for this analysis of the signal is shown in the following picture. The graph only shows the presence of waves with amplitudes higher than 0.5 units. Higher amplitudes have been marked with thicker lines. [As before, a thicker line in this graph just means that there is a higher amplitude for that one particular frequency – it does not mean that there are multiple frequencies centred around that frequency.]

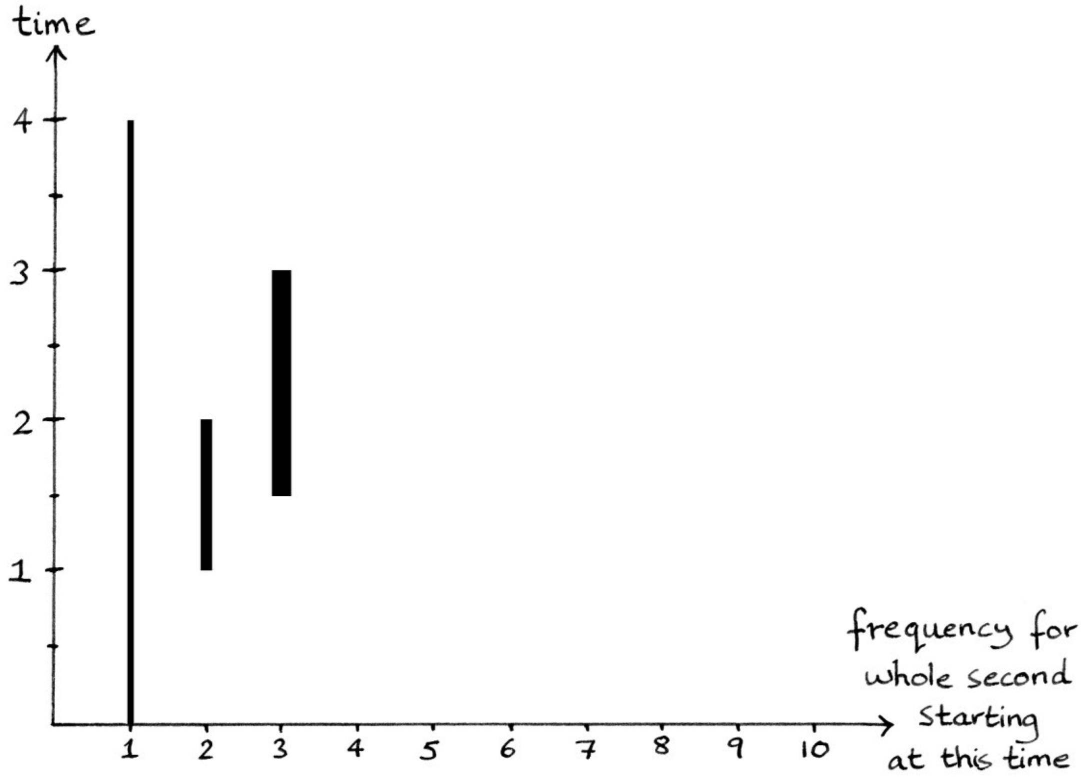
The graph is as so:



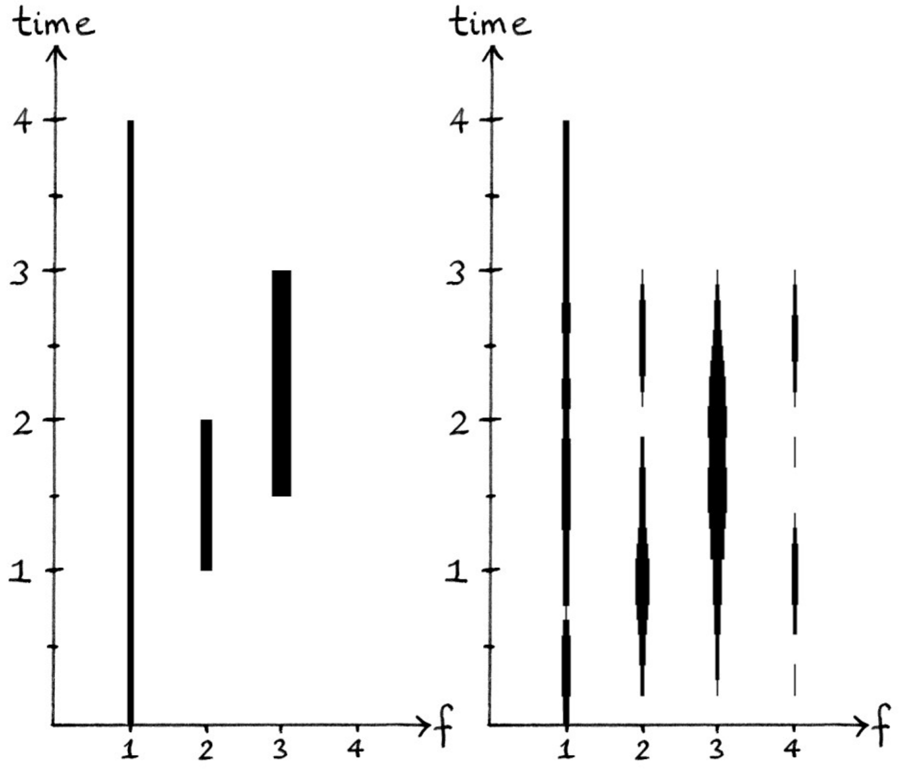
The signal originally consisted of:
"y = sin 360t" for the length of the whole signal.
"y = 2 sin (360 * 2t)" for one second, starting at t = 1 second.
"y = 3 sin (360 * 3t)" for 1.5 seconds, starting at t = 1.5 seconds.

The graph is not correct. However, given that we know what the graph should look like, we can see that the gist of the situation is in the graph. The waves in the graph start too early, and there are waves that should not be there. However, the graph could work as a *very* rough guide.

Using the same layout, if the analysis had been perfect, the graph would have looked like this: [With higher amplitudes marked as thicker lines.]



For comparison, the significant parts of each graph side by side look like this:

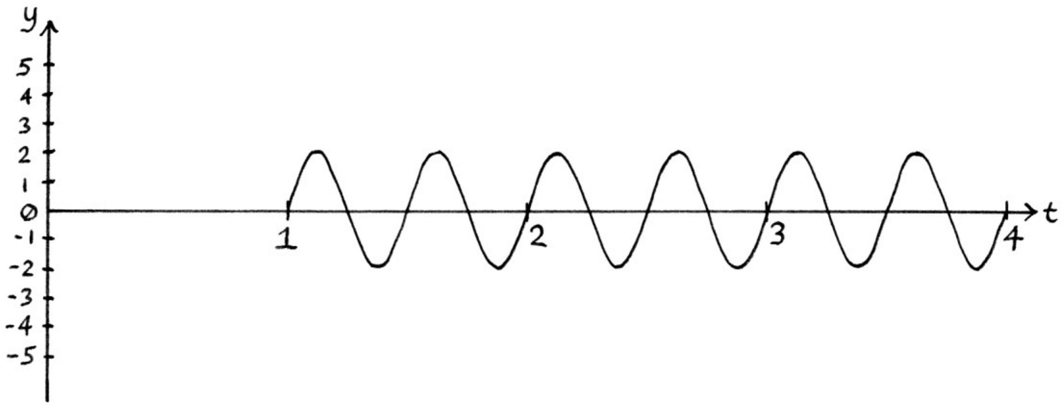


We can see that the start of a new wave in the signal distorts the amplitude of waves in the calculated results.

Flaws with this method

The flaws in this method are similar to the flaws in the first method, except now we are more likely to see the moment when a new constituent wave starts and ends. We still cannot detect waves with frequencies of less than that of the section size (which was 1 cycle per second in this example). We still cannot see waves that exist for less than a full cycle. This overlapping method will result in more situations where there are vertical jumps between cycles, although we know a method for reducing the effects of that. We now have much more noise in the form of spurious calculated constituent waves. Overall, the method is not particularly good, although it could be useful in certain situations.

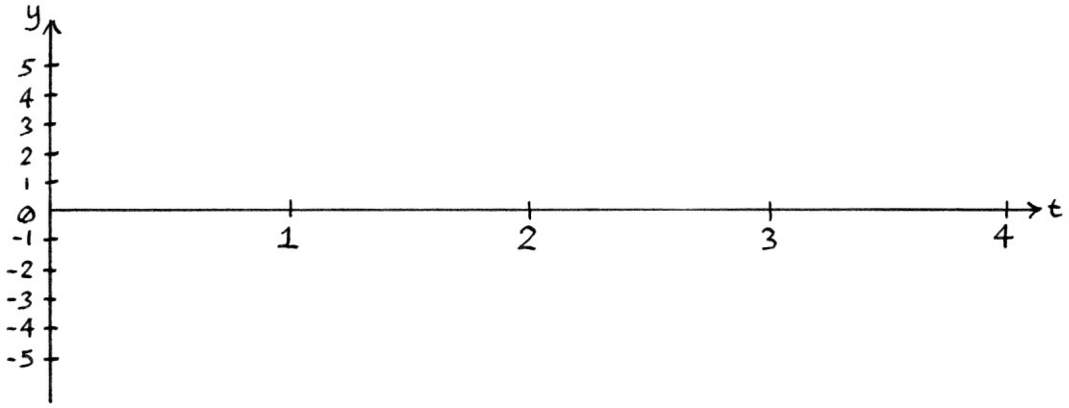
There is a new flaw that distorts how the results appear. To demonstrate this, we will say we have a signal that has y-axis values of zero until 1 second, when a Sine wave with a frequency of 2 cycles per second starts:



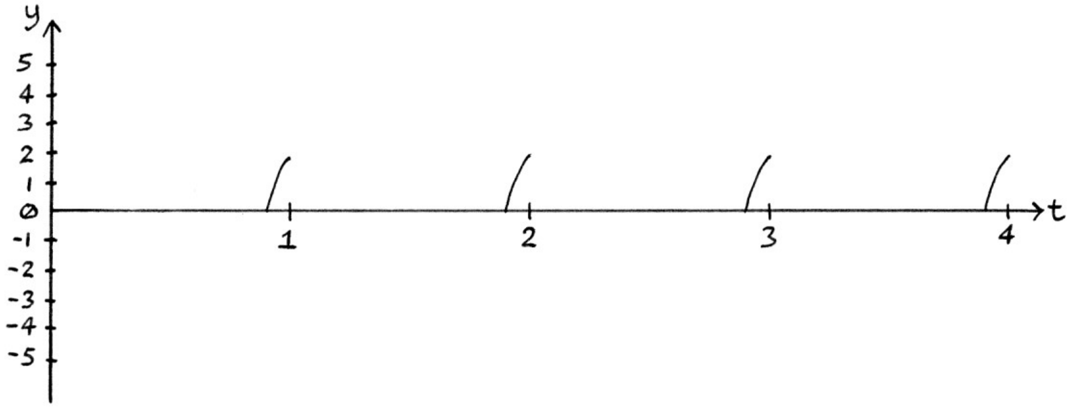
As before, to analyse this signal, we will take sections that are one second long, and that start at 0 seconds, 0.1 seconds, 0.2 seconds and so on.

The child signals look like this:

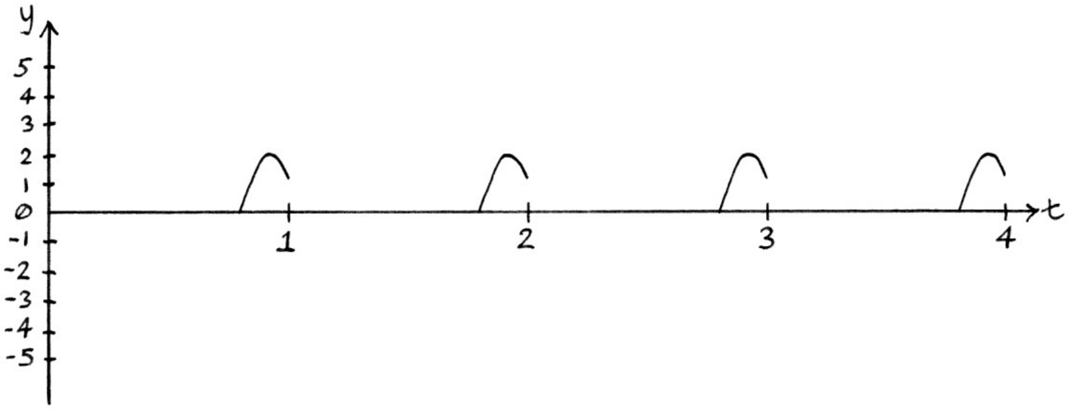
For the second starting at 0 seconds:



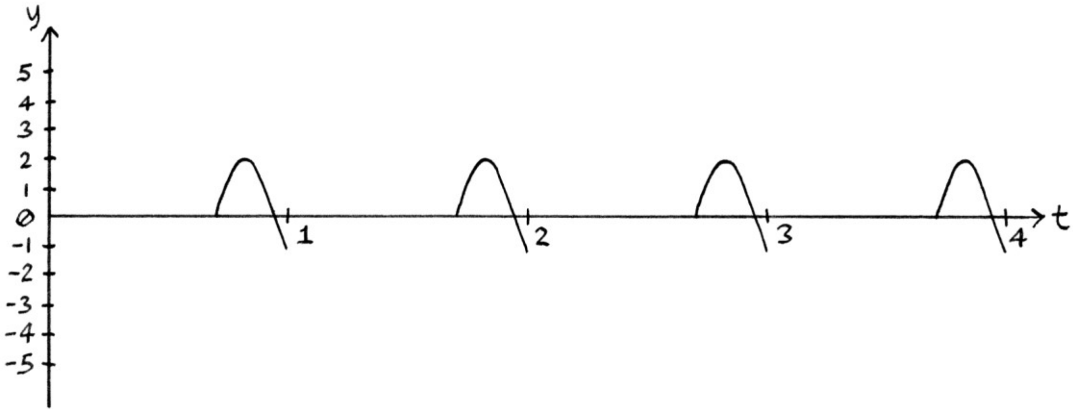
For the second starting at 0.1 seconds:



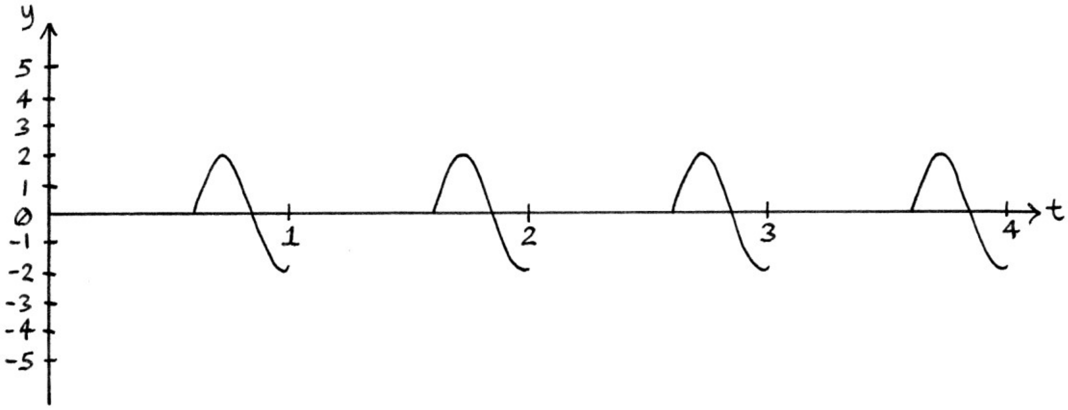
For the second starting at 0.2 seconds:



For the second starting at 0.3 seconds:



For the second starting at 0.4 seconds:



The problem here is that we will be analysing the non-zero part of the signal long before the time when it actually starts. There are no waves for the whole of the first second, and our analysis should indicate that. However, because our sections are one second long, the start of the non-zero part of the original signal is included in our calculations from the second child signal onwards. This leads to misleading results – the boundaries between the blank part of the signal and the non-blank part are blurred.

For the first child signal, the results of the calculations will show (correctly) that there are no constituent waves in it.

The calculations for the second child signal will produce the following list of waves (excluding waves with amplitudes lower than 0.1 units):

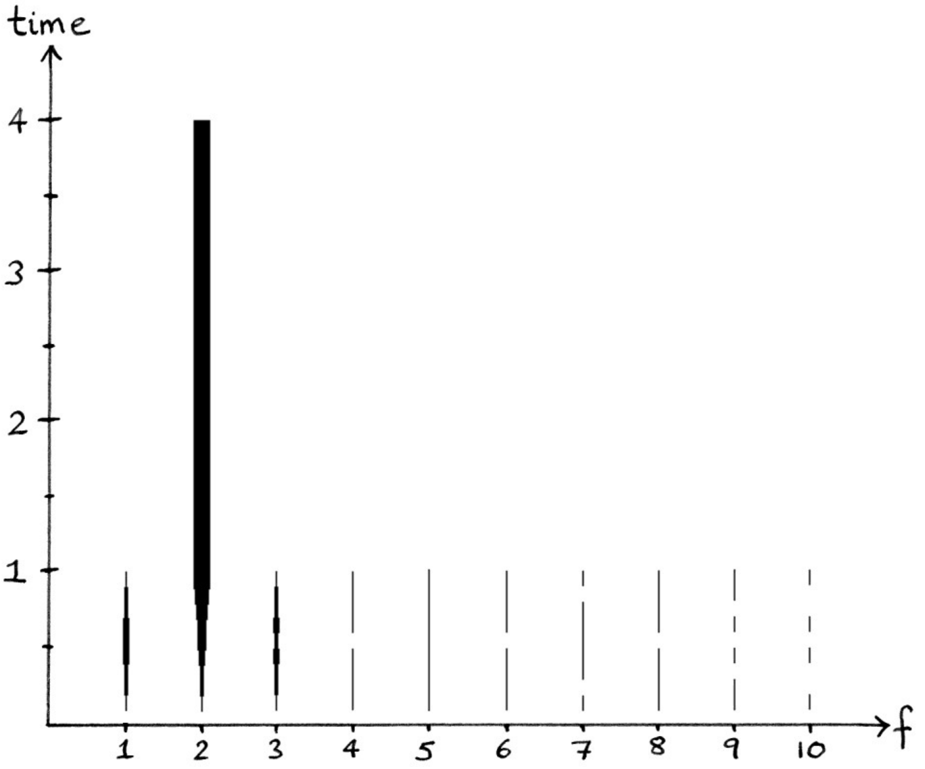
“y = 0.22 sin (360t + 103)”
“y = 0.21 sin ((360 * 2t) + 115)”
“y = 0.20 sin ((360 * 3t) + 128)”
“y = 0.18 sin ((360 * 4t) + 140)”
“y = 0.16 sin ((360 * 5t) + 151)”
“y = 0.14 sin ((360 * 6t) + 162)”
“y = 0.12 sin ((360 * 7t) + 172)”
“y = 0.10 sin ((360 * 8t) + 180)”
... and a mean level of 0.11 units.

This is a good example of how, when we are looking at the frequency domain graph for this type of calculation, we are not seeing the frequencies for a particular time but for the entire second starting at that time.

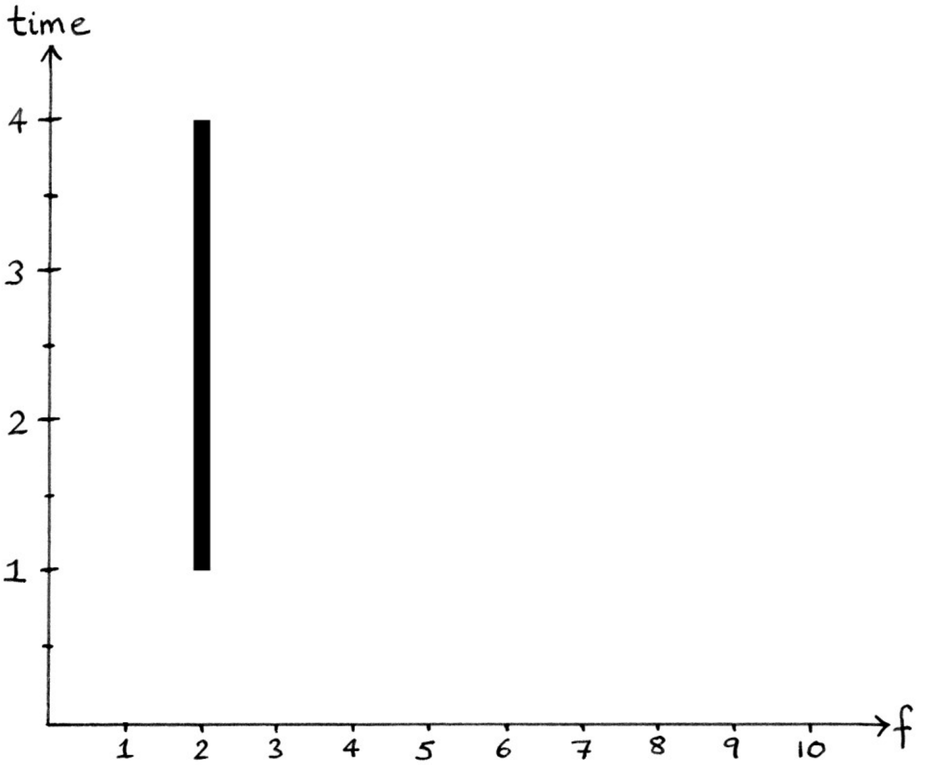
This “blurring” of results could give the impression that the change from having no waves in the signal to having one wave in the signal causes there to be non-zero frequencies long before the wave starts. However, these frequencies only exist due to the way that we have done the calculations. In reality, at 0.999999... seconds, there are no constituent waves, and at 1.000000.... seconds, there is one constituent wave with a frequency of 2 cycles per second. In reality, the only time that there is a frequency other than 0 cycles per second or 2 cycles per second would be if we took the *average* frequency over the brief time that the signal changes from 0 to 2 cycles per second. For example, if we looked at the time from 0.9 seconds to 1.1 seconds, or from 0.999 seconds to 1.001 seconds, the average frequency would be (0 + 2) ÷ 2 = 1 cycle per second. The reason for this becomes clearer if we remember that frequency is analogous to speed, and we imagine a vehicle travelling at 0 kilometres per hour that instantly starts travelling at 2 kilometres per hour. Its average speed over the speed change would be 1 kilometre per hour.

The problem of blurring can be reduced by using smaller sections. The smaller the sections used in the process, the less the problem of blurring will be, but at the expense of introducing other problems.

The frequency domain graph showing *our analysis of this signal* is as follows. [The thickness of the lines is proportional to the amplitude for each frequency.]:



In reality, the frequency domain graph should look like this:



As a side note, the blurring problem is a good example of how flaws in a process can lead to a skewed view of reality. If I had said that the overlapping process was infallible, then the results from the above example might make one believe that in the real world, the start of a new wave alters the other frequencies of a signal some time before it starts. However, that is a side effect of the analysis process, and nothing to do with reality. The frequency domain graph for a signal is based on the *calculations* used to analyse that signal, and not on reality. Mathematical calculations and reality might coincide, but not necessarily. The analysis calculations might be mathematically consistent, and we could create a whole mathematical world based on those calculations, but they still might not reflect reality.

Thoughts

The way that this overlapping process works is that constituent waves with frequencies slower than the frequency of a child signal will produce nonsense results – they are essentially noise. For example, a wave with a frequency of 0.5 cycles per second will not be detected as such if we are only testing for frequencies of integer multiples of 1 cycle per second. If our sections are a second long, then there is no way to detect a 0.5 cycle-per-second wave. If we increase the section size, we can detect slower frequencies, but we will lose the ability to discern when the waves start and finish. The situation is analogous to examining something nearby through either end of a telescope – whichever way around the telescope is held, we will not see all the details of the object being observed. When looking at the results of analysing a signal containing slower frequency waves, there are clues that the slower frequency waves are there, but to know what they are would take more work.

If we could detect partial cycles of waves, we could solve this particular problem. A partial cycle of a particular wave is unique to that wave. The trouble is that recognising that partial cycle within a sum of waves is harder to do than recognising a full cycle within a sum of waves.

We could go through the signal with one particular section size to look for particular frequencies, and then go through it with a different section size to look for slower frequencies. We could then combine the results somehow. This would take much more work, but might possibly give more detailed results. If we were dealing with a wave in real time as it was being received, then we might not be able to do this.

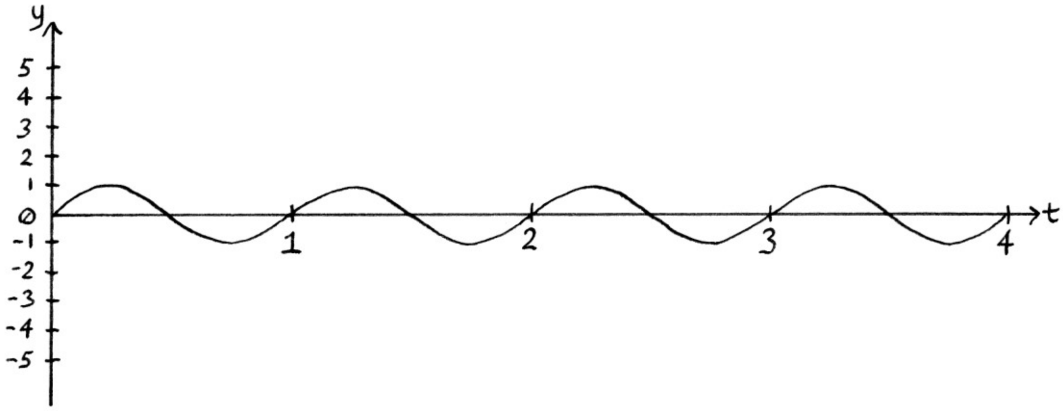
In this example, we went through the signal moving onwards one little piece of time by one little piece of time. Theoretically, there is no minimum amount of time that we could choose to step onwards by. If we were dealing with discrete waves (where the wave is stored as a series of y-axis values at equally spaced moments in time), then there *is* a minimum step – the minimum step would be one y-axis value.

Periodic signals

Ignoring any possible improvements to the "overlapping" process, we will see how it works on various signals. The process works on periodic signals and waves, although it is debatable as to whether there is much advantage in using it in this way.

Periodic waves

For a periodic wave, if the section length matches the period of the wave (or is an integer divisor of it), then the "overlapping" process will find the wave's characteristics, but the phase will be wrong for the sections that do not coincide with the start of a cycle. For example, we will analyse the following wave, which is "y = sin 360t":

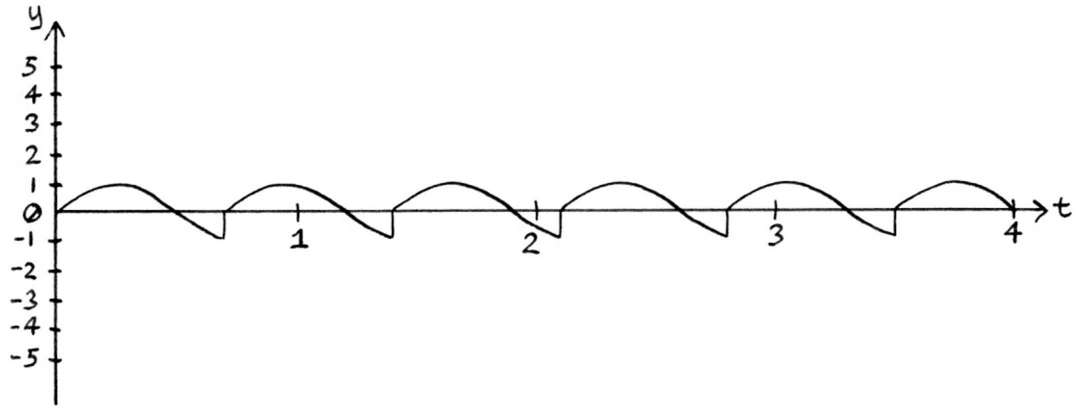


We will use a section length of 1 second. Analysis of this wave will give the following results:

For the section starting at 0 seconds: “y = sin 360t”
For the section starting at 0.1 seconds: “y = sin (360t + 36)”
For the section starting at 0.2 seconds: “y = sin (360t + 72)”
For the section starting at 0.3 seconds: “y = sin (360t + 108)”
For the section starting at 0.4 seconds: “y = sin (360t + 144)”
... and so on.

The results are correct, with the exception of the phase for the later results. However, we could work out what the exact phase is by thinking about the apparent phase at any particular time.

If the section length does not match the period of the wave (and is not an integer divisor of it), then we will end up with other results. Using the same wave of “y = sin 360t”, but a section length of 0.7 seconds, the results will consist of a lot of noise. For one thing, if the section length is 0.7 seconds, then the fundamental frequency is 1 ÷ 0.7 = 1.43 cycles per second. Therefore, we will be testing for frequencies that are integer multiples of 1.43 cycles per second, which means that we will never have correct results. We will also suffer from sudden jumps in the child signals, as is shown in this picture of the first child signal:



These jumps will create more noise in our results. The results are as so (ignoring amplitudes below 0.1 units):

For the section starting at 0 seconds:
“y = 0.82 sin ((360 * 1.43t) + 316)”
“y = 0.19 sin ((360 * 2.85t) + 334)”
“y = 0.11 sin ((360 * 4.29t) + 342)”
... and a mean level of 0.30 units.

For the section starting at 0.1 seconds:

“y = 0.98 sin ((360 * 1.42t) + 347)”
“y = 0.28 sin ((360 * 2.86t) + 354)”
“y = 0.17 sin ((360 * 4.29t) + 356)”
“y = 0.13 sin ((360 * 5.71t) + 357)”
“y = 0.10 sin ((360 * 7.14t) + 357)”
... and a mean level of 0.11 units.

For the section starting at 0.2 seconds:

“y = 0.98 sin ((360 * 1.43t) + 13)”
“y = 0.28 sin ((360 * 2.86t) + 6)”
“y = 0.17 sin ((360 * 4.29t) + 4)”
“y = 0.13 sin ((360 * 5.71t) + 3)”
“y = 0.10 sin ((360 * 7.14t) + 3)”
... and a mean level of −0.11 units.

For the section starting at 0.3 seconds:

“y = 0.82 sin ((360 * 1.43t) + 44)”
“y = 0.19 sin ((360 * 2.86t) + 26)”
“y = 0.11 sin ((360 * 4.29t) + 18)”
... and a mean level of −0.30 units.

For the section starting at 0.4 seconds:

“y = 0.71 sin ((360 * 1.43t) + 90)”
“y = 0.10 sin ((360 * 2.86t) + 90)”
... and a mean level of −0.37 units.

... and so on.

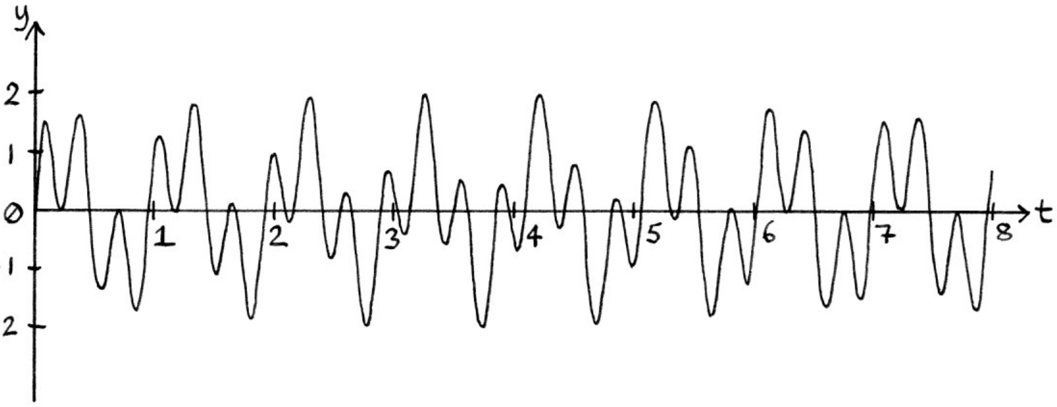
If one were presented with the results and no other information, it might be possible to make some deductions about the underlying wave being analysed. [One could, of course, add the found constituent waves to recreate the actual wave, piece by piece].

Periodic signals

For periodic *signals* that were created by adding pure waves, the effects will be similar to those for a pure wave. If the section length matches the period of the signal, we will have good results with incorrect phases (unless the section starts at the beginning of a cycle, when the phase will be correct). If the section length does not match the period of the signal, then we will have spurious results. If the signal is one that could not be created by adding pure waves, then the results will be slightly worse in each case.

Irrational frequency ratio signals

To demonstrate what happens when we use the "overlapping" process on aperiodic signals that were created by the addition of waves with irrational frequency ratios (and not from adding sections of waves at particular times), we will use the process on the following signal. It is created by adding "y = sin 360t" and "y = sin (360 * πt)". Such a signal was mentioned in Chapter 13 on the addition of waves.



We will use a section length of 1 second and intervals of 0.1 seconds. We will ignore any calculated constituent waves with amplitudes under 0.2 units (which is an arbitrarily chosen value to filter out unwanted noise).

For the first child signal, we will have:
"y = 1.03 sin (360t + 2)"
"y = 0.95 sin ((360 * 3t) + 27)"
... and a mean level of 0.02 units.

For the second child signal, we will have:
“y = 1.02 sin (360t + 40)”
“y = 0.96 sin ((360 * 3t) + 137)”
... and a mean level of 0.03 units.

For the third child signal, we will have:
“y = 0.91 sin (360t + 71)”
“y = 0.99 sin ((360 * 3t) + 252)”
... and a mean level of −0.04 units.

For the fourth child signal, we will have:
“y = 1.00 sin (360t + 106)”
“y = 0.95 sin ((360 * 3t) + 5)”
... and a mean level of 0 units.

For the fifth child signal, we will have:
“y = 1.06 sin (360t + 141)”
“y = 0.98 sin ((360 * 3t) + 117)”
... and a mean level of 0.04 units.

If we continued for longer, we would continue to see results that are based around “y = sin 360t” and “y = sin (360 * 3t)”, but with nearby amplitudes and various phases. We will never find a constituent wave with a frequency of, say, 3.1 cycles per second because we are only testing integer multiples of 1 cycle per second.

As discussed in Chapter 13, a signal made from adding two waves with an irrational frequency ratio resembles a signal made from adding two waves with a rational frequency ratio, where the phases are constantly changing. Our analysis of this signal is consistent with this idea.

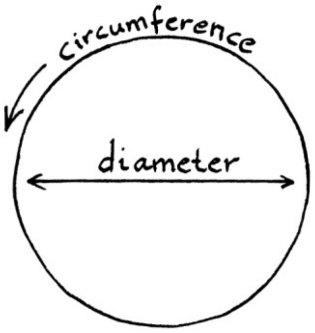
Conclusion

This chapter was intended to introduce the idea of analysing aperiodic signals, and give a brief insight into the difficulties of doing so. It *is* possible to analyse aperiodic signals correctly, but not with the methods described in this chapter. The most common methods for analysing aperiodic signals in signal processing are based on the Fourier Transform, which is more complicated than the methods described in this chapter.

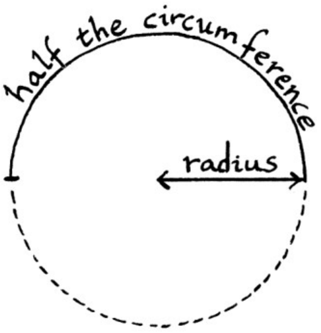
Chapter 21: π

The number "π" or "pi" is frequently used when discussing circles, and given that waves are ultimately based on circles, it plays a significant part in dealing with waves. The symbol "π" (pronounced "pie") is the lower-case Greek equivalent of the Latin letter "p" as in "papa". The symbol is used to represent the number: 3.14159265358979323846264338327950288419716...

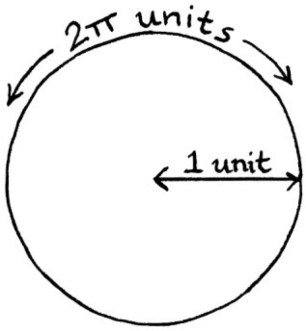
The number π is equal to the length of the circumference of a circle divided by the diameter:



To put this another way, π is the length of half the circumference of a circle divided by its radius:

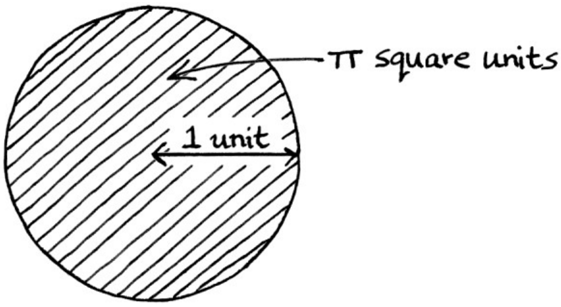


A circle with a radius of 1 unit will have a circumference of exactly 2π units.



A circle with a radius other than 1 unit will have a circumference that is scaled by the same amount that that radius is scaled to 1. In other words, if the radius is 0.5 units, then the circumference will be 0.5 * 2π = π units. If the radius is 2 units, then the circumference will be 2 * 2π = 4π units.

Another thing to know about π is that it is also the *area* of a circle with a radius of 1 unit.



For circles with radiuses other than 1, the area of a circle is π multiplied by the *square* of the radius. For example, if the radius is 4 units long, the area of the circle will be π * 4^2 = 16π square units.

The two main formulas for π relating to circles are:

- The length of the circumference of a circle is 2πr, where “r” is the length of the radius.
- The area of a circle is πr^2, where “r” is the length of the radius.

We can think of these rules in reverse:

- π is equal to the circumference of a circle divided by double the radius.
- π is equal to the area of a circle divided by the square of the radius.

Given that π relates to every circle, it is a universal mathematical constant. Anyone in any civilisation in the universe could stumble across it by studying circles.

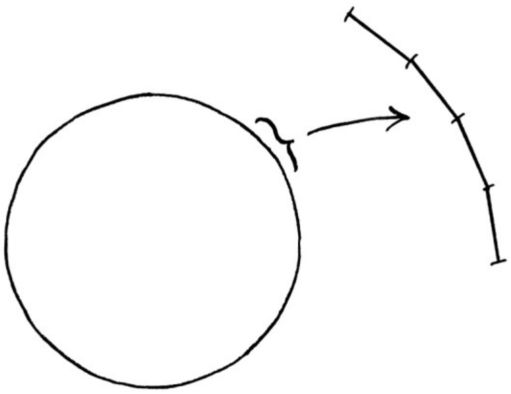
Calculating π

The number π is an irrational number. I explained irrational numbers in Chapter 13 on the addition of waves, in the section on the addition of different frequencies. As a brief reminder, an irrational number is one that cannot be portrayed as a fraction with an integer numerator and an integer denominator. In other words, an irrational number cannot be expressed as one integer divided by a second integer. This means that irrational numbers always have an infinite number of digits after the decimal point. Note that fractions such as one third also have an infinite number of digits after the decimal point, but the difference is that such numbers *can* be expressed as an integer divided by another integer. The term "irrational" refers to how there is no ratio of values that can represent it, with the idea that a fraction can be thought of as a ratio. In this way, "irrational" really means "un - ratio - able".

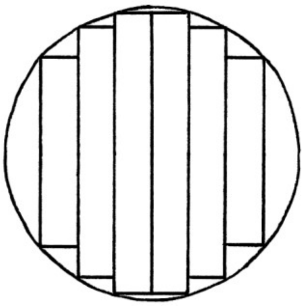
Given how π is an irrational number, its digits continue forever and it cannot be expressed in the form of any fraction. There are also no patterns in its digits that help in calculating it. Not only is π irrational, but any integer multiples of π are irrational too. If there were an integer multiple of π that was not irrational, then π itself could not be irrational.

To find a very rough approximation of π, we can measure the circumference or the area of a circle, and work out π from that. It is difficult to do either of these accurately.

The simplest way to measure the circumference of a circle is to wrap a piece of string around a circle's edge, and measure the length of the string. Another way is treat a circle's edge as if it were made up of equally long tiny straight lines. We then measure one of these lines as accurately as possible with a ruler, and multiply the length by the total number of lines to find the length of the full circumference. The smaller each line is, the more accurate the approximation.



One way to measure the *area* is to divide the circle up into tiny rectangles, calculate the area of each rectangle, and then add them together. The tinier the rectangles, the more accurate the approximation.



The downside to these methods is that they are difficult to do with much precision. For example, when working out the circumference with tiny sections, the smaller the section, the harder it is to measure accurately. When working out the area with tiny rectangles, the tinier the rectangles, the harder it is to measure *them* accurately. Therefore, there is a limit to the accuracy we can achieve with a drawing. If we draw a very large circle to make the tiny rectangles easier to measure, we lose accuracy in the drawing of the circle in the first place.

Sine, Cosine and π

Given that Sine and Cosine relate to the attributes of circles, and that π is an inherent part of circles, one might think that there is an obvious, simple relationship between the three. However, Sine and Cosine taken individually refer to vertical or horizontal values relating to circles – they portray the lengths of the straight edges of triangles at 90 degrees to each other. The number π relates more to the *curve* of a circle. The Sine and Cosine functions can be used to identify any point on a circle's edge using coordinates – they identify the vertical and horizontal distances from the origin of a circle to a particular point on its circumference. Functions involving π can be used to identify the distance between two points along a circle's circumference.

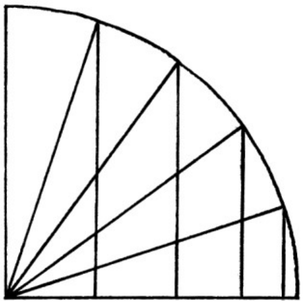
We can use Sine and Cosine to find an approximation of π, but such methods are just variations of the previous examples of calculating the length of tiny pieces of the circumference or calculating the area of tiny rectangles.

Approximating π as an area with Sine and Cosine

We can approximate π using Sine and Cosine with the following method – it is not a particularly good method, but it is obvious how it works. The method shows that the relationship between π and Sine and Cosine is not an obvious one. I am glossing over the details of this method because, in practice, there are much better ways to calculate π.

Imagine that we draw some triangles within a quarter of a unit-radius circle, all with evenly spaced angles between 0 and 90 degrees. As we know, the opposite side of each triangle will be the Sine of that triangle's angle; the adjacent side will be the Cosine of that triangle's angle. We can use Sine and Cosine to calculate the areas of all the triangles. If we then subtract the parts that are counted more than once, we will be close to having the area of that quarter circle. If we multiply the result by 4, we will be close to having an approximation of the area of the whole circle, which will also be equal to π. The more triangles we use, the more insignificant the parts not counted by the triangles will be.

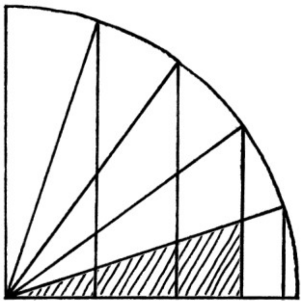
As an example of how this works, we will use 4 triangles, with angles of 72, 54, 36 and 18 degrees:



The first triangle has an area of 0.5 * (sin 72 * cos 72) = 0.1469 square units.
The second triangle has an area of 0.5 * (sin 54 * cos 54) = 0.2378 square units.
The third triangle has an area of 0.5 * (sin 36 * cos 36) = 0.2378 square units.
The fourth triangle has an area of 0.5 * (sin 18 * cos 18) = 0.1469 square units.
The total area is 0.7694 square units.

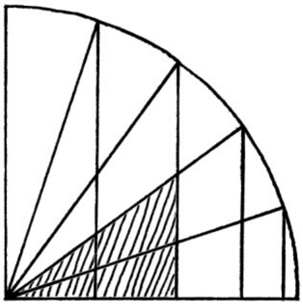
[You might notice that the first two triangles have the same area as the second two triangles. This is because they are the same but mirrored across the 45-degree line. If we have triangles with evenly spaced angles from 0 to 90 degrees, it will always be the case that the first half will be 45-degree mirrored versions of the second half. Knowing this might be useful for a short cut.]

The fourth triangle covers an area that is also included in the areas of the other triangles:



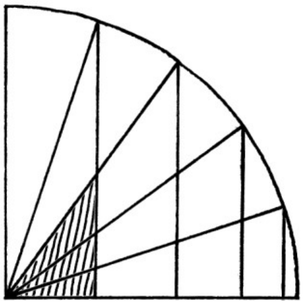
We need to remove this area from the sum or else it will be counted more than once. This area is a triangle with an adjacent side equal to that of the third triangle (cos 36 = 0.8090 units) and an angle equal to the fourth triangle (18 degrees). This triangle is a scaled version of the fourth triangle. As the fourth triangle has an adjacent side of cos 18 = 0.9511 units, and the third triangle has an adjacent side of cos 36 = 0.8090 units, the duplicate triangle is 0.8090 ÷ 0.9511 = 0.8507 of the size of the fourth triangle. Therefore, the duplicate triangle's opposite side will be 0.8507 * sin 18 = 0.2629 units, and its area will be 0.5 * 0.2629 * 0.8090 = 0.1063 square units. We remove this value from the total sum of all four triangles.

The third triangle covers an area that is included in the areas of the first and second triangles. We need to remove this area from the sum:



This duplicate triangle is a scaled version of the third triangle. Its adjacent side is the same as the adjacent side of the second triangle, so it is cos 54 = 0.5878 units. The adjacent side of the third triangle is cos 36 = 0.8090 units. This duplicate triangle is 0.5878 ÷ 0.8090 = 0.7625 of the size of the third triangle. Therefore its opposite side must be 0.7625 * sin 36 = 0.4271 units. The area of the duplicate triangle is 0.5 * 0.5878 * 0.4271 = 0.1255 square units. We subtract this from the total of the area of the triangles.

The second triangle covers an area that is included in the area of the first triangle. We need to remove this area from the sum:

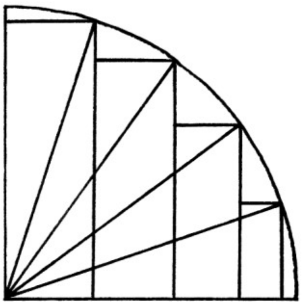


This particular duplicate triangle is a scaled version of the second triangle. Its adjacent side is the same length as the adjacent side of the first triangle. It is cos 72 = 0.3090 units. The adjacent side of the *second* triangle is cos 54 = 0.5878 units. This duplicate triangle is 0.3090 ÷ 0.5878 = 0.5257 of the size of the second triangle. The second triangle's opposite side is sin 54, so the duplicate triangle's opposite side is 0.5257 * sin 54 = 0.4253 units. The area of the duplicate triangle is 0.5 * 0.3090 * 0.4253 = 0.06572 square units. We subtract this from the total area.

The full calculation to find the area of the right-angled triangles is the sum of:
0.5 * sin 72 * cos 72
0.5 * sin 54 * cos 54
0.5 * sin 36 * cos 36
0.5 * sin 18 * cos 18
... minus the sum of:
0.5 * cos 36 * (cos36 ÷ cos 18) * sin 18
0.5 * cos 54 * (cos54 ÷ cos 36) * sin 36
0.5 * cos 72 * (cos72 ÷ cos 54) * sin 54

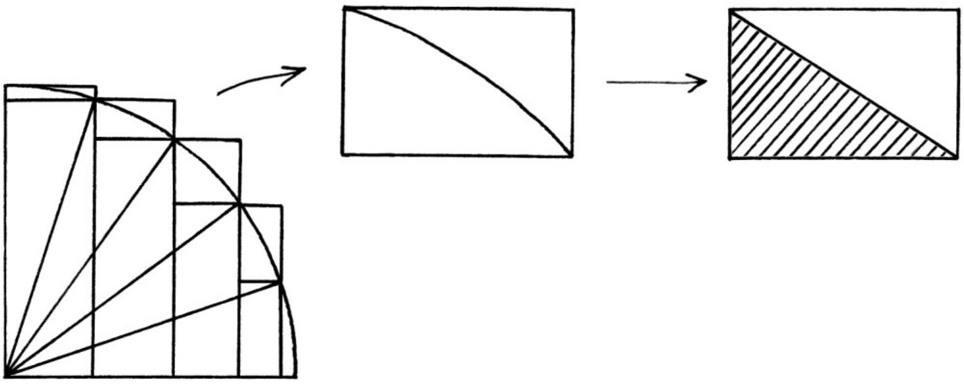
This results in: 0.47186601 square units. We multiply this by 4, and we will have our approximation to π, as so: 0.47186601 * 4 = 1.88746404. This is obviously completely wrong. However, if we use *a lot* more triangles, we will get closer and closer to π. The gaps, which are so obvious in this example, become more and more insignificant as we use more and more triangles. From the final calculation, it is possible to see the pattern of calculations that we would build on to use more triangles. It would not be difficult to write a computer program to use thousands of triangles instead of just 4.

If we make the triangles into rectangles of the same height as the end of the hypotenuses, we can still use Sine and Cosine to calculate π, and, because there will be fewer gaps, we will get closer to having the full area of the quarter circle more quickly.



As the rectangles are based on triangles at evenly spaced angles from 0 to 90 degrees, we can calculate their areas using Sine and Cosine. The more rectangles we use, that is to say the smaller the difference in angle for each triangle, the more accurate the result will be.

If we can include the gaps above the rectangles too, we will achieve more accuracy sooner. We can treat these gaps as being rectangles cut in half diagonally, or in other words, more triangles. Doing this will mean that we will be treating a curve as a straight line. However, if we use small enough steps of angles, this will not be as significant a difference. We can calculate the area of these half-rectangles using Sine and Cosine.



We add up the main rectangles and add the half-rectangles, then multiply the result by four to produce an approximation of π. This method is much quicker than before to achieve the same accuracy. As before, the smaller the angle step for the underlying triangles, the higher the accuracy.

Approximating π as a circumference with Sine and Cosine

One interesting fact about the last method is that the half rectangles along the curve vary in width and height, but the diagonal line through each one remains the same length. That diagonal line is an approximation of that section of the circumference of the circle. As we know that the circumference of the circle is 2π units, we only have to multiply one of the diagonal lengths by the number of underlying triangles we used, then double it, and we will end up with an approximation for π. [We double it as we are dealing with a quarter of a circle, and π is the length of half the circle's circumference.] The smaller the diagonal we use, the more accurate our approximation will be.

The trouble with these methods

The smaller the pieces used to calculate the area or the circumference, the more accurate the result. However, it is also the case that the smaller the pieces, the more accurately one has to calculate Sine and Cosine. That effort would be better spent using a dedicated formula for calculating π.

If the calculations are done on a computer or calculator, there will become a point when the accuracy of π will not improve because the calculations of Sine and Cosine need to be more precise than the computer or calculator can manage.

Programming languages and computer processors generally use radians and not degrees. If the calculations are done on a computer with degrees, they will need to be converted into radians first. To convert into radians, we need to know the value of π. Therefore, if we use degrees, we will need to know the value of π before we start trying to calculate π, which makes the process completely pointless. If, on the other hand, we use radians, we will probably need to know π in order to divide quarter of a circle into evenly spaced angles from 0 to 0.5π.

Calculating π more accurately

As π's digits continue forever, it is impossible to calculate it accurately. However many decimal places we calculate, there will always be more that remain uncalculated. At best, we can only achieve a more accurate approximation.

More accurate approximations of π tend to use never-ending sums that produce more digits with each step, in a similar manner to the Taylor series mentioned in Chapter 2. One very simple sum for calculating π is as follows:

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \frac{1}{11} + \frac{1}{13} \text{ and so on} \ldots$$

The formula takes a very long time to obtain much accuracy. Even after 101 fractions, π is only correct to one decimal place. There are hundreds of formulas for calculating π. Most are much faster than this, but few are as simple.

Tau (τ)

The number π is the result of the circumference of a circle divided by its diameter. However, circles are more usually described by their radius than their diameter. This leads some people to use a number that is the circumference of a circle divided by its *radius*. This number is called “Tau” and is represented by the symbol “τ”. The symbol “τ” is the lower-case Greek letter that is equivalent to the Latin letter “t” for “tango”. Confusingly, it resembles the upper-case Latin letter “T”. The value of “τ” is equal to 2π, and is 6.2831853071795...

The value of 2π is significant when dealing with radians, as we will find out in the next chapter. Therefore, having a symbol that represents 2π makes sense. However, it is not used that much in practice. Generally, formulas using radians, such as “y = sin (2π * 3t)”, are written with 2π, and not with “τ” as “y = sin (τ * 3t)”, despite how it would make things quicker to write. [On the other hand, having “τ” and “t” in the same formula might be confusing, as it is not a particularly good choice of symbol.]

Potential sources of confusion

Note that the symbol “π” is occasionally used to represent entities other than 3.141592653589793... This is relatively rare, so it is safe to assume that if you see π, it is representing 3.141592653589793...

The symbol “τ” is frequently used to represent concepts other than 2π.

Chapter 22: Radians

Dividing circles into portions

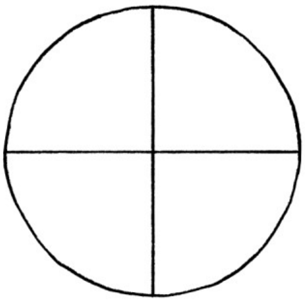
So far, we have measured our angles with degrees. This means that we consider the circle as being divided up into 360 portions. When we say 1 degree, we mean one 360th of a circle.

Whenever we have considered triangles, we have measured the angles in terms of this division of a circle. This leads to the idea that an angle is really a measure that indicates the steepness of a line in terms of a division of a circle.

We do not have to consider a circle as divided up into 360 portions, and we do not have to consider an angle as representing so many 360th divisions of a circle. We could divide a circle into any number of portions, and base our angles on that division.

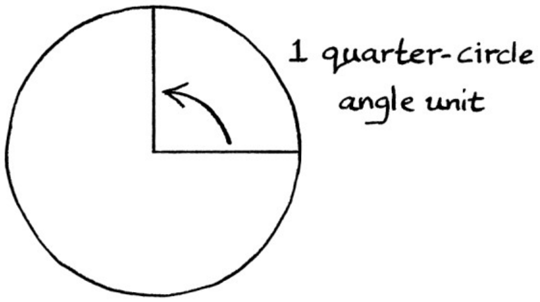
4 piece circle

As an example of a different type of angle, we will divide a circle into 4 pieces as so:

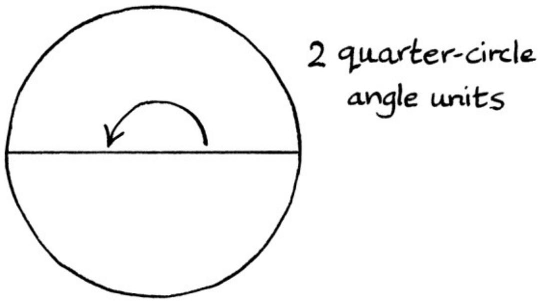


Now we will create an angle system based on this way of dividing a circle. Whereas a degree referred to 1 “angle unit” in a system where a circle was divided into 360 pieces, in this system, 1 “angle unit” will be based on the circle being divided into 4 pieces. We will call this angle unit a “quarter-circle angle unit”.

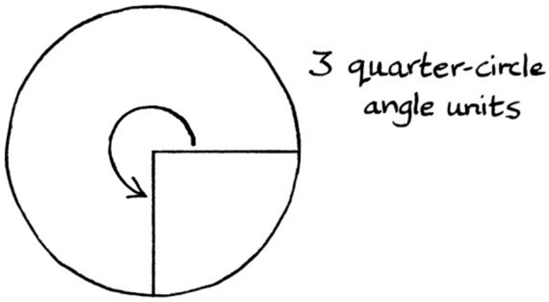
One quarter-circle angle unit moves to quarter of the way around a circle:



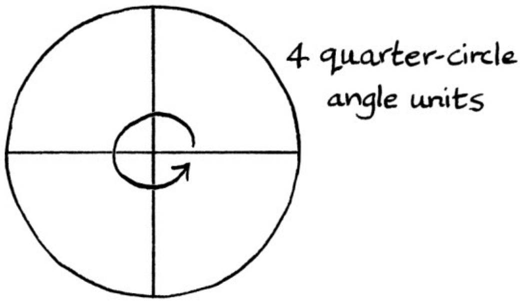
Two quarter-circle angle units move half of the way around a circle:



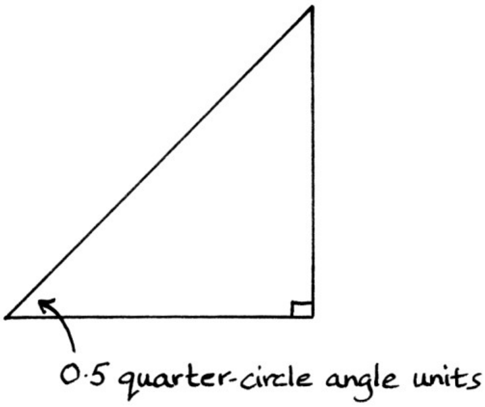
Three quarter-circle angle units move three-quarters of the way around a circle:



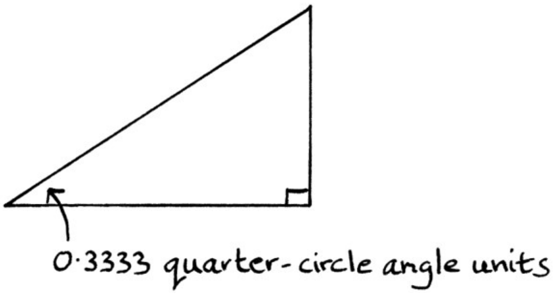
Four quarter-circle angle units move all of the way around a circle:



We can use quarter-circle angle units to measure angles in triangles. For example, this right-angled triangle has an angle of 0.5 quarter-circle angle units:



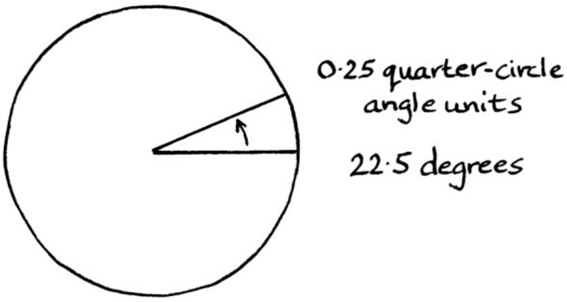
This triangle has an angle of 0.3333 quarter-circle angle units:



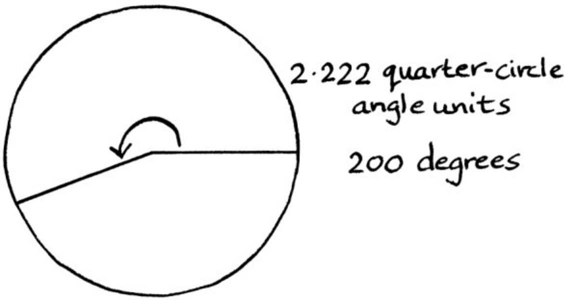
We can convert from quarter-circle angle units to degrees and back quite easily. First, we find out what portion of a circle a given angle represents – this will be the angle divided by the number of such angle units within a circle. Then, we multiply that by the number of divisions a circle has been divided into for the system into which we want to convert.

As an example, we will say we have 0.25 quarter-circle angle units, and we want to find out how many degrees that is. First, we find out the portion of a circle that 0.25 quarter-circle angle units represents. There are 4 quarter-circle angle units in a circle. Therefore, we divide our 0.25 quarter-circle angle units by 4. This is 0.25 ÷ 4 = 0.0625. This is the portion of a circle that our quarter-circle angle represents. We then multiply that by the number of circle divisions for degrees, which is 360. Therefore, we calculate 0.0625 * 360 = 22.5 degrees.

We now know that 0.25 quarter-circle angle units is the same as 22.5 degrees. Both of these ways of measuring angles refer to a sixteenth of the way around a circle.

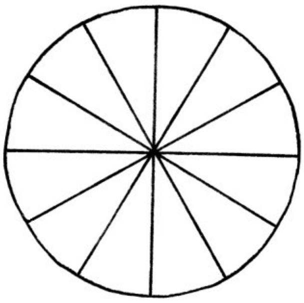


As another example, we will convert the angle of 200 degrees into quarter-circle angle units. First, we calculate the portion of a circle represented by 200 degrees. This is 200 ÷ 360 = 0.5556. Then, we calculate what this portion is if the circle is divided into 4 pieces. Therefore, we calculate 0.5556 * 4 = 2.2222 quarter-circle angle units. We now know that 200 degrees is equal to 2.2222 quarter-circle angle units:

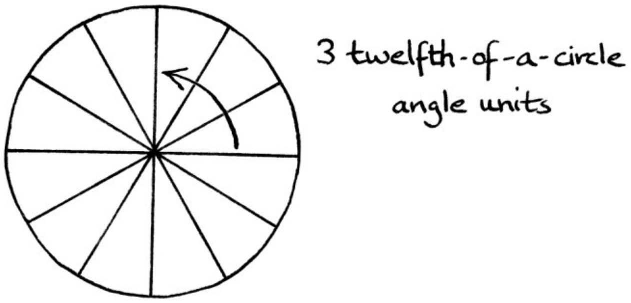


12 piece circle

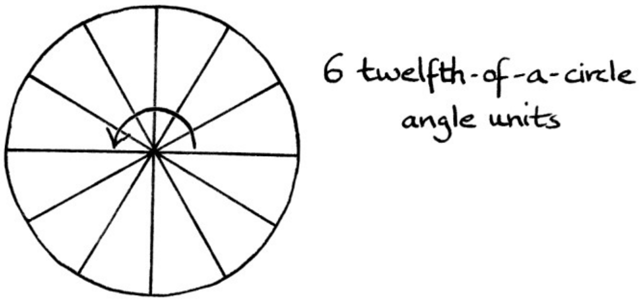
Now, we will make up an angle system based on a circle divided into twelve pieces. In this case, one angle unit is equivalent to one twelfth of a circle.



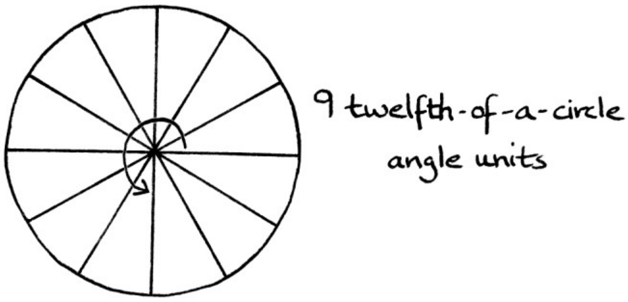
We will call these angle units, "twelfth-of-a-circle" angle units. A quarter of a circle, or 90 degrees, is represented by 3 twelfth-of-a-circle angle units:



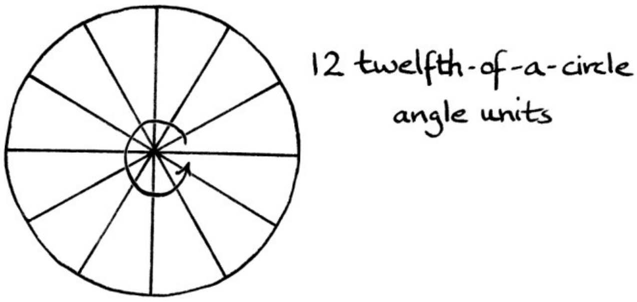
Half a circle is represented by 6 twelfth-of-a-circle angle units:



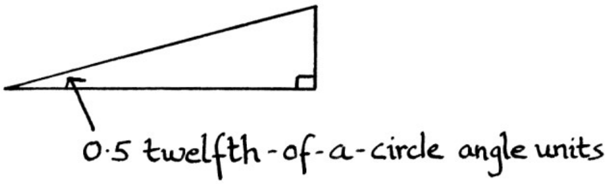
Three-quarters of a circle is represented by 9 twelfth-of-a-circle angle units:



A full circle is represented by 12 twelfth-of-a-circle angle units:



We can measure all angles using the twelfth-of-a-circle angle system. For example, this triangle has an angle of 0.5 twelfth-of-a-circle angle units:

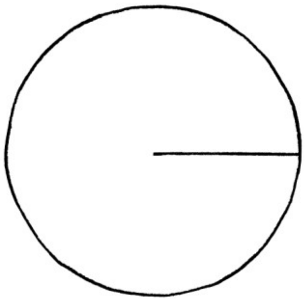


To convert from twelfth-of-a-circle angle units to degrees, we first divide the angle by 12 to find the portion of a circle that that angle represents, and then we multiply by 360. For example, if we had 11.5 twelfth-of-a-circle angle units, it would be equivalent to (11.5 ÷ 12) * 360 = 345 degrees.

To convert from twelfth-of-a-circle angle units to *quarter-circle angle* units, we divide by 12 to find the portion of a circle that that angle represents, and then we multiply by 4. For example, if we had 7 twelfth-of-a-circle angle units, it would represent the same angle as (7 ÷ 12) * 4 = 2.3333 quarter-circle angle units.

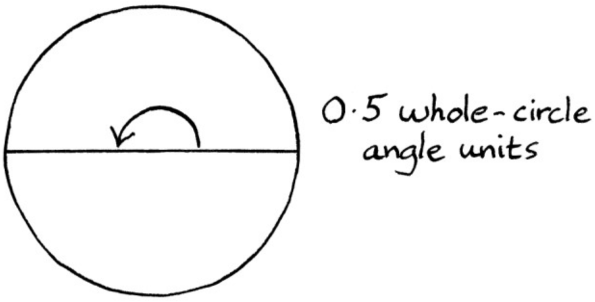
1 piece circle

If we treated a circle as "divided" into just 1 piece, then one angle unit would be equivalent to the whole circle:

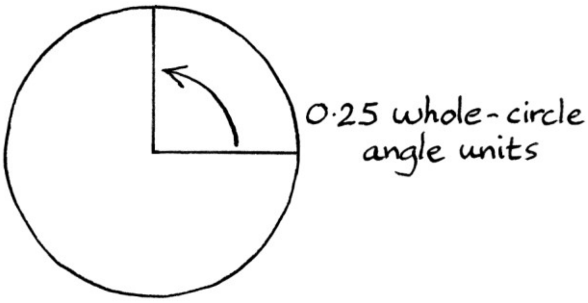


We will call such angles, "whole-circle" angle units. An advantage of this division system is that any value representing a portion of the circle would be that value as an angle unit. In other words, the fraction of a circle is the same as the fraction of the angle unit.

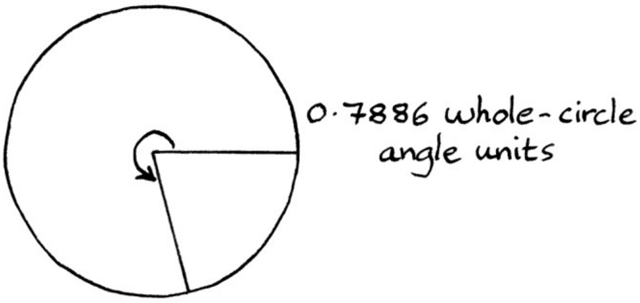
For example, 0.5 of the way around the circle would be the same as 0.5 whole-circle angle units.



0.25 whole-circle angle units would represent 0.25 of the way around a circle (which would be equivalent to 90 degrees).



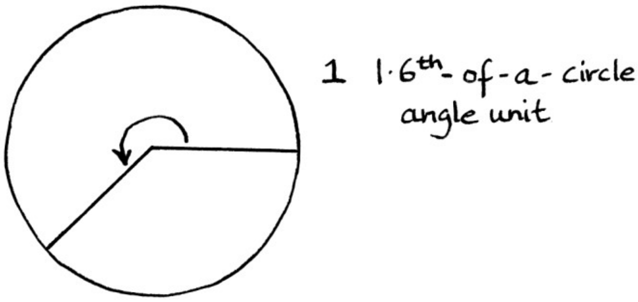
0.7886 angle units would represent 0.7886 of a circle.



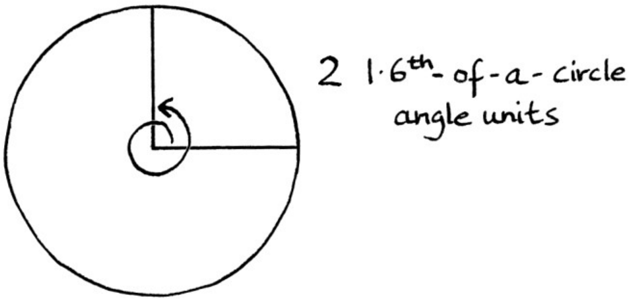
To convert from this angle system to degrees, we just multiply the number of whole-circle angle units by 360. For example, 0.125 whole-circle angle units is 0.125 * 360 = 45 degrees. To convert whole-circle angle units into quarter-circle angle units, we just multiply by 4. For example, 0.1 whole-circle angle units is 0.1 * 4 = 0.4 quarter-circle angle units.

1.6 piece circle

As a more complicated example, we will divide a circle so that there are 1.6 pieces in each circle. This means there are 1.6 angles in one circle. One angle unit will reach 1 ÷ 1.6 = 0.625 of the way around the circle:



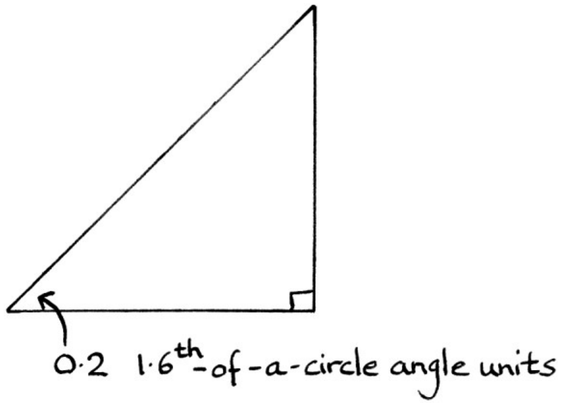
What is significant in this example is that we do not have a whole number of angle units in one circle. One angle unit is 0.625 of a circle. Two angle units would be 1.25 times around the circle:



Three angle units would be 1.875 times around the circle. The units only align with a whole circle at multiples of 5 angle units. Five of these angle units would be exactly 8 times around the circle.

Despite the strangeness of this system, we can still use this system of angles to measure angles in triangles.

The following triangle has an angle of 0.2 in our 1.6th-of-a-circle angle unit system:



To convert from this angle system into degrees, we first divide by the number of angle units in a circle (1.6) to find the portion of a circle represented by that angle. Then, we multiply that by 360. Therefore, 0.2 of these angle units is equal to:
(0.2 ÷ 1.6) * 360 = 45 degrees.

Dividing a circle into a non-integer number of divisions is perfectly valid, but there is little purpose in doing so, except as a step to understanding radians, which we will look at in a moment.

Useful divisions

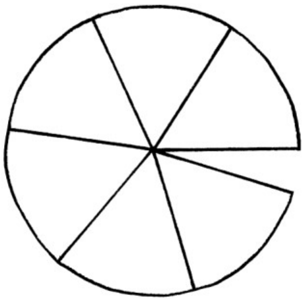
There are countless ways to divide a circle to make an angle system. The reason we might want to divide a circle in a different way is that doing so might make some aspect of maths, learning, or visualisation simpler.

An advantage of having 360 divisions is that we are more likely to have a whole number of degrees as an answer in a simple calculation. The number 360 has many integer divisors: 180, 90, 72, 45, 36, 18, 9, 6, 5, 4, 3, 2, 1. It is easy to do basic maths with degrees, and it is also very easy to recognise patterns. Another advantage is that 360 is a large number, but not so large that we cannot distinguish between the different angles on a circle of a size that we might draw on a piece of paper.

A huge advantage of having 1 division is that the fraction of an angle is the same as the fraction of the circle. In fact, if we are converting from one system of angles to another, we will always have a step where the angle is in this form. However, this method of dividing up a circle has never become popular, which is probably on account of everyone involved in early school education using degrees, and everyone in later education using radians, which I will discuss next.

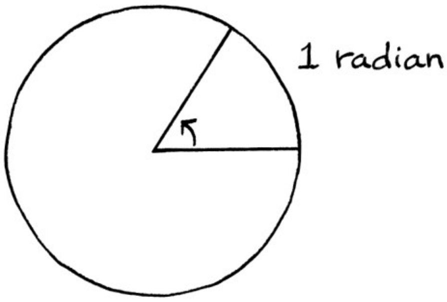
Radians

Radians are another way of dividing a circle, but instead of dividing a circle into 360 pieces, 4 pieces, 12 pieces, 1 piece or 1.6 pieces, radians divide a circle into 2π pieces. In other words, the circle is divided into 6.28318530... portions. This means that 1 radian represents 1 ÷ 6.28318530... of a circle, or 0.15915494... of a circle.

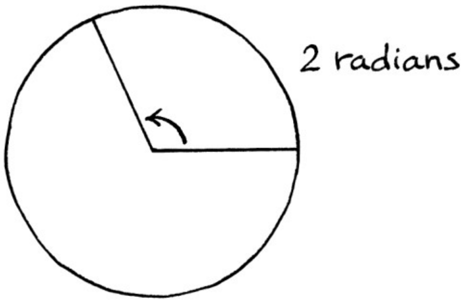


As with our 1.6 division system, dividing a circle into 2π pieces means that there is not an integer number of divisions in a circle. Unlike our 1.6 division system, because 2π is irrational, there will never be a time when a multiple of the units will correspond with a full number of circles. [If a number is irrational then all integer multiples of that number will also be irrational. Given that π is irrational, it must be the case that 2π, 3π, 4π and so on, are all irrational too.]

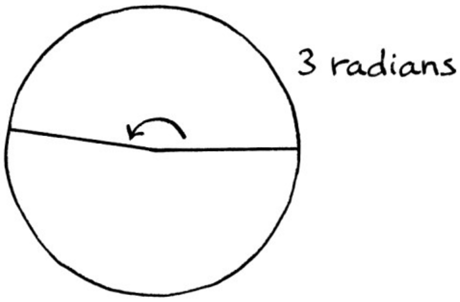
1 radian looks like this:



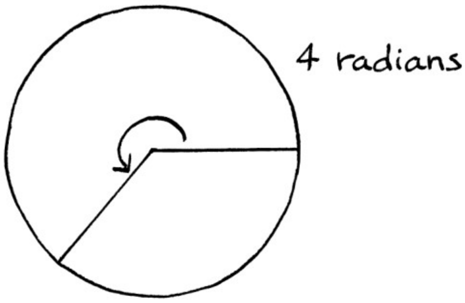
2 radians:



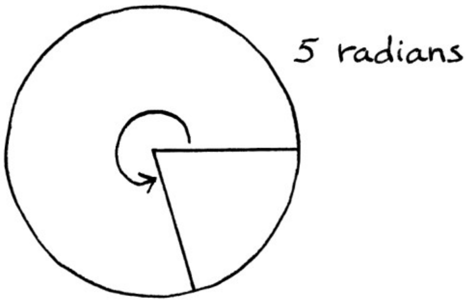
3 radians:



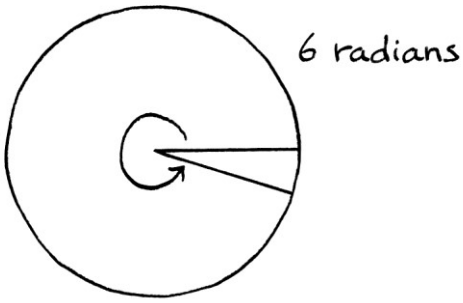
4 radians:



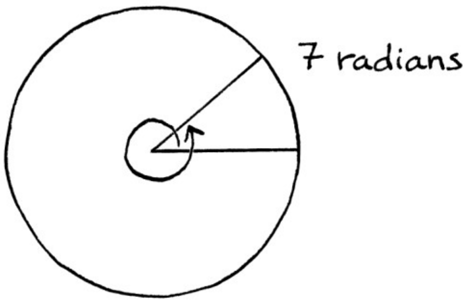
5 radians:



6 radians:

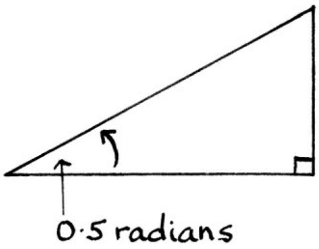


7 radians:



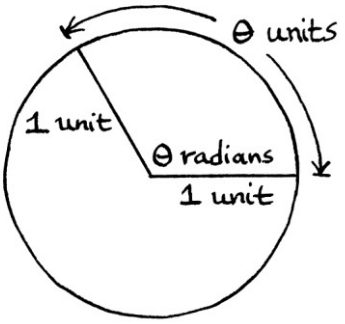
Notice how 7 radians does not line up with a whole circle. Also, notice how 7 radians is not the same angle as 1 radian. [If we were using degrees, then, for example, 1 degree would be the same as 361 degrees.] The only integer number of radians that is equal to 1 radian is 1 radian. The integer multiples of different radian angles can never coincide with each other because they are based on an irrational number (2π).

As an example of radians in use as angles, this right-angled triangle has an angle of 0.5 radians:

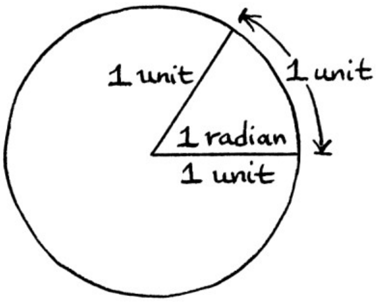


Radians are used almost exclusively as the system of angles in the sort of maths that is used with waves. While it might seem odd to use a number that cannot even be written as a fraction as a basis for dividing a circle, it is not chosen arbitrarily. Radians have an advantage over degrees and other angle systems in that they relate more closely to the properties of circles.

It is easiest to understand one reason for using radians if we think about the circumference of a circle. The circumference of a unit-radius circle is 2π units long. Therefore, if we use an angle system based on dividing the circle into 2π portions, every angle unit on a unit-radius circle will have exactly the same value as the length of the circumference of the circle that that angle unit covers:



This also means that, on a unit-radius circle, the length of the arc (the portion of the circumference) contained in 1 radian is equal to 1 unit:



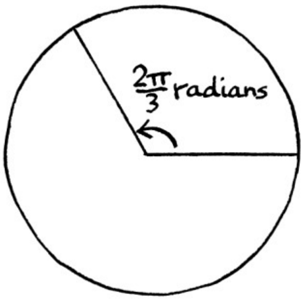
If we were still using degrees, calculating the length of a portion of the circumference would be more effort. For example, to find out the length of the part of the circumference covered by 20 degrees on a unit radius circle, we would first need to find out what portion of the circle 20 degrees is. It is 20 ÷ 360 = 0.05555. Then we would need to multiply that by the length of the whole of the circumference of the circle (2π), so: 0.0555 * 2π = 0.3491 units. A similar calculation done with radians would not have required any maths whatsoever – the angle in radians is the length of that part of the circumference. In other words, 0.3491 radians covers 0.3491 units' length of the circumference.

Dividing a circle into 2π divisions means that there is a direct relationship between the angle units and the circumference. We can think of radians as being an angle system based on the circumference.

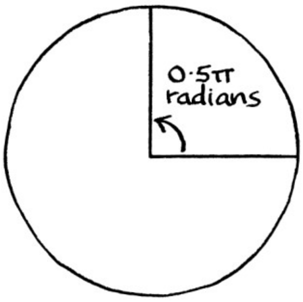
There are mathematical principles relating to circles that only work when the circle is divided into radians – for example, using formulas based on "e^{ix}", which I will explain in Chapter 27. The method of using a Taylor series to calculate Sine or Cosine works, in the most straightforward way, when using radians. Calculus with waves works most simply when using radians. Radians can be thought of as the most natural way to divide a circle.

Radians do complicate matters, though, because π cannot be expressed as an easily written number. An angle such as 45 degrees, which is an easy number to read and remember, becomes 0.785398133974... radians, which is not an easy number to read. However, if we consider that in the real world, when using maths on angles measured in degrees, we might still end up with results such as 23.34234234 degrees, then the main downside to radians is that they are *always* unpleasant values, while degrees only *usually* are. Because radians are based around π, we can sometimes avoid having to write out countless numbers after the decimal point for certain angles by describing the angles as fractions or multiples of π. For example:

The angle of a third of a circle is 2π ÷ 3 radians = 2π/3 radians.



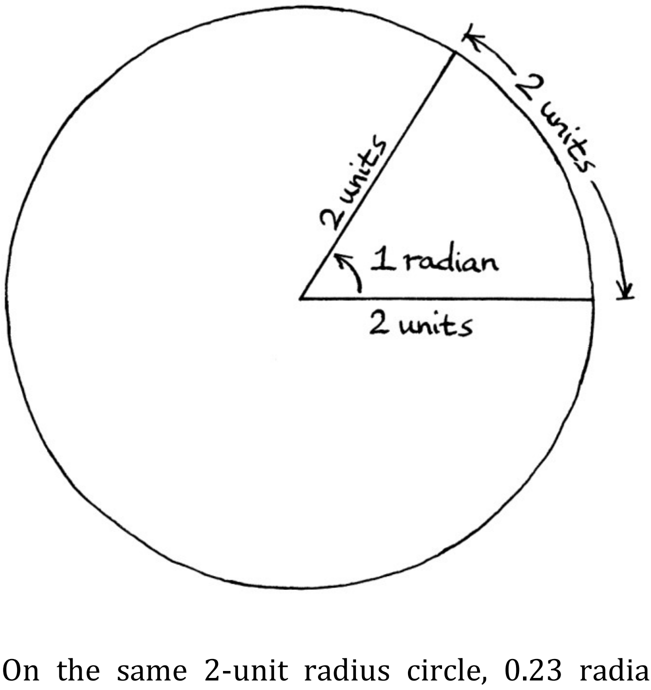
The angle of quarter of a circle is 2π ÷ 4 radians, or π/2 radians, or 0.5π radians.



Writing radians as fractions involving π can be useful at times, but it can also make things harder to read and harder to type on a computer. It also adds an extra level of detachment between the angle and what it actually means on a circle. There are some people who will give radians as a fraction, even when the fraction would be easier to read as a number with a decimal point. For example, sometimes you might see an angle given as $\frac{3\pi}{2}$ radians, when it would be clearer if it were written as 1.5π radians. Unnecessary fractions used in maths in this way can seem very old-fashioned and unscientific. On the other hand, some values can be given exactly with fractions of π, while they could only be approximated as a number with a decimal point. For example, $\frac{2\pi}{3}$ gives the exact angle, while 0.66666667π is only an approximation.

Scaling

For circles with radiuses of lengths other than 1 unit, the angle as measured in radians will mark out a length of the circumference that is scaled in proportion to the radius. For example, if a circle's radius is 2 units long, then one radian will cover an arc of the circumference that is 2 units long.



On the same 2-unit radius circle, 0.23 radians will mark out a length of the circumference that is 2 * 0.23 = 0.46 units long.

If a circle has a radius of 0.5 units, 1 radian will mark out a length of the circumference that is 0.5 units long. On the same 0.5 unit radius circle, 1.24 radians will mark out a length of the circumference that is 1.24 * 0.5 = 0.62 units long.

Conversion

Converting from radians to degrees and back again works in the same way as if we were dealing with any other angle system. To convert from degrees to radians, we divide the angle in degrees by 360 to find the portion of a circle that that angle represents, and then we multiply that by 2π. For example, 27 degrees is (27 ÷ 360) = 0.075 of a circle, which is 0.075 * 2π = 0.4712 radians. By luck, in this case, the result can also be given as a fraction of π to make things look tidier: $\frac{3\pi}{20}$

If we want to convert from radians to degrees, we divide the angle by 2π to find the portion of a circle that it represents, and then we multiply it by 360 degrees. As an example, we will convert 0.5 radians to degrees. First, we find the fraction of a whole circle: 0.5 ÷ 2π = 0.07976. Then, we multiply that by 360, so: 0.07976 * 360 = 28.6479 degrees.

One radian is equivalent to about 57.295779513082 degrees.

In my opinion, for more advanced maths, there is seldom much need to convert between radians and degrees. When using radians, it is generally simpler to stay in the world of radians and not convert out. It is also easier to become used to radians if you do not convert them to degrees all the time. However, it pays to know 5 different points on the circle in terms of both radians (as multiples of π) and degrees. All of these you can figure out yourself by imagining a circle (2π radians or 360 degrees) and dividing it into pieces.

- A quarter of a circle = 90 degrees = $\frac{2\pi}{4}$ radians = $\frac{\pi}{2}$ radians = 0.5π radians.
- Half a circle = 180 degrees = $\frac{2\pi}{2}$ radians = π radians.
- Three quarters of a circle = 270 degrees = $\frac{6\pi}{4}$ radians = $\frac{3\pi}{2}$ radians, = 1.5π radians.
- A full circle = 360 degrees = 2π radians.
- An eighth of a circle = 45 degrees = $\frac{2\pi}{8}$ radians, or $\frac{\pi}{4}$ radians = 0.25π radians.

To summarise the above, it is helpful if you can remember 0.25π, 0.5π, π, 1.5π and 2π radians, which are 45, 90, 180, 270 and 360 degrees respectively.

Sine and Cosine and radians

When the Sine and Cosine functions operate on a number, they treat that number as an angle. This is another way of saying that they treat that number as a portion of a circle. The Sine of, for example, "half a circle" will always be the same whether that half circle is measured in radians, degrees, quarter-circle angles, twelfth-of-a-circle angles or, in fact, any angle unit. The Sines of "quarter of a circle", "eleven twelfths of a circle" or "0.58382 of a circle" will also be the same whether that portion of a circle is measured in radians, degrees, or other angle system units. By using an angle system other than degrees, we are just changing the measurement we use to describe "half a circle", "quarter of a circle", "eleven twelfths of a circle" or "0.58382 of a circle".

It is important to note that the Sine or Cosine of, say, 10 will give a different answer depending on whether that 10 represents degrees (in which case it is a small part of a circle) or radians (in which case it is a full circle and a bit more). Using radians means that Sine and Cosine give a different result to before for any particular *number*, but they still give the same answer for the *portion* of the circle that that number represents. [Another way of understanding this is by realising that the dimensions of a particular triangle will be the same whether the angle is measured in degrees, radians, or any other angle system.]

We can think of this idea as if there were a "radians version" of Sine and Cosine, a "degrees version" of Sine and Cosine, a "quarter-circle angle unit" version of Sine and Cosine, and so on. If we use the *radians* version of Sine on a portion of a circle described using *degrees*, we will end up with the wrong result because the fraction of a circle described with degrees will be different from that described with that same number of radians.

If you understand that Sine and Cosine actually work on fractions of circles, then radians and all angle systems can be much simpler to understand.

Here are some examples of the results of Sine in degrees and radians:

- A full circle: sin (360 degrees) = sin (2π radians) = 0 units.
- Half a circle: sin (180 degrees) = sin (π radians) = 0 units.
- Quarter of a circle: sin (90 degrees) = sin (0.5π radians) = 1 unit.
- An eighth of a circle: sin (45 degrees) = sin (0.25π radians) = 0.7071 units.
- Eleven twelfths of a circle: sin (330 degrees) = sin ((11/12) * 2π radians) = −0.5 units.
- 0.58382 of a circle: sin (210.1752 degrees) = sin (3.6682 radians) = −0.5026 units.

Cosine works in exactly the same way as Sine, in that it works on the fraction of a circle.

When we dealt in degrees, the result of the Cosine of a number was always the same as the Sine of that number plus 90 degrees. In other words:

“cos (θ degrees) = sin (θ degrees + 90 degrees)”.

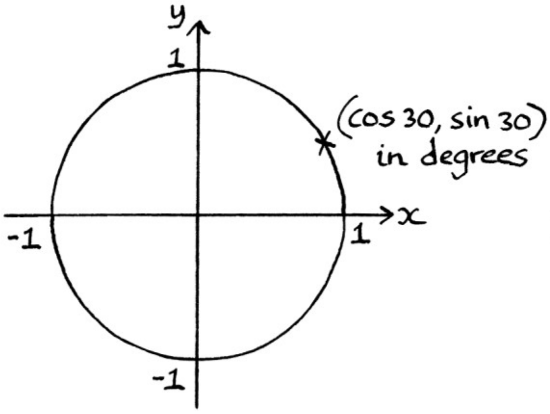
This is really the same as saying that the Cosine of a number is the same as the Sine of that number plus quarter of a circle. More specifically, this is saying that the Cosine of *a fraction of a circle* is the same as the Sine of that same fraction of a circle plus quarter of a circle. If we are working in radians, this is still true, but a quarter of a circle is now measured as 0.5π radians. Therefore:

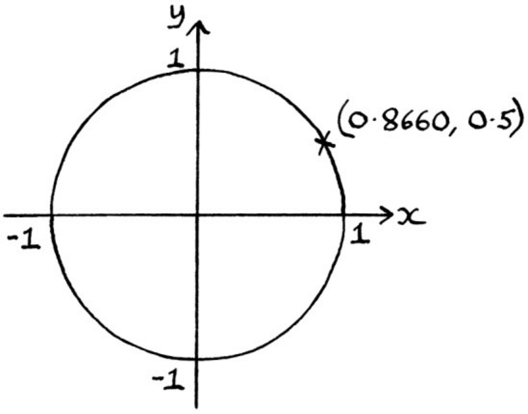
cos (θ radians) = sin (θ radians + 0.5π radians).

To summarise everything that I have just said: The Sine and Cosine functions give results based on the portion of a circle that is represented by an angle. If that portion is being measured in degrees, then the Sine and Cosine functions must be working in degrees to treat that angle as the correct portion of a circle. If that portion is being measured in radians, then the Sine and Cosine functions must be working in radians to treat that angle as the correct portion of a circle. If that portion is being measured in a system where the circle is divided into 4 parts, then the Sine and Cosine functions must be working within that system too. Given that there are countless ways to divide a circle, there are also countless ways in which the Sine and Cosine functions can operate to work with these systems. The Sine function on any particular portion of a circle will give the same result whether that portion is measured in degrees, radians or any other system of angles. The same is true for the Cosine function.

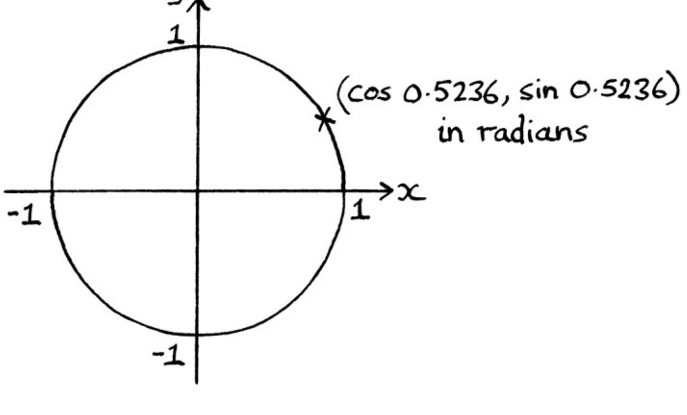
Side note: coordinates

When we were working in degrees, it was the case that any point on a unit-radius circle that was centred on the origin of the axes could have its coordinates given as the Cosine of the angle of that point and the Sine of the angle of that point. For example, the point on the circumference of a unit-radius circle at 30 degrees can be given by the coordinates: (cos 30, sin 30), where 30 is an angle in degrees. When written out as normal coordinates, this becomes (0.8660, 0.5).





If we are working in radians, then it is still the case that any point on a unit-radius circle, centred on the origin of the axes, can have its coordinates given in terms of the Cosine of the angle and the Sine of the angle. For example, the point on the circumference of a unit-radius circle at 0.5236 radians has the coordinates: (cos 0.5236, sin 0.5236), where 0.5236 is an angle in radians, *and Sine and Cosine are working in radians*. When written out in full, this point is also at (0.8660, 0.5).



If we were working in quarter-circle angle units, then the same point would be at an angle of 0.3333 quarter-circle angle units, and its coordinates would be given by (cos 0.3333, sin 0.3333), where 0.3333 is an angle in quarter-circle angle units, and *Cosine and Sine are working in quarter-circle angle units.* When written out in full, this point is also at the coordinates of (0.8660, 0.5).

[Calculators generally let Sine and Cosine work in degrees and radians, and it is rare to find one that allows the use of other angle units such as quarter-circle angle units. However, we know how to convert from any angle unit to radians or degrees and back to see how other angle-units work, so this is not too much of a problem.]

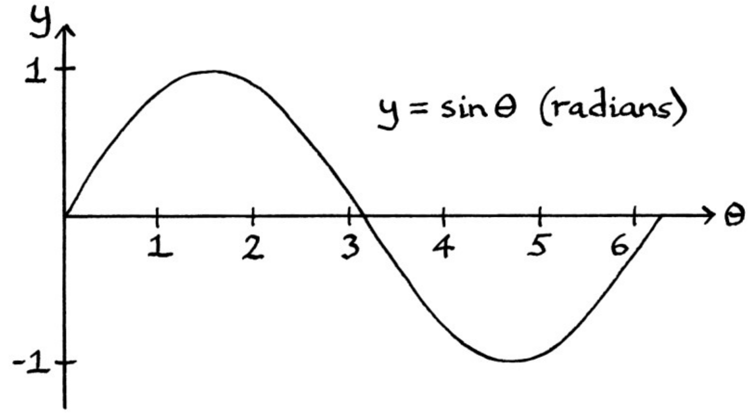
As another example, the point on a unit-radius circle with the coordinates (0, 1) can be given in terms of:

- (cos 90, sin 90), where 90 is an angle in degrees, and Cosine and Sine are working in degrees.
- (cos 0.5π, sin 0.5π), where 0.5π is an angle in radians, and Cosine and Sine are working in radians. This could also be given as (cos 1.5708, sin 1.5708), because 0.5π = 1.5708 to 4 decimal places), but often it is simpler to give the angle as a multiple of π.
- (cos 1, sin 1), where 1 is an angle in quarter-circle angle units, and Cosine and Sine are working within that system of angles.
- (cos 3, sin 3), where 3 is an angle in twelfth-of-a-circle angle units, and Cosine and Sine are working within that system of angles.
- (cos 0.25, sin 0.25), where 0.25 is an angle in whole-circle angle units, and Cosine and Sine are working in whole-circle angle units.

Wave graphs

Angle-based Sine wave graphs

When using radians, the “y = sin θ” wave graph looks like this:

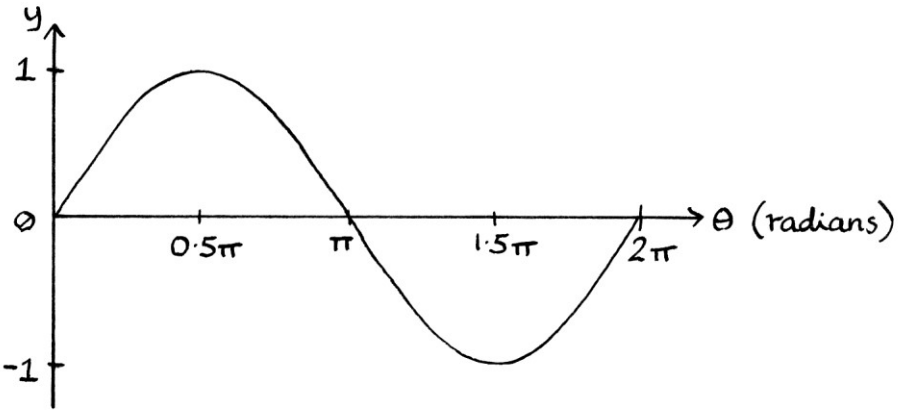


“θ” is the angle in radians and Sine is working in radians. Instead of the θ-axis showing angles in degrees from 0 up to 360 degrees, it now shows angles in radians from 0 up to 2π radians (where 2π is 6.28318531). The θ-axis still shows all the angles for one circle. All that has changed is the unit that we are using to measure those angles.

The graph is identical in shape to the graph of “y = sin θ” when “θ” is an angle measured in degrees and Sine is working in degrees. This is because the graph still represents the y-axis values of points around a circle’s edge at equally spaced angles from its centre. Those y-axis values will be the same, no matter how the angles of the points are described. The graph will be the same as long as the Sine function is working in the same system as that used to measure the angles. [Having said that, the graph’s length will vary depending on how long the units for the θ-axis are in the drawing of the graph.]

The maximum and minimum points on the y-axis are still +1 and −1, because they relate to the heights of points on the circumference of a circle, and the circle has not changed – only the way that we measure the angles.

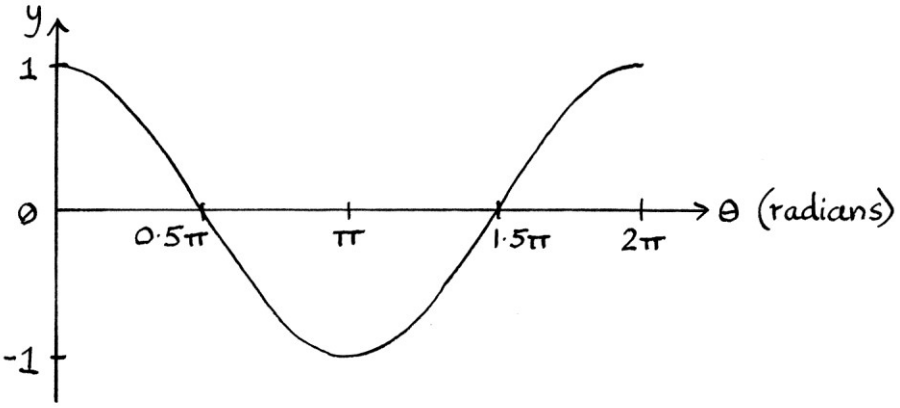
The places where the rises and falls of the wave occur are clearer if the θ-axis is numbered according to multiples of π as so:



Although this graph makes it harder to know the exact values at each point, it is a more useful way of portraying a wave in radians. It lets us know the equivalent places for 0, 90, 180, 270 and 360 degrees, which are generally the most significant places on the graph. In this book, I will usually label angle-based waves in radians in this way.

Angle-based Cosine wave graphs

When using radians, the “y =cos θ” wave graph looks like this:



This radian-based Cosine wave graph is identical in shape to that of one using degrees, with the exception that the θ-axis is labelled from 0 radians up to 2π radians. The Cosine wave graph is showing the x-axis values of points on the edge of a unit-radius circle at equally spaced angles. It does not matter which system of angles is used to measure those angles as long as Cosine is working within the same system.

Time-based Sine waves

When we had a time-based Sine wave using degrees, we needed to multiply the time by the number of degrees in a circle, so that we could use the "degrees" Sine function directly on the time itself. As there are 360 degrees in a circle, this meant that we multiplied the time by 360. Our basic time-based Sine wave in degrees had the formula:
"y = sin (360 * t)"

If we want to use radians, then we need to multiply the time by the number of *radians* in a circle, so that we can use the "radians" Sine function on the time itself. Therefore, as there are 2π radians in a circle, we multiply the time by 2π. Our time-based Sine wave formula in radians looks like this:
"y = sin (2π * t)"

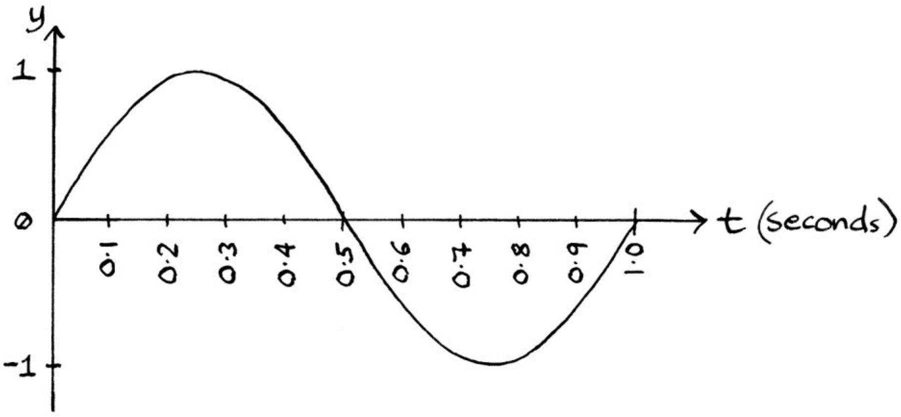
The most obvious difference between the time-based Sine wave formula for degrees and the time-based Sine wave formula for radians is that instead of a multiplication by 360, there is a multiplication by 2π. There is also the *huge* difference that Sine is working in radians and not in degrees. It is important to note that there is no written indication as to the system in which Sine is working. The system is implied by the context, and for most of the time, it will be unambiguous. If there is a formula such as "y = sin 360t", then we can presume that Sine is working in degrees. If there is a formula such as "y = sin 2πt", then we can presume that Sine is working in radians. For more advanced maths, radians are used almost exclusively, so there will be little chance of confusion. Ideally, there would be some indication as to the system of angles in which a particular Sine function were working, however, as far as I know, there is not such a thing. It would be possible to use suffixes or subscripts such as "sin-d" or "$\sin_d$" to indicate Sine was working in degrees, and "Sin-r" or "$\sin_r$" to indicate that Sine was working in radians. [Another idea would be for the number of divisions in the circle to act as a suffix such as sin360, sin2π, sin4 and so on.] As it is, for most of the time, such a thing is not needed. Such an indicator would be fairly useful in this chapter explaining radians. It would remove ambiguity for formulas such as "y = sin 10", which, without any further explanation, could mean the Sine function working in radians on the angle of 10 radians, or it could mean the Sine function working in degrees on the angle of 10 degrees. It could also mean the Sine function working in quarter-circle angle units or any other system of angle units. These would all produce completely different results. In this book, I will not use suffixes or subscripts, but will instead mention the system of angles that Sine and Cosine are working in for each example.

A full formula for a time-based Sine wave using radians is as so:
“$y = h_s + A \sin((2\pi * ft) + \phi)$”
... where:

- “h_s” is the mean level in units.
- “A” is the amplitude in units.
- Sine is working in radians.
- “f” is the frequency in cycles per second or hertz (being two names for the same thing).
- “t” is the time in seconds.
- “ϕ” is the phase in *radians*.

It is important to realise that if Sine is working in radians, then the phase must be an angle in radians. Therefore, the phase in this formula must be in radians.

A time-based Sine wave graph using radians is identical to a time-based Sine wave graph using degrees. Whether we are using the formula for degrees or the formula for radians, every single moment in time will produce the exact same y-axis value.



This is because the y-axis values on the wave graph refer to the y-axis values on the circle chart of an object rotating around a circle at particular moments in time. That object will be at the same place on the circle’s edge at a particular time, whether we measure the angles in degrees or radians. When we are dealing with time, it does not matter whether we deal in degrees, radians or any other system of angles, as long as the Sine function is working in that same system of angles, and the appropriate time correction is applied.

The time-based Sine wave graph is a good example of how the change from using degrees to using radians is fairly straightforward. For time-based graphs, there is no difference in the actual results. For time-based formulas, we just replace 360 with 2π and put the phase in radians.

[Remember to make sure that your calculator is set up to use radians instead of degrees.]

Time-based Cosine waves

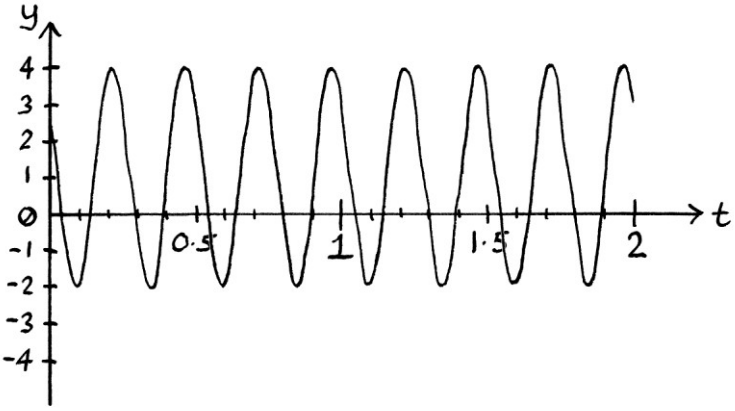
A time-based Cosine wave using radians is as straightforward as a time-based Sine wave using radians. When we were dealing with degrees, we multiplied the time by 360. Now we are using radians, we multiply the time by 2π.

The formula "y = cos (360 * t)" in degrees becomes "y = cos (2π * t)" in radians.

A full formula for a time-based Cosine wave using radians is as so:
"$y = h_c + A \cos ((2\pi * ft) + \phi)$"
... where:

- "h_c" is the mean level in units.
- "A" is the amplitude in units.
- Cosine is working in radians.
- "f" is the frequency in cycles per second or hertz.
- "t" is the time in seconds.
- "ϕ" is the phase in radians.

If we had the formula "y = 1 + 3 cos ((2π * 4t) + 0.25π)", where Cosine is working in radians, its graph would look like this:



Note that the phase is 0.25π radians. A full circle is 2π radians, so 0.25π radians is an eighth of a circle, which is same as 45 degrees. We could just as easily have given the formula as 0.78539816 radians (to 8 decimal places), but it is much easier to see that the angle is an eighth of a circle if we write it as 0.25π radians. This is a good example of how the system of radians is far less suitable for recognising patterns than degrees. [If we were using an angle system based on "dividing" a circle into 1 piece, then it would be obvious from any value how much of a circle it represented. Any value would be equal to the fraction of a circle it represented. With degrees, we can recognise fractions of circles reasonably easily, but with radians, it is much, much harder.]

The above time-based Cosine wave graph is identical to the graph for this formula:
"y = 1 + 3 cos ((360 * 4t) + 45)"
... where Cosine is working in degrees, and the phase is 45 degrees.

Angular frequency in radians

We first looked at the concept of angular frequency in Chapters 4 and 6. The two basic intertwined ideas behind angular frequency, when dealing with *degrees*, are:

- For the Sine and Cosine functions to be performed directly on the time as a value in seconds, and for them to result in a basic frequency of one cycle per second, it is necessary to scale that value first by multiplying it by 360.
- By multiplying the time in seconds by 360, we are really working with degrees per second or angles per second. In this way, the frequency can be thought of as an angle frequency or "angular frequency".

When we are working with radians, the basis behind these ideas is still true. For the Sine and Cosine functions to be performed directly on the time in seconds, and for them to result in a basic frequency of one cycle per second, it is necessary to scale that time by multiplying it by 2π if we are working in radians. Once we have done this multiplication, we are really dealing in angles per second, or more specifically, radians per second.

The concept of angular frequency is *much* easier to understand when using degrees. Degrees are a conveniently small unit of angle. The idea of a rotating object completing so many degrees in a set amount of time is fairly simple. If you are first introduced to angular frequency when using radians, it is much harder to visualise. For one thing, a radian is a sizeable chunk of a circle – there are only about 6 of them in a circle. Another thing is that a radian is an untidy quantity – it is represented by an irrational number. Without understanding the background to angular frequency in degrees, you might wonder what the purpose of it is. [One might say that there is no "purpose" – it is just a consequence of having the default wave repeat one cycle once every second.]

If you understand the concept of angular frequency in degrees, then the step to radians is straightforward.

Angular frequency in radians is still represented by the lower-case Greek letter Omega: "ω". Before now, "ω" was an abbreviation for "360 * f", but now it is an abbreviation for "2π * f". Nearly everyone who uses the symbol "ω" is using it with radians in mind.

The formulas for angular frequency in radians look the same as the formulas for angular frequency in degrees:
$y = h_s + A \sin (\omega t + \phi)$
$y = h_c + A \cos (\omega t + \phi)$
They *look* the same, but there are the differences that "ω" is now short for "2π * f" instead of "360 * f", Sine and Cosine are working in radians, and ϕ represents an angle in radians.

It is a matter of debate as to whether there is any great advantage in writing the formula with "ω" instead of "2πf". It is no quicker to type unless you have a Greek keyboard with "ω" as one of the keys. However, there are some academic subjects where "ω" is used exclusively in formulas, so it can pay to become used to it. It is also a sign of understanding the basics of waves if you can use either formula without caring.

Positive and negative frequencies

When we were working in degrees, we could change a positive-frequency formula into a negative-frequency formula, and vice versa, while keeping the meaning of the formula the same. Doing this involved altering the phase while changing the sign of the frequency. This was discussed in Chapter 11. When calculating the new phase for a Sine wave *in degrees*, we have three options:

- We can imagine the circle from which the wave is derived, and mirror it left or right. The new phase will be where the phase point ends up.

- We can see how many degrees above or below 90 degrees the phase is, and then change the phase to that number of degrees below or above 90 degrees. For example, if the phase is 91 degrees, we set the phase to 89 degrees and change the direction of the frequency.

- We can see how many degrees the phase is above or below 270 degrees, and set it that same number of degrees below or above 270 degrees. For example, if the phase is 271 degrees, we set it to 269 degrees and change the direction of the frequency.

When working in radians, the circle mirroring method still works. For the other two methods, the angles we count from will not be 90 degrees and 270 degrees, but instead, their radian equivalents: 0.5π radians and 1.5π radians. Therefore, if we have a phase of "0.5π + 1" radians and a negative frequency, we change the formula to have a phase of "0.5π − 1" radians and a positive frequency. It is usually hard to know how far an angle is from 0.5π or 1.5π radians without using a calculator. However, if the phase is exactly 0.5π or 1.5π radians, we know that it will not need changing, and we can just change the direction of the frequency.

When we working in degrees, to calculate the new phase for a *Cosine* wave, we have four options:

- We can imagine the circle from which the wave is derived and flip it upwards or downwards. The new phase will be where the phase point ends up.

- We can see how many degrees above or below 0 degrees the phase is, and then change the phase to that same number of degrees below or above 0 degrees. For example, if the phase is 1 degree and the frequency is negative, we set the phase to −1 degrees, which is +359 degrees, and make the frequency positive.

- We can just negate the phase. This is essentially the same as seeing how far the phase is from 0 degrees. If the phase is 200 degrees, we set the phase to −200 degrees and change the direction of the frequency.

- We can see how many degrees the phase is above or below 180 degrees, and set it that same number of degrees below or above 180 degrees. For example, if the phase is 181 degrees, we set it to 179 degrees and change the direction of the frequency.

When working in radians, the circle flipping method will still work. We can also just negate the phase. For the other two methods, the angles we count from will not be 0 degrees and 180 degrees, but instead, their radian equivalents: 0 radians and π radians. Therefore, if we have a Cosine wave formula with a phase of "$\pi + 0.5$" radians and a negative frequency, we can change the formula to have an angle of "$\pi - 0.5$" radians and a positive frequency, and it will still have the same curve.

Multiplication of waves

In Chapter 16, we created rules that gave the sum of waves that would be equivalent to a multiplication of two waves, while working in *degrees*. Here we will look at their radian equivalences.

Sine waves

The rule for finding the sum of four Sine waves and the mean level that are equivalent to the multiplication of two Sine waves, when using degrees, is as so:

If we multiply:
"$y = h_1 + a_1 \sin ((360 * f_1 * t) + \phi_1)$"
... by:
"$y = h_2 + a_2 \sin ((360 * f_2 * t) + \phi_2)$"
... where f_1 is faster than, the negative of, or equal to f_2, then the equivalent sum of waves and a mean level will be:
"$y = 0.5 * (a_1 * a_2) * \sin ((360 * (f_1 - f_2)) + (\phi_1 - \phi_2 + 90)$"
"$y = 0.5 * (a_1 * a_2) * \sin ((360 * (f_1 + f_2)) + (\phi_1 + \phi_2 - 90)$"
"$y = (h_2 * a_1) \sin ((360 * f_1 * t) + \phi_1)$"
"$y = (h_1 * a_2) \sin ((360 * f_2 * t) + \phi_2)$"
"$h_1 * h_2$"

The equivalent rule when using radians is as so:

If we multiply:
“$y = h_1 + a_1 \sin ((2\pi * f_1 * t) + \phi_1)$”
... by:
“$y = h_2 + a_2 \sin ((2\pi * f_2 * t) + \phi_2)$”
... where f_1 is faster than, the negative of, or equal to f_2, then the equivalent sum of waves and a mean level will be:
“$y = 0.5 * (a_1 * a_2) * \sin ((2\pi * (f_1 - f_2)) + (\phi_1 - \phi_2 + 0.5\pi)$”
“$y = 0.5 * (a_1 * a_2) * \sin ((2\pi * (f_1 + f_2)) + (\phi_1 + \phi_2 - 0.5\pi)$”
“$y = (h_2 * a_1) \sin ((2\pi * f_1 * t) + \phi_1)$”
“$y = (h_1 * a_2) \sin ((2\pi * f_2 * t) + \phi_2)$”
“$h_1 * h_2$”

The main difference is that the Addition Waves have 0.5π radians added and subtracted instead of 90 degrees. Each formula now has a multiplication by 2π instead of 360.

Cosine waves

The rule for finding the sum of four Cosine waves and the mean level that are equivalent to the multiplication of two Cosine waves, when working in degrees, is as follows:

If we multiply:
“$y = h_1 + a_1 \cos ((360 * f_1 * t) + \phi_1)$”
... by:
“$y = h_2 + a_2 \cos ((360 * f_2 * t) + \phi_2)$”
... where f_1 is faster than, the negative of, or equal to f_2, then the equivalent sum of waves and a mean level will be:
“$y = 0.5 * (a_1 * a_2) * \cos ((360 * (f_1 - f_2)) + (\phi_1 - \phi_2)$”
“$y = 0.5 * (a_1 * a_2) * \cos ((360 * (f_1 + f_2)) + (\phi_1 + \phi_2)$”
“$y = (h_2 * a_1) \cos ((360 * f_1 * t) + \phi_1)$”
“$y = (h_1 * a_2) \cos ((360 * f_2 * t) + \phi_2)$”
“$h_1 * h_2$”

The equivalent rule when working in radians is as so:

If we multiply:
“$y = h_1 + a_1 \cos ((2\pi * f_1 * t) + \phi_1)$”
... by:
“$y = h_2 + a_2 \cos ((2\pi * f_2 * t) + \phi_2)$”
... where f_1 is faster than, the negative of, or equal to f_2, then the equivalent sum of waves and a mean level will be:
“$y = 0.5 * (a_1 * a_2) * \cos ((2\pi * (f_1 - f_2)) + (\phi_1 - \phi_2)$”
“$y = 0.5 * (a_1 * a_2) * \cos ((2\pi * (f_1 + f_2)) + (\phi_1 + \phi_2)$”
“$y = (h_2 * a_1) \cos ((2\pi * f_1 * t) + \phi_1)$”
“$y = (h_1 * a_2) \cos ((2\pi * f_2 * t) + \phi_2)$”
“$h_1 * h_2$”

The only difference is that there are multiplications by 2π instead of by 360.

Taylor series for Sine and Cosine

In Chapter 2, I gave formulas for calculating the Sine and Cosine of angles. These formulas required treating the angles as radians. The fact that the simplest formula for calculating Sine and Cosine uses radians is an example of how radians can be thought of as one of the “natural” divisions of circles.

The symbol for radians

When using degrees, it is customary to use the little raised up circle symbol: ° to indicate that a value represents an angle in degrees. Therefore, 45 degrees is written as 45°. In this book, I just say “degrees” instead of using the symbol so as to make everything visually clearer. One generally accepted way of indicating that a value represents an angle in radians is to use the word “rad” after the value, as so: “2π rad”. Older methods include using a superscripted (as in raised up) “r”, as so: “π^r”, which suffers from the problem that the “r” can be confused with “r” for radius in certain situations. In this book, I just say “radians” after a value as it seems the simplest way. As I said earlier, the context will usually indicate whether a value is in degrees or radians. In more advanced mathematical books and explanations, it is rare to see angles given in any form but radians.

τ (Tau)

The number "τ", as mentioned in the previous chapter, is equal to 2π. Given that, there is a good reason for using "τ" instead of 2π when working with radians. By doing this, any angle given in radians will be prefixed by the fraction of a circle that the angle represents. For example, 0.75 of the way around the circle is given by 0.75τ radians instead of 1.5π radians. As another example, 0.45 of the way around the circle is given by 0.45τ radians instead of 0.9π radians. It is easier to see how much of a circle is represented when using "τ" than when using 2π. Using "τ" is closer to the idea of "dividing" a circle into 1 piece. In practice, "τ" is rarely used to specify angles in radians, although there are people who advocate its use. The biggest problem with "τ" is how it is easily mistaken for the Latin letter "T". Note that the symbol "τ" is frequently used to represent values other than 2π, which is probably another reason its use has never become popular.

Thoughts on radians

In most maths, radians are used all the time instead of degrees. In fact, in some maths, it only makes sense to use radians. Whether radians need to be used *exclusively* in the mathematics of waves is a different matter entirely. There are good reasons for using radians. For example:

- Using Imaginary powers of the number "e" to identify points on the circumference of a unit-radius circle only works in radians.

- The Taylor series for Sine and Cosine work in the most straightforward way in radians.

- Calculus with waves works most simply when using radians.

- Computer programming languages and computer processors tend to use radians.

Conversely, there are many situations when using radians unnecessarily complicates matters. In Chapter 16 on the multiplication of waves, recognising patterns in the results would have been much harder if we had used radians instead of degrees. It is easy to tell that, say, 15 degrees is quarter of a circle less than 105 degrees, and how these angles look on a circle. It is much harder to tell that 0.26179939 radians is a quarter of a circle less than 1.83259571 radians. Similarly, it is obvious that 89.9999 degrees is just under quarter of a circle, and that 90.0001 degrees is just over quarter of a circle. It is far from obvious that 1.57079807 radians is just under quarter of a circle and that 1.57079458 is just over quarter of a circle.

In some situations, radians are more useful, and in some situations, degrees are more useful. I think it is a mistake to use radians exclusively once a certain level of maths has been reached. You should use whichever is best for the task that is being done. It is illogical to make something more complicated than it needs to be just because you feel obliged by social convention to use radians.

From the point of view of learning about waves, radians add an unnecessary level of complication to matters. If we want to describe some small angles, it makes more sense to use a whole number of degrees, rather than some unwieldy fraction. It is easier to buy a protractor with degrees on it, and it is easier to visualise a number of degrees. It is for that reason that I am only introducing radians now. Having said all of that, radians are the angle system that is generally used in more advanced maths, so it pays to become used to them. The more you use them, the easier they will become.

Conclusion

Some maths is much more straightforward when using radians and π, than when using degrees. This is because π is a significant number in the world of circles: the circumference of a circle is 2πr and the area is πr^2. An angle system based on 2π can, at times, be more useful than one based on an arbitrary division such as 360.

From this point onwards in this book, I will be mostly using radians in formulas as that is the best way to become used to standard mathematical conventions. I will use degrees if I think it makes an example easier to understand, and to start with, I will use both at the same time. When it comes to describing whole circles, half circles, and quarter circles, I will tend to say 360 degrees, 180 degrees, and 90 degrees, instead of 2π radians, π radians, and 0.5π radians. These are more natural ways of expressing these concepts in English, even if they are not more natural ways of expressing them in maths.

If you want to find radians easy to use, then it is useful to know the following angles in radians and degrees by heart:

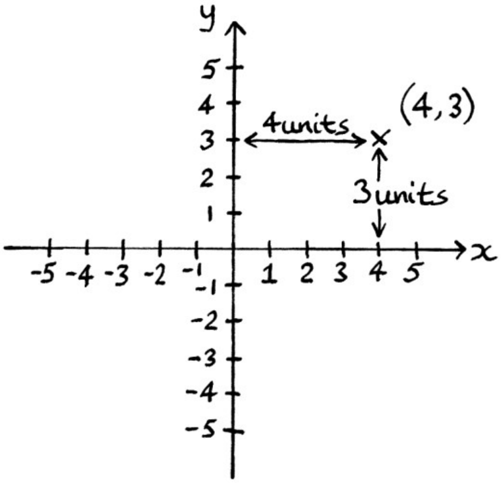
0.25π radians = 45 degrees
0.5π radians = 90 degrees
π radians = 180 degrees
1.5π radians = 270 degrees
2π radians = 360 degrees

Chapter 23: Complex numbers

Explaining Complex numbers is difficult to do without introducing concepts and then explaining them in more detail later on. Once you aware of all the concepts, you can put them together and everything should make sense. For this reason, while you read this chapter, you might not understand what I am trying to say until later on. However, you should keep reading and do not be worried if it does not make sense to start with.

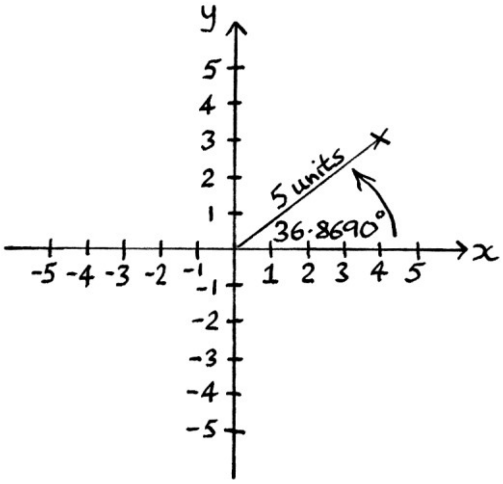
Coordinates

One way of identifying the position of any point on a graph with "x" and "y" axes is by using coordinates. For example, if the x-axis position of a point is at 4, and the y-axis position of a point is at 3, the coordinates are given as (4, 3).



This way of identifying the position of a point uses what are called "Cartesian coordinates". The system is named after the French mathematician René Descartes who used the idea in his works. Cartesian coordinates are also sometimes called "rectangular coordinates", or just "coordinates".

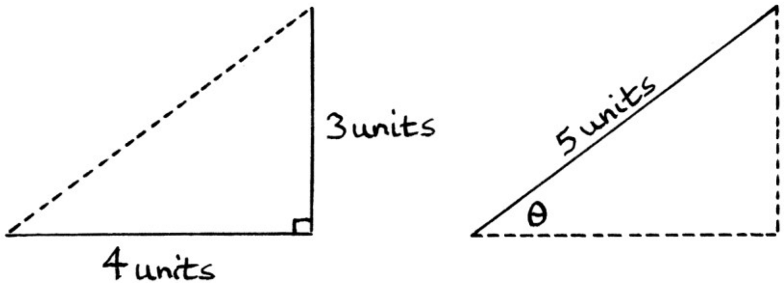
Another way to describe any point on the axes is to use a line of a particular length at a particular angle. For example, the point at (4, 3) is 5 units from the origin of the axes at an angle of 36.8690 degrees. One way of writing this is by using the angle symbol "∠" as so: 5∠36.8690 degrees. This means that there is a point 5 units from the origin at an angle of 36.8690 degrees. This system uses what are called "Polar coordinates".



We now have two main ways of writing coordinates for the same point:

- As a grid reference based on the origin of the axes (Cartesian coordinates).
- As an angle and distance from the origin of the axes (Polar coordinates).

If we think of a point as being at the end of the hypotenuse of a right-angled triangle, the Cartesian coordinates are describing its position using the length of the adjacent and opposite sides, and the polar coordinates are describing its position using the angle and the length of the hypotenuse.

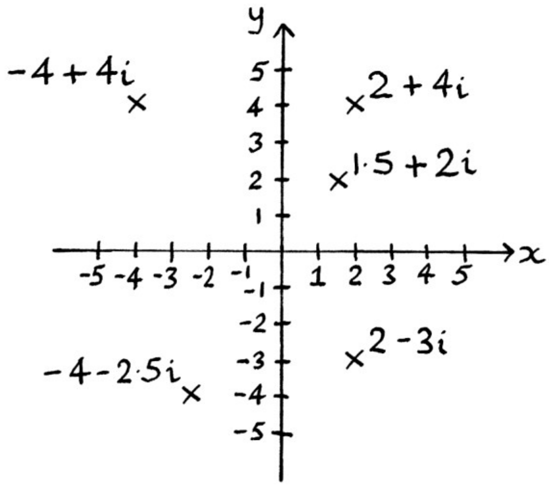


Having the two systems makes mathematical calculations on the coordinates easier. We can choose whichever system fits best with the sort of calculation we want to do. As a trivial example, if we wanted to calculate what would happen if our point were shifted horizontally or vertically, it would be easier to use Cartesian coordinates; if we wanted to calculate what would happen if our point were rotated around the axes, it would be easier to use polar coordinates.

Complex numbers

Complex numbers are, in essence, Cartesian coordinates that are treated as one entity to make it easier to perform mathematical operations on them. Whereas before, we gave Cartesian coordinates in the form (3, 4), when using Complex numbers, we give them in the form "3 + 4i". The first number, in this case 3, represents the x-axis value of the coordinate. The second number, in this case 4 followed by the letter "i" represents the y-axis value of the coordinate. The numbers are called Complex numbers because they involve two components. In this sense, the word "complex" means "consisting of different parts" as opposed to "difficult to understand".

Some examples of Complex numbers as drawn on x and y-axes are:



A Complex number is really a sum, which might be apparent if you think of how it is two parts separated by a plus or minus sign, such as with "12 + 7i". The first part is just a typical run-of-the-mill number; the second part is a typical run-of-the-mill number multiplied by a symbol called "i". The symbol "i" represents the square root of −1. As the square root of −1 is an unsolvable number, this means the two parts can never be solved as a sum. Therefore, they remain separated and that allows them to remain as coordinates. Although this might not make any sense at the moment, it will become clearer as this chapter progresses.

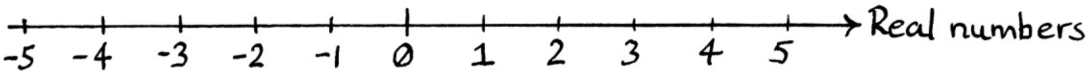
By convention, Complex numbers are usually written with the "i" part second, as in "7 + 8i" instead of "8i + 7", although mathematically, it does not matter in which order they appear.

Number lines

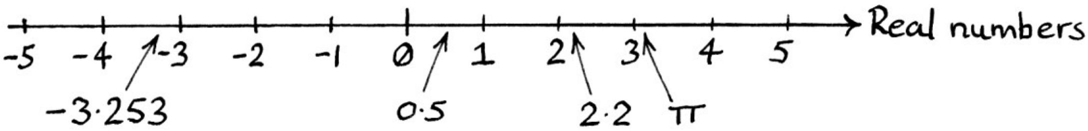
Complex numbers in their most basic form identify points on two number lines at 90 degrees to each other. They act as coordinates identifying a point in a world of numbers. In this section, I will explain what this means.

The Real number line

“Real numbers” are typical run-of-the-mill everyday numbers. In other words, they are positive and negative multiples and fractions of 1 such as 2, 27, −100, 0.323, 838, a million, −11.7 and so on. Real numbers are what most people would just call “numbers”. The list of every Real number can be laid out in a line on a chart as so:



This is called a “number line” because it is a line of numbers. Every number is on this line. The positive numbers are to the right of zero, and the negative numbers are to the left of zero. The whole numbers are marked on the line, but every possible non-integer is there too, between them.



The line is identical to an x-axis on a graph, or to put it another way, an x-axis on a graph is really a number line listing every possible number. We generally only draw the part of an x-axis that is relevant to what we are trying to show, and the same is true for a number line. If we were to draw all of a number line, it would reach forever to the left and to the right.

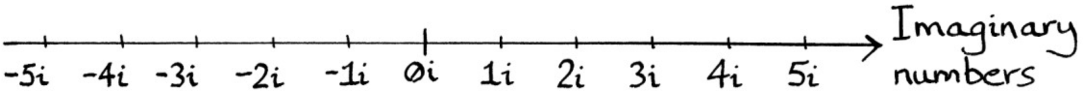
It is best to think of the word “Real” in the phrase “the Real number line” as just a randomly chosen word used as a label to *identify* a type of numbers. The numbers in the line are not particularly more or less real than anything else. In that sense, the word “Real” does not *describe* the numbers. Any word could have been chosen to identify the number line, but unfortunately the worst one possible was picked, and one that already has a meaning that conflicts with the entity it is being used to describe. [For example, negative numbers appear on the Real number line, but do not appear in real life. We could use Real numbers to count pebbles on a beach (which exist), and also to count unicorns living on the moon (which do not exist).]

Real numbers do not have to refer to something that is real. Every explanation of Complex numbers complains about the word “Real”, but no one ever changes it.

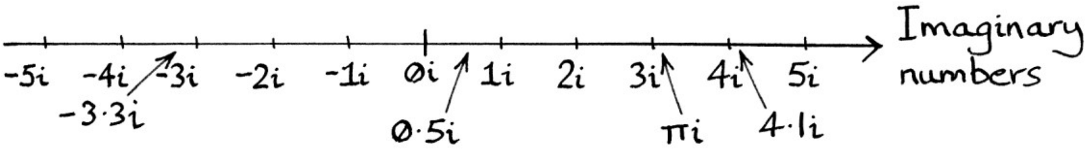
The Imaginary number line

Imaginary numbers are similar to real numbers, except that they are positive and negative multiples and fractions of the *square root of −1*. The square root of −1 is usually represented by the letter “i”, although it is sometimes represented by the letter “j”. Of course, the square root of −1 (or “i”) is an impossible number – such a calculation cannot be solved. However, “i” is a useful contrivance that helps solve other mathematical calculations that would be impossible otherwise. It is as if someone said, “We do not know what the square root of −1 is, but what if we pretended that we did?” As long as we never actually need to solve the square root of −1, we can use the number, and numbers based on it, in calculations. With luck, the “i” parts will cancel out leaving us with solutions that do not involve “i”. If they do not cancel out, we end up with more multiples of “i”, which is fine because such numbers still have *meaning* even if they cannot be reduced to a nice Real number.

The Imaginary number line is similar to the Real number line, but instead of being a list of multiples and fractions of 1, it is a list of multiples and fractions of 1 multiplied by “i” (so in other words 1 multiplied by the square root of −1). The number 1 multiplied by “i” is more usually written as just “i” on its own.



All multiples and fractions of 1 multiplied by “i” appear on the Imaginary number line:

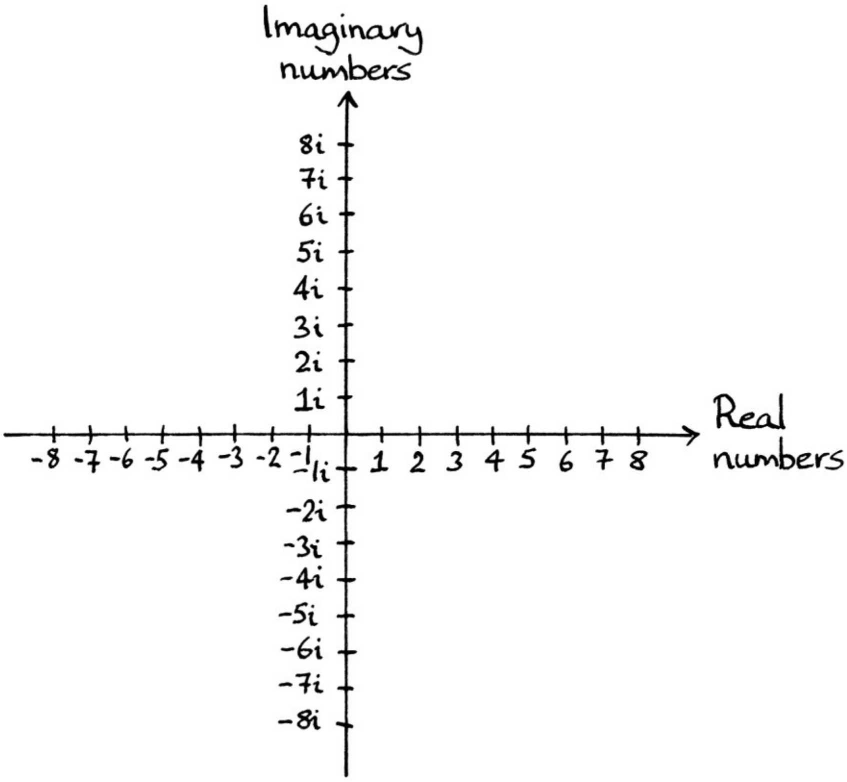


As with the word “Real”, it is best to think of the word “Imaginary” in the phrase “Imaginary number line” as just a randomly chosen word used as a label to *identify* a type of numbers. Although the numbers cannot be reduced mathematically to simple run-of-the-mill numbers, anything represented by the numbers in this line is not particularly more or less imaginary than anything else. As with the word “Real”, the word “Imaginary” does not *describe* the numbers. Any word could have been chosen to identify the Imaginary number line, but we ended up with a terrible one instead.

It is very common to find people trying to learn who do not understand that “Real” and “Imaginary” are words that *identify* the types of numbers, but do not *describe* the types of numbers.

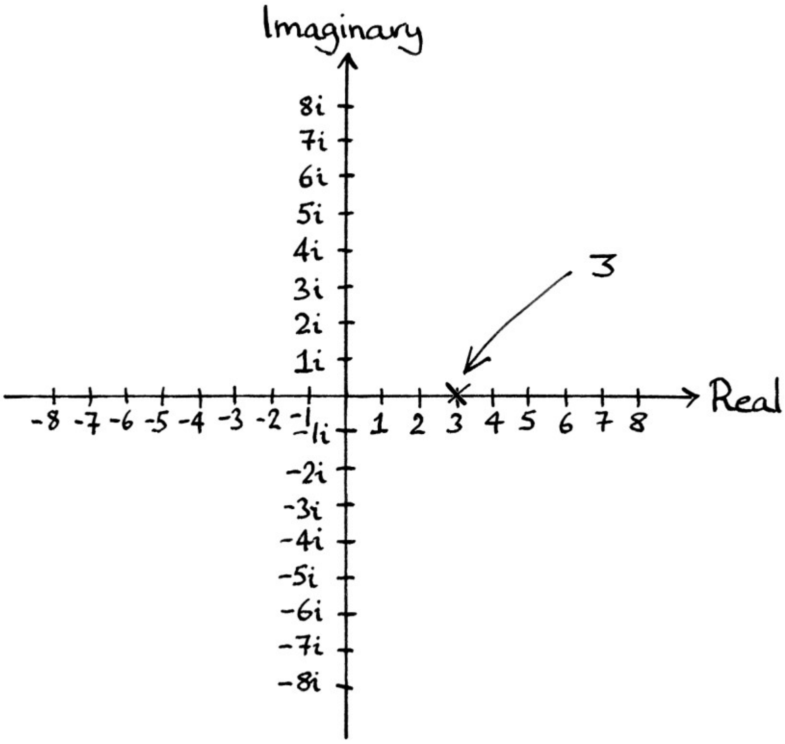
Real and Imaginary number lines

There is a relationship between Real numbers and Imaginary numbers that means it makes mathematical sense to have the two number lines on the same chart, with the Imaginary number line shown vertically, and the Real number line shown horizontally. The Imaginary number line becomes the y-axis and the Real number line becomes the x-axis.

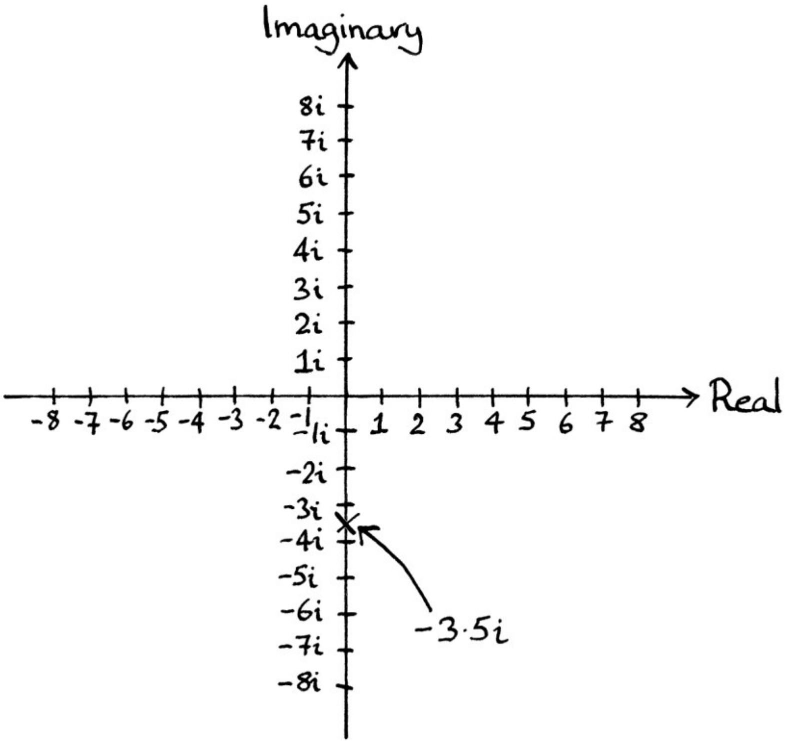


The axis of Imaginary numbers is called the “Imaginary axis”, and the axis of Real numbers is called the “Real axis”. These are often abbreviated to “I” and “R” when drawn on graphs.

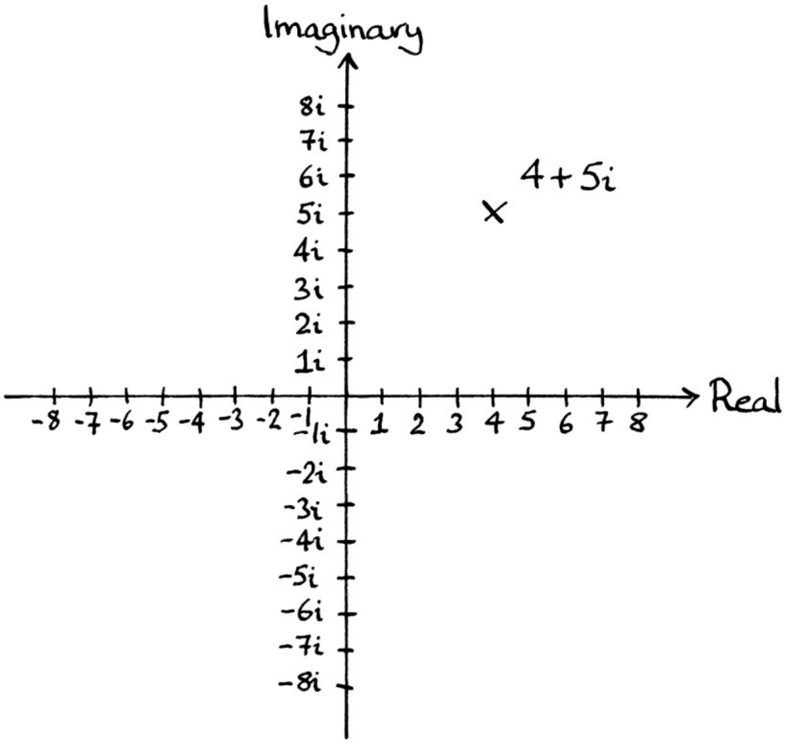
Any point lying on the x-axis (the Real axis) indicates the Real number at that point.



Similarly, any point lying on the y-axis (the Imaginary axis) indicates the Imaginary number at *that* point:



Everything becomes more interesting when we can pick points that are not directly on the axes. In doing so, we are identifying something that combines parts from both the Real and the Imaginary number lines. This "something" can be represented with what is called a "Complex number".



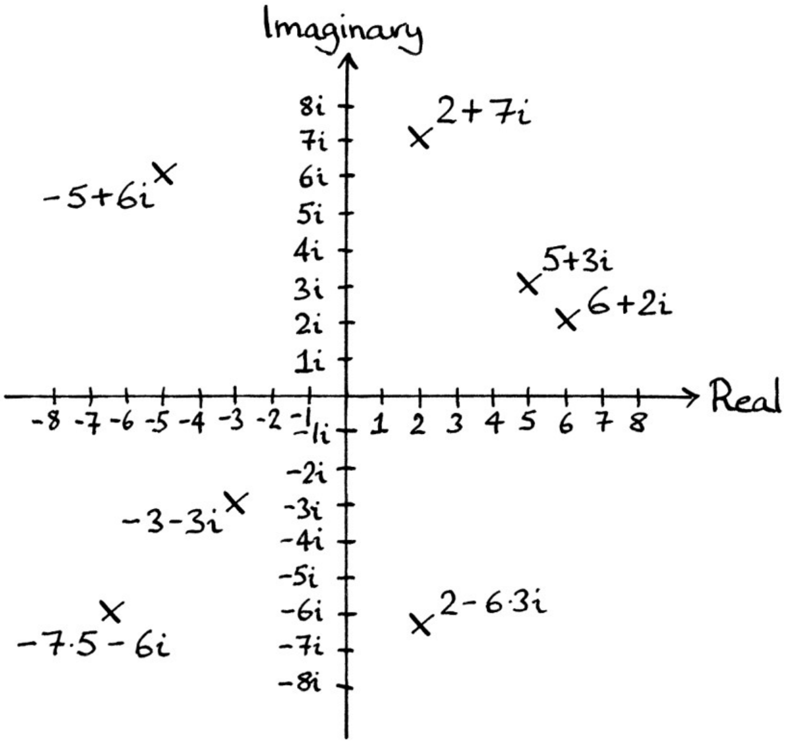
A Complex number is an entity with two parts: a Real part and an Imaginary part. The two parts identify a place on the Real and Imaginary number chart. In that sense, a Complex number is really a coordinate. A normal, non-Complex, run-of-the-mill Real number, on its own, describes how far, and in which direction, a value is situated along the Real number line. An Imaginary number, on its own, describes how far, and in which direction, a value is situated along the Imaginary number line. A Complex number describes how far, and in which direction, a value is situated along *both* number lines. Another way of thinking about this is that a Complex number gives the coordinates of where in the "number world", a particular point of interest lies.

The Real and Imaginary number chart is more usually called "the Complex plane".

The Complex number system treats all numbers as having a Real part (that goes along the Real axis or x-axis) and an Imaginary part (that goes along the Imaginary axis or y-axis). For example, the number "12 + 8i" has a Real component of 12, which is to say its x-axis value is 12, and it has an Imaginary component of 8, which is to say its y-axis value is 8. A number that does not seem to have an Imaginary part, such as a run-of-the-mill everyday Real number, can be said to have an

Imaginary part of zero. Therefore, the everyday number 7 is really, if we think of it as a Complex number, “7 + 0i”. Similarly, an Imaginary number that does not have a Real part can be said to have a Real part of zero. Therefore, the Imaginary number 23i is really the Complex number, “0 + 23i”.

Some examples of Complex numbers are:



To summarise things so far: the Real and Imaginary number chart (or what is properly called the Complex plane), is really a chart that shows all the possible Real and Imaginary numbers, and all their possible combinations. We can identify particular combinations of Real and Imaginary numbers by marking their position on the chart. We can identify these positions using Complex numbers, which can be thought of as another way of writing coordinates.

Complex numbers are a sum

The notation for a Complex number is really a sum. The number "10 + 7i" means: "10" added to "7 multiplied by the square root of −1"

Of course, this sum cannot be solved, as $\sqrt{-1}$ is incalculable. However, this means that the sum stays as a sum, so the coordinates stay as coordinates. One of the beauties of Complex numbers is that two different aspects are stored as one entity.

Complex number maths

Complex numbers lend themselves to mathematical calculations as if they were just typical run-of-the-mill Real numbers. We can do calculations on them involving Real numbers, Imaginary numbers, and even other Complex numbers.

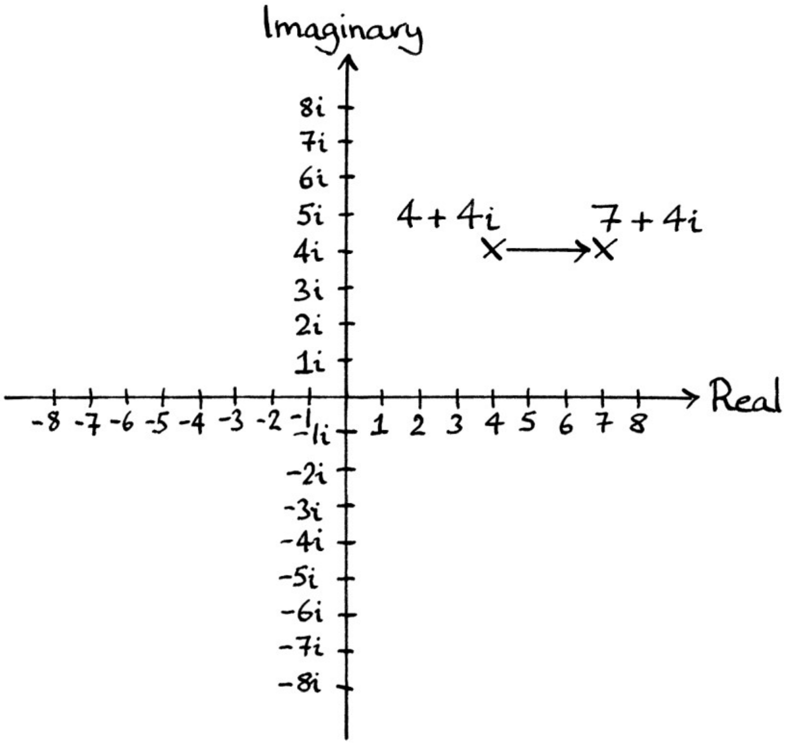
Complex number maths is easier to understand if you think of Complex numbers as coordinates of points on a grid.

Addition and subtraction

To add a Real number to a Complex number, we just add the Real number to the Real part of the Complex number. For example, 3 added to "4 + 6i" is "7 + 6i".

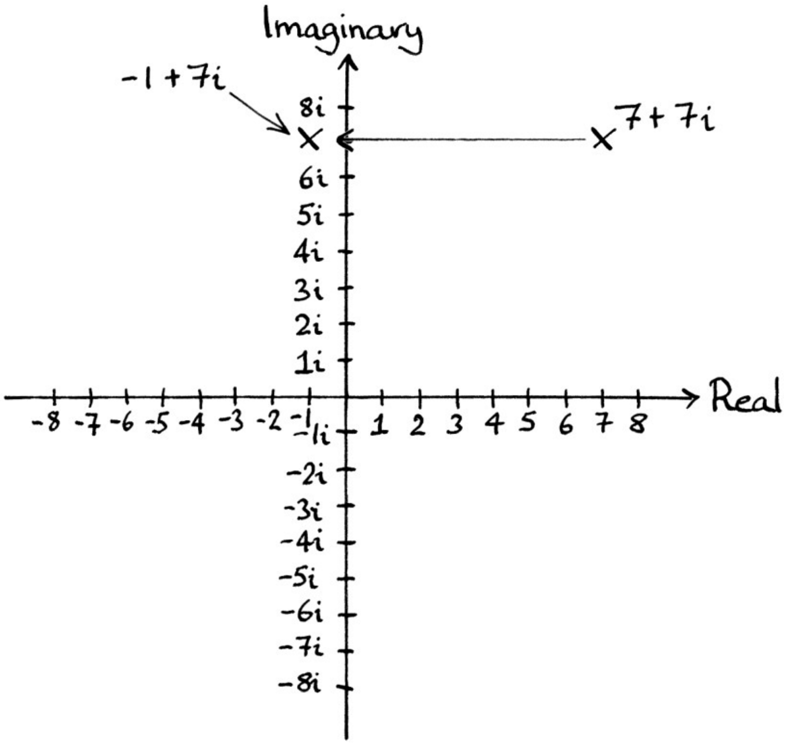
On the Complex plane, that is to say the graph, adding a positive Real number to a Complex number results in the point indicated by the Complex number being shifted to the right. This is more obvious if we think about how we are adding or subtracting to the Complex number's x-axis value.

For example, 3 added to "4 + 4i" equals "7 + 4i". The point becomes shifted to the right along the Real axis by 3 units. This works in the same way as adding 3 to a normal run-of-the-mill Real number – it ends up further down the number line.

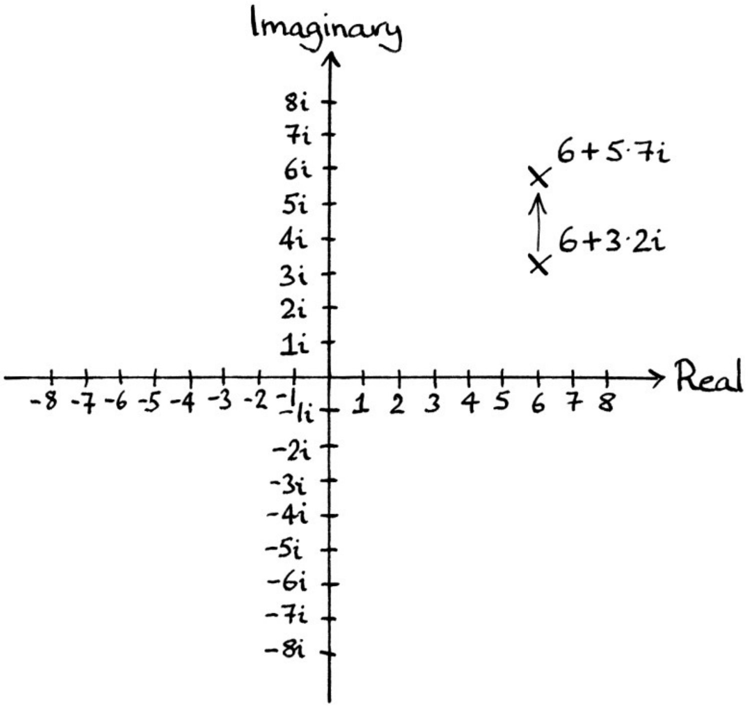


Subtraction works in the same way – if we subtract a Real number from a Complex number, we just subtract that value from the Real part of the Complex number. On the graph, the point described by the Complex number becomes shifted to the left.

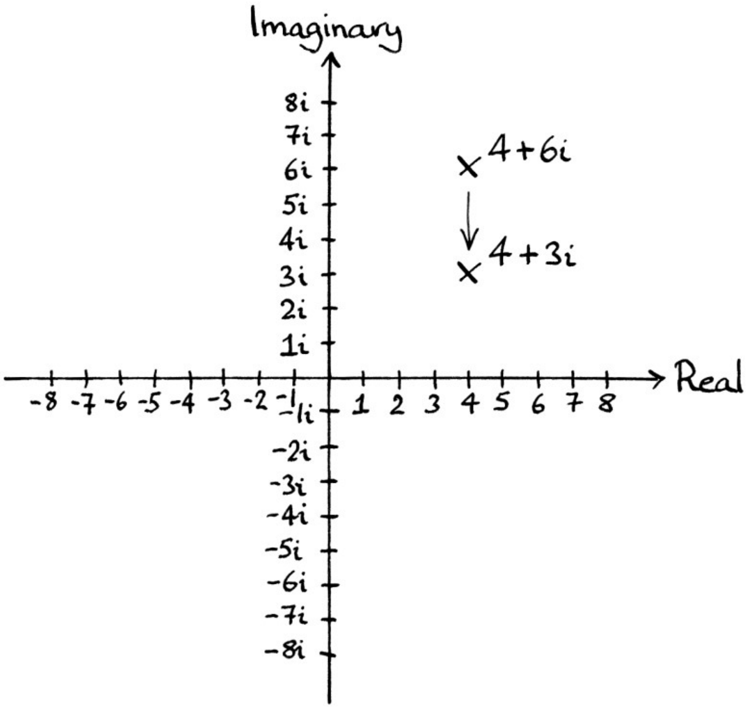
For example, 8 subtracted from “7 + 7i” results in “−1 + 7i”.



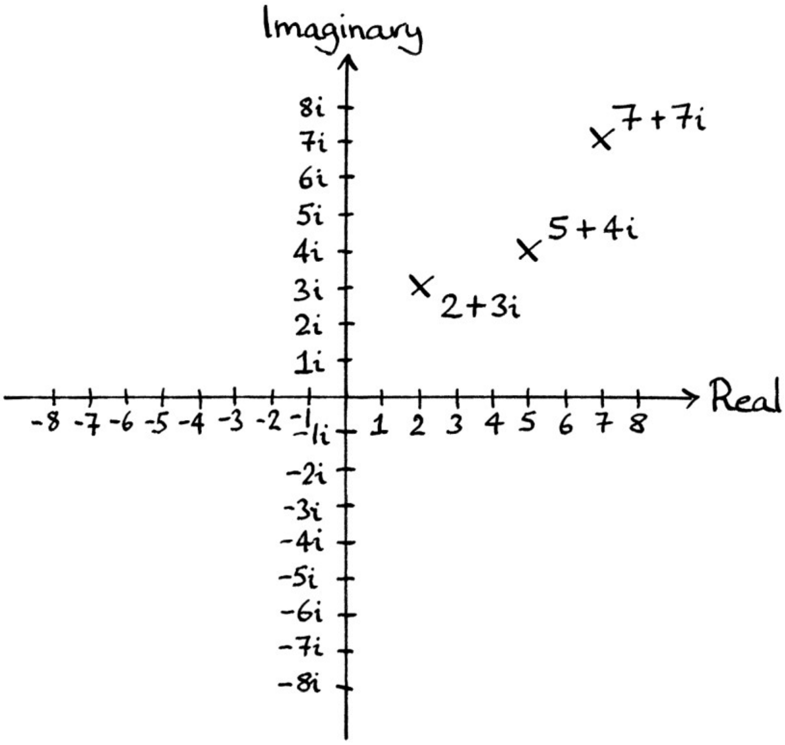
Adding an *Imaginary* number to a Complex number involves adding it to the Imaginary part of the Complex number. Adding a positive Imaginary number will result in the point indicated by the Complex number moving up the Imaginary axis. If we add 2.5i to “6 + 3.2i”, we end up with “6 + 5.7i”. The point moves up the Imaginary axis by 2.5 units.



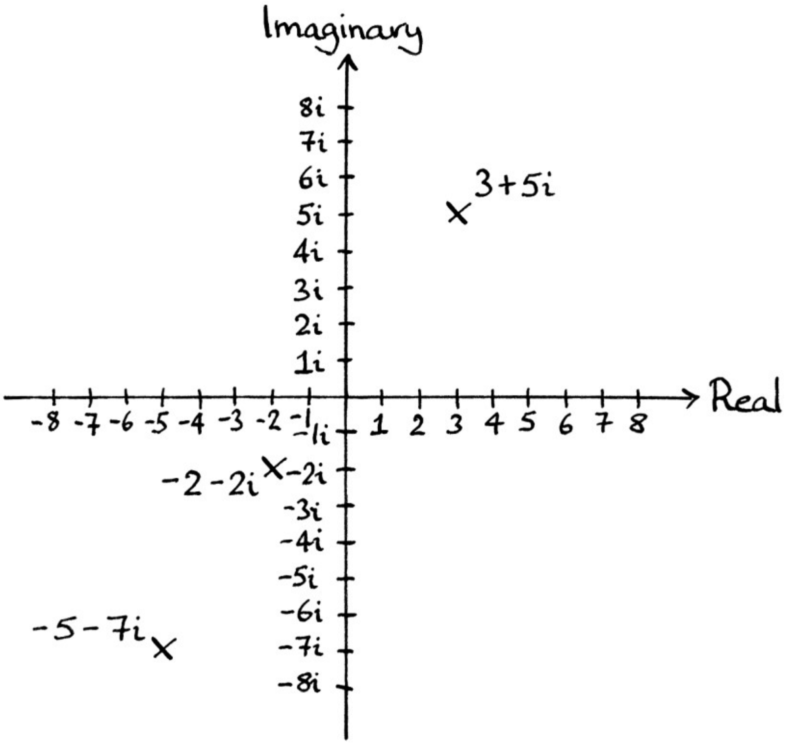
Subtracting an Imaginary number from a Complex number works in the same way. For example, 3i subtracted from “4 + 6i” results in “4 + 3i”.



Adding two Complex numbers together just involves adding each individual part together separately. In other words, we add the two Real parts and we add the two Imaginary parts. For example, “5 + 4i” added to “2 + 3i” is equal to “7 + 7i”.



The Complex number “−5 − 7i” added to “3 + 5i” is “−2 − 2i”

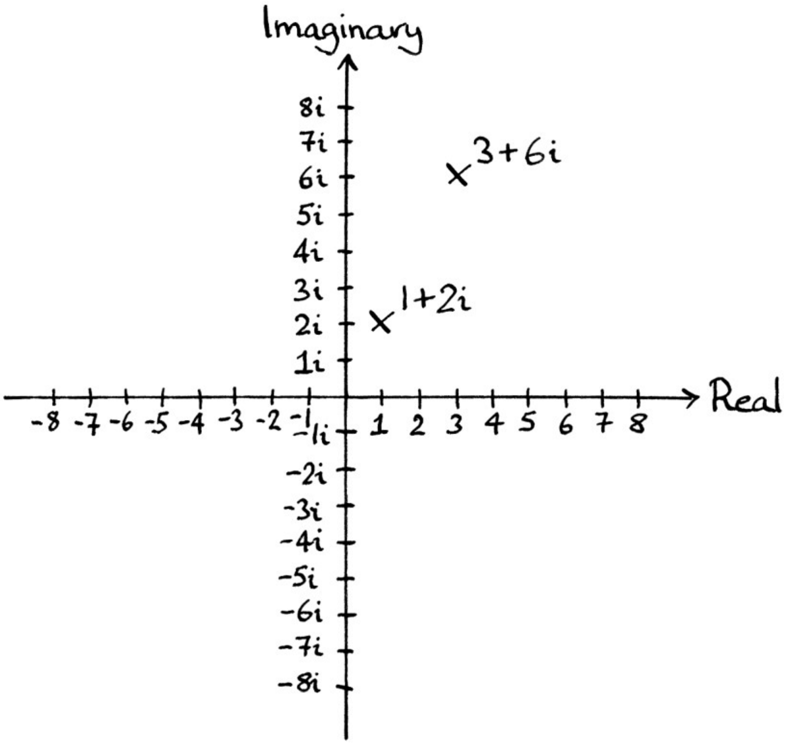


Subtraction works in the same way.

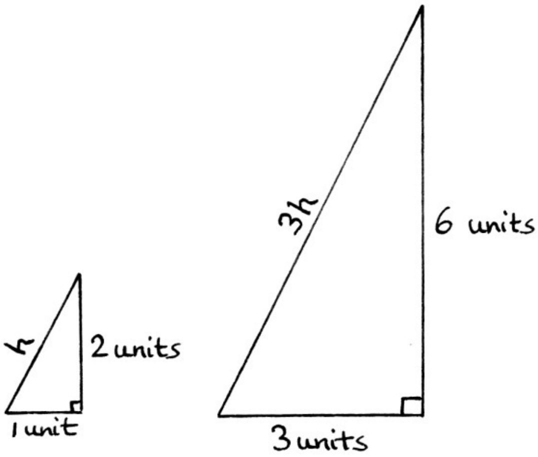
When using addition and subtraction with a Real or Imaginary number and a Complex number, it is easiest to treat the Real or Imaginary number as a Complex number. In other words, if we have the Real number, “6”, we can treat it as “6 + 0i”. Similarly, if we have the Imaginary number “3.45i”, we can treat it as being “0 + 3.45i”. This can sometimes make things easier to visualise.

Multiplication

Multiplying a Real number by a Complex number is reasonably straightforward. We multiply each part of the Complex number by the Real number. For example, “3” multiplied by “1 + 2i” becomes “3 + 6i”. On the Complex plane graph, this results in a point that is 3 times as far away from the origin.



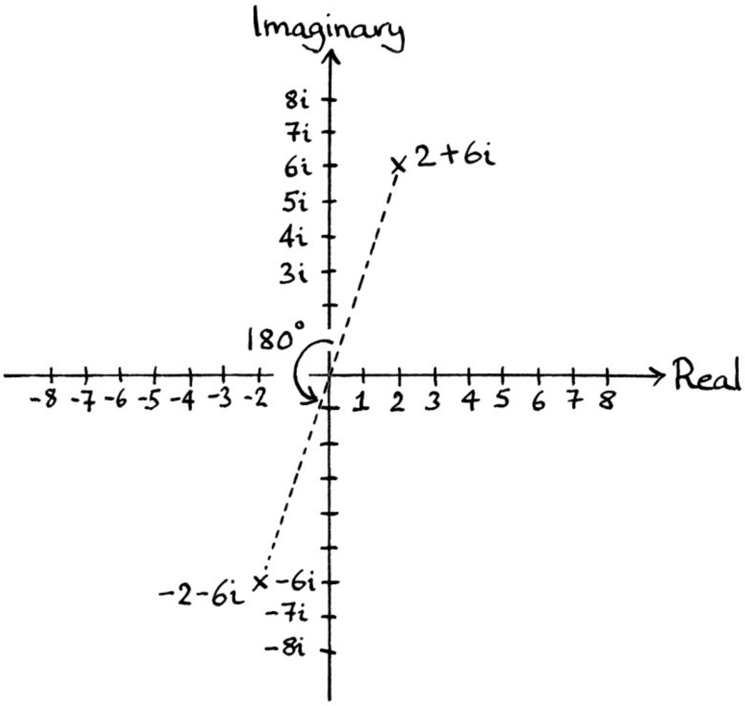
We can know this is true by thinking of the relevant right-angled triangle – its opposite and adjacent sides have been scaled by 3, so its hypotenuse also becomes scaled by 3.



As another example: "−4 – 5i" multiplied by 2 is "−8 – 10i".

If we multiply a Complex number by +1, it makes no difference. As an example, "5 + 8i" multiplied by 1 = "5 + 8i". This is obvious, but it pays to make this observation because of the next few examples.

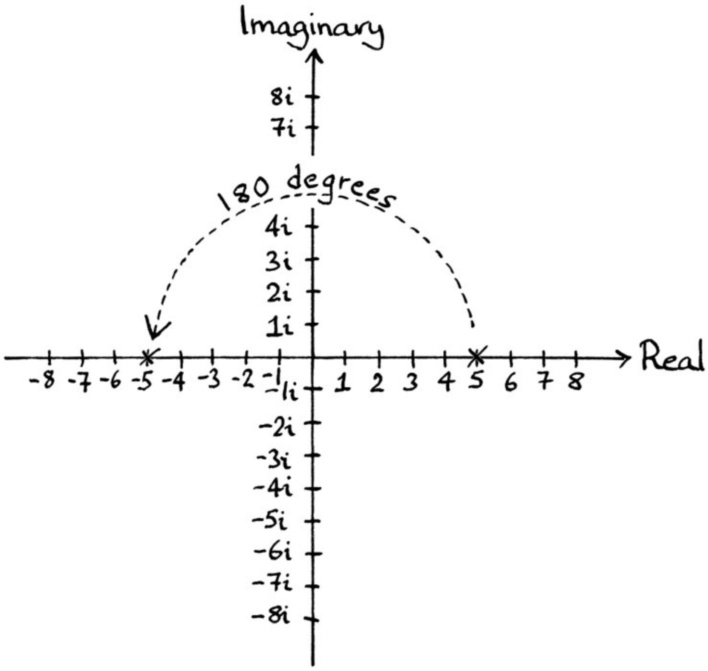
If we multiply a Complex number by −1, the result is still straightforward: "2 + 6i" multiplied by −1 is "−2 – 6i". Interestingly, on the graph, the point has become rotated by 180 degrees.



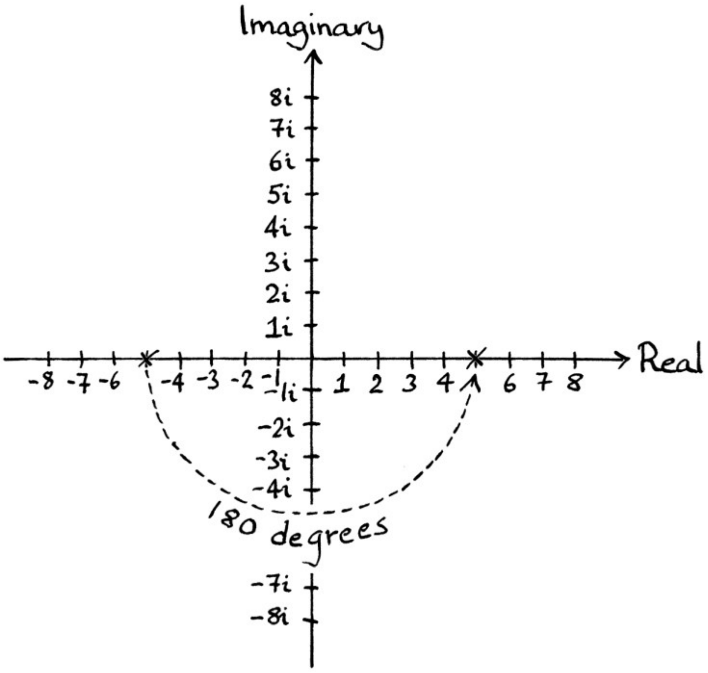
Note that the point is still the same distance from the origin.

Multiplying anything by −1 always rotates a point by 180 degrees. This is easiest to see if we multiply a Real number by −1. It becomes rotated by 180 degrees to end up on the other side of the number line.

For example, 5 multiplied by −1 is −5.



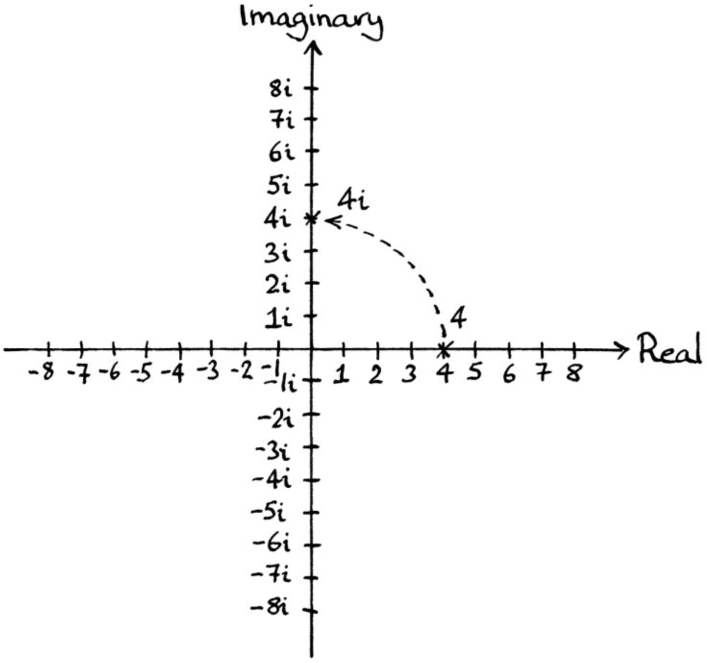
And −5 * −1 = 5. The point becomes rotated by 180 degrees again.



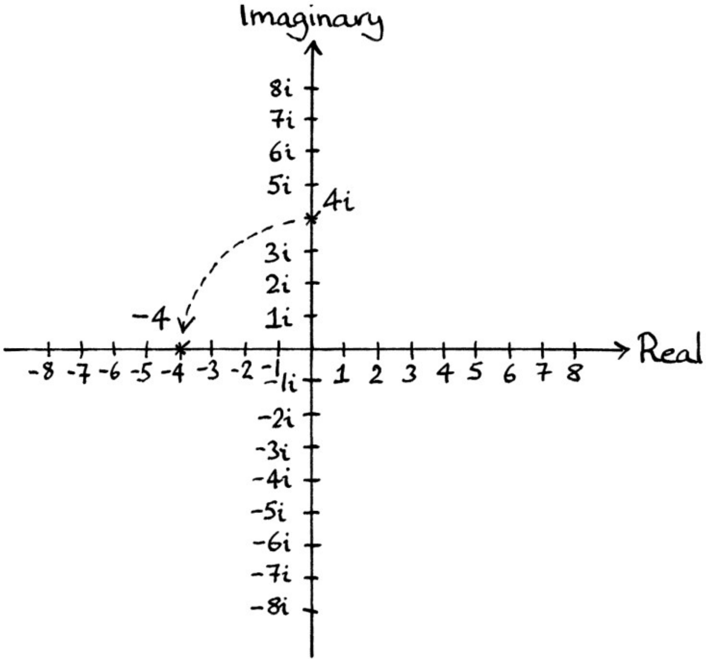
If we multiply any point twice by −1, it is the same as multiplying that point by $(-1)^2$, which is the same as multiplying it by +1.

Possibly the most interesting thing about Complex numbers is what happens when we multiply a number, whether Complex, Imaginary or Real, by "i" (the square root of −1). Whereas a multiplication by −1 rotates a point 180 degrees, a multiplication by "i" rotates a point by +90 degrees.

For example, if we multiply the Real number "4" by "i", we end up with the Imaginary number "4i". The number has gone from being an entirely Real number to being an entirely Imaginary number. On the graph, it is clear that this multiplication has rotated the point by +90 degrees.

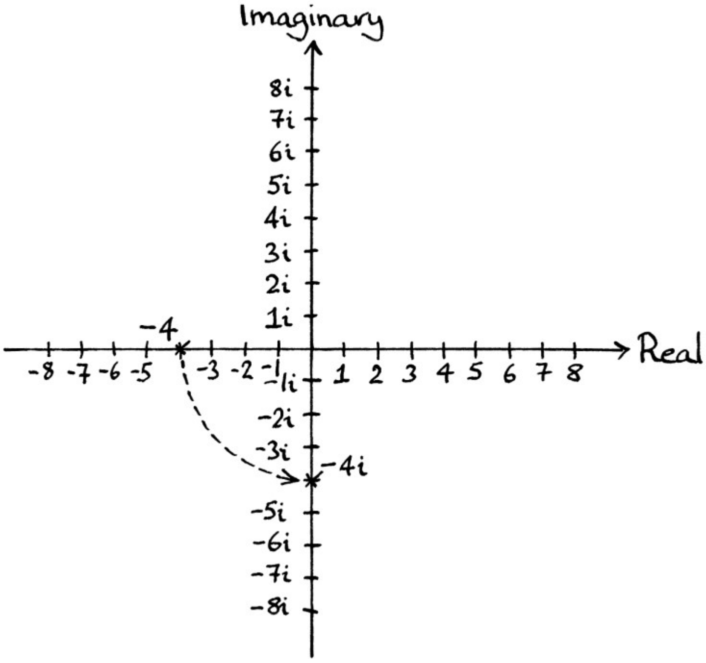


If we then multiply 4i by "i", we end up with $4i^2 = -4$ (because "i^2" is $(\sqrt{-1})^2$, which is −1). Again, this is a rotation of +90 degrees.

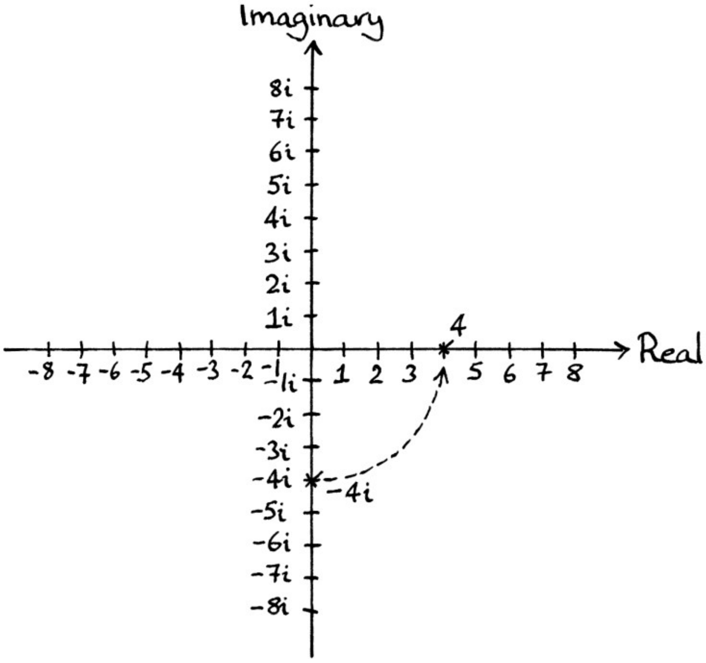


From this, we can see that a multiplication by "i" is equivalent to half the "rotational amount" of a multiplication by −1.

If we multiply −4 by "i", we end up at −4i:



If we multiply −4i by "i" we end up back at +4:



Multiplying any number by "i" results in the point it represents being rotated by +90 degrees. For example:

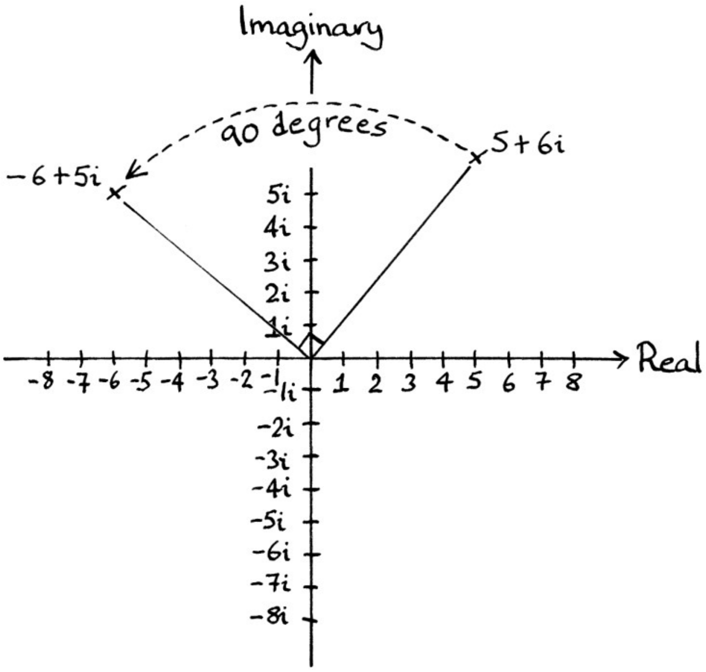
(5 + 6i) * i

= 5i + 6i^2

... and because i^2 is really $(\sqrt{-1})^2$, which is −1, this ends up as:

= 5i – 6, which would normally be written as:

–6 + 5i



Given that "i" is the square root of −1, it makes sense that two multiplications by it would be identical to a single multiplication by −1. Two multiplications by "i" are the same as a multiplication by the *square* of "i", which is −1. From this we can see that the rotation effect of a multiplication by "i" is half that of a multiplication by −1. In this way, it is possible to have an insight into Imaginary numbers. It is as if Imaginary numbers are in some "half-way dimension" that does not really make sense if we only ever looked at the Real number line.

It is the +90 degree rotation aspect of a multiplication by an Imaginary number that explains why the Imaginary number line is the y-axis to the Real number line's x-axis. We can see how someone's pursuit of a "half way number dimension" would have led them to inventing the concept of Imaginary numbers.

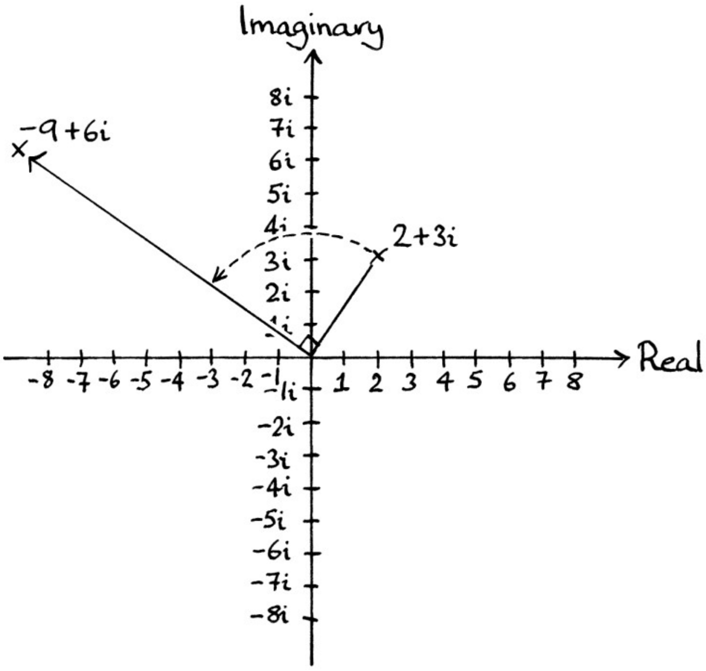
Note that multiplication by just "i" does not alter the point's distance from the origin. In this sense, the distance from the origin of the result is similar to that of a multiplication by "1". In fact, you can think of "i" here as being "1i".
Multiplying by "i" four times is the same as multiplying by:
$(\sqrt{-1})^4 = (-1)^2 = +1$

Any number multiplied by "i" four times has the same result as if it were multiplied by 1. Multiplying by "i" two times is the same as multiplying by $(\sqrt{-1})^2 = -1$.

Given that a multiplication by "i" results in a +90 degree rotation, and a rotation by −1 results in a 180 degree rotation, we could say that a multiplication by +1 results in a 360 degree rotation. Of course, in everyday life, it probably makes more sense to think of a multiplication of +1 as having no effect. However, being aware that we *could* say that a multiplication by +1 results in a rotation of 360 degrees leads on to useful ideas when beginning to understand the meaning behind "e" raised to Imaginary powers.

As we now know, if we multiply a Complex number by "i", it rotates the point by +90 degrees. For example:
$(3 + 4i) * i = 3i + 4i^2 = 3i - 4$, which we can write as $-4 + 3i$.

If we multiply a Complex number by a *multiple* of "i", it rotates the point by +90 degrees at the same time as scaling the point's distance from the origin. For example:
$(2 + 3i) * 3i$
$= 6i + 9i^2$
$= 6i - 9$
$= -9 + 6i$



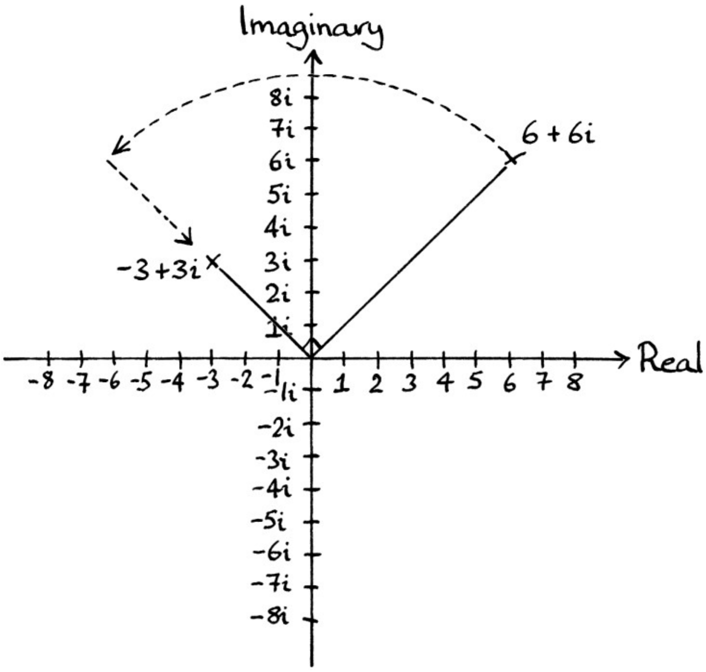
As another example:

(6 + 6i) * 0.5i

= 3i + 3i^2

= 3i + −3

= −3 + 3i



Multiplication of two Complex numbers

Things become slightly more complicated when we multiply two Complex numbers together. In such a case, we have to break the calculation into parts. The idea is easiest to visualise if we multiply the following variables, which represent ordinary, run-of-the-mill numbers: (a + b) * (c + d)

First, we multiply: a by (c + d) to produce: ac + ad
Then, we multiply: b by (c + d) to produce: bc + bd
Then we add the two results together to produce: ac + ad + bc + bd

This is basic maths, but it pays to see that the multiplication of Complex numbers works in the same way.

If we want to multiply: (3 + 7i) * (2 + 6i), then we will have:
3 * (2 + 6i) = 6 + 18i
... added to:
7i * (2 + 6i) = 14i + 42i^2

The result of this sum is:
6 + 18i + 14i + 42i^2
... which, because i^2 is −1, becomes:
6 + 18i + 14i + −42
... which is:
−36 + 32i

One of the good things about Complex numbers is that we never end up with unsolvable squares or cubes because "i" is not an unknown variable – it is the square root of −1. Therefore, any power of "i" will either be a Real number or a multiple of "i".

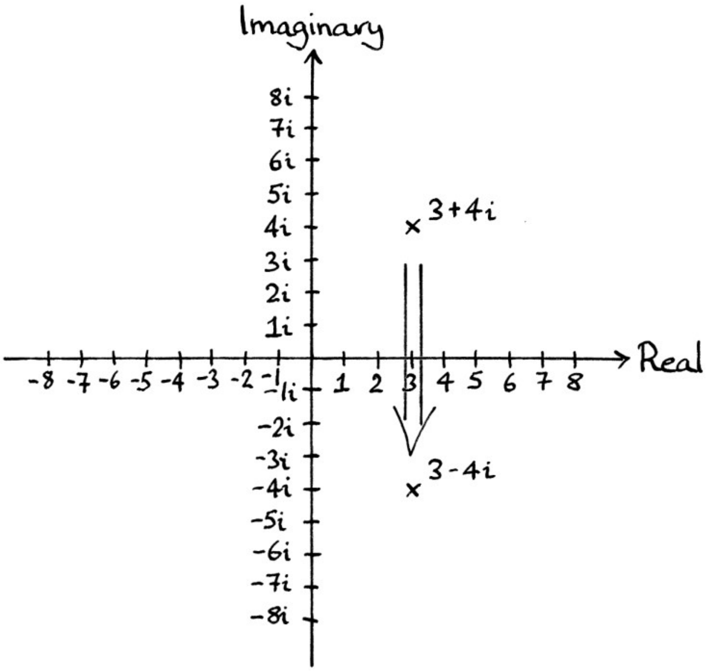
Division with two Complex numbers

Division of one Complex number by another is the more complicated of the basic mathematical calculations, although it is not particularly difficult once you know how it works.

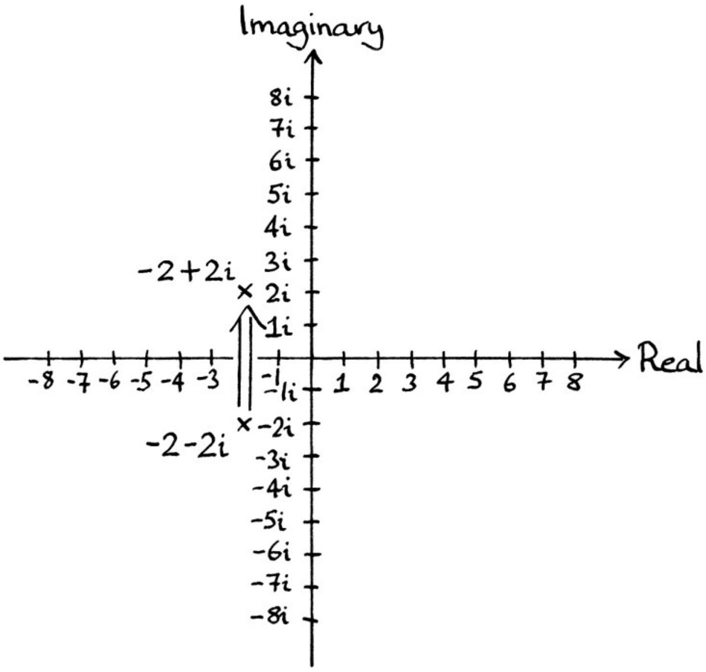
Before we look at division, we have to introduce a new concept called "the Complex Conjugate".

The Complex Conjugate

The Complex Conjugate is just an overly wordy term for the coordinates of a point on the Complex plane after it has been mirrored vertically over the Real axis (the x-axis). For example, if we have a point at 3 + 4i, then its Complex Conjugate is the point on the other side of the Real axis (the x-axis), which is 3 – 4i.



The points 3 + 4i and 3 – 4i are the same point but mirrored vertically across the Real axis. If we have the point at −2 – 2i, then its Complex Conjugate is the point at −2 + 2i.



The points −2 – 2i and −2 + 2i are the same point mirrored vertically across the Real axis. They are the same distance from the Real axis (the x-axis), but in different directions. The Real value part of each Complex number stays the same – it is only the Imaginary part that changes when the point is flipped up or down.

Another way of thinking about the Complex Conjugate of a Complex number is that it is the same number but with the sign of the Imaginary part switched – if the original Imaginary part is positive, it becomes negative; if it is negative, it becomes positive.

The term "Complex Conjugate" makes the concept seem as if it should be difficult to understand, but really it is just a badly chosen term for something extremely simple. A more descriptive name might be something such as the "vertically flipped version of the number" or the "vertically mirrored number". From an etymological point of view, the term "Complex" in "Complex Conjugate" refers to how it relates to Complex numbers. The term "Conjugate" ultimately derives from the Latin word "coniungere" or "conjungere" meaning to join together or connect, with the idea that the Complex numbers in the top half of the axes are somehow connected to those in the bottom half. The term "Complex Conjugate" can, in this way, be thought of as being "the associated Complex number in the other half of the axes", or the "corresponding Complex number in the other half of the axes".

Ignoring the badly chosen name for the concept, here are some examples of Complex Conjugates to illustrate how simple the idea is:

A Complex number	**The Complex Conjugate of that number**
1 + 1i	1 – 1i
1 – 1i	1 + 1i
2 + 20i	2 – 20i
2 – 20i	2 + 20i
−7 – 8i	−7 + 8i
−45.8789 + 0.00005i	−45.8789 – 0.00005i
11 + 23.5i	11 – 23.5i
23 [which is also "23 + 0i"]	23 [which is also "23 – 0i"]
9i [which is also "0 + 9i"]	−9i [which is also "0 – 9i"]

The idea is very simple, but it can also be very useful.

One property of Complex Conjugates is that if we multiply a Complex number by its Complex Conjugate, we will end up with just a Real number on its own because the Imaginary parts will always cancel out.

For example, we will multiply “2 – 20i” by “2 + 20i”. This produces:
2 * (2 + 20i) = 4 + 40i
... added to:
−20i * (2 + 20i) = −40i + −400i^2 = −40i + (−400 * −1) = −40i + 400
... which all ends up as:
4 + 40i + −40i + 400
... which is:
404

We have ended up with just one Real number and no Imaginary numbers.

[Note how when we were converting positive-frequency Cosine wave formulas as portrayed on a circle to negative-frequency Cosine wave formulas, and vice versa, we flipped the phase point up or down across the x-axis. This is the same concept as the Complex Conjugate, but a different use of the idea]

Division with two Complex numbers continued

Now, we can return to the topic of division. To divide one Complex number by another, we first multiply *both* numbers by the Complex Conjugate of the divisor. [The divisor is the number we are dividing by.] Two things to note from doing this are:

- The calculation will still produce the same result – we are just scaling the top and bottom of the division by the same amount. This is equivalent to calculating 4 ÷ 2, and calculating 8 ÷ 4. The results are the same.
- The divisor will end up as a Real number because if we multiply a Complex number by its Complex Conjugate, the Imaginary part becomes cancelled out. This makes the calculation much easier.

As an example, we will divide “3 + 4i” by “2 + 1i”:

$$\frac{3 + 4i}{2 + 1i}$$

First, we calculate the Complex Conjugate of the divisor “2 + 1i”, which is “2 – 1i”. Then, we multiply both the top and the bottom of the division by this number:

$$\frac{(3 + 4i) * (2 - 1i)}{(2 + 1i) * (2 - 1i)}$$

This will still produce the same result because we have scaled the top and bottom by the same amount. However, this will make the calculation easier as the divisor will end up as a Real number with no Imaginary part.

The top half of the calculation is:
3 * (2 – 1i) = 6 – 3i
... added to:
4i * (2 – 1i) = 8i – 4i^2 = 8i – (4 * −1) = 8i + 4
... which ends up as:
6 – 3i + 8i + 4
... which is:
10 + 5i

The bottom half of the calculation is:
2 * (2 – 1i) = 4 – 2i
... added to:
1i * (2 – 1i) = 2i – i^2 = 2i – −1 = 2i + 1
... which ends up as:
4 – 2i + 2i + 1
... which is:
5

Therefore, we can rewrite the division as:

$$\frac{10 + 5i}{5}$$

... which is easy to solve. The result is “2 + 1i”. Therefore, we can say that:

$$\frac{3 + 4i}{2 + 1i} = 2 + 1\mathrm{i}$$

By coincidence, this also means that “3 + 4i” is the square of “2 + 1i”. Normally, such a Complex number would be written as: “2 + i”, but I am leaving in the 1 to make things clearer.

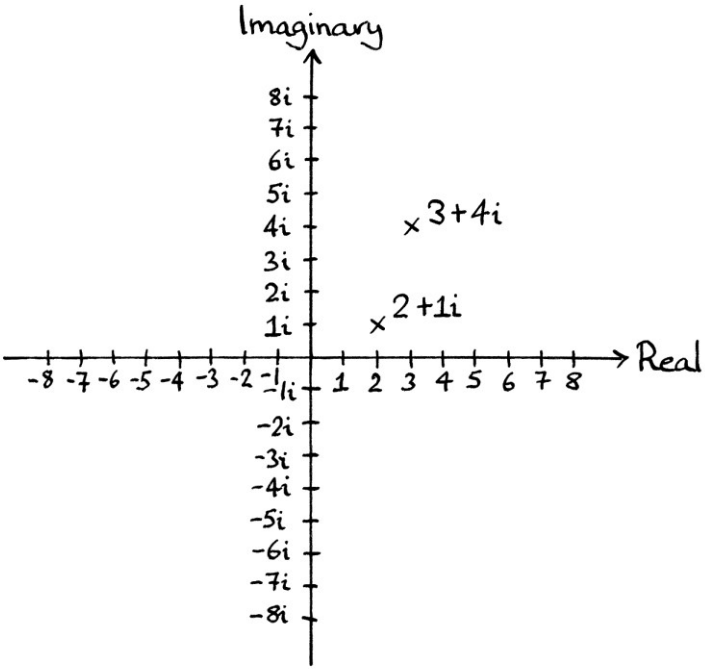
Supposing the last part of the calculation had not been as simple, we could have split it up into two parts as so:

$$\frac{10}{5} + \frac{5i}{5}$$

... or rewritten it as 0.2 * (10 + 5i). Both of these would have given us another way to solve it.

On a good modern calculator app, it is possible to do maths with Complex numbers and save a lot of effort, but it pays to understand how to do it by hand.

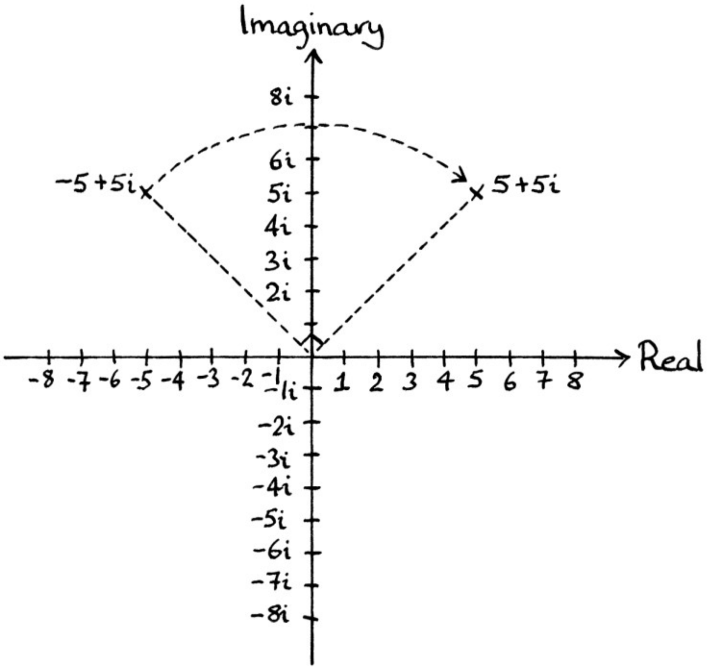
On the Complex plane, the points in the calculation, “3 + 4i” and “2 + 1i”, look like this:



Division by “i”

We saw how multiplication of a number by “i” rotates the point it represents by +90 degrees around the Complex plane. When we divide a number by “i”, the point it represents is rotated by −90 degrees around the Complex plane. In other words, a division by “i” rotates the point *clockwise* by 90 degrees.

As a simple example, “−5 + 5i” divided by “i” will be that point rotated by −90 degrees. This means that the point is rotated *clockwise* by 90 degrees. From thinking of the point on the axes, we know that the result will be “5 + 5i”. We can know this from using our imaginations, or we could draw some axes and use a protractor to calculate the result.



To calculate the result using maths, we would proceed as follows:
(−5 + 5i) ÷ i
= (−5 ÷ i) + (5i ÷ i)
= (−5 ÷ i) + 5
= 5 + (−5 ÷ i)
... and we would be stuck with this division.

To calculate "−5 ÷ i", we can use the Complex conjugate, while remembering that "i" is really "0 + 1i":

$$\frac{-5}{i} * \frac{0-1i}{0-1i} = \frac{5i}{-i^2} = \frac{5i}{-(-1)} = \frac{5i}{+1} = 5i$$

Therefore, the result of the original division is "5 + 5i", which is −90 degrees around the axes, which we already knew.

A quick rule for instantly knowing values divided by "i" is:

$$\frac{a}{i} = -ai$$

... where "a" is any number.

We can test the rule to see how it saves us time:
2 ÷ i = −2i. [This is the point at "2 + 0i" being rotated by −90 degrees to end up at "0 – 2i".]
11 ÷ i = −11i
−0.567 ÷ i = 0.567i [The Real number was negative, so it becomes positive.]

The rule works for Complex numbers too, but the solution requires more work. In such situations, it is clearer to give the formula as:

$$\frac{a}{i} = -ia$$

... with the "i" in front of the "a". This is because we will be multiplying the "–i" by something within brackets, and it is easier to do the negation and multiplication in one go than to have the calculation as "−(....)i"

We will use this rule in an example:

$$\frac{3+4i}{i} = -i * (3+4i) = -3i - 4i^2 = -3i + 4 = 4 - 3i$$

The point "3 + 4i" rotated by −90 degrees (as in clockwise) ends up at "4 – 3i".

Not only does the rule help us in calculating divisions by "i", but it also tells us that a division by "i" is the same as a multiplication by "−i". This means that a multiplication by "−i" also rotates by −90 degrees. A division by "i" and a multiplication by "−i" are the same thing. Interestingly, this implies that a division by "−i" is the same as a multiplication by "i". We can list the various rules:

- A multiplication by "i" rotates a point by +90 degrees.
- A division by "i" rotates a point by −90 degrees.
- A multiplication by "−i" rotates a point by −90 degrees.
- A division by "−i" rotates a point by +90 degrees.

If we start with the point "4 – 4i", and then multiply it by "i", we will end up with the point "4 + 4i", which is a rotation of +90 degrees. If we then *divide* that by "−i", we will end up with the point at "−4 + 4i", which is another rotation by +90 degrees. If we multiply that by "−i", we will end up with the point at "4 + 4i", which is a rotation by −90 degrees. If we divide that by "i", we will end up with the point at "4 – 4i", which is another rotation by −90 degrees, and is the point that we started with.

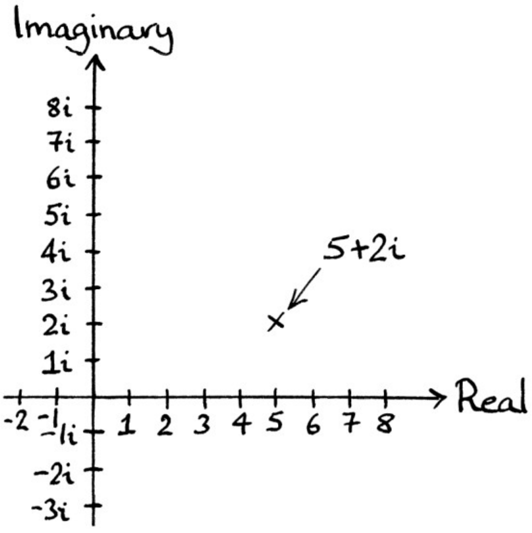
Although it is good to know these four ways of rotating a point, in this book we will generally concentrate on *multiplications* of "i", which are anticlockwise rotations of 90 degrees.

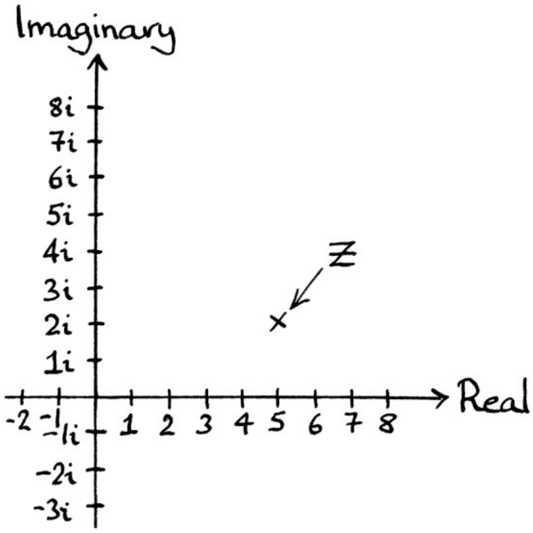
A subtle distinction

In all discussions of Complex numbers and the Complex plane, it is important to know that it is the *position* of a point in the number plane that is the significant thing. Complex numbers are just a way of identifying where that point is in the Complex plane. Strictly speaking, that point is not a Complex number – that point is just being *described* by a Complex number. There are several ways to identify that point's position, and those ways do not necessarily use Complex numbers.

This distinction is similar to how a place on a map is not a pair of coordinates, but we can use coordinates to identify that place. When it comes to Complex numbers the distinction between the place on the Complex plane that we are identifying and the Complex number itself are often blurred.

The letter "z" is used to identify the actual place in the Complex plane that is being referenced by a Complex number. In other words, we might say something such as, "z = 5 + 2i", where the position of "z" is at "5 + 2i". Using the map analogy, this is like saying a building is at the coordinates (5, 2). The point of interest is identified using the letter "z", and the position of "z" is described using Complex numbers.

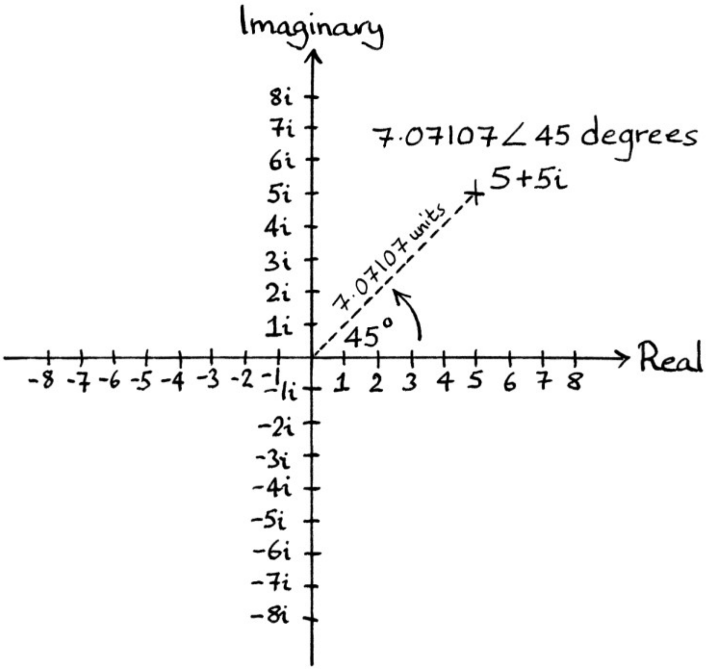




Converting Complex numbers to Polar coordinates

Given that Complex numbers are ultimately a variation of Cartesian coordinates referring to points in the Complex plane, it is also possible to refer to the same points using Polar coordinates. In this way, we would be describing the points using their angle and distance from the origin.

We can easily convert Complex numbers to Polar coordinates. We just have to treat the Complex numbers as adjacent and opposite sides of a right-angled triangle. For example, "5 + 5i" refers to the point that is 7.07107 units away from the origin at an angle of 45 degrees. We can write such a concept as "7.07107∠45 degrees" (or "7.07107∠0.25π radians") within the Complex number plane.



Although "7.07107∠45 degrees" or "7.07107∠0.25π radians" is not a number – it is really a description – it still refers to the same point within the Complex plane. Its meaning is the same, but it is a different way of thinking about it.

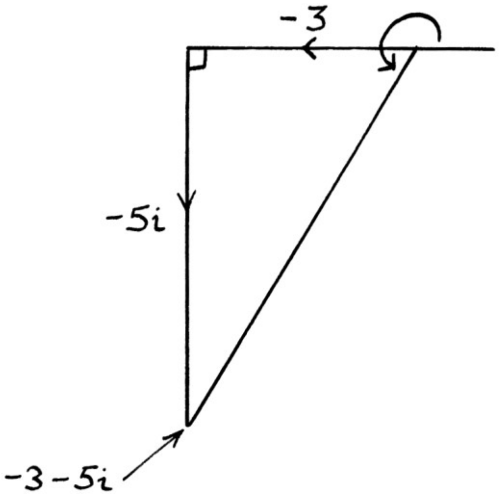
The distance from the point indicated by a Complex number to the origin of the Real and Imaginary axes is the same as the length of the hypotenuse of a right-angled triangle. We can calculate that length with Pythagoras's theorem:
$\text{hypotenuse}^2 = \text{opposite}^2 + \text{adjacent}^2$

The angle of the point's position with reference to the origin of the axes is the same as the arctan of the gradient of the right-angled triangle's hypotenuse:
$\theta = \arctan (\text{opposite} \div \text{adjacent})$

[As always, after using arctan, check if we have the correct result of the two possible ones.]

We will look at some examples of converting Complex numbers to Polar coordinates.

First, we will convert the number “−3 − 5i”. We will treat the number as indicating the point at the end of the hypotenuse of a right-angled triangle:



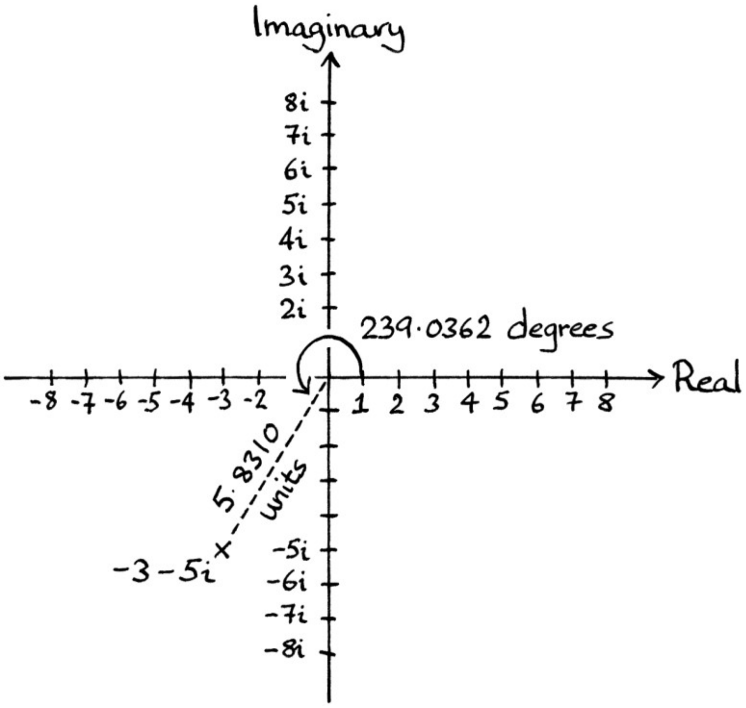
To calculate the dimensions of the triangle, we can ignore how the y-axis is Imaginary, and just treat the triangle as having sides of −5 and −3 units.

The hypotenuse is $\sqrt{-5^2 + -3^2}$ = $\sqrt{25 + 9}$ = 5.8310 units.

The angle is arctan (−5 ÷ −3) = arctan (1.6667) = 59.0362 degrees or 1.0304 radians. As the angle is in the bottom left hand quarter of the axes (the bottom left hand quarter of the Complex plane), we add 180 degrees (π radians) to it. The result is 239.0362 degrees or 4.1720 radians.

Therefore, the polar coordinate is:
5.8310∠239.0362 in degrees
... or:
5.8310∠4.1720 in radians.

The position of the point is 5.8310 units away from the origin at an angle of 239.0362 degrees or 4.1720 radians.



Next, we will convert this Complex number: “2.4 − 11.2i”.

The hypotenuse is $\sqrt{2.4^2 + -11.2^2}$ = 11.4543 units.

The angle is arctan (−11.2 ÷ 2.4) = −77.9052 degrees or −1.3597 radians. We will turn this into a positive angle. For the answer in degrees, we add 360 degrees: −77.9052 + 360 = 282.09476 degrees. For the answer in radians, we add 2π radians: −1.3597 + 2π = 4.9235 radians. The original point is in the lower right hand quarter of the axes (in other words of the Complex plane), and the angle of 282.09476 degrees is clearly in that quarter, so it is the correct answer for arctan.

[4.9235 radians is also in the lower right quarter of the axes, but it is harder to know that unless you have some experience in recognising values of radians. This is a good example of how degrees are a better type of angle for recognising patterns.]

The point of interest is 11.4543 units from the origin and at an angle of 282.09476 degrees, which is 4.9235 radians. We could give the Polar coordinates as:
11.4543∠282.09476 in degrees
... or:
11.4543∠4.9235 in radians.

Converting from Polar coordinates to Complex numbers is straightforward as we just think of right-angled triangles and use Sine and Cosine.

We will convert the result of the previous example (11.4543∠282.09476 in degrees or 11.4543∠4.9235 in radians) back to a Complex number.

In degrees, the Real part of the Complex number is:
11.4543 cos 282.09476 in degrees.

The Imaginary part of the Complex number is:
"i" multiplied by 11.4543 sin 282.09476 degrees
... which would normally be written as so:
11.4543i sin 282.09476

Using radians, the Real part of the Complex number is:
11.4543 cos 4.9235 radians.

The Imaginary part of the Complex number is:
"i" multiplied by 11.4543 sin 4.9235 radians
... which would normally be written as so:
11.4543i sin 4.9235

Whether we use degrees or radians, we end up with:
2.4 − 11.2i

Note that a Polar coordinate pointing to something in the Complex plane is not technically a "Complex number". Polar coordinates can refer to the same point in the Complex plane as a Complex number, but they, themselves are not Complex numbers.

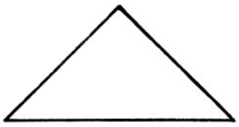
The Complex plane as general axes

The Complex plane contains the position of points in the realm of two number lines: the Real number line and the Imaginary number line. A Complex number refers to the position of a point within the Complex plane. In practice, Complex numbers and the Complex plane do not have to be used to refer specifically to just the domain of numbers. Instead, they can be treated as axes to refer to concepts other than just Real and Imaginary numbers.

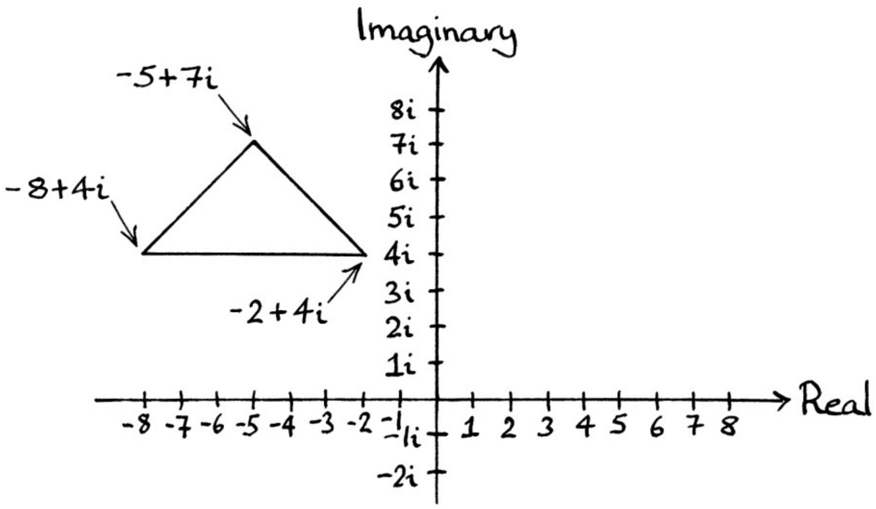
We can take any two-dimensional grid where the “x” and “y” axes represent anything, and treat the points as being on a Complex plane. Doing so enables us to perform mathematical operations on points described on those axes.

Using the Complex plane in this way is not particularly strange, as the “x” and “y” axes of *any* graph are really just made up of the Real number line drawn in two different directions. When using the Complex plane in this way, everything is the same as using the Complex plane normally. The number “i” is still used, and all the maths continues to work in the same way.

As an example, we will use a simple picture of a triangle. It is made up of three points connected by straight lines:

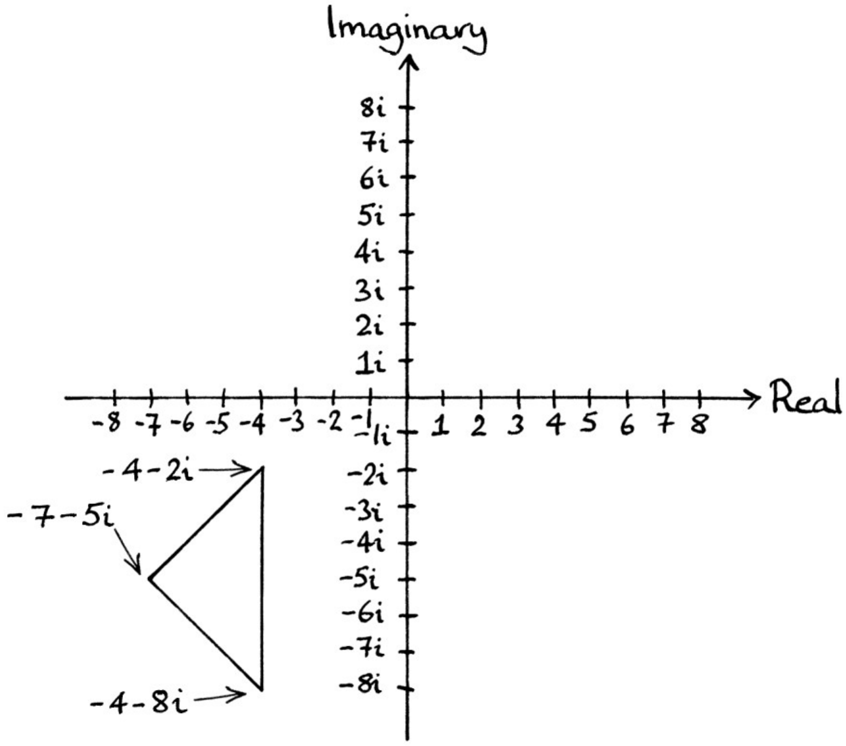


We can decide to treat the picture as if it were in the Complex plane, and so position it on a graph with Real and Imaginary axes:



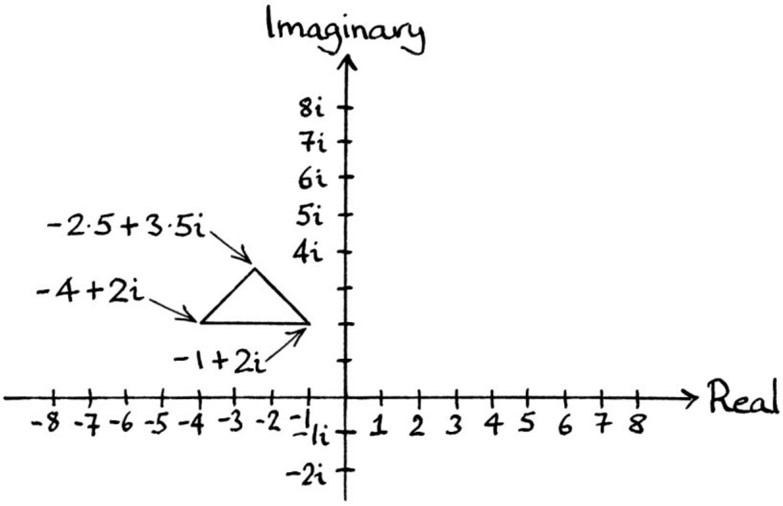
We can identify any point on the triangle using Complex numbers. To keep things simple, we will just look at the corners of the triangle, which are at:
−2 + 4i
−5 + 7i
−8 + 4i

If we multiplied the three points on the picture by “i”, we would end up rotating the triangle 90 degrees anticlockwise. The points would become:
(−2 + 4i) * i = −2i + 4ii = −2i – 4, which is: −4 – 2i
(−5 + 7i) * i = −5i + 7ii = −5i – 7, which is: −7 – 5i
(−8 + 4i) * i = −8i + 4ii = −8i – 4, which is: −4 – 8i



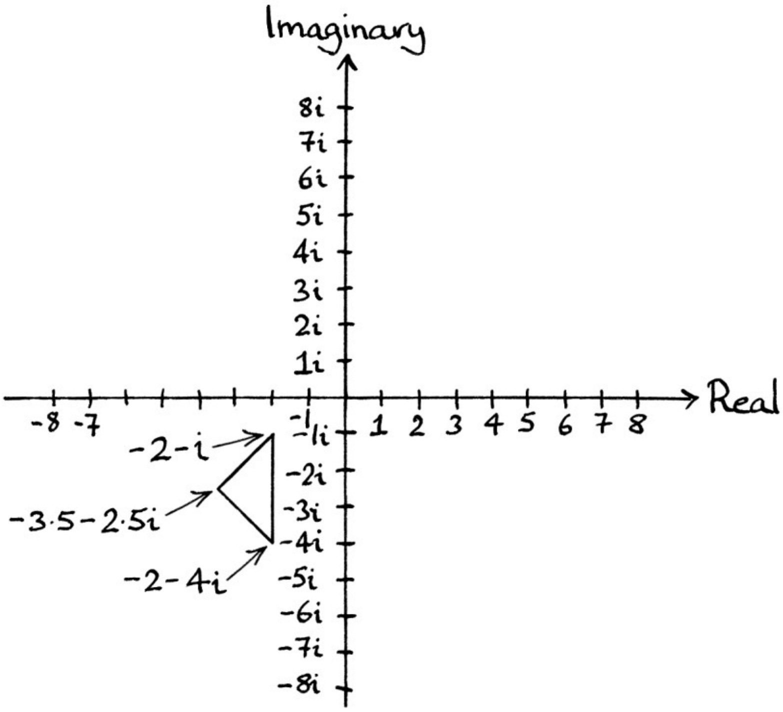
If we multiplied the original points by 0.5 (a Real number), we would end up scaling the whole picture by 0.5. Every single point would become half the distance from the origin. The points would become:

−1 + 2i
−2.5 + 3.5i
−4 + 2i



If we multiplied the original points on the picture by 0.5i, we would rotate the picture and scale it at the same time. The points would be as so:

−2 – 1i
−3.5 – 2.5i
−2 – 4i



If we referred to the points on the triangle with Polar coordinates, some other transformations would be easy. For example, if we wanted to rotate the triangle by just +1 degree, we would add 1 degree to the Polar coordinate for every point. The triangle is still in the Complex plane, but we are using Polar coordinates to alter it.

Thinking of the triangle as being in the Complex plane does not change the nature of the triangle. It is still a triangle. However, having it in the Complex plane allows us to think of it in a more mathematical way, and to perform mathematical operations on its points.

Square roots of negative numbers

As I have said before, "i" is the square root of −1. It is also possible to give the square roots of other negative numbers in terms of "i".

The square root of +2 is 1.4142. The square root of −2 is 1.4142i.
The square root of +3 is 1.7321. The square root of −3 is 1.7321i.
The square root of +8 is 2.8284. The square root of −8 is 2.8284i.
The square root of +2.5 is 1.5811. The square root of −2.5 is 1.5811i.

From these examples, you might be able to see that the square root of a negative value is the same as the square root of that value made positive multiplied by "i". A more mathematical way of expressing this is:
$\sqrt{-x} = i * \sqrt{+x}$

Potential sources of confusion

The names given to concepts in maths are often terrible, and this is very apparent with Complex numbers.

Complex, Real, Imaginary, and complex, real and imaginary

One of the most confusing matters to do with Complex numbers is how some people will use the term "complex" to refer to Complex numbers, but also use the term to refer to "complicated" things or things that have more than one component. For example, in a book that mentions Complex numbers, you might read the phrases: "a complex situation", "a complex formula", "a complex sound". If you are new to Complex numbers, it can be difficult to know if an author is referring to something complicated, or if they are referring to some new concept relating to Complex numbers.

The most common confusing term is "a complex waveform", which means a signal created from adding two or more pure waves, or what I would call an impure signal. In this sense, "complex" means "made up of more than one part". A "complex waveform" has nothing to do with Complex numbers. To make this more confusing, you will see the terms "Complex signal" or "Complex sinusoid", which mean a signal or pure wave existing in the Complex plane, or in dimensions related to the Complex plane. The terms "Complex signal" and "Complex sinusoid" *do* relate to Complex numbers, and the "Complex" part of the name refers to Complex numbers. If one wished to be completely specific, one might use the term "Complex-plane signal" or "Complex-plane sinusoid" to emphasise the difference.

Personally, I think it is confusing to use the word "complex" in anything remotely related to Complex numbers, unless it specifically refers to Complex numbers. That is why I do not do that in this book. I also write "Complex", when it refers to Complex numbers, with a capital "C" to emphasise the difference. Most people would write the word "Complex" when referring to Complex numbers with a lower-case "c". The Oxford English dictionary spells it with a lower-case "c".

Given that not everyone realises that the terms "Complex" and "complex" can be ambiguous, do not be surprised to see variations of "complex wave" or "Complex signal" that are intended to mean whatever the writer wants them to mean.

Similar to the ambiguity of the word "Complex", you will often find explanations that use the word "Real" to mean the x-axis aspect of the Complex number plane, and then use the word "real" to mean "existing in nature". For example, "the change in amplitude is real, and not the result of a mistake" or "this signal cannot be real". The same goes for the word "imaginary". I always write the words "Real" and "Imaginary" with initial capital letters when I am talking about Complex numbers to avoid any ambiguity. Most people would not capitalise these words, and the Oxford English dictionary spells them with lower-case initial letters. If you do not want to confuse people, it pays to be careful when using the words "real" and "imaginary" when you are not referring to Complex numbers.

Complex numbers are not complicated

Complex numbers are not "complex" in the sense of being complicated. The term "complex" in the name refers to how they consist of multiple elements, or specifically two elements – a Real part and an Imaginary part. Complex numbers are really just another way of writing coordinates. Once you realise that, they are fairly simple. It is easy to be put off by their name and presume that they require great intelligence to understand, which is far from the truth. The greatest difficulty you will ever have from Complex numbers is from overthinking them.

"Real" and "Imaginary" are not literally real and imaginary

Another source of confusion is thinking that the Complex number term "Real" literally refers to items that can exist in the real world, and thinking that the term "Imaginary" refers to items that cannot exist in the real world. As I have explained before, "Real" and "Imaginary" are just badly chosen words used to *identify* the axes. They are just names. The words do not *describe* the axes. Any words could have been picked instead of "Real" and "Imaginary", and to be honest, any other words would have been better.

Depending on what entity a Complex number is being used to portray, the Real and Imaginary components might both refer to things that exist in nature, or they might both refer to things that do not exist in nature. They might both refer to things that can partially exist in nature. One might refer to something that exists in nature, and the other one might not.

If we were to use Complex numbers to refer to, and process, points on a map, then it would be obvious that the Real and Imaginary axes both refer to genuine real-world existing entities. If we used Complex numbers to refer to a picture of a triangle, then it becomes a philosophical point as to whether the axes refer to a real-world existing entity or not.

It is common to see people give more importance to the Real axis than the Imaginary axis because they mistakenly believe that it is the axis that describes real-world entities.

More on Complex numbers

i and j

So far I have used the symbol "i" to represent the square root of −1. It is also common to see people use the symbol "j" to mean the same thing. Generally mathematicians use the letter "i", while electrical engineers tend to use "j", to avoid confusing the symbol "i" with the "i" intended to mean electrical current. Everyone else uses whichever they feel like using.

Unless you have a good reason for using one or the other of "i" and "j", it is just a matter of choice which one you use. Personally, I think "j" looks better than "i", but I have been using "i" here as it is more commonly used when people are introduced to Complex numbers. Therefore, in this book, I will continue to use the letter "i".

Order of numbers and i

Sometimes, you will see Imaginary numbers written with the "i" or "j" after the number, and sometimes with it before. As the "i" or "j" in a Complex number is being multiplied by the number, strictly speaking, the order does not matter. However, it is more usual to see "i" placed after a number, such as 2i, and "j" placed before a number. If the order will be clearer with the "i" or "j" placed the other way around, then it can be better to do that.

Examples are as follows:

3 + 4i
3 + j4
2 + i sin θ [This is clearer than writing “2 + (sin θ)i”]
2 + 4i sin θ
5 + i $\sqrt{2}$ [This is clearer than writing “$\sqrt{2}$ i”, which might be confused with the square root of “2i”].

Individual components of a Complex number

It is common to see people refer to, or isolate, the Real or Imaginary part of a Complex number by saying something similar to the following:
Re {4 + 8i} = 4
... or:
Im {4 + 8i} = 8.

What these mean is that the Real part of “4 + 8i” is 4, and the Imaginary part of “4 + 8i” is 8. Of course, there is no point in doing such a thing with a simple Complex number such as “4 + 8i”. However, later on, you might come across complicated formulas referring to points in the Complex plane, where it is not obviously clear what the Real and Imaginary components really are. In such a case, isolating the Real or Imaginary parts can be useful.

If the Complex number is given as a formula involving “z” such as “z = 4 + 8i”, then you might see:
Re {z} = 4.
Im {z} = 8.
... which mean exactly the same thing.

Another way of expressing the same thing is to use the symbols $\Re$ and $\Im$. These are the letters “R” and “I” in the Fraktur font, which is a type of Gothic font. Depending on how the font is displayed, these symbols can go from being difficult to read to being completely illegible. Unless you are good at calligraphy, they are very difficult to write by hand. The symbols are used in the same way as “R” and “I”:
$\Re$ {4 + 8i} = 4.
$\Im$ {4 + 8i} = 8.

Names for the Complex plane

The Complex plane is also sometimes called the "z plane", on account of points on the Complex plane being referred to with the symbol "z". There is also another type of plane called the "z plane", which is a different concept.

A drawing of the Complex plane is sometimes called an Argand diagram after Jean-Robert Argand, who was the first person to receive credit for the idea.

Unneeded ones and zeroes

Generally, with Complex numbers, unneeded ones and zeroes are ignored. For example, the number "0 + 3i" would more commonly be written as just "3i". The Complex number "5 + 0i" would normally be written as just "5". The number "3 + 1i" would be written as "3 + i".

Some people strongly believe that a Complex number such as "3 + 0i" *must* be written as "3" without the zero Imaginary part. I think this comes from a culture of school education where arbitrary rules are enforced and obeyed without thought as to their usefulness. In my opinion, if we are working in the realm of Complex numbers, it is sensible to keep the Imaginary part in the number, even if it is zero. By leaving it in, it informs us that we are dealing with Complex numbers – it gives us the setting for the maths that we are using. Knowing that a number was produced by Complex maths tells us more about the nature of the number and what we should expect from further calculations. There are often situations in which knowing that we are working in the world of Complex numbers changes the nature of how we should visualise what we are doing. This is especially true when it comes to Imaginary exponents of "e" and calculus with Complex numbers.

In this book, I will often leave the unneeded ones and zeroes in Complex numbers to clarify what I am trying to explain. Whether you want to do the same will depend on whether you are required to conform to the wishes of someone else. If there are no consequences for doing otherwise, you can do as you like.

Other terminology

Magnitude: The magnitude of a Complex number is the distance of the point it represents from the origin of the axes. In other words, it is the length of the hypotenuse of the right-angled triangle portrayed by the coordinates of that point. It can be calculated with Pythagoras's theorem.

Absolute value: If we were dealing with everyday Real numbers, the absolute value of a number is that number made positive, whether it is positive or negative to start with. For example, the absolute value of −4 is +4. The absolute value of +4 is also +4. The absolute value of an everyday Real number can also be thought of as how far away it is from zero in either direction. Both the numbers −4 and +4 are +4 units from zero. With Complex numbers, the meaning of absolute value is the same. The absolute value is how far the point represented by the Complex number is from zero, or to put it more succinctly, how far it is from the origin of the axes. In this way, the absolute value of a Complex number is identical to the magnitude of a Complex number. To calculate the absolute value, we use Pythagoras's theorem.

The circle chart and the Complex plane

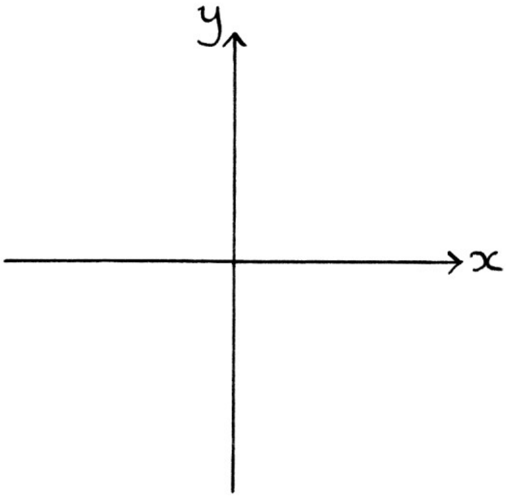
What I have been calling the circle chart in previous chapters is usually thought of as being the Complex plane. Circles that relate to waves are usually drawn and thought about as if they were in the Complex plane. Therefore, coordinates on their edges are given, not in terms of coordinates, but in terms of Complex numbers. Essentially, everything is the same as before, but instead of giving the coordinates of points on the circumference in the form of, say, (1, 2), they are given in the form of "1 + 2i". This is a very minor difference, but allows us to perform mathematical procedures slightly more easily.

When using Complex numbers, the helix chart is also altered to use axes of Real, Imaginary and time, instead of x, y and time. When we are thinking in terms of Complex numbers, I will call the helix chart, the "Complex helix chart". The only actual difference between the standard helix chart and the Complex helix chart will be the labelling of the axes and the fact that we are will identify points on it in terms of Complex numbers and times, instead of coordinates and times.

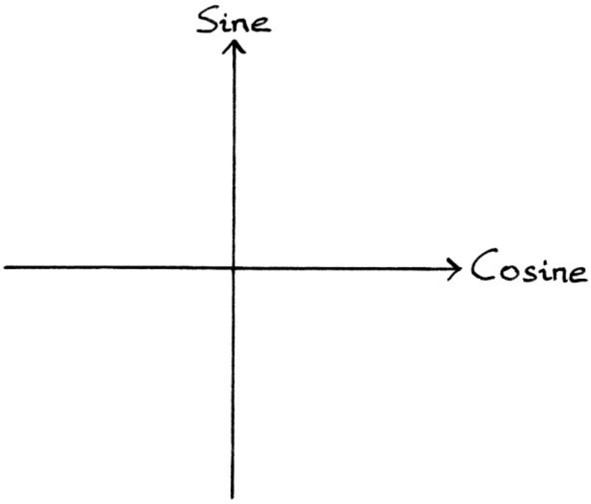
The variety of axes

Now that I have introduced Complex numbers, we can see that there are multiple ways of treating the axes on which we might draw a circle.

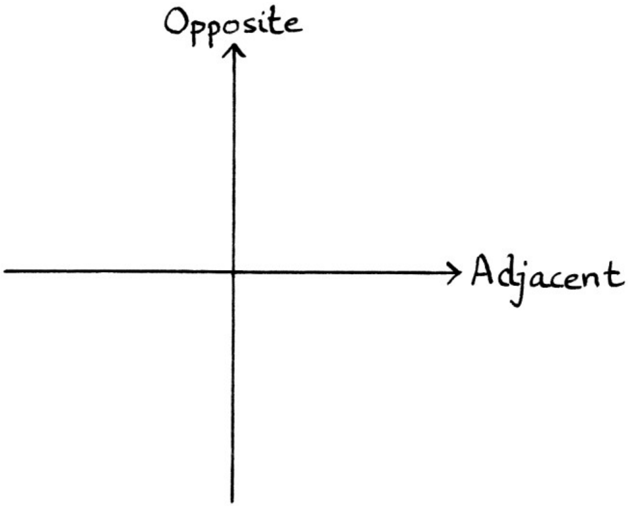
The simplest way is to label the axes with “x” and “y”. These indicate the horizontal and vertical distances from the origin of the axes, and are generic enough that they can be applied to any graph.



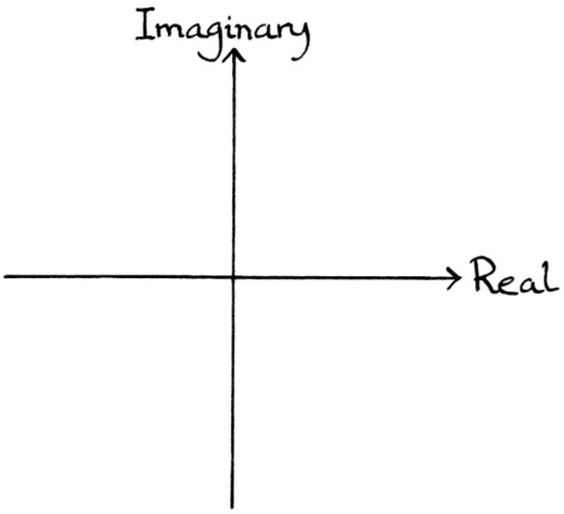
We could think of every possible point on the chart in terms of its angle and distance from the origin. In this way, a point’s horizontal distance from the origin will be a result of the Cosine function, and its vertical distance from the origin will be a result of the Sine function. We are really treating every point as being at the end of the hypotenuse of a right-angled triangle. If we do this, then we could label the axes the “Cosine axis” and the “Sine axis”:



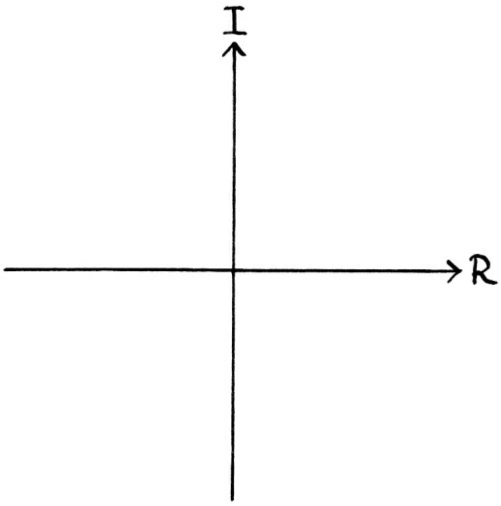
In practice, you are unlikely to see such labelling, but it is an example of another way of thinking about the axes. A similar idea is to treat every point on the graph in terms of the opposite and adjacent sides of right-angled triangles. We could then label the axes the “adjacent axis” and the “opposite axis”.



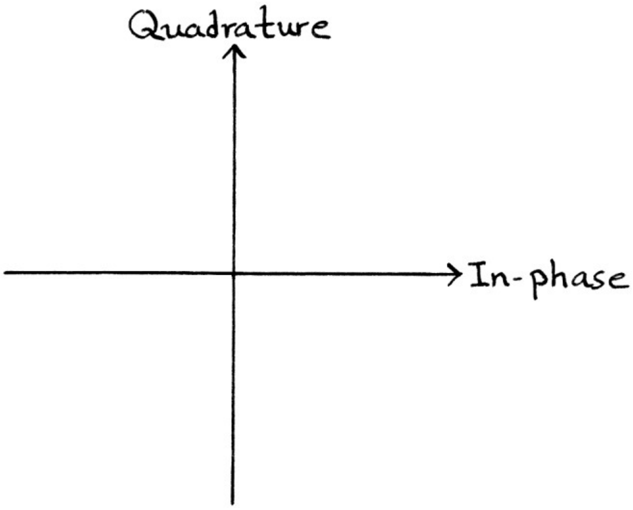
Again, you are unlikely to see axes labelled in this way. If we consider all the points on the chart as referenced by Complex numbers, which is the same as them being in the Complex plane, then we can label the axes the “Real axis” and the “Imaginary axis”. This is likely to be the most common labelling that you will see, after “x” and “y”.



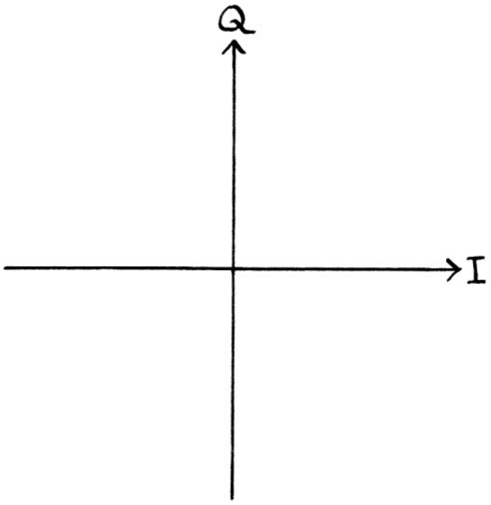
To save space, the words “Real” and “Imaginary” are often abbreviated to “R” and “I”:



In the field of signal processing, the Complex plane is often labelled with the x-axis as the “In-phase” axis, and the y-axis as the “Quadrature” axis. In Chapter 7 on phase, I discussed the idea of constellation diagrams where the phases of different waves can be compared. The names “In-phase” and “Quadrature” make most sense with constellation diagrams, as the In-phase axis can be used to signify the phase with which other phases are being compared. A wave with a phase ninety degrees higher than the reference wave would be “in quadrature” with the reference wave.



The labels "Quadrature" and "In-Phase" are often abbreviated to "I" and "Q":



Note how the labelling of "I" for In-Phase, which is the x-axis in this case, conflicts with the labelling of "I" for the Imaginary axis (the y-axis) when the graph is being used for Complex numbers. The context and the presence or absence of "Q" will indicate what is meant, and strictly speaking, if you know that you are dealing with the Complex plane, you will know what the axes are without needing to read the labels.

We can also have three-dimensional graphs with the same labelling and a third axis of time.

I have shown several different ways of labelling the axes. They all essentially amount to exactly the same thing, but their names relate to the different ways in which we think about the points. The axes do not alter the content drawn on the graphs in any way – a circle or shape drawn on any of these axes would still be the same. We could just refer to all graphs as having an x-axis and a y-axis, but different labels help distinguish the context in which we are treating the contents of a graph.

Conclusion

Complex numbers are useful because they provide a consistent way of performing maths on entities that exist in two dimensions.

If Complex numbers still seem confusing, it is sufficient for a reasonable understanding of them to ignore everything in this chapter, and just treat them as if they were an eccentric way of writing normal “x” and “y” coordinates on a grid. In fact, they *are* coordinates, but ones where the “x” and “y” coordinates appear as what is really a single number, and that number lends itself to mathematical operations. The more you experience Complex numbers, the more straightforward they will appear. It can be easy to overthink Complex numbers and think that there is more to them than there really is.

Chapter 24: Complex numbers and waves

Sine and Cosine and Complex numbers

We can think of the point represented by a Complex number as being at the end of the adjacent and opposite sides of a right-angled triangle within the Complex plane. Such an idea essentially uses Cartesian coordinates. For example, “2 + 3i” is a point at the end of the hypotenuse of a right-angled triangle with an adjacent side of 2 units and an opposite side of 3 units.

We can also think of that point as being described by the angle and hypotenuse of a right-angled triangle within the Complex plane. Such an idea uses Polar coordinates. For example, we could identify a point as being 2 units from the origin at an angle of 135 degrees. We could phrase this as 2∠135 degrees.

We can also think of the point as being at the end of the adjacent and opposite sides of a right-angled triangle, but with those sides given in terms of the Cosine and Sine of the angle. If we were using a normal set of axes, such as the circle chart, we could give coordinates as so: (cos 60, sin 60) in degrees, which is (cos 0.3333π, sin 0.3333π) in radians. These are Cartesian coordinates, but ones that use Cosine and Sine. On the Complex plane, we can do the same thing, but instead of using coordinates, we can indicate the position of the point using a Complex number:
“cos 60 + i sin 60” in degrees
... or:
“cos 0.3333π + i sin 0.3333π” in radians.

As another example, the point two-units away from the origin at an angle of 270 degrees can be identified as so:
“2 cos 270 + 2i sin 270” in degrees
... or:
“2 cos 1.5π + 2i sin 1.5π” in radians.

Any point on a set of axes can be treated as being at the end of the hypotenuse of a right-angled triangle, and any point can be treated as being on the circumference of a circle. The two ideas are ultimately the same thing.

Using Cosine and Sine in a Complex number is really a combination of a Complex number and a Polar coordinate.

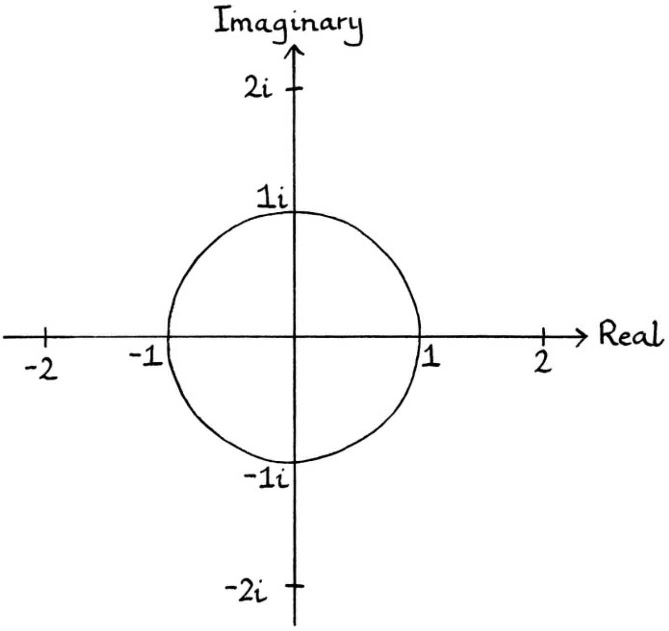
Supposing we had the Complex number:
"4.3301 – 2.5i"
... then its Polar coordinate would be:
"5∠120" in degrees
... or:
"5∠2.09440" in radians.

Its Complex number using Cosine and Sine would be:
"5 cos 120 + 5i sin 120" in degrees
... or:
"5 cos 2.09440 + 5i sin 2.09440" in radians.

Sine and Cosine and circles

The most interesting aspect of using Complex numbers with the Cosine and Sine of the angle occurs when we turn the Complex number into a general formula. By this, I mean when we have something such as the following:
"z = cos θ + i sin θ"
... where either "θ" is an angle in degrees, and Cosine and Sine are working in degrees, or "θ" is an angle in radians, and Cosine and Sine are working in radians.

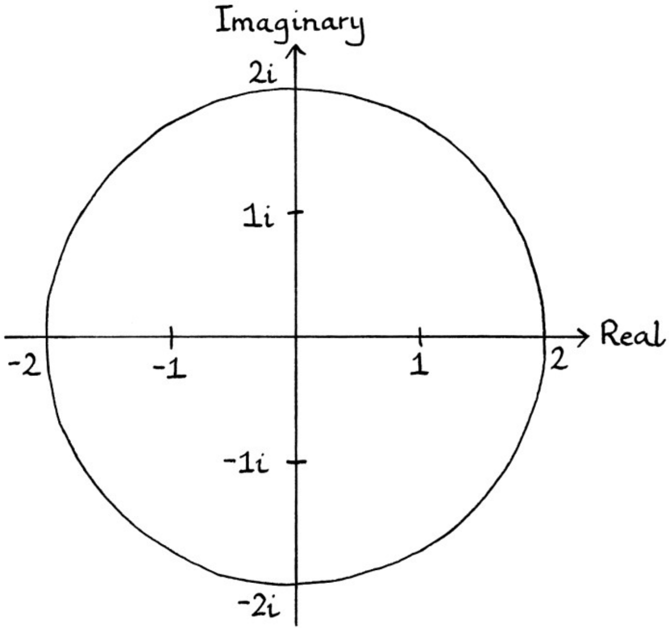
If we plot the points referenced by this Complex number formula for every value of "θ" from 0 degrees up to 360 degrees, or from 0 radians up to 2π radians, it results in a unit-radius circle.



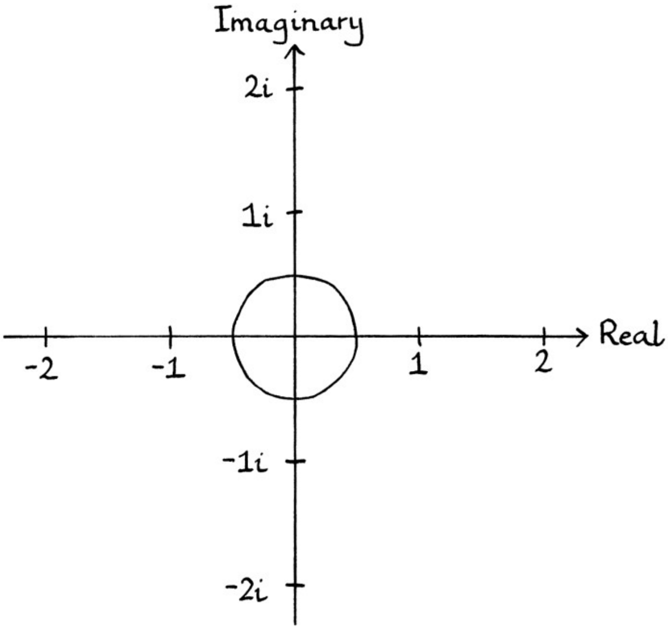
We could have achieved the same thing with coordinates on x and y-axes, in which case the coordinates would have been (cos θ, sin θ) for values of "θ" from 0 degrees up to 360 degrees, or for values of "θ" from 0 radians up to 2π radians, depending on the system in which Cosine and Sine are working.

This idea is really based on those in the very first chapter of this book. The Sine and Cosine functions give the y-axis and x-axis values of the edge of a unit-radius circle at particular angles. Therefore, by using the Sine and Cosine of every angle of a circle, we will draw a circle.

We can change the radius of the circle by scaling the formula. If we want a radius of 2 units, we just multiply the formula by 2:
z = 2 * (cos θ + i sin θ)
... which becomes:
z = 2 cos θ + (i * 2sin θ)
... which we can write most clearly as:
z = 2 cos θ + 2i sin θ

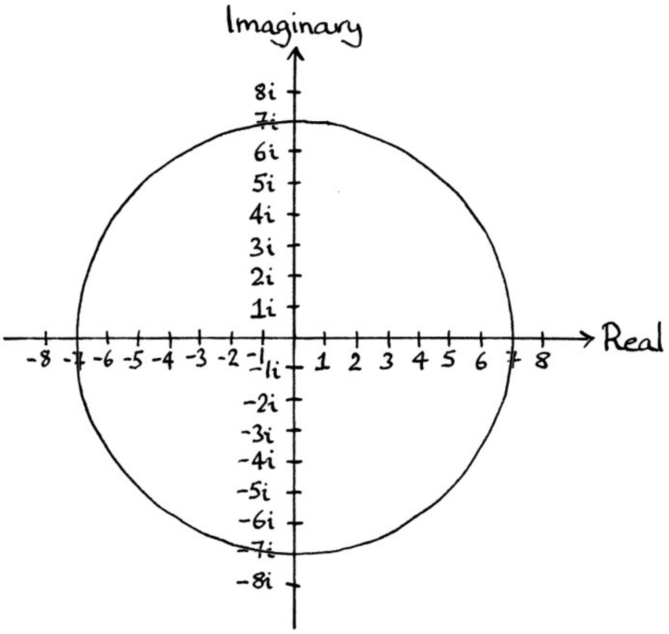


If we want a radius of 0.5 units, we multiply the formula by 0.5:
z = 0.5 * (cos θ + i sin θ)
... which is:
z = 0.5 cos θ + i * 0.5 sin θ
... which we can write as:
z = 0.5 cos θ + 0.5i sin θ



Despite the essence of this idea being explained in the very first chapter of this book, this idea's importance cannot be overstated. We have found a way of drawing our circle graph using maths, or to put it another way, we have a formula for our circles. We now have a mathematical way of portraying the circles from which our Sine and Cosine wave graphs are derived, and now we can also perform certain mathematical processes on the circles and the points on those circles.

If we have a circle of radius 7 units, then the Sine and Cosine waves derived from that circle will have an amplitude of 7 units. The circle, as portrayed with a Complex number formula, will be:
"z = 7 * (cos θ + i sin θ)"
... which can also be written as:
"z = 7 cos θ + 7i sin θ".



The Sine wave derived from this circle will have a formula of "y = 7 sin θ". The Cosine wave will have a formula of "y = 7 cos θ". In other words, the derived waves have identical formulas to the Real and Imaginary parts of the Complex number formula for the circle (except that we ignore the "i" for the Sine wave).

[Note that this is all true whether we are working in degrees or radians as long as Sine and Cosine are working in the same system as the type of angle.]

Previously in this book, when trying to recreate the circle from a Sine wave and a Cosine wave, we used the y-axis values of each wave as coordinates for every point on the circle's edge. Now, we are doing the same thing, but with the minor difference that now we are *defining* the circle as being made up from coordinates of points from each wave.

This is one of the most important aspects of Complex numbers. It is why Complex numbers are useful with waves. It is also the basis for using the number "e" in formulas, which is something I introduce in Chapter 27. One could say that this is the stepping stone to the next level of understanding waves and signals.

Sine and Cosine and time

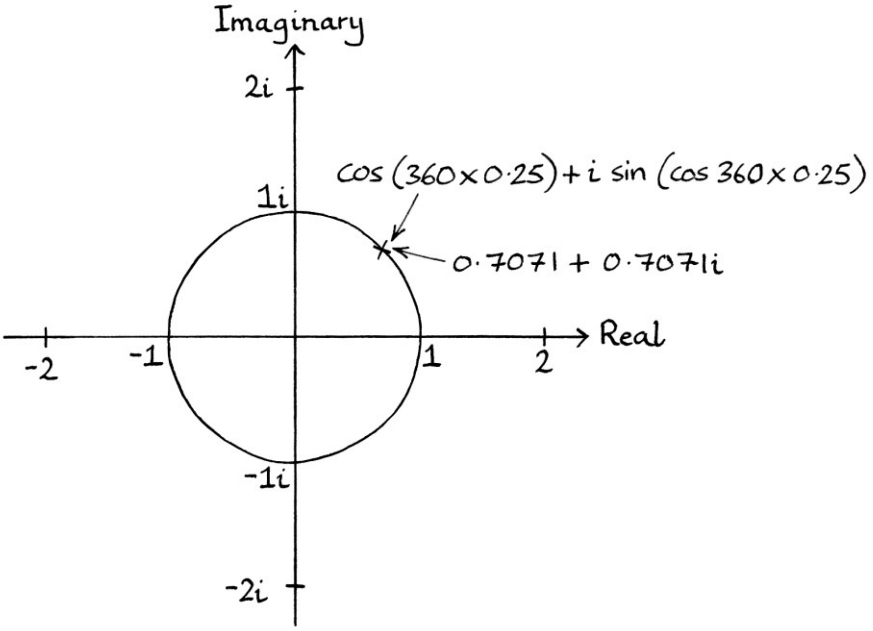
For increasing angles of "θ", the formula "cos θ + i sin θ" draws out a circle. We can plot the derived Sine wave graph, being the y-axis values for each angle, and we can plot the derived Cosine wave graph, being the x-axis values for each angle. We actually have the Sine of the angle and the Cosine of the angle in the formula, so calculating the wave graphs is very straightforward.

Given that we normally deal with Sine and Cosine waves changing over time, we can alter the "cos θ + i sin θ" circle formula to relate to time too. Doing this requires correcting the time in seconds, so that the Cosine and Sine functions work on the time itself. If we are working in degrees, our formula becomes:
"z = cos 360t + i sin 360t".

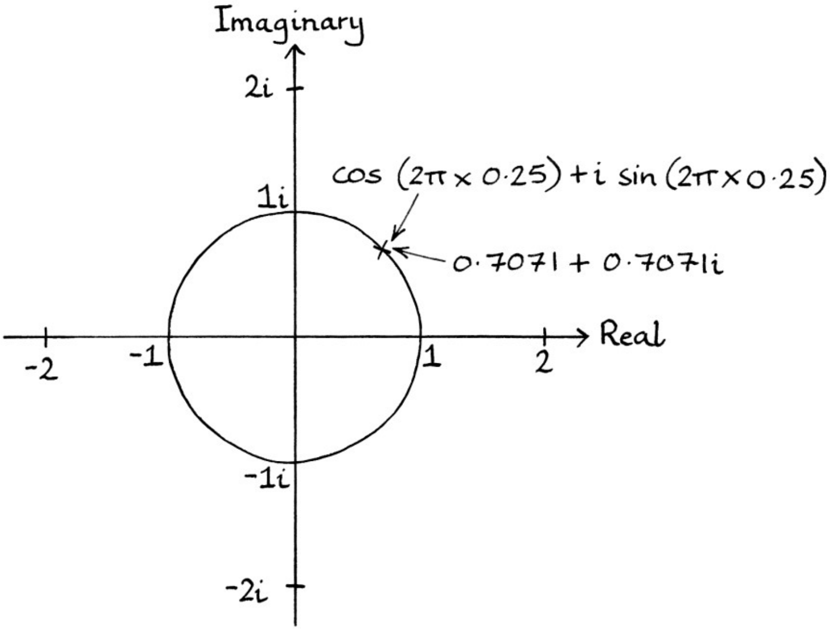
If we are working in radians, our formula becomes:
"z = cos 2πt + i sin 2πt".

The Sine and Cosine waves derived from the circle formula in *degrees* have the formulas "y = sin 360t" and "y = cos 360t". The Sine and Cosine waves derived from the circle formula in *radians* have the formulas "y = sin 2πt" and "y = cos 2πt". It does not matter whether we use degrees or radians, the time-based wave graphs will look exactly the same. We saw this in Chapter 22 on radians.

If we are dealing in degrees, for an object rotating around the circle described by "z = cos 360t + i sin 360t", at 0.25 seconds, the object will be at the point described by: "cos (360 * 0.25) + i sin (360 * 0.25)". This is the same as "0.7071 + 0.7071i".



If we are dealing in radians, for an object rotating around the circle described by "z = cos 2πt + i sin 2πt", at 0.25 seconds, the object will be at the point described by: "cos (2π * 0.25) + i sin (2π * 0.25). This is also "0.7071 + 0.7071i".



These are the same point, which should be expected as the object is in the same place, no matter how we measure its angle at that time. The Complex number that identifies its position is "0.7071 + 0.7071i", and the coordinates of the point are: (0.7071, 0.7071).

Sine and Cosine and the wave attributes

Given that the Complex number formula for the circle is just the circle's derived waves used as coordinates, it might be apparent that we can adjust the formula to take into account every aspect that can be portrayed on a wave. In other words, we can portray the amplitude, frequency, phase and the two mean levels in the circle formula.

Amplitude

Changing the amplitude involves scaling the Complex number. Therefore, if we want a radius of 2 units, which equates to an amplitude of 2 units, we just double the Complex number.

For degrees, this works out as so:
2 * (cos 360t + i sin 360t) = 2 cos 360t + 2i sin 360t.

For radians, this works out as so:
2 * (cos 2πt + i sin 2πt) = 2 cos 2πt + 2i sin 2πt.

Each wave has double the amplitude, and therefore the circle has double the radius.

Note that both waves in the Complex number *must* have the same amplitude or else the Complex number will not refer to a circle. The amplitude of a Sine or a Cosine wave is the same as the radius of the circle from which the waves are derived, and as a circle only has one radius, the waves must have matching amplitudes.

Frequency

Changing the frequency, as always, just involves scaling the time being treated as an angle in the formula. Therefore, if we want a frequency of 10 cycles per second, we adjust the frequency for the Real wave and the Imaginary wave accordingly.

For degrees, this works out as so:
cos (360 * 10t) + i sin (360 * 10t)

For radians, this works out as so:
cos (2π * 10t) + i sin (2π * 10t)

Note that both waves *must* have the same frequency. It makes no sense for the frequencies to be different. The frequency of a single object rotating around a circle dictates what the frequency of the Sine and Cosine waves derived from that circle will be. As they both relate to the frequency of the same object, they cannot be different.

Phase

Changing the phase, as always, just involves adding an angle on to the value being Sined or Cosined, so that the object rotating around the circle starts at an angle other than zero degrees (zero radians).

If we want a phase of 90 degrees (0.5π radians), we just add 90 degrees (0.5π radians) to the value being Sined or Cosined.

For degrees, this works out as so:
cos (360t + 90) + i sin (360t + 90)

For radians, this works out as so:
cos (2πt + 0.5π) + i sin (2πt + 0.5π)

Note that the phase of both waves *must* be identical. It makes no sense for the phases to be different. The phase of a Sine or Cosine wave is identical to the starting point of a single object about to rotate around a circle. As they both refer to the same point, they cannot be different.

Mean levels

Changing the mean level is slightly more complicated than changing the other characteristics. Normally, if we were changing the mean level for a wave, we would just add the mean level to the result of the Sine or Cosine function. We can still do this for the Cosine wave – in which case, the circle becomes moved left or right along the Real axis. However, for the Sine wave, the mean level needs to be an Imaginary number – the Sine wave's mean level means that the circle moves up or down the Imaginary axis, so it must be an Imaginary number. Therefore, if we wanted a mean level of 7 units for the Cosine wave and a mean level of 10 for the Sine wave, we would add 7 to the result of the Cosine wave function, and 10i to the result of the Sine wave function.

For degrees, this looks like this:
(7 + cos 360t) + (10i + i sin 360t).
... which we can write as:
7 + 10i + cos 360t + i sin 360t.

For radians, this looks like this:
(7 + cos 2πt) + (10i + i sin 2πt).
... which we can rewrite as:
7 + 10i + cos 2πt + i sin 2πt.

If you have read this book from the start, then it should be obvious that, unlike with amplitude, frequency and phase, the mean levels for each wave *can* be different. The mean level for Sine relates to the y-axis (Imaginary axis) position of the centre of the circle. The mean level for Cosine relates to the x-axis (Real axis) position of the centre of the circle. The two mean levels are completely independent of each other.

Note that usually, you will not see formulas with an Imaginary mean level. This is because most work on waves is done with radio or sound waves, and they are usually, arbitrarily, treated as Cosine waves. Therefore, in such cases, there will only be a mean level for the Cosine wave part of a circle.

Amplitude, frequency, phase and mean levels

If we wanted, we could alter the amplitude, frequency, phase and both mean levels for the Complex number circle formula. For example, if we wanted an amplitude of 3 units, a frequency of 12 cycles per second, a phase of 270 degrees (1.5π radians), a Sine wave mean level of 5, and a Cosine wave mean level of 6, we would have the following formulas:

For degrees:
6 + 3 cos ((360 * 12t) + 270) + 5i + 3i sin ((360 * 12t) + 270)
... which is:
6 + 5i + 3 cos ((360 * 12t) + 270) + 3i sin ((360 * 12t) + 270)

For radians:
6 + 3 cos ((2π * 12t) + 1.5π) + 5i + 3i sin ((2π * 12t) + 1.5π)
... which is:
6 + 5i + 3 cos ((2π * 12t) + 1.5π) + 5i + 3i sin ((2π * 12t) + 1.5π)

The full formula

The full formula for a circle in the Complex plane that is being described with Sine and Cosine waves is as follows:

In degrees:
$z = h_c + ih_s + A \cos (360ft + \phi) + i A \sin (360ft + \phi)$

In radians:
$z = h_c + ih_s + A \cos (2\pi ft + \phi) + i A \sin (2\pi ft + \phi)$

... where:

- “z” is the range of places on the Complex plane specified by the formula.
- “h_c” is the mean level for the Cosine wave, which is the Real-axis value of the centre of the circle.
- “i” is the square root of −1.
- “h_s” is the mean level for the Sine wave, which is the Imaginary-axis value of the centre of the circle.
- “A” is the amplitude. It is the same for both the Cosine and Sine waves.
- “f” is the frequency. Similarly, it is the same for both waves.
- “t” is the time in seconds.
- “ϕ” is the phase in degrees or radians, depending on the angle system in which Sine and Cosine are working. The value will be the same for both waves.

Negative frequencies

We can account for negative frequencies using waves and Complex numbers. In other words, we can portray an object rotating clockwise around a circle.

For example, an object rotating clockwise around a unit-radius circle at 5 cycles per second would have this formula in degrees:
“$z = \cos (360 * -5t) + i \sin (360 * -5t)$”
... and it would have this formula in radians:
“$z = \cos (2\pi * -5t) + i \sin (2\pi * -5t)$”

At, say, 0.01 seconds, the object would be at the coordinates: (0.9511, −0.3090), which means it would be at “0.9511 – 0.3090i” on the Complex plane. This is in the lower right quarter of the axes, which shows that the object is rotating anticlockwise.

Rephrasing the formula

Technically, because a negative-frequency Cosine wave *with zero phase* is the same as a positive-frequency Cosine wave *with zero phase*, the formula could be rewritten as:
“z = cos (360 * 5t) + i sin (360 * −5t)”.

Although this would still produce the same circle with the object moving clockwise, I would say that this is not a good way to write the formula. The formula will still work, but it goes against the rules for describing a circle. In my opinion, the frequencies of the Sine wave and the Cosine wave should always be the same as each other because they derive from the same circle. The result of this formula will still be the same as the original formula, but I think it is better not to refer to the movement of the object in this way. This method of phrasing the formula makes it harder to see what is happening. It is not clear that the object is rotating clockwise around the circle.

Again, technically, because a negative-frequency Sine wave *with zero phase* is the same as a positive-frequency Sine wave with a phase of 180 degrees, the formula could be rephrased to be:
“z = cos (360 * 5t) + i sin ((360 * 5t) + 180)”

Again, although this would produce identical results to the original formula, I would say it is a bad way to write the formula. The phases should be the same as each other because they refer to the starting point of an object about to rotate around a circle. Although the formula will work, it is much harder to see that it refers to an object rotating clockwise around a circle, and it is much harder to know the phase point of the object from the formula. In my opinion, it is better not to write negative-frequency formulas in this way.

[To show that this formula still works, at 0.01 seconds, the object would still be at the place on the Complex plane described by “0.9511 – 0.3090i”. The object is still rotating clockwise, but it is harder to tell this from the adjusted formula.]

Rephrasing with a negative amplitude

Given that a Sine wave with a positive amplitude, a negative frequency, and *zero phase* is the same as that Sine wave with the amplitude made negative, its frequency made positive and *zero phase*, in the example above, we could rephrase the equation to be:
“z = cos (360 * 5t) − i sin (360 * 5t)”

You will frequently see negative frequencies portrayed in this way. However, in my opinion, this way is likely to cause mistakes because we can only rephrase a formula in this way if the Sine wave has zero phase. If the phase is not zero, then we have to pay attention to the phases – we cannot just swap the signs of the amplitude and the frequency for the Sine wave part.

As an example, we will look at the following negative-frequency formula, where Cosine and Sine are working in degrees:
“z = cos ((360 * −5t) + 45) + i sin ((360 * −5t) + 45)”

We will turn it into a formula as an addition with positive frequencies, and then into a formula with a subtraction of positive frequencies.

The first thing to realise is that we *cannot* rephrase the first part as:
“cos ((360 * 5t) + 45)”
This is because a negative-frequency Cosine wave with a non-zero phase is not the same as that Cosine wave with the frequency made positive. We have to adjust the phase. For Cosine waves, we imagine the phase point on the circle, and then flip the circle up or down. Therefore, the phase for the Cosine wave made positive will be −45 degrees, which is +315 degrees. The Cosine part should, therefore, be:
“cos ((360 * 5t) + 315)”

For the second part to be rephrased as a Sine wave with a positive frequency and a *positive* amplitude, we have to adjust the phase for that too. For Sine waves, we imagine the phase point on the circle, and then mirror the circle left or right. This is also the same as seeing how many degrees *above or below* 90 degrees the phase is, and then choosing the angle that is that same number of degrees *below or above* 90 degrees. The phase of 45 degrees is 45 degrees below 90, so the phase we will want will be 45 degrees above 90 degrees, which is 135 degrees. The Sine wave part of the formula becomes:
“i sin ((360 * 5t) + 135)”

The whole formula becomes:
“z = cos ((360 * 5t) + 315) + i sin ((360 * 5t) + 135)”

The first thing to note is that although this refers to the same movement of an object around a circle, it is much harder to understand what is happening from this formula. For one thing, we cannot tell what the phase point is without some thought. We also cannot immediately tell that the object will be rotating clockwise around the circle.

It is easiest to leave the formula as an addition, but supposing we wanted to make it into a subtraction instead, we would proceed as follows. To turn a positive-amplitude Sine wave into a negative-amplitude Sine wave, we change its sign and add 180 degrees to the phase. Therefore, the Sine wave’s phase becomes 135 + 180 = 315 degrees. The Sine part becomes:
“−i sin ((360 * 5t) + 315)”
[Note that the “i” is really acting as the amplitude of the wave.]

The whole formula phrased as a subtraction is:
“z = cos ((360 * 5t) + 315) − i sin ((360 * 5t) + 315)”

When presented with the original formula:
“z = cos ((360 * −5t) + 45) + i sin ((360 * −5t) + 45)”
... if we had performed the maths incorrectly, and just made the frequencies positive, and negated the amplitude of the Sine part, we would have ended up with an incorrect result:
“z = cos ((360 * 5t) + 45) − i sin ((360 * 5t) + 45)”

This refers to a different shape altogether because the phases are both wrong.

We can check that the:
“z = cos ((360 * 5t) + 315) − i sin ((360 * 5t) + 315)”
... formula is the same as the original formula:
“z = cos ((360 * −5t) + 45) + i sin ((360 * −5t) + 45)”

At 0.01 seconds, the object represented by either formula will be at the place on the Complex plane described by “0.8910 + 0.4540i”. [The object starts at 45 degrees (0.7071 + 0.7070i), so “0.8910 + 0.4540i” is further clockwise around the circle].

As before, although the “$z = \cos((360 * 5t) + 315) - i \sin((360 * 5t) + 315)$” formula is mathematically the same as the original formula, we cannot easily tell what the phase point is or that the object is rotating clockwise. The object actually starts at 45 degrees, but this is not shown in the formula.

Personally, I think it is better not to rearrange a formula describing an object’s movement around a circle unless it is for a good reason. [Other people might disagree with me.]

Rephrasing rule

There is a quick way to convert a negative-frequency Complex formula to a positive-frequency formula with a negative-amplitude Sine wave. We just negate the phases of both the Cosine and Sine waves. This saves the effort of having to think about flipping circles and converting amplitudes.

If we start with this formula:
$z = \cos((360 * -ft) + \phi) + i \sin((360 * -ft) + \phi)$
... then it has the same meaning as this formula:
$z = \cos((360 * +ft) - \phi) - i \sin((360 * +ft) - \phi)$

For radians, the rule is essentially the same. This formula:
$z = \cos((2\pi * -ft) + \phi) + i \sin((2\pi * -ft) + \phi)$
... has the same meaning as this formula:
$z = \cos((2\pi * +ft) - \phi) - i \sin((2\pi * +ft) - \phi)$

As a summary of the rule: the frequencies are made positive, the amplitude of the Sine wave is made negative, and the phases of each wave are negated. If we want positive phases in the result, then we have another step in which we add 360 degrees (or 2π radians) to the phases.

As an example of the rule working in practice, if we have this formula:
“z = cos ((360 * −2t) + 27.456) + i sin ((360 * −2t) + 27.456)”
... then it is the same as this formula:
“z = cos ((360 * +2t) − 27.456) − i sin ((360 * +2t) − 27.456)”
... which would normally be written without the unnecessary plus signs as:
“z = cos ((360 * 2t) − 27.456) − i sin ((360 * 2t) − 27.456)”
... and, if we want positive phases, we add 360 degrees to each phase, and the formula becomes:
“z = cos ((360 * 2t) + 332.544) − i sin ((360 * 2t) + 332.544)”

We can go through each step the long way to see how this result is correct. We start with:
“z = cos ((360 * −2t) + 27.456) + i sin ((360 * −2t) + 27.456)”
... then change the Cosine wave to a positive-frequency one by imagining its circle and flipping it upwards or downwards:
“z = cos ((360 * +2t) − 27.456) + i sin ((360 * −2t) + 27.456)”

Next, we turn the Sine wave into a positive-frequency one by imagining its circle and mirroring it left or right (or by seeing how many degrees above or below 90 degrees it is, and finding the angle the same distance below or above 90 degrees:
“z = cos ((360 * +2t) − 27.456) + i sin ((360 * +2t) + 152.544)”

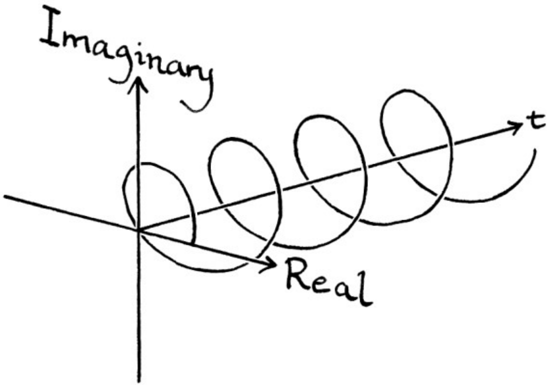
Next, we turn the Sine wave into a negative-amplitude one by adding 180 degrees to, or subtracting 180 degrees from, its phase:
“z = cos ((360 * +2t) − 27.456) − i sin ((360 * +2t) + 332.544)”

Then, we give the Cosine wave a positive phase to match that of the Sine wave (by adding 360 degrees), and remove any unnecessary plus signs:
“z = cos ((360 * 2t) + 332.544) − i sin ((360 * 2t) + 332.544)”

This is the same result as before, which confirms that the rule works.

The time helix

As always when dealing with circles, the amplitude, phase, and mean levels can be portrayed on a drawing of the circle, but the frequency cannot. However, we can portray all the attributes if we draw everything as a helix. In this way, it is really a Complex plane time helix on what I will call the "Complex helix chart". While the object rotates around the circle at so many cycles per second, it also moves outwards down the time axis. As I have explained before, it is best to use a helix chart just as a way of visualising concepts. A helix on the Complex helix chart appears the same as if it were drawn on the normal helix chart, except the x and y-axes are now the Real and Imaginary axes.



Amplitude and phase as a Complex number

In Chapter 12 on recreating the circle from which a Sine wave or Cosine wave were derived, I explained that one pair of coordinates was enough to indicate both the radius of the circle (the amplitude of its waves), and the phase of the circle (the phase of its waves) – but *only* if the circle is centred on the origin of the axes. It cannot indicate the frequency. To say this another way, if the circle from which a wave is derived is centred on the origin of the axis, then we only need to know the coordinates of its phase point to know its radius and phase, which will be the amplitudes and phases of the two derived waves.

Now that we are considering the circle on the Complex plane, in the sense that we are now using Complex numbers to identify points around the chart, we can give the coordinates of such a point with a Complex number. We are not doing anything that we could not do before, but using Complex numbers for the coordinates makes the idea fit in with the general mathematical methods being used here.

As an example, we can use the Complex number “12 + 6i” to indicate the phase point of a circle. This is enough information to calculate the radius of the circle and the amplitude of the derived waves. We can use Pythagoras’s theorem to find out that they are all 13.4164 units. We can also know that the angle of the phase point of the circle and the phases of the two derived waves will be: arctan (6 ÷ 12) = 26.5651 degrees (0.4636 radians). [As always, we should check the result of arctan to see if it is the one we want.]

Note that the Complex number does not tell us the frequency of the object rotating around the circle or its waves.

Adding circles with Complex numbers

If we identify circles using Complex number formulas such as:
“z = cos 360t + i sin 360t”
... then we can perform addition with circles.

For example, if we wanted to add the circles (in degrees):
“z = 2 cos (360 * 3t) + 2i sin (360 * 3t)”
... and:
“z = 3 cos (360 * 3t) + 3i sin (360 * 3t)”
... we would end up with:
“z = 2 cos (360 * 3t) + 2i sin (360 * 3t) + 3 cos (360 * 3t) + 3i sin (360 * 3t)”
... which is:
“z = 5 cos (360 * 3t) + 5i sin (360 * 3t)”

Describing circles with Complex numbers does not necessarily make the addition any easier than if we were describing circles with just their two derived waves. The methods are the same as before, and we still cannot reduce sums of circles with different frequencies to anything else.

Multiplying circles

One important drawback with treating circles as Complex numbers in the way described in this chapter is that multiplication does not work in the expected way. By this, I mean that the result of a multiplication of circles as Complex numbers ceases to represent the multiplication of the underlying waves. We can explore this idea by imagining a circle made up of the waves:
"y = 1.5 sin (360 * 2t)" and "y = 1.5 cos (360 * 2t)"

This circle can be represented by the Complex number:
"z = 1.5 cos (360 * 2t) + 1.5i sin (360 * 2t)"

We will multiply the circle by itself. There are two ways we can do this:

- We can square each of the derived waves, and use the results as the coordinates of a new shape.
- We can square the Complex form of the circle.

These will produce different results as we will see.

Squaring the derived waves

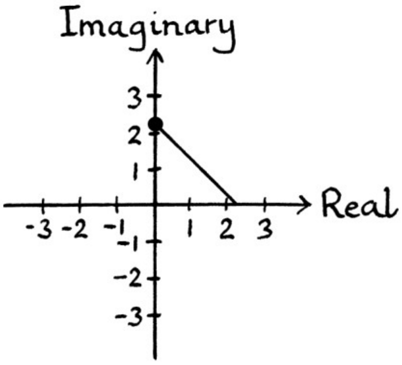
If we square each derived wave, we will end up with a shape with x-axis coordinates indicated by:
$(1.5 \cos (360 * 2t))^2$
... and y-axis coordinates indicated by:
$(1.5 \sin (360 * 2t))^2$

The square of the Cosine wave ends up as:
"1.125 + 1.125 cos (360 * 4t)"

The square of the Sine wave ends up as:
"1.125 + 1.125 sin ((360 * 4t) + 270)"

The coordinates of an object rotating around a circle at any moment in time will be:
(1.125 + 1.125 cos (360 * 4t), 1.125 + 1.125 sin ((360 * 4t) + 270))

The resulting shape looks like this:



From a basic mathematical point of view, if the coordinates of the object rotating around the original circle at any moment in time were:
(x, y)
... then the coordinates of the object moving along the resulting shape at that time would be:
(x^2, y^2)

Squaring the Complex number

The Complex number for the circle is:
"z = 1.5 cos (360 * 2t) + 1.5i sin (360 * 2t)"

To square this, we will proceed as if we were squaring any sum of the form (a + b). The result of such a squaring will be:
(a + b) * (a + b)
= aa + ab + ba + bb
= aa + ab + ab + bb
= $a^2 + 2ab + b^2$

Therefore, our result will be:
$(1.5 \cos (360 * 2t))^2$
+
2 * 1.5 cos (360 * 2t) * 1.5i sin (360 * 2t)
+
$(1.5i \sin (360 * 2t))^2$

The Cosine part in the first line ends up as: 1.125 + 1.125 cos (360 * 4t)

As the Sine wave has an Imaginary amplitude, to solve the next two lines, it is helpful to remember the rules for multiplying Sine waves from Chapter 16:

If we multiply:
"$y = h_1 + a_1 \sin ((360 * f_1 * t) + \phi_1)$"
... by:
"$y = h_2 + a_2 \sin ((360 * f_2 * t) + \phi_2)$"
... where f_1 is faster than, the negative of, or equal to f_2, then the equivalent sum of waves and a mean level will be:
"$y = 0.5 * (a_1 * a_2) * \sin ((360 * (f_1 – f_2)) + (\phi_1 – \phi_2 + 90)$"
"$y = 0.5 * (a_1 * a_2) * \sin ((360 * (f_1 + f_2)) + (\phi_1 + \phi_2 + 270)$"
"$y = (h_2 * a_1) \sin ((360 * f_1 * t) + \phi_1)$"
"$y = (h_1 * a_2) \sin ((360 * f_2 * t) + \phi_2)$"
"$h_1 * h_2$"

For the second line of our squaring calculation, we have:
2 * 1.5 cos (360 * 2t) * 1.5i sin (360 * 2t)
... which if we make both waves into Sine waves is:
2 * 1.5 sin ((360 * 2t) + 90) * 1.5i sin (360 * 2t)
... which is:
3 sin ((360 * 2t) + 90) * 1.5i sin (360 * 2t)

This becomes:
(0.5 * 3 * 1.5i) sin ((360 * 0t) + 180)
+
(0.5 * 3 * 1.5i) sin (360 * 4t)

... which is:

2.25i sin (180)
+
2.25i sin (360 * 4t)
... which is:
0 + 2.25i sin (360 * 4t)
... which is:
2.25i sin (360 * 4t)

For the third line, we end up with:

0.5 * $(1.5i)^2$ * sin ((360 * 0t) + 90)
+
0.5 * $(1.5i)^2$ * sin ((360 * 4t) + 270)

... which is:

0.5 * −2.25 * sin (90)
+
0.5 * −2.25 * sin ((360 * 4t) + 270)

... which is:

−1.125 * sin (90)
+
−1.125 * sin ((360 * 4t) + 270)

... which is:

−1.125 – 1.125 sin ((360 * 4t) + 270)

All three lines of the squaring become:

1.125 + 1.125 cos (360 * 4t)
+
2.25i sin (360 * 4t)
+
−1.125 – 1.125 sin ((360 * 4t) + 270)

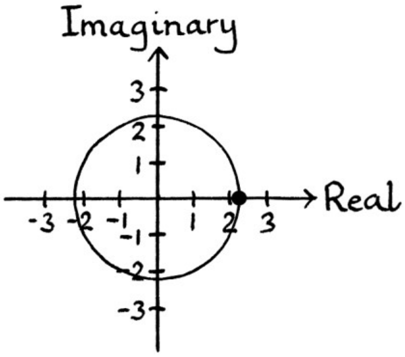
As the Cosine part and the non-Imaginary Sine part have the same frequency, we can add them together. We will rephrase them as Cosine waves, and calculate this:
1.125 + 1.125 cos (360 * 4t) + −1.125 – 1.125 cos ((360 * 4t) + 180)
= 0 + 1.125 cos (360 * 4t) – 1.125 cos ((360 * 4t) + 180)
= 1.125 cos (360 * 4t) – 1.125 cos ((360 * 4t) + 180)

As the second wave is an upside down version of the first, this can be rephrased as:
1.125 cos (360 * 4t) + 1.125 cos (360 * 4t)
= 2.25 cos (360 * 4t)

After all that, we can say that the result of squaring our circle *in its Complex form* is:

2.25 cos (360 * 4t) + 2.25i sin (360 * 4t)

The resulting shape looks like this:



The differences

When we multiplied the circle by itself by squaring the derived Cosine and Sine wave, we ended up with every x-axis coordinate of the resulting shape as:
1.125 + 1.125 cos (360 * 4t)
... and every y-axis coordinate as:
1.125 + 1.125 sin ((360 * 4t) + 270)

These are two pure waves with the same amplitude and non-zero mean levels, but with different phases. The shape described by these coordinates is a straight line at a 135-degree angle.

When we multiplied the circle by itself by squaring the Complex number for the circle, we ended up with this:
2.25 cos (360 * 4t) + 2.25i sin (360 * 4t)

This shape is a perfect circle.

As we can see, we achieve different results depending on whether we treat the circle as consisting of its derived waves, or as consisting as a Complex number.

We can rephrase the "derived wave" result as a Complex number to emphasise the difference, as so:

1.125 + 1.125i + 1.125 cos (360 * 4t) + 1.125i sin ((360 * 4t) + 270)

[We have a Real and an Imaginary mean level. The Imaginary Sine wave part has a 270 degree phase, while the Real Cosine part has zero phase.]

When we perform a multiplication, the connection between the "Complex number" form of a circle and the "derived wave" form of a circle breaks down. The Complex number form cannot be used for multiplication, unless it is just a scaling by a number. [For example, we could multiply a "Complex number" circle by 2, and it would scale the circle's radius by 2, and the derived waves would also have their amplitudes scaled by 2.] In this example, we multiplied a circle by itself, but the connection breaks with any multiplication of circles.

That the connection breaks in this way is important because it is not particularly intuitive, and it can lead to confusion when we have powers of "i" or Imaginary powers of "e" in later chapters.

A closer look

We can confirm that the connection breaks by checking where an object moving around a shape will be at a particular time. We will use the circle and shape from before, and we will see where the object is at t = 0.0123 seconds. At this time, the object will not have moved far from its starting position.

On the original circle, *as defined by its derived waves*, at 0.0123 seconds, the object will be at the coordinates:
(1.5 cos (360 * 2 * 0.0123), 1.5 sin (360 * 2 * 0.0123))
... which is:
(1.4821, 0.2309)

On the original circle, *as defined by the Complex number*, at 0.0123 seconds, the object will be at the Complex number:
1.5 cos (360 * 2 * 0.0123) + 1.5i sin (360 * 2 * 0.0123)
... which is:
1.4821 + 0.2309i

[The two results match, as they should do, given the connection between the "derived wave" form and "Complex number" form of the circle.]

On the squared circle, *as defined by its derived waves*, at 0.0123 seconds, the object will be at the coordinates:
(1.125 + 1.125 cos (360 * 4 * 0.0123), 1.125 + 1.125 sin ((360 * 4 * 0.0123) + 270))
... which is:
(2.5717, 0.05333)
[Each coordinate is the square of the coordinate from the "derived wave" form of the circle.]

On the squared circle, *as defined by the Complex number*, at 0.0123 seconds, the object will be at the Complex number:
2.25 cos (360 * 4 * 0.0123) + 2.25i sin (360 * 4 * 0.0123)
... which is:
2.1433 + 0.6845i

This is a completely different place [as expected, given how the formulas for the squared circles are different].

Reason

The connection between the "derived wave" form and the "Complex number" form breaks when we perform multiplication because the "Complex number" form involves multiplication on both coordinates as one combined number. We are performing calculations on a different type of number. The underlying maths is different.

Potential sources of confusion

What the circle is

No matter how we treat it, the circle from which waves are derived is a circle. If we consider a circle drawn on the circle chart or any set of axes, then we can describe it using normal coordinates. If we consider a circle drawn on the Complex plane, then that circle can be described using Complex numbers. However it is described, it is still a circle, and it is still a circle drawn on a graph. Without understanding the relationship between Sine and Cosine and the circle, some people assign the circle drawn on a Complex plane a higher level of significance than a circle drawn anywhere else. The circle on the Complex plane is the same as any other circle.

Forgetting i

It is easy to be forgetful and think of the circle as being "z = cos θ + sin θ", or "z = A cos (360ft + ϕ) + A sin (360ft + ϕ)". In other words, it is easy to forget the Imaginary aspect. On the Complex plane, a circle is not made up of the sum of a Cosine wave and a Sine wave unless those waves are treated as Complex numbers. If the waves are not provided as Complex numbers, we would be just adding two waves of the same amplitude and frequency but with a different phase, which just creates another pure wave. This is not what we want, so it is important to remember that the waves are being treated as Complex numbers.

Conversely, once you become used to seeing circles described with Complex numbers, it is very easy to misread an actual sum of pure waves as a Complex number.

Helices

Some people would consider the Sine and Cosine waves in a Complex number as only portraying a helix. In other words, they would not think of it as portraying a circle. That thinking is wrong though, because whether it is a helix or a circle depends on how many axes we want to consider. If we draw a third axis of time, it is a helix; if we only have the Real and Imaginary axes, it is a circle. Sometimes, it can be useful to think of it as a circle; sometimes, it can be useful to think of it as a helix.

Reality

It is common in signal processing for received radio or sound signals to be treated as if they were Cosine waves or the sum of Cosine waves, as opposed to being Sine waves or the sum of Sine waves. One common but incorrect justification for using Cosine waves is that, because they represent the "Real" aspect of the circle, they are in some way "real" themselves. In other words, it is thought that Cosine waves are representative of the real world, while Sine waves are not. However, this is not correct. Both the Sine and Cosine waves derived from a circle are equally valid – the dimension on which they are portrayed is irrelevant to their place in reality. Cosine waves are probably preferred over Sine waves due to a baseless tradition.

Complex number or Polar coordinate

The number "cos 12 + i sin 11" is a Complex number. It has a Real part and an Imaginary part. How it differs to the most basic examples of Complex numbers is that the number is based on the Sine and Cosine of the angle of the identified point, instead of being the actual "x" and "y" values. However, the Sine and Cosine of the angle still produce the coordinates of the identified point. Similarly, "cos 360t + i sin 360t" is a Complex number, or at least, a Complex number formula.

Some people would say that a Complex number based on Cosine and Sine is really a Polar coordinate because it mentions angles. I would say that this is wrong because a Polar coordinate is a way of identifying a point by mentioning its distance and angle from the origin of the axes, with no mention of Imaginary numbers. A Complex number is a way of identifying a point by mentioning its horizontal and vertical distances from the origin of the axes, with the vertical distance given as a multiple of "i". Although angles are mentioned in a number such as "cos 12 + i sin 11", it is still a Complex number.

Complex sinusoid

The helix formed by the formula "z = cos 360t + i sin 360t" or "z = cos 2πt + i sin 2πt" is often called "a Complex sinusoid". This term is easily confused with the term "complex waveform", which has nothing to do with Complex numbers, and just means a signal made up of two or more pure waves. Personally, I think it makes sense to avoid using both terms. The term "complex waveform" sounds as if it involves Complex numbers, when it does not. The term "a Complex sinusoid" implies the shape is a sinusoid, when it is actually a helix.

Multiplication breaks the connection

As I have already explained, multiplying two or more circles that are encapsulated with Complex number formulas breaks the connection between the circles and their derived waves. The result of multiplying two circles as Complex number formulas is not the same as the result of multiplying the derived waves of those circles, and then using the results as coordinates to create a shape. The reason for this is clearer at this level of maths, but becomes less obvious as the maths becomes more complicated. When we use Imaginary powers of "e" to represent circles (as described in Chapter 27), the problems of multiplication become more subtle.

Conclusion

If you had started reading this book at this chapter, this chapter would seem fairly complicated. As it is, the ideas explained here are really just variations of the ideas explained in Chapter 3 of this book – a circle's edge can be said to be made up of coordinates based on Sine and Cosine waves. In this chapter, we saw that those waves can be combined together as a Complex number.

Chapter 25: Powers of i

Notes on this chapter

This chapter uses exponentials, which are numbers raised to a power such as 3^2. You do not need to know much about exponentials to understand this chapter, but it pays to realise that something such as "$2a^b$" means "$2 * (a^b)$" and not "$(2a)^b$". I will explain exponentials in Chapter 26. In an exponential such as "a^b", "a" is called the "base", and "b" is called the "exponent". I have not explained exponentials before this chapter because it is easier to learn something once you have a need to learn it.

This chapter mentions calculators that can work with Complex numbers. You do not need to have such a calculator. As the chapter progresses you will see that you can perform the maths here without one.

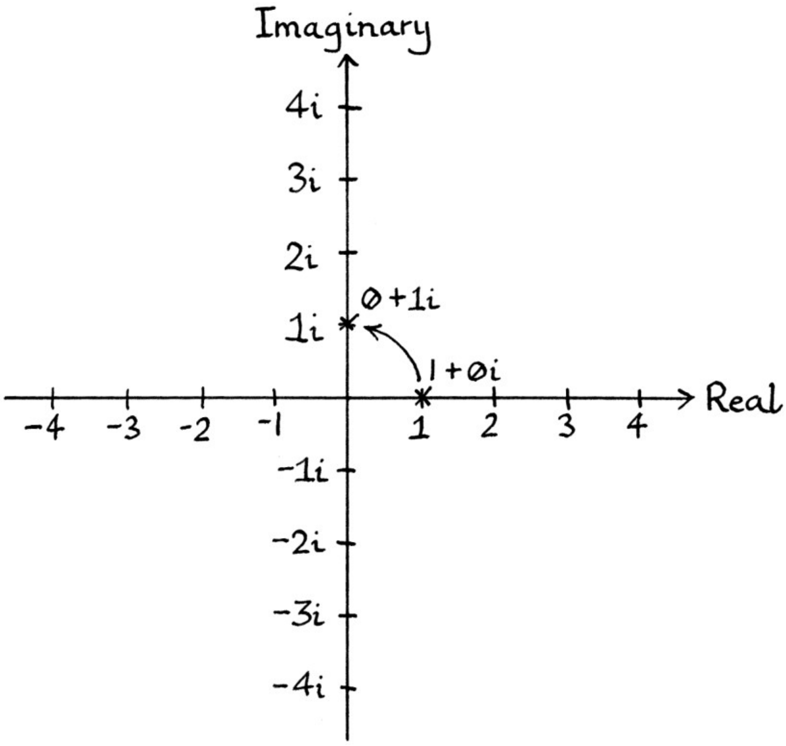
Note how I frequently include the unneeded ones and zeroes in Complex numbers. For example, I might say "2 + 0i", when most people would say "2" instead. Normally in maths, unneeded ones and zeroes are ignored in Complex numbers. I am keeping them in to make everything clearer.

Multiplication by i

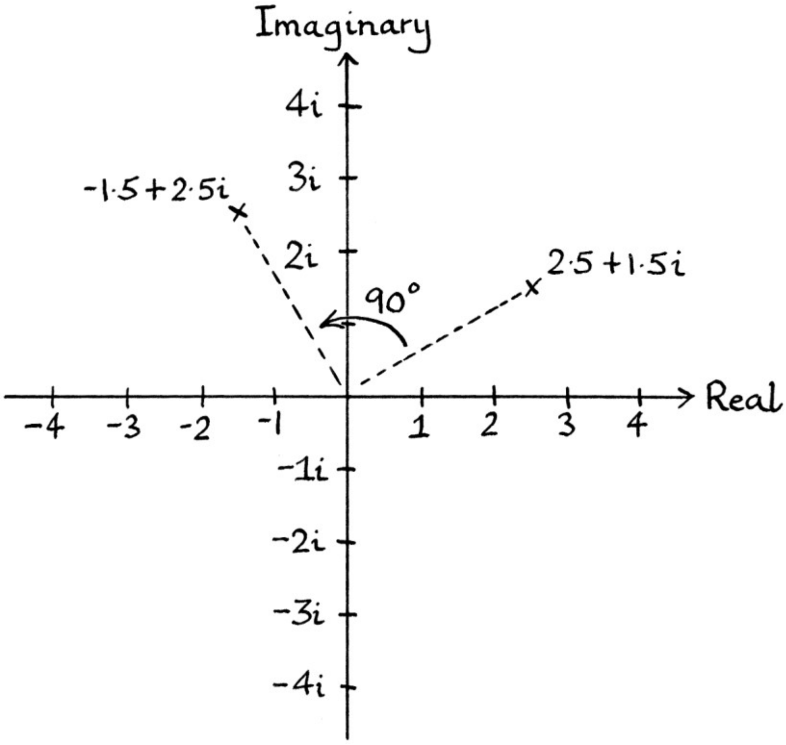
This chapter is designed as a stepping stone to understanding the meaning of the number "e" raised to Imaginary powers. Outside of this book, you might never need to use anything in this chapter, it pays to understand everything here.

In Chapter 23, I explained how a multiplication by "i" results in a coordinate on the Complex plane being rotated by 90 degrees. [It is also the case that a division by "−i" also results in a coordinate being rotated by 90 degrees, a multiplication by "−i" results in a rotation of −90 degrees, and a division by "i" results in a rotation of −90 degrees, but these will not be relevant in this chapter].

As an example, “1 + 0i” multiplied by “i” results in a point 90 degrees further around the origin of the axes at “0 + 1i”. This is shown in the following picture:



The number “2.5 + 1.5i” multiplied by “i” also results in a point 90 degrees further around the origin: “−1.5 + 2.5i”.

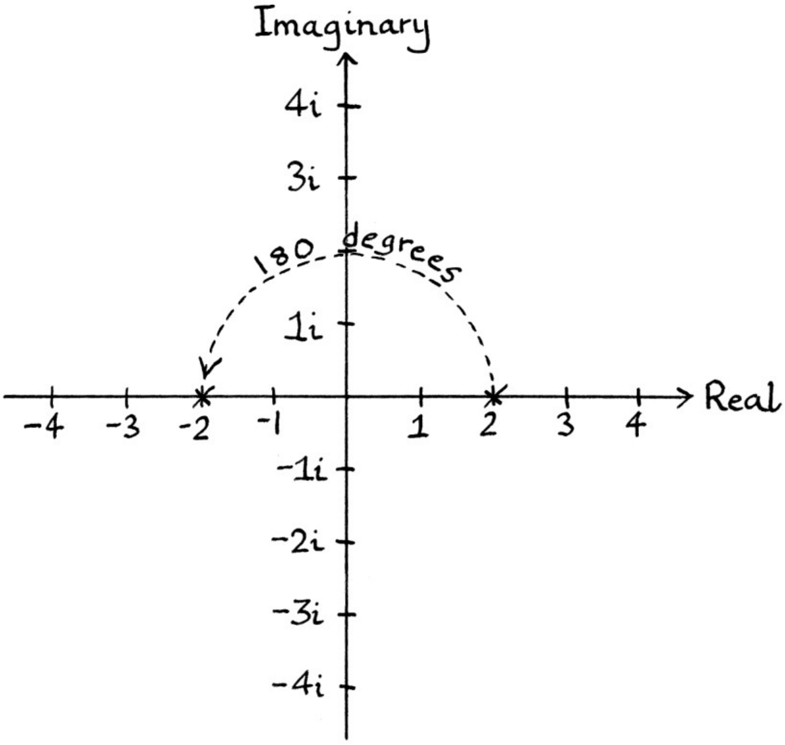


Knowing this property of multiplications by “i” might be useful if we want to check if two coordinates equidistant from the origin on a graph are at 90 degrees to each other with respect to their angle from the origin. For example, if we have the coordinates (2.3, 4.5) and (−4.5, 2.3) but do not have time to draw a graph, we can convert them both to Complex numbers, “2.3 + 4.5i” and “−4.5 + 2.3i”, and then multiply the first by “i”:
(2.3 + 4.5i) * i
= 2.3i – 4.5
= −4.5 + 2.3i, which matches the second Complex number. This converted back to coordinates is (−4.5, 2.3), so the two points are at 90 degrees to each other from the origin.

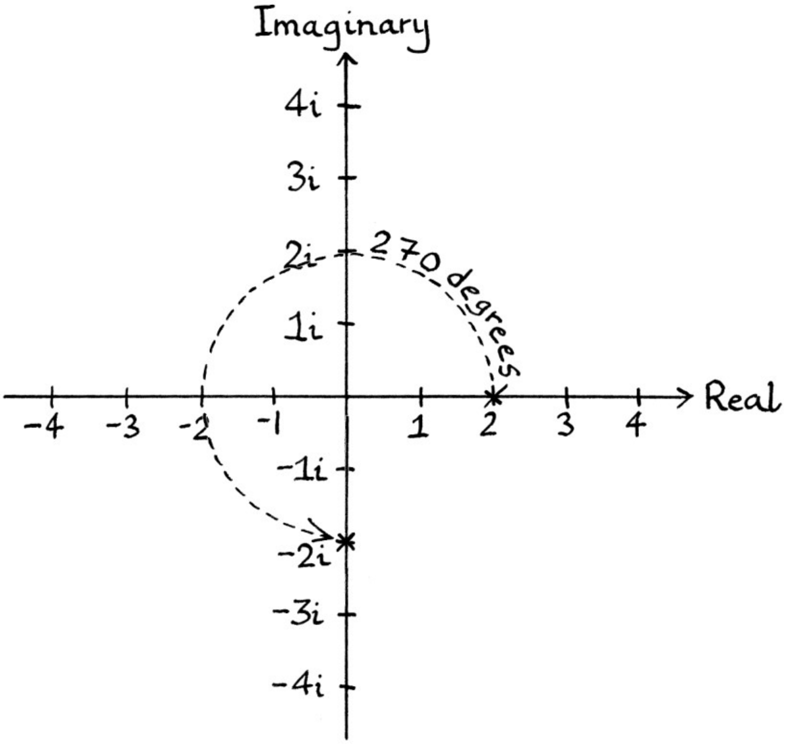
In this example, it was obvious from thinking about the coordinates that they were at 90 degrees to each other, but the underlying idea is still useful.

Multiplication by integer powers of i

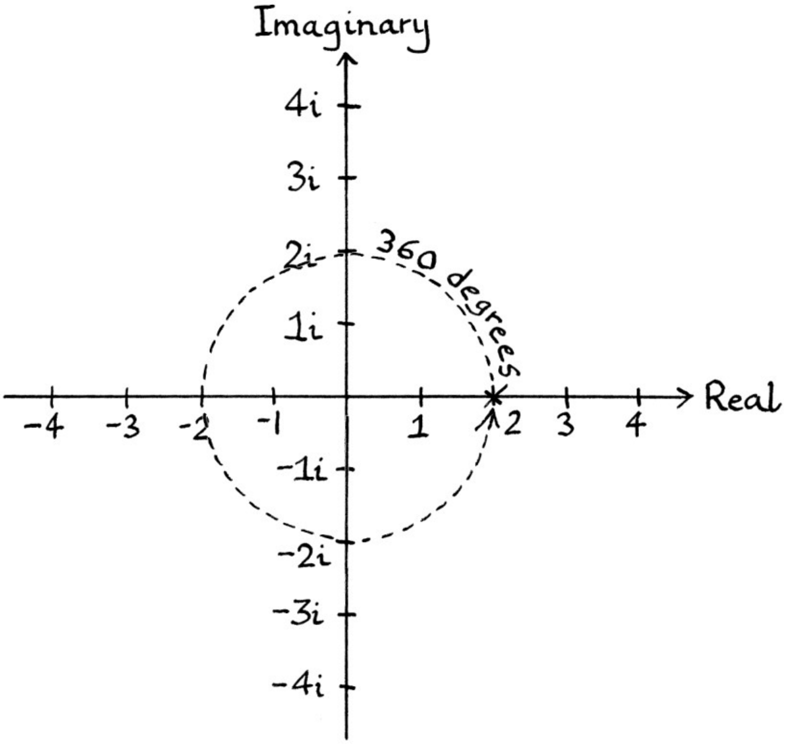
Given that a multiplication by “i” rotates by 90 degrees, it is clear that two multiplications by “i” will result in a rotation by 180 degrees:



3 multiplications by “i” will rotate by 270 degrees:



4 multiplications will rotate by 360 degrees.



We can put all of the above more mathematically with powers of “i”:
A multiplication by i^1 rotates by 90 degrees.
A multiplication by i^2 rotates by 180 degrees.
A multiplication by i^3 rotates by 270 degrees.
A multiplication by i^4 rotates by 360 degrees.

We can also say that a multiplication by i^0 (or in other words, 1) rotates by 0 degrees. [Any number raised to the power of zero results in 1.]

We know that:
i^0 is: 1
i^1 is: i or $\sqrt{-1}$
i^2 is: $i * i = \sqrt{-1} * \sqrt{-1} = -1$
i^3 is: $i * i * i = \sqrt{-1} * \sqrt{-1} * \sqrt{-1} = -1 * \sqrt{-1} = -i$
i^4 is: $i * i * i * i = \sqrt{-1} * \sqrt{-1} * \sqrt{-1} * \sqrt{-1} = +1$

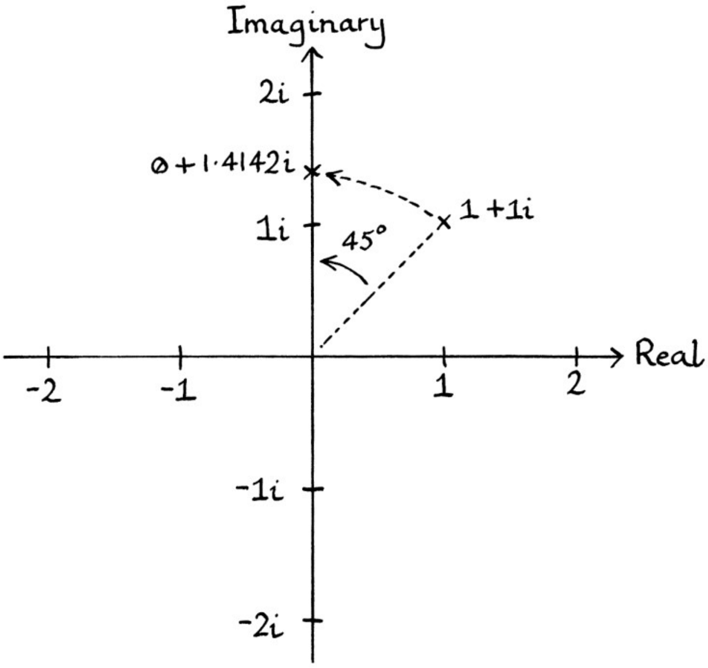
In effect, we are saying that a multiplication by −1 rotates a point by 180 degrees; a multiplication by +1 rotates a point by 360 degrees. Of course, normally, we would say that a multiplication by +1 does not rotate at all. However, for the purposes of this explanation, we will say that a multiplication by “i^4” rotates by 360 degrees.

If we wanted to know whether two coordinates were, say, exactly 270 degrees apart, we could convert them to Complex numbers, multiply the first by “i^3”, and see if the result matched the second. In practice, it would probably be obvious just by looking at the coordinates.

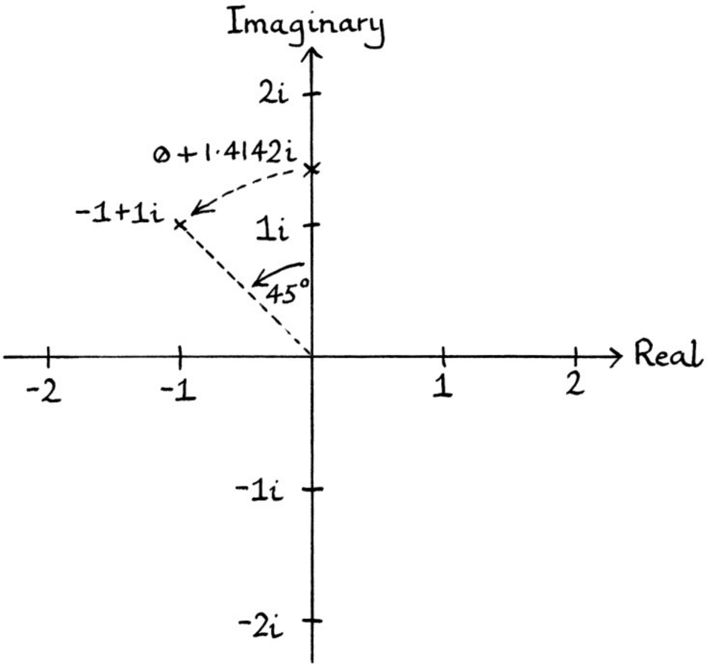
Multiplication by fractional powers of i

We can extend the “powers of i” idea to incorporate fractional powers of “i”. For example, if we multiply a Complex number by “$i^{0.5}$” (which is the square root of i), it will rotate the point by 45 degrees.

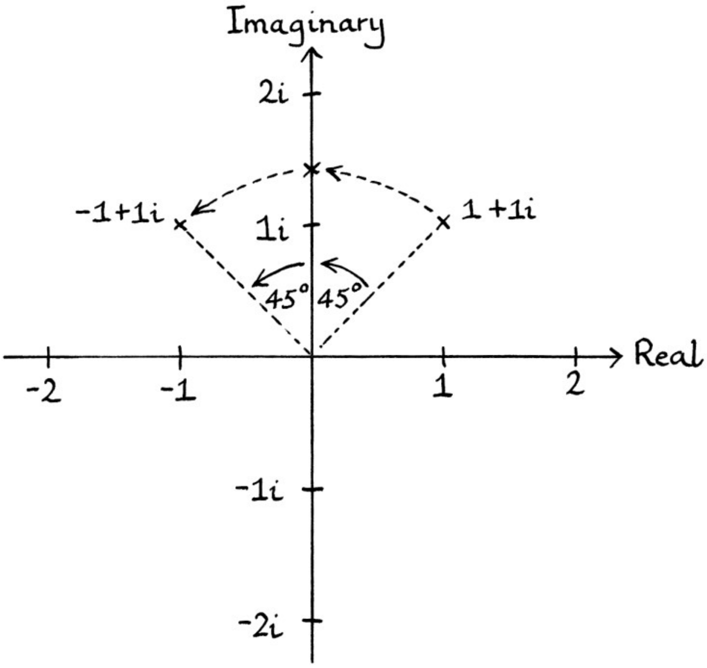
As an example, “1 + 1i” multiplied by “$i^{0.5}$” results in a point at “0 + 1.4142i” [which would normally be written as just “1.4142i”]:



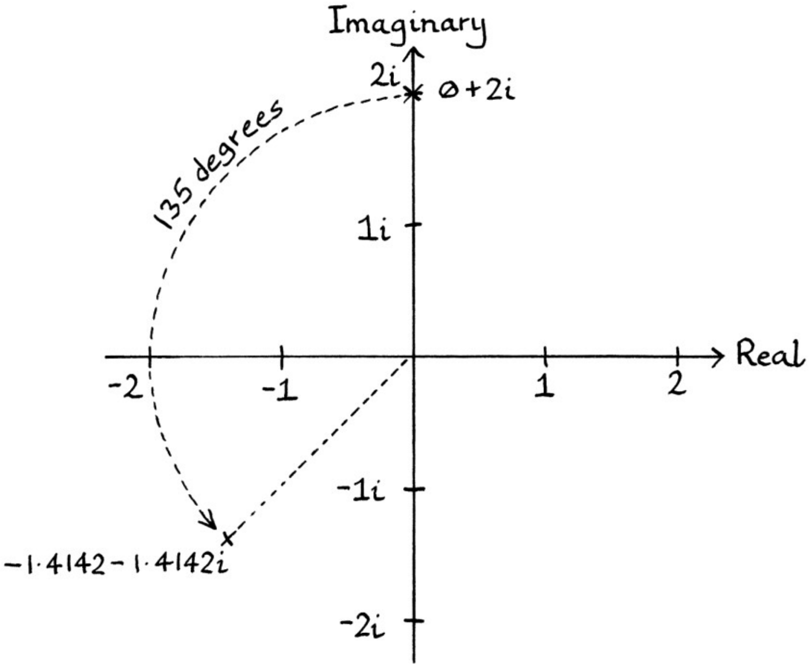
If we multiply that resulting point by $i^{0.5}$, we will have a result that is rotated 45 degrees again: “−1 + 1i”.



The Complex number “−1 + 1i” is the point we would have ended up with if we had multiplied “1 + 1i” by i^1, which makes sense as $i^{0.5} * i^{0.5} = i^1 = i$. [To multiply two exponentials with the same base, we just add the exponents.]



If we multiply a Complex number by $i^{1.5}$, we will end up with a point that is that Complex number rotated by 135 degrees. For example, “0 + 2i” multiplied by $i^{1.5}$ results in the point at “−1.4142 – 1.4142i”:



If we multiply a Complex number eight times by $i^{1.5}$, we will end up with a point at the same place. This is because rotating a point by 135 degrees eight times is the same as rotating it by 1080 degrees, which is 3 complete revolutions around the origin of the axes. We can also see this is true by seeing that $i^{1.5}$ multiplied by itself 8 times is $i^{(1.5 * 8)} = i^{12}$. The number i^{12} is:

$\sqrt{-1} * \sqrt{-1} * \sqrt{-1} * \sqrt{-1} * \sqrt{-1} * \sqrt{-1} * \sqrt{-1} * \sqrt{-1} * \sqrt{-1} * \sqrt{-1} * \sqrt{-1} * \sqrt{-1}$
$= -1 * -1 *-1 * -1 *-1 * -1$
$= 1 * 1 * 1$
$= 1$

Therefore, multiplying by i^{12} is the same as multiplying by 1. The point ends up at the place where it started.

Multiplying by $i^{2.5}$ rotates by 225 degrees.
Multiplying by $i^{3.5}$ rotates by 315 degrees.
Multiplying by $i^{0.25}$ rotates by 22.5 degrees.
Multiplying by $i^{0.125}$ rotates by 11.25 degrees.

If we have the coordinates (0.5, 2) and (1.7678, 1.0607), it is hard to tell from just looking at the coordinates that they are at 315 degrees to each other with respect to their angle from the origin. However, we can use our knowledge of fractional powers of "i" to find out. We convert the coordinates to Complex numbers:
"0.5 + 2i"
... and:
"1.7678 + 1.0607i"

Then, we multiply the first number by the equivalent of 315 degrees, which is $i^{3.5}$.
$(0.5 + 2i) * i^{3.5}$
$= 0.5i^{3.5} + 2i^{4.5}$

This result does not particularly help us though, as we do not yet have a method of calculating a fractional power of "i". I will explain a simple method later in this chapter. For now, we can use a calculator that can work with Complex numbers to find the result. It will confirm that $(0.5 + 2i) * i^{3.5}$ is equal to "1.7678 + 1.0607i", so the two coordinates are at 315 degrees to each other.

Summary of this section

By multiplying the Complex numbers identifying points by fractional powers of “i”, we can rotate a point by *any* angle. We have gone from only being able to rotate by multiples of 90 degrees to being able to rotate by any amount that we wish. [Note that if we *divided* the Complex number by a fractional power of “i”, we would rotate the point *clockwise*].

Knowing which power to use

The powers of “i” that are used for rotation by multiplication go from i^0 (which is 1) up to i^4 (which is also 1). A multiplication by i^0 is really a multiplication by 1, and therefore has no rotational effect on a point. A multiplication by i^4 is *technically* also a multiplication by 1, but it is better to think of the result of such a multiplication as equivalent to a rotation by 360 degrees. If we think of it in this way, we can say that:

A multiplication by i^0 rotates by 0 degrees.
A multiplication by i^1 rotates by 90 degrees.
A multiplication by i^2 rotates by 180 degrees.
A multiplication by i^3 rotates by 270 degrees.
A multiplication by i^4 rotates by 360 degrees.

To calculate the required power of “i” to multiply against a point to rotate it by a chosen number of degrees, it is important to notice that the “powers of i” system divides a circle into 4 portions. From this, we can then see that one portion of a circle in this system will be 90 degrees, two portions will be 180 degrees, three portions will be 270 degrees, and four portions will be 360 degrees. This is an identical way of dividing a circle as the “quarter-circle angle” system described in Chapter 22. This means that the “i to a power” system of rotation is based on quarter-circle angle units. If we have an angle in degrees or radians by which we want to rotate a Complex number by using “i” raised to a power, we have to convert that angle into quarter-circle angle units first. We then put the result as the exponent of “i”, and multiply the Complex number by that.

To calculate the required power of “i”, we take the angle by which we want to rotate, and calculate the portion of a circle that that angle represents. Then we multiply that by 4 to find out how many quarter-circle angle units that is. Any point multiplied by “i” raised to that result will rotate the point by the angle we wanted.

For example, if we want to know what power of “i” will rotate a point by 250 degrees, then we divide 250 by 360 to find the portion of a circle represented by that angle, and then we multiply that by 4, to find out how many quarter-circle angle units that is. In other words:

- The portion of an entire circle represented by 250 degrees is:
  250 ÷ 360 = 0.69444444 [to 8 decimal places]
- The number of quarter-circle angle units that this amounts to is:
  0.69444444 * 4 = 2.77777778 quarter-circle angle units.

Therefore, to rotate a point by 250 degrees, we need to multiply it by “$i^{2.77777778}$”.

To see what power of “i” is required to rotate a point by 1.3π radians, we first calculate the portion of a circle represented by an angle of 1.3π radians. This is 1.3π ÷ 2π = 0.65. We then multiply that by 4 to see how many quarter-circle angle units that is: 0.65 * 4 = 2.6 quarter-circle angle units. Therefore, we need to multiply a point’s Complex number by “$i^{2.6}$” to rotate it by 1.3π radians.

Example

We will say that we want to rotate the point at “2 + 1.7i” by 100 degrees. First, we see how much of a circle 100 degrees represents. It is 100 ÷ 360 = 0.27777778 of a circle. We multiply this by 4 to see how many quarter-circle angle units this is:
0.27777778 * 4 = 1.11111111 quarter-circle angle units.

Therefore, to rotate “2 + 1.7i” by 100 degrees, we must multiply it by “$i^{1.11111111}$”.

The calculation will be:
$i^{1.11111111} * (2 + 1.7i)$
... which is the same as:
$(2 * i^{1.11111111}) + (1.7 * i * i^{1.11111111})$
... which, for reasons that we will learn in Chapter 26 on exponentials, is:
$(2 * i^{1.11111111}) + (1.7 * i^{2.11111111})$
... which can be written as:
$2i^{1.11111111} + 1.7i^{2.11111111}$

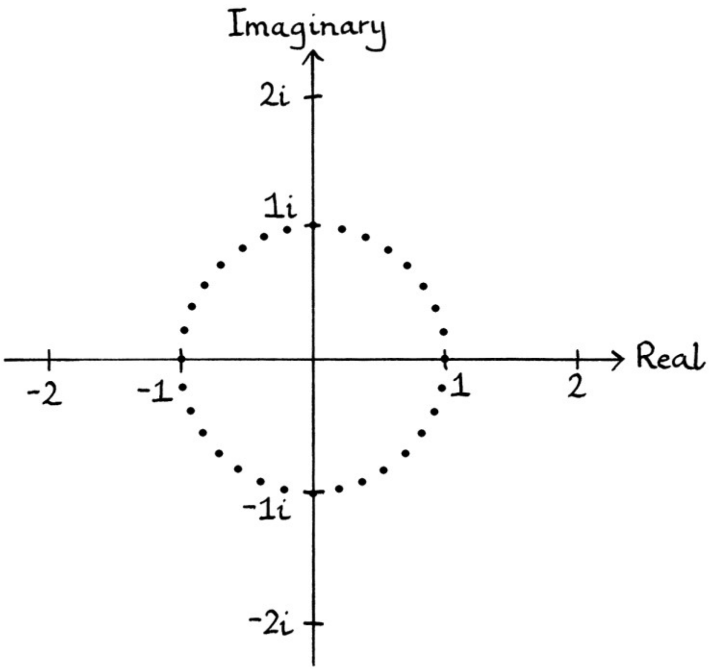
A calculator that can work with Complex numbers will give the result of this as: −2.02147 + 1.6744i [to 4 decimal places].

[I will explain how to solve such calculations without a calculator later in this chapter.]

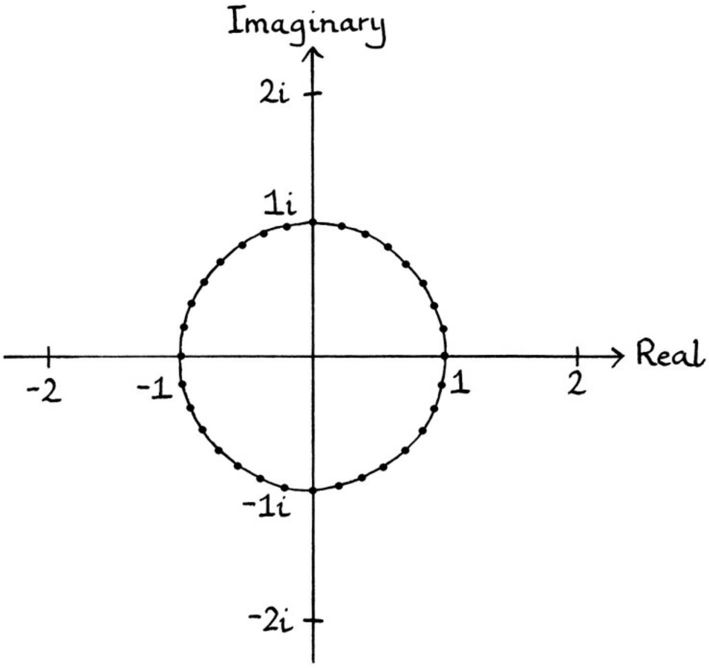
We can check that the above result is correct. Using Pythagoras's theorem and arctan, we can tell that "2 + 1.7i" is 2.6249 units away from the origin at an angle of 40.3645 degrees. We can also write this as 2.6249∠40.3645 degrees. Using Pythagoras's theorem and arctan (and choosing the correct arctan answer), we can tell that "−2.02147 + 1.6744i" is also 2.6249 units from the origin, and at an angle of 140.3645 degrees. We can also give this as 2.6249∠140.2645 degrees. These two points are 100 degrees apart, which shows that our method of rotation works.

Identification with powers of i

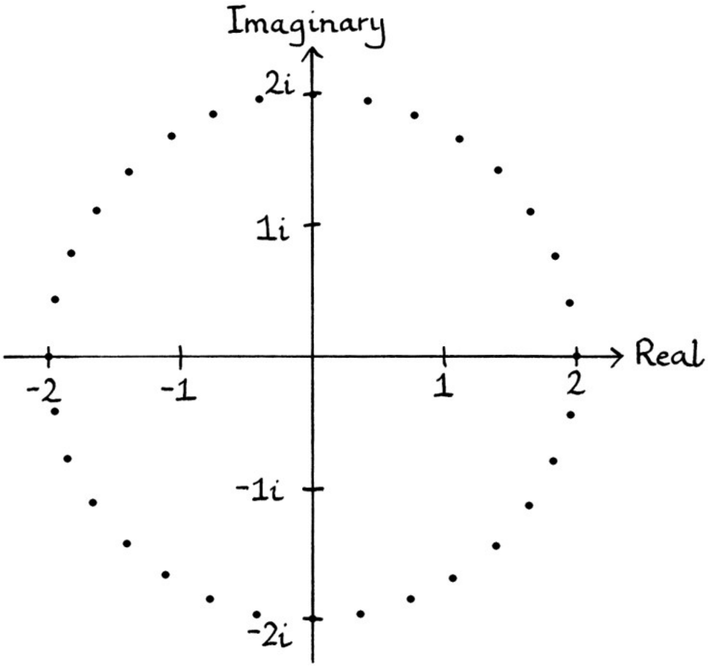
Given that multiplication of a point by a power of "i" rotates it, it is possible to draw a circle by repeatedly rotating a point by small angles. For example, if we repeatedly multiply the point "1 + 0i" by $i^{0.125}$, we will have a series of points around a unit circle spaced at 11.25 degrees. [0.125 quarter-circle angle units is the same as 11.25 degrees].



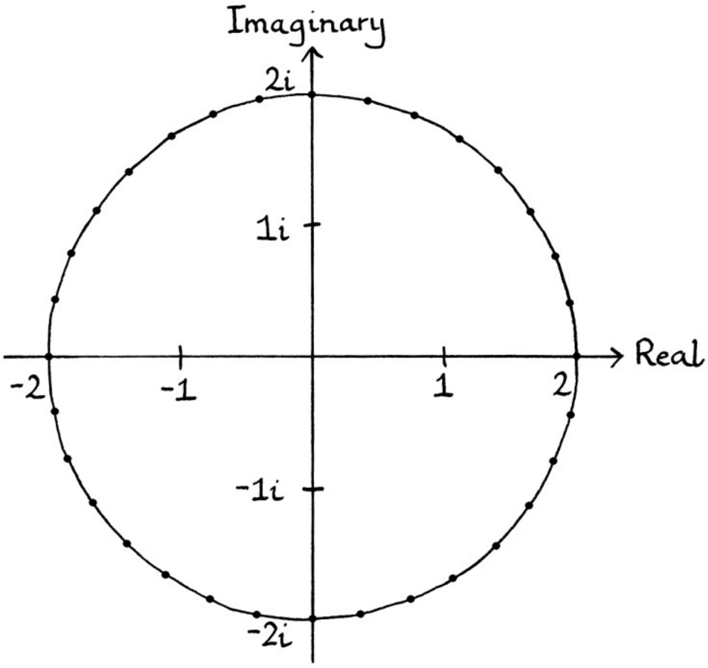
We can join up the points to make a circle:



Supposing we repeatedly multiplied the point at “2 + 0i” by $i^{0.125}$, we would end up with a series of evenly spaced points around a circle with a radius of 2 units.



We can join these points up to make a circle:



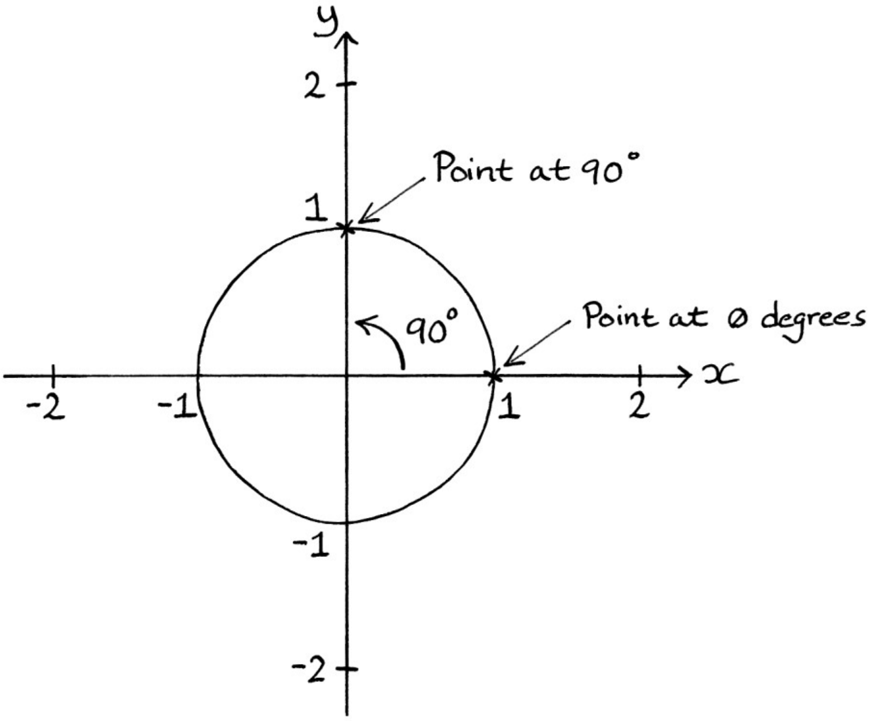
From these examples, we can see that there is a relationship between powers of “i” and circles. In both examples, if we had used a smaller angle, we would have had a smoother circle.

Identification

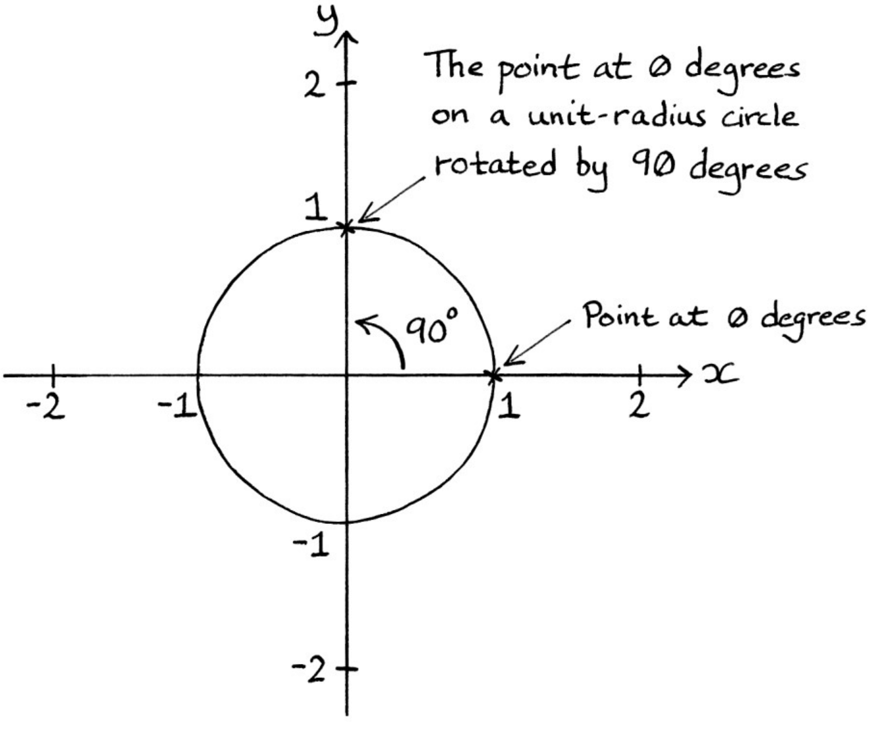
A useful idea is that we can actually *identify* a point on a circle in terms of powers of “i”. This is easiest to do if we look at circles with a radius of 1 unit.

Outside the world of Complex numbers, we could say that we can identify the position of a point on the circumference of a unit-radius circle by saying by how much a point at 0 degrees on that circle would have had to have been rotated to get there.

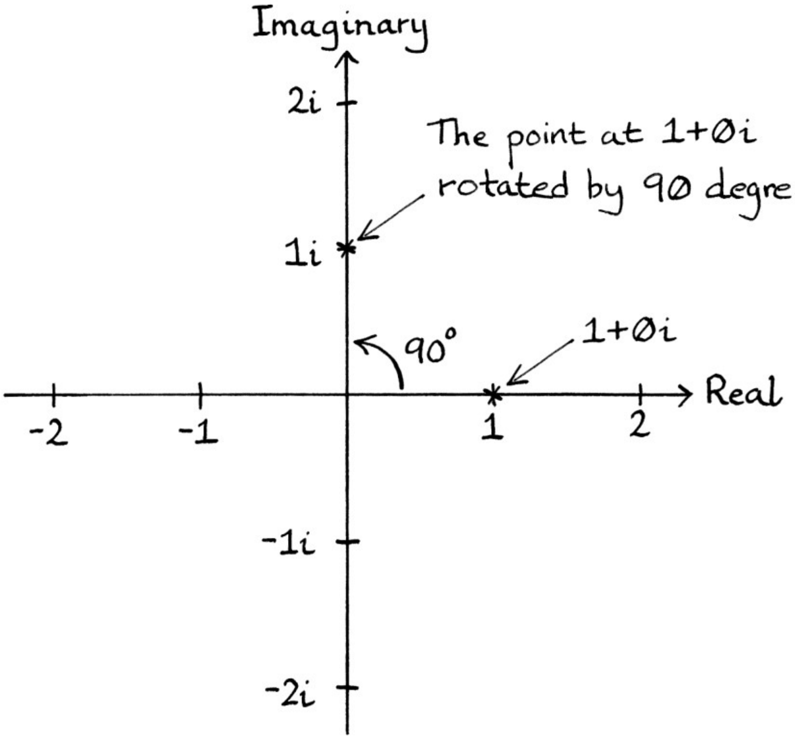
As an example, we can say that the point at 0 degrees on a unit-radius circle would need to be rotated by 90 degrees to reach the position at the very top of the circle.



This means that we could describe the point at the very top of the circle by referring to it as "the point at 0 degrees on a unit-radius circle rotated by 90 degrees". Of course, describing a point in this way is more long winded than describing it normally as an angle. However, the idea is relevant to this section.

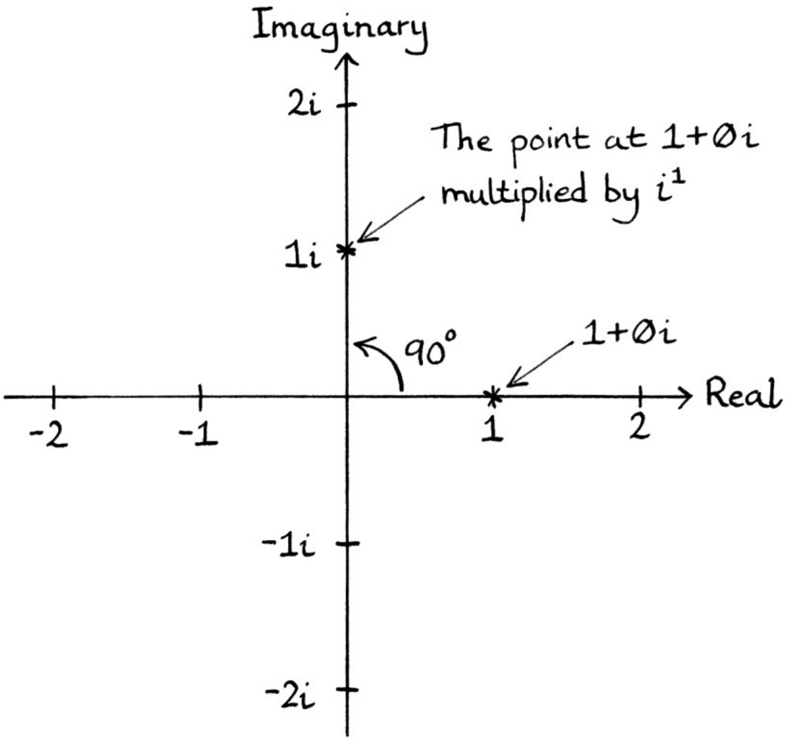


On the Complex plane, we could describe the same point ("0 + 1i"), by referring to it as "1 + 0i" rotated by 90 degrees". [We do not actually need to draw the unit-radius circles when describing a point in this way.]

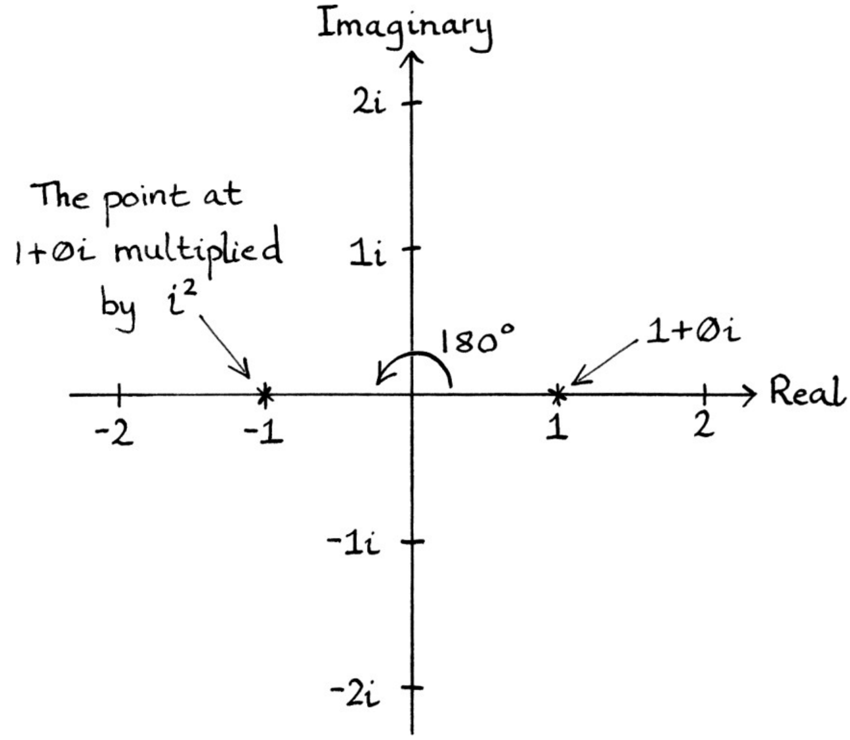


If we can identify the position of a point by saying by what angle the point at "1 + 0i" would need to be rotated to get there, then we can also identify its position in terms of what power of "i" would need to be multiplied by "1 + 0i" to rotate it there.

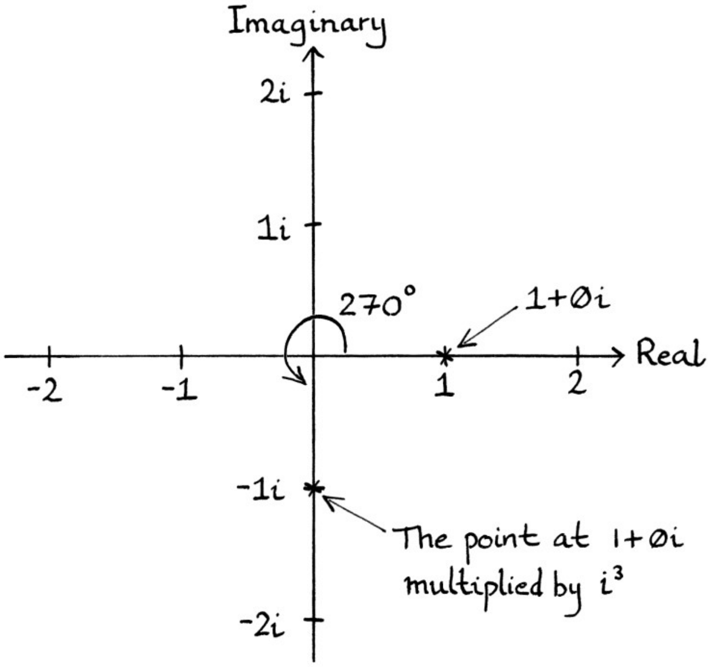
As an example, the point at "1 + 0i" would need to be multiplied by "i^1" to end up at 90 degrees (in other words, at "0 + 1i"). Therefore, we could call the point at 90 degrees, "the point at '1 + 0i' multiplied by 'i^1'".



The point at 0 degrees (“1 + 0i”) would need to be multiplied by “i^2” to end up at 180 degrees (“−1 + 0i”). Therefore, we can call the point “−1 + 0i”, “the point at ‘1 + 0i’ multiplied by ‘i^2’”.



The point at 0 degrees (“1 + 0i”) would need to be multiplied by “i^3” to end up at 270 degrees (“0 – 1i”). Therefore, we can call the point “0 – 1i”, “the point at ‘1 + 0i’ multiplied by ‘i^3’”.



A significant conclusion

To express the above ideas more succinctly:

- We can identify the position of any point on a unit-radius circle by saying by how much the point at 0 degrees (“1 + 0i”) would need to be rotated to get there.

- We can portray rotations around a unit-radius circle in terms of multiplications by powers of “i”.

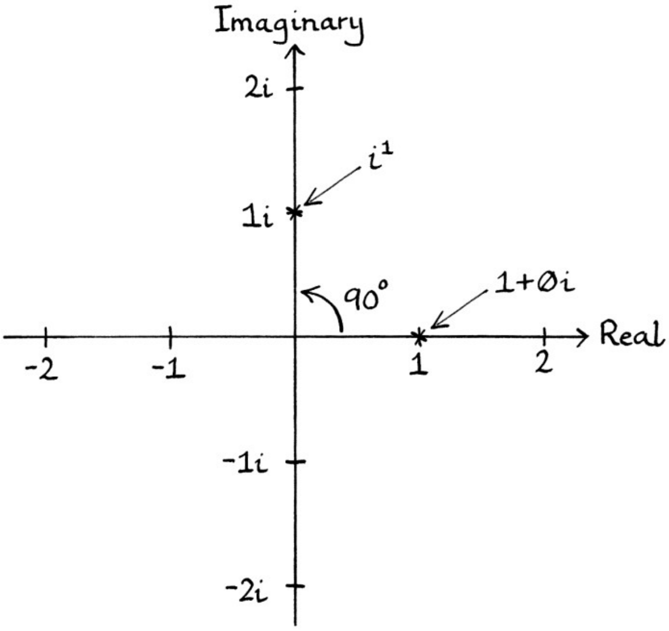
These two ideas mean that we can identify the position of any point on a unit-radius circle by saying by what power of “i” the Complex number “1 + 0i” would need to be multiplied to end up there.

As an example, if we wanted to identify the point at 45 degrees, we could say it is “1 + 0i” multiplied by $i^{0.5}$. If we wanted to identify the point at 315 degrees, we could say that it is “1 + 0i” multiplied by $i^{3.5}$.

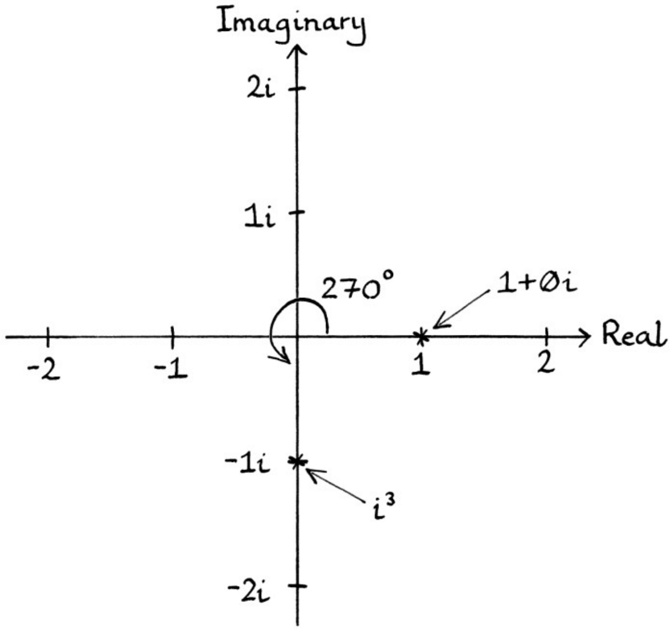
A more concise rule

Given that the point "1 + 0i" would more usually be written as just "1", the result of such a multiplication will always be just 1 multiplied by the power of "i", which can be written as just the power of "i" on its own. In other words, if we wanted to identify the point at 45 degrees, we could just say that it is "$i^{0.5}$". If we wanted to identify the point at 315 degrees, we could just say that it is "$i^{3.5}$".

This means that any point on a unit radius circle can be identified in terms of just "i" raised to a power. We can identify the point at "0 + 1i" with just "i^1":



We can identify the point at "0 − 1i" with just "i^3":

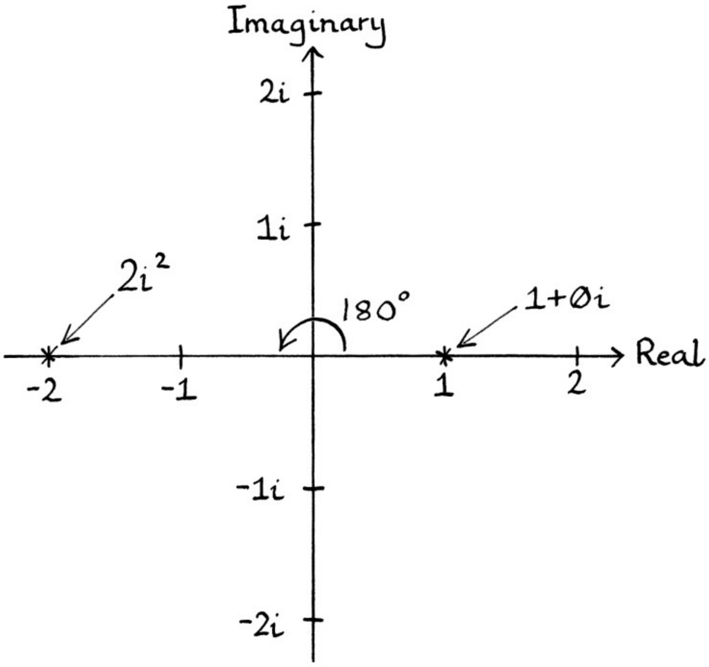


Extending the idea

As might be expected, we can identify points on circles with radiuses other than 1, using the “powers of i” idea.

For example, if we have a circle with a radius of 2 units, we can identify any point on its circumference by seeing by how much the point at “1 + 0i” would need to be rotated *and scaled* to get there, and give this rotation in terms of a multiplication by a *multiple* of “i” raised to a power.

Supposing we wanted to identify the point at “−2 + 0i” (which is at 180 degrees from “1 + 0i”), we would need the point at “1 + 0i” to be multiplied by 2i^2. Therefore, we could identify the point at “−2 + 0i” by saying it is:
(1 + 0i) * 2i^2
... which is:
2 * i^2
... which is:
2i^2



We are giving the position of “−2 + 0i” in terms of the multiplication that would rotate and scale “1 + 0i” to put it there. The Complex number “1 + 0i” needs to be multiplied by “2i^2” to end up at “−2 + 0i”. This means we can refer to “−2 + 0i” with just the number “2i^2”.

Any point on the Complex plane

It might be apparent that we can identify *any* point on the Complex plane by thinking of it as a scaled and rotated version of "1 + 0i", and by doing the scaling and rotating by multiplying "1 + 0i" by a multiple of a power of "i".

As an example, the point at "−10 + 24i" is 26 units from the origin and at an angle of 112.6199 degrees. We could express this as 26∠112.6199 degrees, or we can express it in terms of multiples of "i" raised to a power. The angle 112.6199 degrees is 112.6199 ÷ 360 = 0.3128 of a circle. This is 0.3128 * 4 = 1.2513 quarter-circle angle units. Therefore, we can express "−10 + 24i" as "$26i^{1.2513}$". The point at "−10 + 24i" is the point at "$26i^{1.2513}$". A calculator that can work with Complex numbers will confirm that these numbers refer to the same point, although it will not know the significance of what it is doing. [Later in this chapter, I will show ways of confirming the results of these calculations without needing a calculator that can work with Complex numbers.] We can think of "$26i^{1.2513}$" as being a rotated and scaled version of the point at "1 + 0i".

[Remember that "$26i^{1.2513}$" is "26 * ($i^{1.2513}$)" and not "$(26i)^{1.2513}$"]

As an easier example, the point at "0 − 1234i", which would normally be written as "−1234i", can be expressed as "$1234i^3$". This example is easy to check because "i^3" is i * i * i, which is the square root of −1 multiplied by itself 3 times. This ends up as "−1 * i", which is "−i". Therefore, "$1234i^3$" is −1234i. We can also tell, from knowing how quarter-circle angle units work that 3 quarter-circle angle units is the same as 270 degrees. Therefore, the point must be 1234 units away from the origin and at an angle of 270 degrees.

As a more in-depth example, we will start with the point at "23.7 + 89i" and try to find its multiple-and-power of "i". The point can be thought of as being the result of a point at 0 degrees and some distance away from the origin, being rotated by a particular amount to get there. We can work out the distance from the origin of the chosen point if we treat it as being at the end of the hypotenuse of a right-angled triangle with an adjacent side of 23.7 units and an opposite side of 89 units. The hypotenuse of this triangle, and hence, the distance of the point from the origin will be: $\sqrt{23.7^2 + 89^2}$ = 92.1015 units. Therefore, our chosen point will be given in terms of a scaling of the point at "1 + 0i" by 92.1015 units, and a rotation of a yet-to-be-determined angle.

The angle of our chosen point can be found using arctan:
θ = arctan (89 ÷ 23.7)
θ = 75.0886 degrees [or 1.3105 radians]
As we used arctan, we need to check that this answer is the one out of the two possible ones that we actually require, which it is.

We now know that our chosen point will be one that is equal to "1 + 0i" having been scaled by 92.1015 and rotated by 75.0886 degrees.

We work out the power in the usual way. The portion of a circle represented by 75.0886 degrees is 75.0886 ÷ 360 = 0.2086. The number of quarter-circle angle units that this represents is 0.2086 * 4 = 0.8343 quarter-circle angle units.

Therefore, after all that, we can say that "1 + 0i" multiplied by 92.1015, and then multiplied by "$i^{0.8343}$" results in the point at "23.7 + 89i". Therefore, we can represent the point "23.7 + 89i" with the value: "$92.1015i^{0.8343}$".

If you have a calculator that can work with Complex numbers, you can check that this is correct by entering 92.1015 * $i^{0.8343}$ and seeing the result. [The result will be close enough given the accuracy we have been using].

Summary of the rule

An outline of the rule that we have discovered is that *any* point on the Complex plane can be identified in terms of a scaled and rotated “1 + 0i” [which would normally be given as just “1”]. Given that we can perform scaling and rotating by multiplying a Complex number by multiples of powers of “i”, we can say that:

> *“Any point on the Complex plane can be identified in terms of a multiple of a power of i”.*

This is a significant observation and is the basis for using Imaginary powers of “e” in later chapters.

We could put this more mathematically by saying:
“$z = ai^x$”
... where:

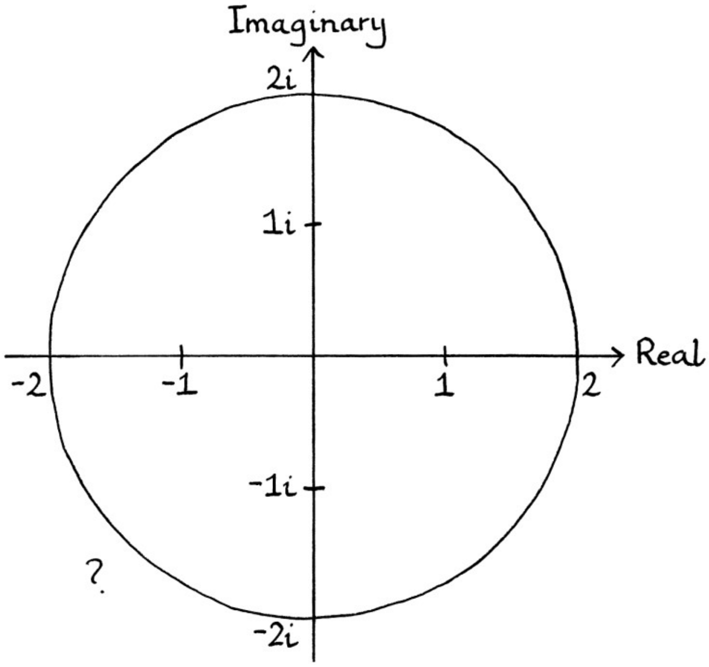
- “z” is the position of any point on the Complex plane.
- “a” is the value that scales the rest of the equation. It is equivalent to the radius of a circle, or the hypotenuse of a right-angled triangle.
- “i” is the square root of −1, as always.
- “x” is the power to which “i” is raised, which is an angle in quarter-circle angle units.

Calculating i raised to a power

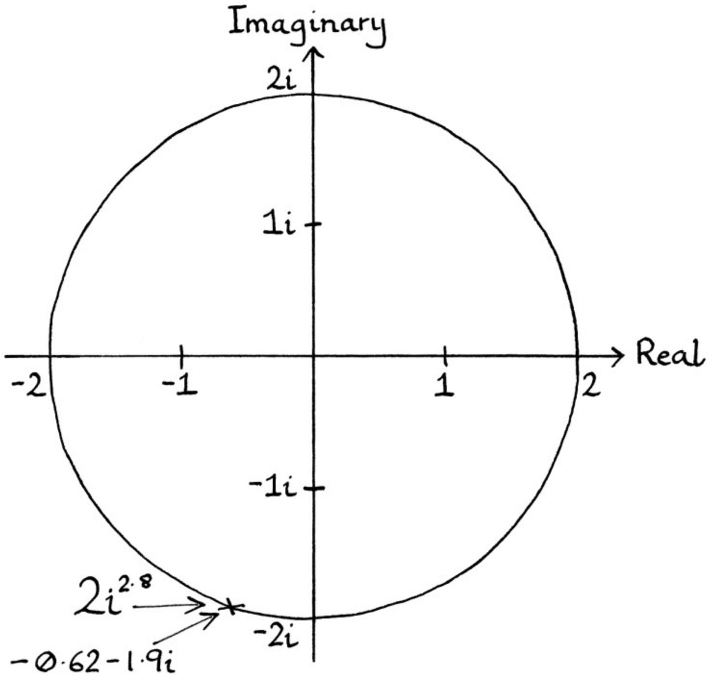
As we now know, we can identify any point on the Complex plane using a multiple of “i” raised to a power. It is also possible, and straightforward, to calculate the Complex number for a point identified by a multiple of “i” raised to a power. Doing this amounts to calculating the result of a power of “i”.

For example, suppose we want to know the Complex number represented by “$2i^{2.8}$”. We know that this represents “1 + 0i” (or just “1”) rotated by a particular amount around the origin of the axes, and then scaled by 2. To make things easier, in this case, we can think of this as the point “2 + 0i” rotated by a particular amount around the circle. Therefore, we can draw a circle with a radius of 2 units. The point “$2i^{2.8}$” will be somewhere on the circumference of this circle.

Knowing how powers of "i" split the circle up into 4 parts, we know it will be somewhere in the third quarter:



The 2.8 in the power of "i" represents the number of quarter-circle angle units by which "2 + 0i" is rotated to reach the point of interest. Therefore, we convert the 2.8 into a portion of a circle: 2.8 ÷ 4 = 0.7. We then convert that into degrees: 0.7 * 360 = 252 degrees. [If we had converted it into radians, we would have used: 0.7 * 2π = 4.3982 radians]. On our circle, we mark the point at 252 degrees. This is the point indicated by $2i^{2.8}$. We measure the x-axis value of this point: it is −0.62. Then we measure the y-axis value of this point: −1.90. The coordinates of the point are (−0.62, −1.9). The Complex number for this is "−0.62 – 1.9i". We can therefore say that the point identified by "$2i^{2.8}$" has the Complex number "−0.62 – 1.9i". This is another way of saying that "$2i^{2.8}$" reduced to a Complex number is "−0.62 – 1.9i".



We can confirm this is correct by typing $2i^{2.8}$ into a calculator that can work with Complex numbers. A calculator will give the result as "−0.6180 – 1.9021"

In reality, we did not need to draw and measure a circle, but doing so helps us visualise exactly what we are calculating. We could have just calculated the coordinates using Cosine and Sine, once we had worked out the angle: the x-axis value is "2 cos 252"; the y-axis value is "2 sin 252" (where Cosine and Sine are working in degrees).

If we had done this using radians, the angle would have been 4.3982 radians, so the coordinates would have been "2 cos 4.3982" and "2 sin 4.3982" (where Cosine and Sine are working in radians). The result is identical.

While drawing a circle or using Sine and Cosine to find the Complex number represented by a multiple of a power of "i" is useful for dealing with the matters in this chapter, they are also useful for understanding why a power of "i" should be a Complex number. If you did not know much about "i", then this should help remove one of the mysteries surrounding it. In a way, this chapter is revealing two different ideas: first, that we can identify any point using multiples of a power of "i", and second, that any power of "i" will be a Complex number (or maybe a Real number or an Imaginary number).

A formula for calculating i raised to a power

Given what we now know, we can make a formula for solving powers of “i”.

If we had “$2i^2$”, we know it would refer to the point at “−2 + 0i”. Another way of identifying the point is by saying it is at “2 cos 180 + 2i sin 180” when Sine and Cosine are working in degrees. We can make a general formula based on this:

“ai^x = a cos ((360 * x) ÷ 4) + ai sin ((360 * x) ÷ 4)”

... which we can rephrase to be:

“ai^x = a cos 90x + ai sin 90x”
... where:

- “a” is a value scaling the power of “i”.
- “x” is the exponent of “i”.
- Cosine and Sine are working in degrees.

This formula means that we do not need to have a calculator that works with Complex numbers. We only need a calculator that can calculate Sine and Cosine in degrees.

As an example of the formula working, we will test it with “$11.3i^{0.8}$”. This should be the same as:
“11.3 cos (90 * 0.8) + 11.3i sin (90 * 0.8)”
... which is:
“11.3 cos 72 + 11.3i sin 72”
... which is:
“3.4919 + 10.7469i”
[A calculator that can work with Complex numbers will confirm this to be correct.]

The formula to solve powers of “i” that works with radians is as so:
“ai^x = a cos ((2π * x) ÷ 4) + ai sin ((2π * x) ÷ 4)”

... which we can rephrase to be:

“ai^x = a cos 0.5πx + ai sin 0.5πx”
... where:

- “a” is scaling the power of “i”.
- “x” is the exponent of “i”.
- Cosine and Sine are working in radians.

We can test this radians formula with the previous example: "11.3$i^{0.8}$". We would have:
"11.3 cos (0.5 * π * 0.8) + 11.3i sin (0.5 * π * 0.8)"
... which is:
"11.3 cos 1.2566 + 11.3i sin 1.2566"
... which is:
"3.4919 + 10.7469i"
... which is the answer we calculated before.

The two formulas show that we do not need a calculator that can work with Complex numbers, and it also shows that we do not need anything other than Sine and Cosine to calculate powers of "i". We knew this already from how we could measure the result of a power of "i" by drawing a circle, but the formulas make a tidy summary of the fact.

Sine, Cosine and i

Sine and Cosine are functions that work with circles. "i" raised to a power marks out the position of points around a circle. Therefore, there is a connection between Sine and Cosine and "i" raised to a power.

When we raise "i" to a power, which we can portray with "i^x", we are really dividing the circle into quarters. Every point on the circumference of a unit-radius circle can be identified with a value of "x" from 0 to 4. In this sense, the "x" in "i^x" is really an angle in a system where circles are divided into 4 portions: "x" is referring to quarter-circle angle units. We looked at this idea in Chapter 22. Given that "x" is an angle unit, it is more appropriate to use the symbol "θ" instead of "x" to emphasise that. Therefore, from now on, we will give "i" raised to powers in terms of "i^θ".

As another clue to how the power is actually an angle, when we were converting from degrees to powers of "i", we converted first to a fraction of a circle, by dividing by 360, and then multiplied by 4 to find the number of quarter circles that that fraction represents. [This is similar to how if we had been converting from degrees to radians, we would have converted first to a fraction of a circle by dividing by 360, and then multiplied by 2π to find the number of radians that fraction represents.] We were treating the power as an angle long before now.

We know that Sine and Cosine, if they are working in a system that divides a circle up into 360 pieces, find the y-axis and x-axis values of a point on a unit-radius circle at a particular angle in *degrees*. [Strictly speaking, it is not so much that Sine and Cosine find those values, but that Sine and Cosine, when working in degrees, are *defined* as the functions that find those values]. We also know that Sine and Cosine, if they are working in a system that divides a circle up into 2π pieces, find the y-axis and x-axis values of a point on a unit-radius circle at a particular angle in *radians*. [Again, strictly speaking Sine and Cosine, when working in radians, are defined as the functions that find those values]. Therefore, Sine and Cosine, if they are working in a system that divides a circle up into 4 pieces, can find the y-axis and x-axis values of a point on a unit-radius circle at a particular angle in quarter-circle angle units. [Again, strictly speaking, Sine and Cosine, for this angle system, are *defined* as finding the y-axis and x-axis values of a point].

This means that if Sine and Cosine are working in the quarter-circle angle system, "θ" is an angle in quarter-circle angle units, and "i^{θ}" indicates the position of a point on the Complex plane, then "Sine θ" indicates that point's y-axis value, and "Cosine θ" indicates that point's x-axis value. To put this another way, "i^{θ}" gives the position of the point at (cos θ, sin θ) when Cosine and Sine are working in a system where a circle is divided into 4 pieces.

To put this in terms of Complex numbers, the position of the point given by "i^{θ}" can also be given by "cos θ + i sin θ", when Cosine and Sine are working in quarter-circle angle units.

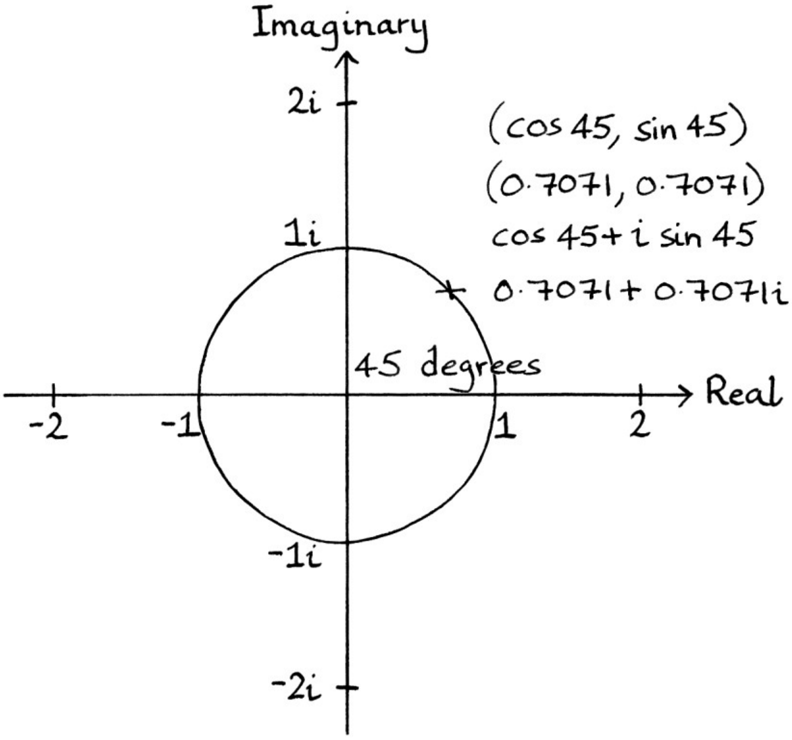
To put this more succinctly:
"$i^{\theta} = \cos \theta + i \sin \theta$"
... when "θ" is an angle in quarter-circle angle units, and Cosine and Sine are operating in the quarter-circle angle system.

This is a fairly significant conclusion. When dealing with "e^{ix}" in later chapters, this will become more relevant.

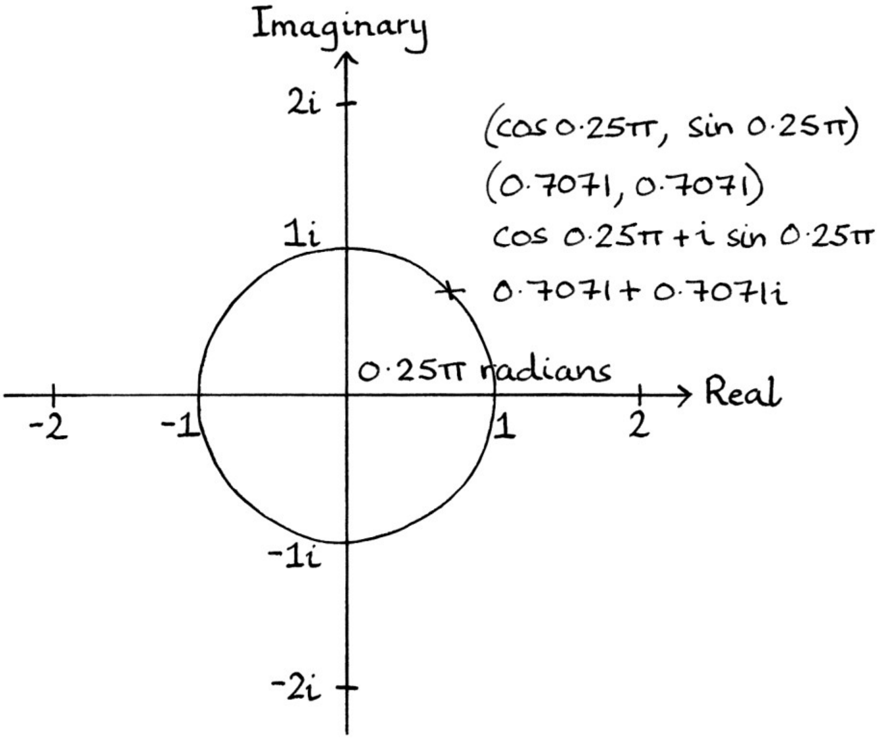
Calculators tend to work with Sine and Cosine only in degrees or radians, and it is unlikely that many work with quarter-circle angle units. However, we can see how "$i^{\theta} = \cos \theta + i \sin \theta$" is true by remembering that Sine and Cosine really work on the portion of a circle represented by an angle. A point on a circle's circumference will have the same x-axis and y-axis value regardless of the angle system used to describe its position. Therefore, the Sine of the angle and the Cosine of the angle will still refer to the same y-axis and x-axis value, no matter in which system Sine and Cosine are working, as long as that system is the same one as the type of angle being used.

It is easiest to illustrate how “i^{θ}” is equal to “cos θ + i sin θ” in a quarter-circle angle system by thinking of circles.

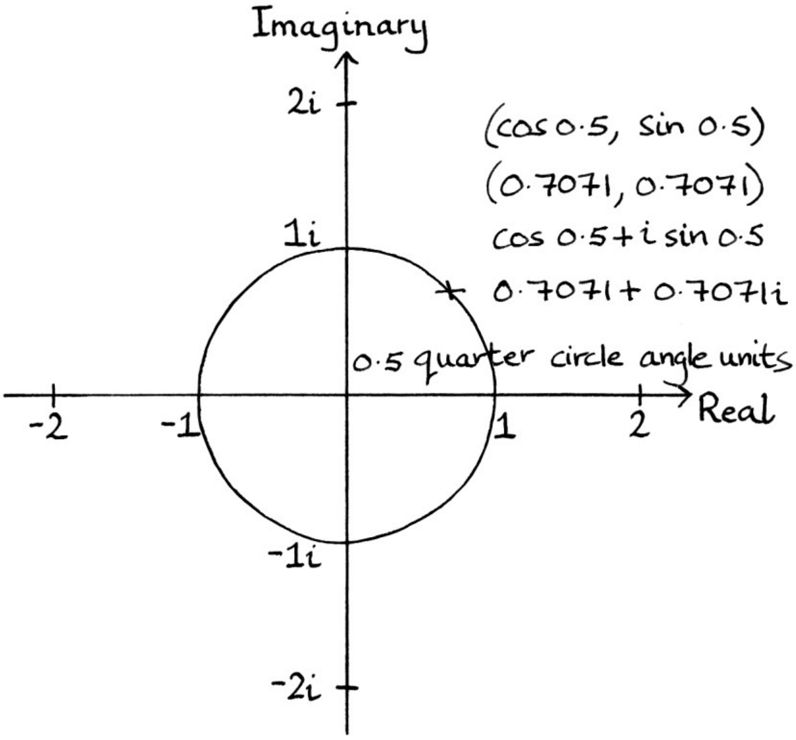
A point that is one eighth of the way around the circumference of a unit-radius circle can be said to be, in the system of dividing a circle into 360 portions, at an angle of 45 degrees. Its coordinates are (cos 45, sin 45) in degrees, which ends up as (0.7071, 0.7071). The Complex number for this is “cos 45 + i sin 45” in degrees or “0.7071 + 0.7071i”.



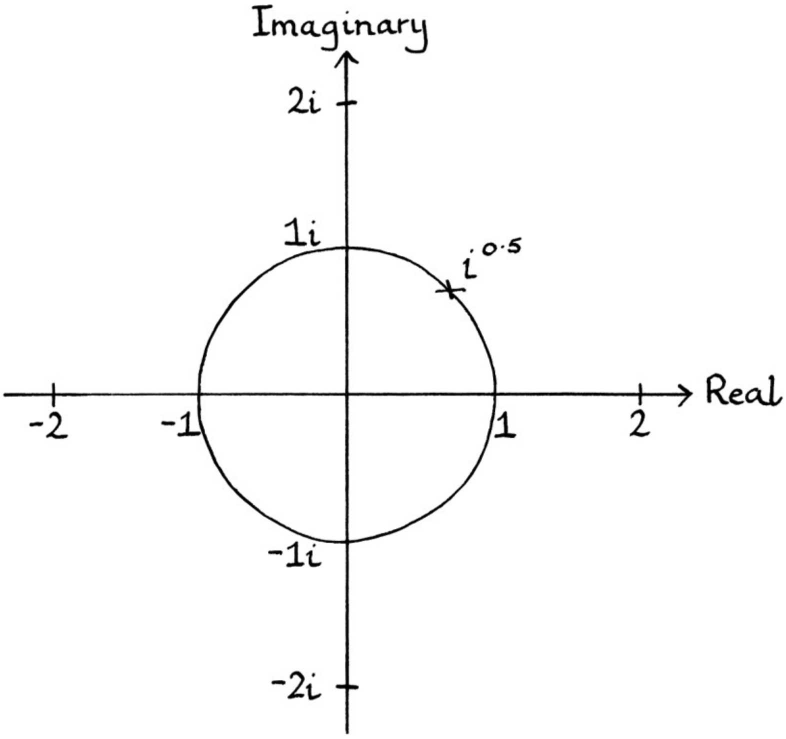
That same point is also at angle of 0.25π radians in an angle system that divides the circle into 2π portions. Its coordinates are, in that case, (cos 0.25π, sin 0.25π) in radians, which ends up as (0.7071, 0.7071). The Complex number for this is: “cos 0.25π + i sin 0.25π” in radians or “0.7071 + 0.7071i”.



That same point is also at an angle of 0.5 quarter-circle angle units, in an angle system that divides the circle into 4 portions. Its coordinates are, in that case, (cos 0.5, sin 0.5) in the quarter-circle angle system, which is (0.7071, 0.7071). The Complex number for this is "cos 0.5 + i sin 0.5" in quarter-circle angle units or "0.7071 + 0.7071i".



That same point's position can also be described as "$i^{0.5}$". Therefore, because in the quarter-circle angle system, the point's position is also "cos 0.5 + i sin 0.5", it is the case that "$i^{0.5}$ = cos 0.5 + i sin 0.5" in quarter circle-angle units. This is consistent with the original formula "i^{θ} = cos θ + i sin θ".



Degrees and radians

So far, the “θ” in “i^{θ}” refers to an angle in the quarter-circle angle system. There are 4 quarter-circle angle units in a circle. We can make the formula work with “θ” referring to a value in degrees or radians by altering it slightly.

Degrees

If we want to use the formula with a particular angle in degrees, we first need to calculate what portion of a circle that angle represents by dividing it by 360. We then need to calculate how many quarter-circle angle units that is by multiplying the result by 4.

We can alter the formula to take this all into account, and it will become:
$i^{(\theta \div 360)*4}$
$= i^{(4\theta \div 360)}$
$= i^{(\theta \div 90)}$
... which can also be written as:
$i^{(\theta/90)}$

[It is clearer using “/” in exponentials than using “÷”.]

With this new formula, we can enter an angle in degrees as “θ”, and the power of “i” as a whole will indicate the position of the chosen point. The solution to the formula will be the Complex number that indicates that point. Most significantly, the position of the point will be at the x-axis value of Cosine θ and at the y-axis value of Sine θ, when Cosine and Sine are working in *degrees*. In other words, we can say that:
$i^{(\theta/90)} = \cos\theta + i \sin\theta$
... when “θ” is in degrees, and both Cosine and Sine are working in degrees.

As an example of this working in practice, we will take the point on the unit-radius circle at an angle of 120 degrees. We can put this into the formula as so:
$i^{(120/90)}$

The exponential is the same as “cos 120 + i sin 120” when everything is working in degrees.

The number "$i^{(120/90)}$" is "$i^{1.3333}$". Note that the 1.3333 is really an angle in a system that divides the circle up into 4 parts, but the full exponential, $i^{1.3333}$ still refers to exactly the same point. We can check the position of the point: 1.3333 is 1.3333 ÷ 4 = 0.3333 of the way around a circle. The result of 0.3333 * 360 is 120, so it is still the same point.

The result of "cos 120 + i sin 120"(in degrees) is "−0.5 + 0.08660i".

If we use a calculator that can work with Complex numbers [or some other method], we can solve $i^{1.3333}$ and the result will be "−0.5 + 0.8660i". Therefore, this all confirms that $i^{(120/90)}$ = cos 120 + i sin 120 (when Cosine and Sine are working in degrees, and 120 is an angle in degrees), and that our method for using degrees works.

Note that entering an angle in degrees into "$i^{(\theta/90)}$" instantly converts the angle into quarter-circle angle units. In the above example, the "120" in "$i^{(120/90)}$" is in degrees, but the "120/90" part as a whole is in quarter-circle angle units.

Radians

To use the "i" formula with a particular angle in radians, we first need to calculate what portion of a circle that angle represents by dividing it by 2π. We then need to calculate how many quarter-circle angle units that is by multiplying the result by 4.

We can alter the original formula to take this all into account, and it will become:
$i^{(\theta \div 2\pi) * 4}$
$= i^{(4\theta \div 2\pi)}$
$= i^{(2\theta \div \pi)}$
... which we will write as:
$i^{(2\theta/\pi)}$

With this new formula, we can enter an angle in radians as "θ", and the exponential as a whole will indicate the position of a chosen point. The solution to the formula will be the Complex number that identifies the position of that point. The point will be at the x-axis value of Cosine θ and at the y-axis value of Sine θ, when Cosine and Sine are working in radians. In other words:

$i^{(2\theta/\pi)} = \cos\theta + i \sin\theta$
... when "θ" is an angle in radians, and Cosine and Sine are working in radians.

As an example, the point on a unit radius circle at an angle of 1.1π radians can be expressed as a power of "i" in the form:

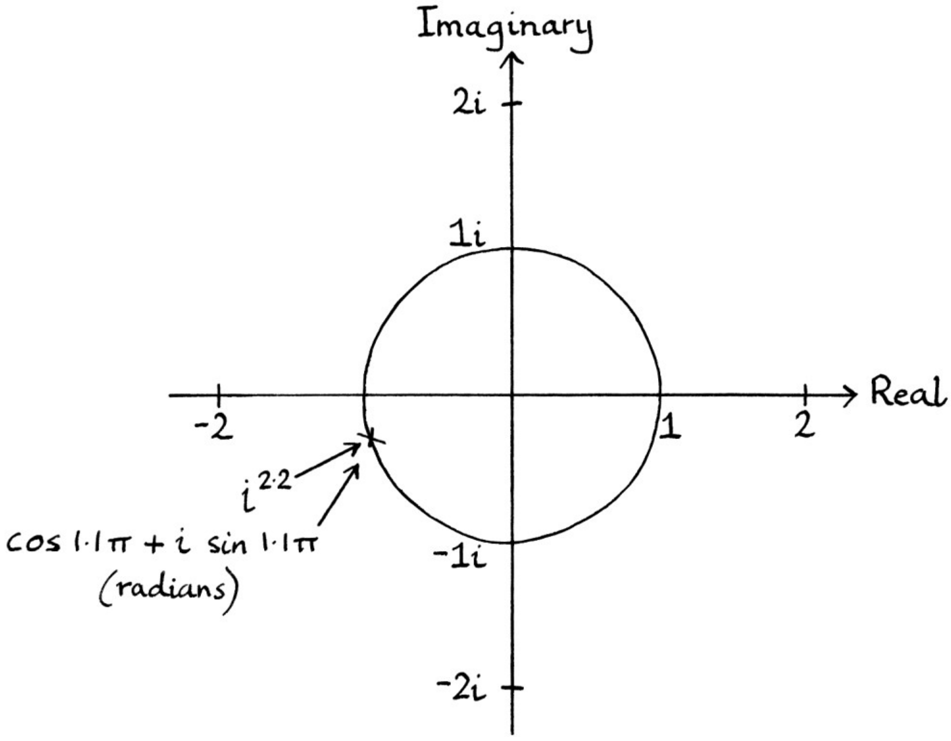
$i^{(2 * 1.1\pi)/\pi}$

... which ends up as:

$i^{2.2}$

That point's position can also be referred to as "cos 1.1π + i sin 1.1π" (in radians).

We can draw a unit-radius circle and we will see that the point at $i^{2.2}$ is the same point as that at 1.1π radians. The circle also reveals that the x-axis and y-axis values of the point are cos 1.1π and sin 1.1π, (in radians) [which they would have to be if the point is at an angle of 1.1π radians]. Using a calculator to solve $i^{2.2}$ would confirm this.



Note that entering an angle in radians into "$i^{(2\theta/\pi)}$" instantly converts the angle into quarter-circle angle units. In the above example, "1.1π" in "$i^{(2 * 1.1\pi)/\pi}$" is in radians, but the "(2 * 1.1π)/π" part as a whole is in quarter-circle angle units. Any exponent of "i" will always be treated as an angle in quarter-circle angle units. In the above examples, we are just converting the angles into quarter-circle angle units.

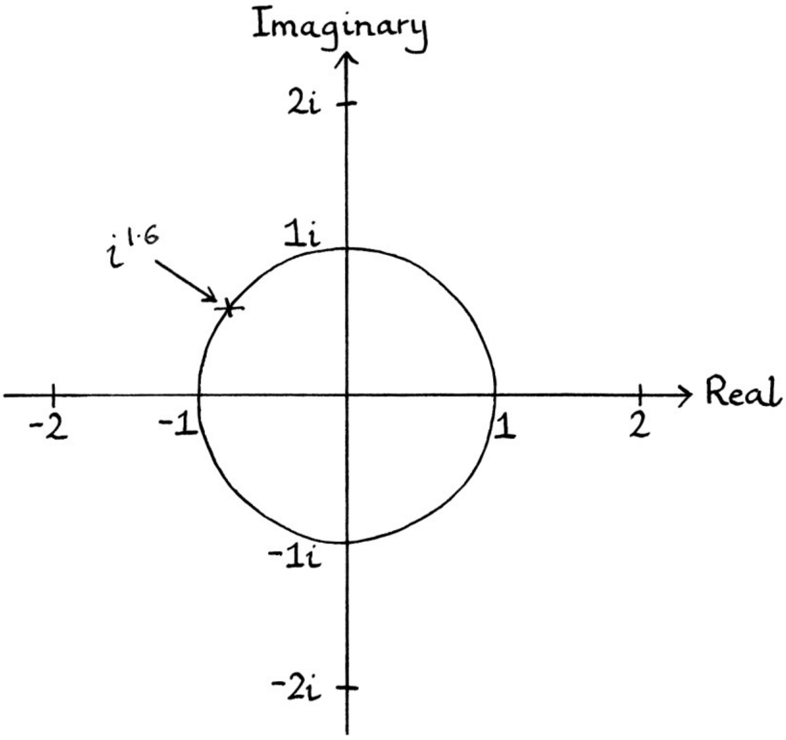
One division

For interest's sake, we could use the formula with an angle system where every angle is a fraction of a circle. In this system, 0.5 whole-circle angle units is half a circle or 180 degrees; 0.25 whole-circle angle units is quarter of a circle; 0.75 whole-circle angle units is three-quarters of a circle and so on. To use the formula with such an angle system, we would just need to know how many quarter-circle angle units a fraction of a whole circle represents. To do this, we just multiply the angle unit by 4.

We can alter the formula to take this all into account, and it will become:
$i^{4\theta}$

We can say that "$i^{4\theta} = \cos \theta + i \sin \theta$", when "θ" is an angle in a system where angles are just fractions of a circle, and Cosine and Sine are both working in that system.

As an example of this in use, we will look at a point on a unit-radius circle that is at 0.4 whole-circle angle units. This value represents a point on the unit-radius circle that is 0.4 of the way around. If we put this into the formula, we will have:
$i^{4 * 0.4}$
... which is equal to:
$i^{1.6}$
This point appears as so:



The y-axis value of this point will be Sine θ when Sine is working in the system of whole-circle angle units, and the x-axis value of this point will be Cosine θ. Strictly speaking, this should be obvious because the Sine and Cosine functions are defined as being the processes that find the y-axis and x-axis values of an angle of a point on a unit-radius circle, in whichever angle system they are operating. The significant idea here is not so much that Sine and Cosine are finding the x-axis and y-axis values, but that the "i" formula is referring to the same point as "cos θ + i sin θ".

In the above example, 0.4 is an angle in whole-circle angle units, but when it is put into the "$i^{4\theta}$" formula, it becomes converted into quarter-circle angle units. The value "4 * 0.4" or 1.6 is an angle in quarter-circle angle units.

Waves and circles

Given that "$i^{(\theta/90)}$" is equal to "cos θ + i sin θ" in degrees, and given that "$i^{(2\theta/\pi)}$" is equal to "cos θ + i sin θ" in radians, it might be apparent that we can use "$i^{(\theta/90)}$" and "$i^{(2\theta/\pi)}$" as a way of dealing with waves in exactly the same way that we did in the last chapter with "cos θ + i sin θ". However, instead of thinking of two separate waves – the Cosine and Sine waves – we are now thinking in terms of the circles from which those waves are derived. In other words, "$i^{(\theta/90)}$" and "$i^{(2\theta/\pi)}$" can be used to refer to entire circles and helices.

As we know, we can identify any point on a unit-radius circle in terms of the Cosine of its angle from the origin and the Sine of its angle from the origin. The position of any point can be given as the Complex number: "cos θ + i sin θ". What is more, if we mark the points indicated by "cos θ + i sin θ" for every value of "θ" from 0 degrees up to 360 degrees, or from 0 radians up to 2π radians, we draw out a unit radius circle. This should be obvious from how the points around a unit-radius circle are how we calculate the Sine and Cosine of an angle in the first place.

If want to draw a circle with a radius other than 1, we can scale the whole of the "cos θ + i sin θ" formula. For example, a circle with a radius of 2 can be described using "2 * (cos θ + i sin θ)" or "2 cos θ + 2i sin θ". We can draw any circle using a scaled version of "cos θ + i sin θ".

We can also draw Sine and Cosine wave graphs from the circle. This is very easy as we have the Sine and Cosine formulas in the Complex number to start with.

Given that “$i^{(\theta/90)}$” is equal to “cos θ + i sin θ” in degrees, we can actually refer to the circle described by “cos θ + i sin θ” using just the formula “$i^{(\theta/90)}$” when “θ” is in degrees. Given that “$i^{(2\theta/\pi)}$” is equal to “cos θ + i sin θ” in radians, we can refer to the circle described by “cos θ + i sin θ” using just the formula “$i^{(\theta/90)}$” when “θ” is in radians. We will look at these in turn.

Degrees circle

As “$i^{(\theta/90)}$” is equal to “cos θ + i sin θ” in degrees, everything that applies to “cos θ + i sin θ” in degrees also applies to “$i^{(\theta/90)}$”. In other words, “$i^{(\theta/90)}$” draws out a circle for values of “θ” from 0 to 360. The formula “$i^{(\theta/90)}$” gives us a concise way of describing a circle.

If we plot the y-axis (Imaginary) values of the points identified by “$i^{(\theta/90)}$” on an angle-based graph, we will end up with a Sine wave graph. If we plot the x-axis (Real) values on an angle-based graph, we will end up with a Cosine wave graph. [As “$i^{(\theta/90)}$” makes a circle based on degrees, this should really be expected]. We had all of this when we were using the Complex number “cos θ + i sin θ” in degrees, but now we are using “$i^{(\theta/90)}$”, we have another way of expressing it. This way might allow us to perform some calculations more easily.

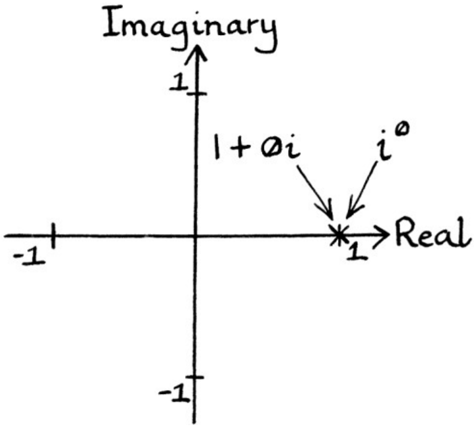
Radians

Everything I just said about “$i^{(\theta/90)}$” in degrees is also true for “$i^{(2\theta/\pi)}$” in radians. As “$i^{(2\theta/\pi)}$” is equal to “cos θ + i sin θ” in radians, it is the formula for a circle. The formula draws out a circle for values of “θ” from 0 to 2π. The y-axis (Imaginary) values of this circle drawn on an angle-based graph will draw out a Sine wave graph; the x-axis (Real) values will draw out a Cosine wave graph.

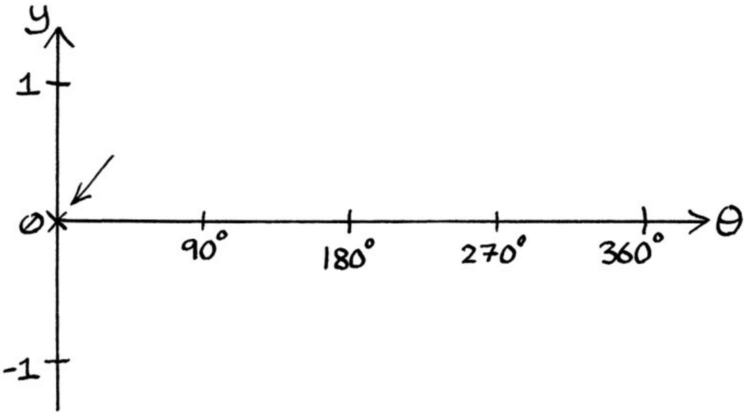
Example

We will put the angles 0 to 360 degrees into "$i^{(\theta/90)}$" and see how it results in a circle, and how that circle produces its Cosine and Sine waves.

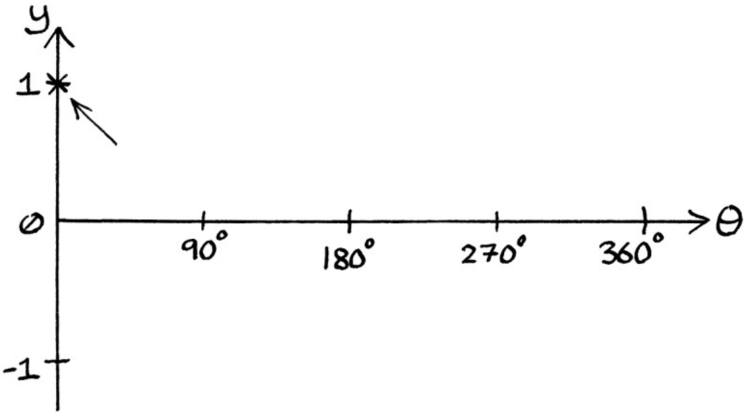
When "θ" is 0, the formula "$i^{(\theta/90)}$" will be i^0, which is 1. As we are using the Complex plane, this is actually "1 + 0i". We can mark the point on the graph:



... and we can mark the y-axis (Imaginary) value, 0, on our Sine wave graph:



... and the x-axis (Real) value, 1, on our Cosine wave graph:



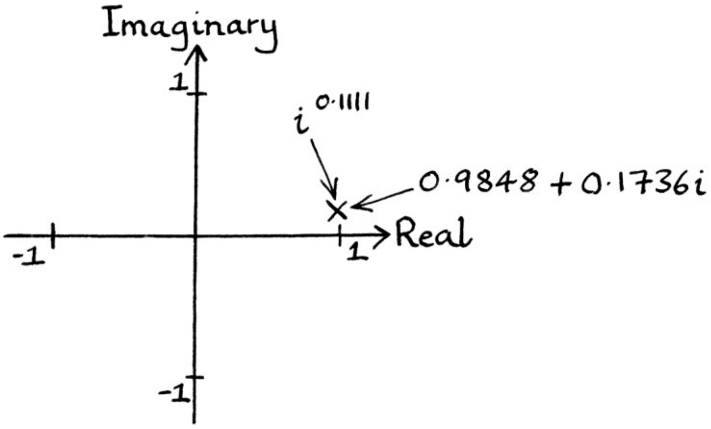
When “θ” is 10 degrees, the formula “$i^{(\theta/90)}$” will be:
$i^{10/90}$

... which is:
$i^{1/9}$

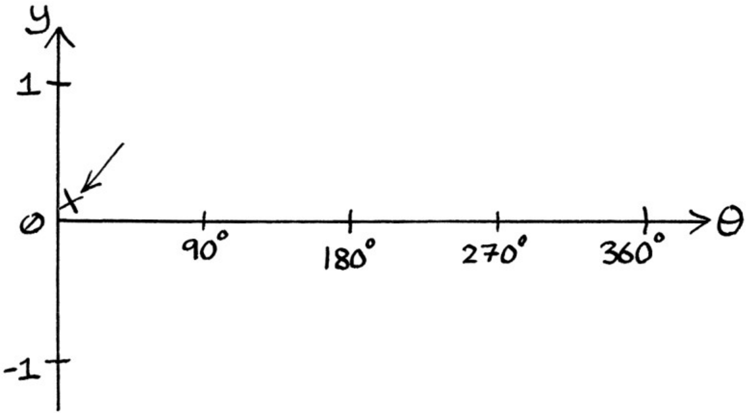
... or:
$i^{0.1111}$

To calculate this point, we will use a calculator, which will tell us that the point is at “0.9848 + 0.1736i”. [We could have calculated this by measuring around a unit-radius circle or by using Sine and Cosine, but we are trying to demonstrate the relationship between the waves and the circle, so for the purpose of this particular explanation, it is more convenient to use a calculator].

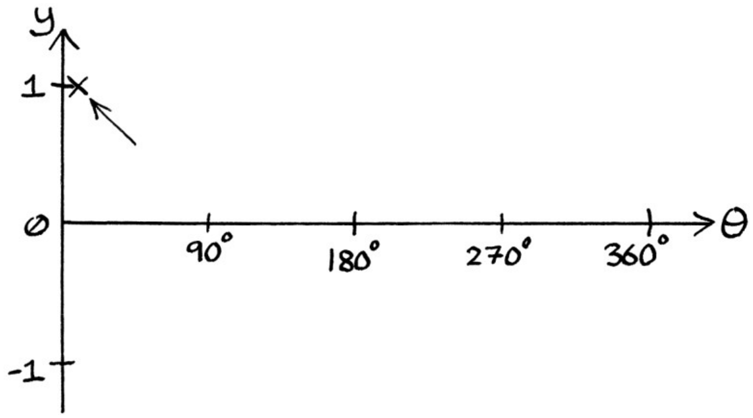
We mark that point on the Complex plane:



... and we can mark the y-axis (Imaginary) value of this point, 0.1736, on the Sine wave graph:



... and the x-axis (Real) value of this point, 0.9848, on the Cosine wave graph:



When “θ” is 20 degrees, the formula “$i^{(\theta/90)}$” will be:

$i^{20/90}$

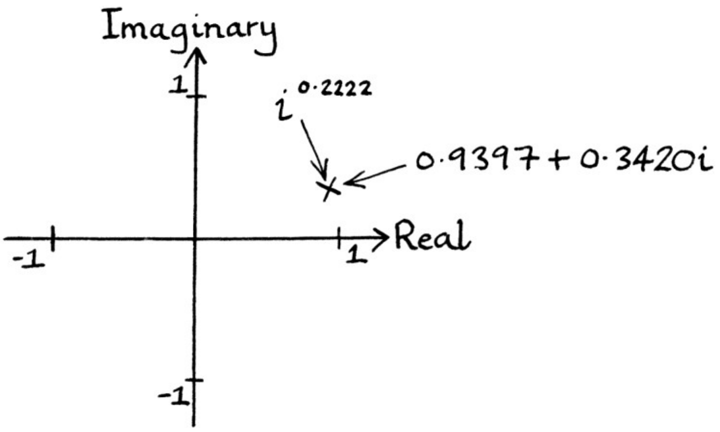
... which is:

$i^{2/9}$

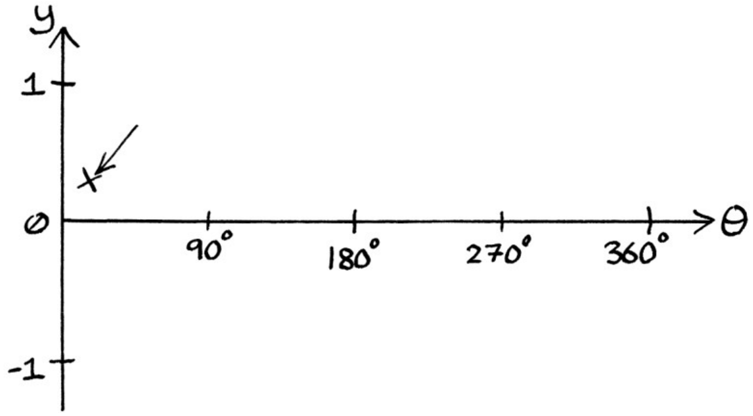
... which is:

$i^{0.2222}$

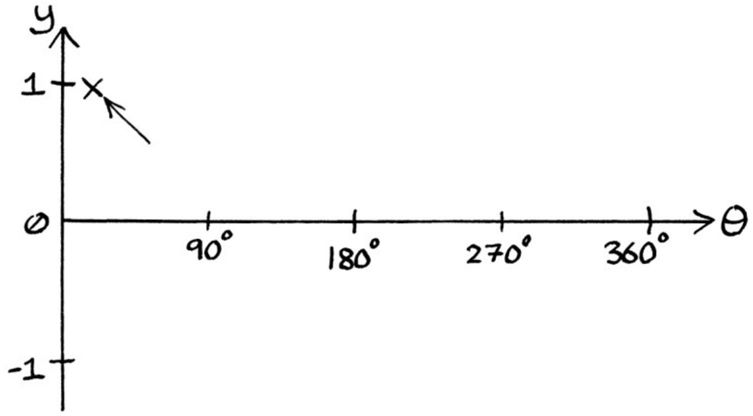
This point is at “0.9397 + 0.3420i”. We plot this point on the circle:



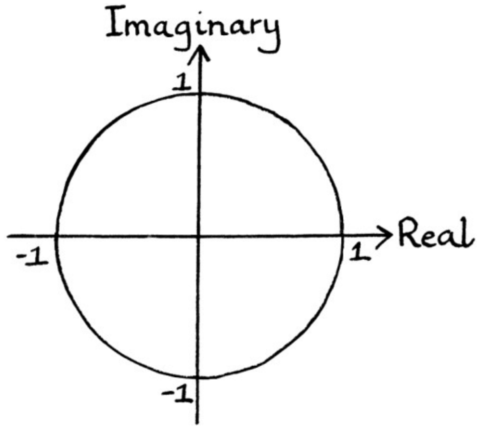
... and the Sine wave graph:



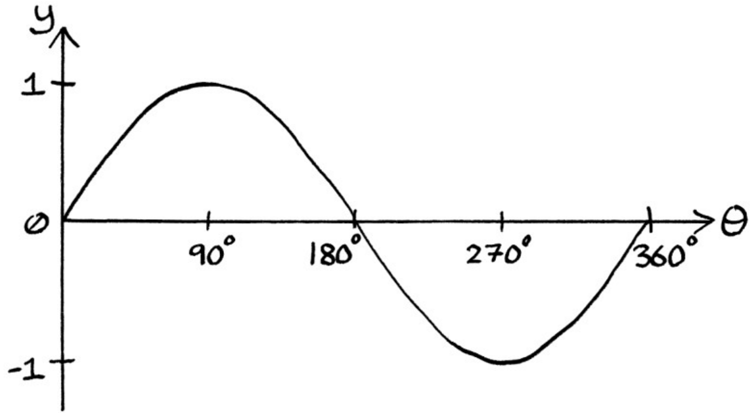
... and the Cosine wave graph:



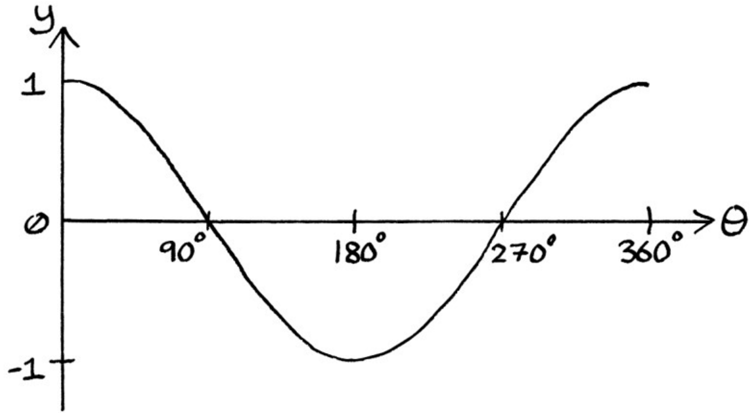
We continue plotting points around the circle and on the wave graphs until we have done one revolution. If we do this with very small angle intervals, we will end up with a full circle and one full cycle of each wave. Unsurprisingly, these appear as so:



... and the Sine wave:



...and the Cosine wave:



Waves with time

Things become more interesting, and slightly more complicated, when we introduce time into the "i^θ" formula – in other words, when we use the "i^θ" formula to indicate the position of an object rotating around a circle.

The "i^θ" formula treats "θ" as an angle in quarter-circle angle units. There are 4 quarter-circle angle units in one circle. To have the "i^θ" formula work on the time and have a default frequency of 1 cycle per second, it will be necessary to multiply the time in seconds by 4 before it is subjected to being a power of "i". This is similar to how we needed to multiply the time in seconds by 360 before it was subjected to the Sine or Cosine functions in degrees, and how we needed to multiply the time by 2π before it was subjected to the Sine or Cosine functions in radians. The multiplication can be thought of as a time correction.

Our formula for an object rotating around a unit-radius circle at a frequency of 1 cycle per second is: “i^{4t}”.

Supposing we had adjusted the “$i^{(\theta/90)}$” formula for degrees to work with time, we would have ended up with the same result. The time would have needed to be multiplied by 360, so the formula would be:
$i^{((360*t)/90)}$
... which ends up as:
i^{4t}

Supposing we had adjusted the “$i^{(2\theta/\pi)}$” formula for radians to refer to time, we would have ended up with the same result again. The time would need to be multiplied by 2π. Therefore, our time formula would be:
$i^{(2*2\pi*t/\pi)}$
... which ends up as:
$i^{(4\pi t/\pi)}$
$= i^{4t}$

The formulas end up as “i^{4t}” because powers of “i” are treated as quarter-circle angle units. Therefore, the time on its own, when being a power of “i”, will be treated as an angle in quarter-circle angle units. For a default frequency of 1 cycle per second, the time must be multiplied by the number of quarter-circle angle units in a circle.

We can say that:
$i^{4t} = \cos 4t + i \sin 4t$
... when “t” is in seconds and Cosine and Sine are working in quarter-circle angle units.

It is also true that:
$i^{4t} = \cos 360t + i \sin 360t$
... when “t” is in seconds and Cosine and Sine are working in degrees.

It is also true that:
$i^{4t} = \cos 2\pi t + i \sin 2\pi t$
... when “t” is in seconds and Sine and Cosine are working in radians.

It is also true that:
$i^{4t} = \cos t + i \sin t$
... when “t” is in seconds and Sine and Cosine are working in the whole-circle angle system.

All of these are true because “i^{4t}” indicates the position of a point rotating around a circle at any particular moment in time. The Cosine and Sine equivalences do the same, but with the differences that Cosine and Sine are working in different angle systems. Another way of explaining this is to say that all of the following are identical. For any time in seconds, each of the following will indicate the same point:

- cos 4t + i sin 4t, when Cosine and Sine are working in quarter-circle angle units.
- cos 360t + i sin 360t, when Cosine and Sine are working in degrees.
- cos 2πt + i sin 2πt, when Cosine and Sine are working in radians.
- cos t + i sin t, when Cosine and Sine are working in whole-circle angle units.

If it still seems confusing how “i^{4t}” can be equivalent to four (and technically countless) variations of Cosine and Sine, it helps to remember that “i^{4t}” indicates the position of a point on the edge of a circle at time “t”. The position of that point is the same however we want to describe it. If we want to describe that point’s position using Cosine and Sine, then we have to remember that Cosine and Sine are functions that operate on values that are treated as angles. There are different angle systems. We can use different angle systems to identify exactly the same point, but Cosine and Sine will need to be working within that particular angle system to produce the relevant y-axis and x-axis values.

We will look at some examples of “i^{4t}” in use.

Example 1

We will say that we want to know the position of an object rotating around a unit-radius circle at 1 cycle per second at the time of 0.75 seconds. We put this into the formula “i^{4t}” and we will have: $i^{4*0.75} = i^3$. Even without putting this into the formula, we know that at 0.75 seconds, the object rotating around the circle will be 0.75 of the way around the circle. After putting it in the formula, we also know by now that “i^3” refers to the point at 270 degrees. The Complex number that identifies this point is “0 – 1i”. Therefore, we know that at 0.75 seconds, the object will be at “0 – 1i”.

The object’s position at “0 – 1i” can also be described using the Cosine and Sine of an angle. If we wanted to indicate the object’s position using coordinates with an angle in degrees, then its coordinates would be (cos 270, sin 270), where Cosine and Sine are operating in degrees. If we wanted to indicate its position with a Complex number, we would have “cos 270 + i sin 270” in degrees.

If we wanted to indicate the object's position with an angle in radians, then its coordinates would be (cos 1.5π, sin 1.5π), where Cosine and Sine are operating in radians. As a Complex number, this would be "cos 1.5π + i sin 1.5π" in radians.

If we wanted to describe the object's position using whole-circle angle units, its coordinates would be (cos 0.75, sin 0.75), where Cosine and Sine are operating in this system. The Complex number would be "cos 0.75 + i sin 0.75" in whole-circle angle units.

Example 2

We will consider the same object rotating around the unit circle at 1 cycle per second. This time we want to know its position at 3.14 seconds. We put this into the formula "i^{4t}" and we will have: $i^{4*3.14} = i^{12.56}$. In other words, it will be 12.56 quarter-circle angle units around the circle at 3.14 seconds. Using a calculator that can work with Complex numbers, we will see that "$i^{12.56}$" is "0.6374 + 0.7705i".

We could also have solved this by knowing that at 3.14 seconds, it would be in the same place as at 0.14 seconds [because it completes one revolution of the circle every second]. At 0.14 seconds, it would be 0.14 of the way around the circle. This is the same as being at 0.14 * 360 degrees, which is 50.4 degrees. We can then use Cosine and Sine in degrees to find out its position.

This point can also be described as:
"cos (360 * 3.14) + i sin (360 * 3.14)", in degrees, which, as expected, is also equal to "6.374 + 0.7705i".

It can also be described as:
"cos (2π * 3.14) + i sin (2π * 3.14)" in radians, which is also "6.374 + 0.7705i".

It can also be described as:
"cos 3.14 + i sin 3.14" in whole-circle angle units, which is also "6.374 + 0.7705i".

Example 3

If the same object has been travelling for 0.6 seconds, we can portray that fact using "$i^{(4*0.6)}$", which is "$i^{2.4}$". We can calculate where this is in several ways. One is by converting 2.4 quarter-circle angle units into degrees and using Cosine and Sine to calculate the x and y-axis coordinates. Another is just to solve it on a calculator that can work with Complex numbers. Either way will result in the Complex number "−0.8090 – 0.5878i".

This point can also be described using:
"cos (4 * 0.6) + i sin (4 * 0.6)", when Cosine and Sine are working in quarter-circle angle units
...or:
"cos (360 * 0.6) + i sin (360 * 0.6)", when Cosine and Sine are working in degrees
... or:
"cos (2π * 0.6) + i sin (2π * 0.6)", when Cosine and Sine are working in radians
... or:
"cos 3.14 + i sin 3.14", when Cosine and Sine are working in whole-circle angle units.

Advantages of thinking of circles in this way

Previously, when we had an object rotating around a circle, we had to describe it using both Sine and Cosine. We always needed to use the two corresponding waves derived from a circle to describe that circle. For example, in earlier chapters, when I described a circle, I had to describe it by saying something such as, "The circle based on the waves 'y = sin 360t' and 'y = cos 360t'." Later, when we had learnt about Complex numbers, we could describe it by saying something such as "z = cos 360t + i sin 360t."

Now we can describe the circle and an object moving around it using a more concise formula: "i" raised to a power. This formula encapsulates both waves. Although it might not appear to use angles, it always uses values that are being treated as quarter-circle angle units.

As the formula is also an exponential (in other words, a number raised to the power of another number), there are certain mathematical processes that are easier to perform on it than if it were in a different form. I explain more about exponentials in Chapter 26.

Amplitude, frequency, phase and mean level

In its current form, the formula “i^{4t}” relates to an object moving around a unit-radius circle at 1 cycle per second, with no phase, and the circle is centred on the origin of the axes. The Cosine and Sine waves derived from the circle have an amplitude of 1, a frequency of 1, a phase of zero, and a mean level of zero. However, it is possible to incorporate all the attributes of a time-based wave into the formula.

Radius or amplitude

If we want to change the radius of the circle, we can just scale the circle by multiplying the formula by a number.

For example, if we want to have a radius of 2 units, and therefore amplitudes of 2 units in the derived waves, we multiply the formula by 2, and have: 2 *i^{4t}, which is $2i^{4t}$.

[If you are not used to how exponentials are written out, it is important to be aware that $2i^{4t}$ is really 2*(i^{4t}) and not $(2i)^{4t}$].

In degrees, the derived waves from this circle would be “2 cos 360t” and “2 sin 360t”. In radians, the waves would be “2 cos 2πt” and “2 sin 2πt”. The position of an object rotating around the circle at any particular moment in time would be “2 cos 360t + 2i sin 360t” in degrees, or “2 cos 2πt + 2i sin 2πt” in radians.

A general formula for circles that takes into account radius (which is the amplitude of the derived waves) is:
$z = a * i^{4t}$
... which can also be written as:
$z = ai^{4t}$
... where “a” is the radius (or amplitude), and “z” is the specific point on the Complex plane indicated by the formula at one instant in time.

[The use of “z”, in this case, is similar to the use of “y” when we are talking about waves. We might describe a wave as “y = sin 360t”, in which case “y” refers to the position on the y-axis. As the Complex plane has two axes (Real and Imaginary, or x and y), we cannot use “y”, and we have to use a symbol that means the position along two axes.]

Frequency

If we want to have a frequency other than 1 cycle per second, we simply scale the value of “t” in the formula.

For example, if we want a frequency of 5 cycles per second, we would use the formula:
“$z = i^{(4 * 5t)}$”
... which could also be written as:
“$z = i^{20t}$”

However, it is easier to know what the frequency is if we keep the “4” separate from the frequency.

In degrees, the derived waves from the “$z = i^{(4 * 5t)}$” circle would be:
“cos (360 * 5t)”
... and:
“sin (360 * 5t)”

In radians, the waves would be:
“cos (2π * 5t)”
... and:
“sin (2π *5t)”

The position of an object rotating around the circle at any particular moment in time would be:
“cos (360 * 5t) + i sin (360 * 5t)” in degrees
... or:
“cos (2π * 5t) + i sin (2π * 5t)” in radians.

A general formula for circles that take into account frequency is:
“$z = i^{4ft}$”
... where “f” is the frequency in cycles per second.

Phase

Phase in this situation is slightly more complicated than amplitude or frequency. When we included phase in the formulas for Sine and Cosine waves, we included it as an angle in degrees or radians. With the "i^{4t}" formula, we are working in a system that divides a circle up into 4 portions, and therefore, it uses quarter-circle angle units. Therefore, the phase needs to be given in this quarter-circle system. In this case, a full circle is 4 quarter-circle angle units; three-quarters of a circle is 3 quarter-circle angle units; half a circle is 2 quarter-circle angle units; and quarter of a circle is 1 quarter-circle angle unit.

To indicate phase in the formula, we add the phase angle to the corrected time. For example, if we want a phase of 1.5 quarter-circle angle units, the formula would be:
"$i^{(4t + 1.5)}$"

If we want a phase equivalent to a number of degrees, we have to convert it into the quarter-circle angle system first. To do this we divide it by 360 to give the portion of a circle represented by that angle, and then we multiply it by 4. Therefore, if we wanted a phase of 23 degrees, this would be: (23 ÷ 360) * 4 = 0.2556. In the formula, this would appear as:
"$i^{(4t + 0.2556)}$"

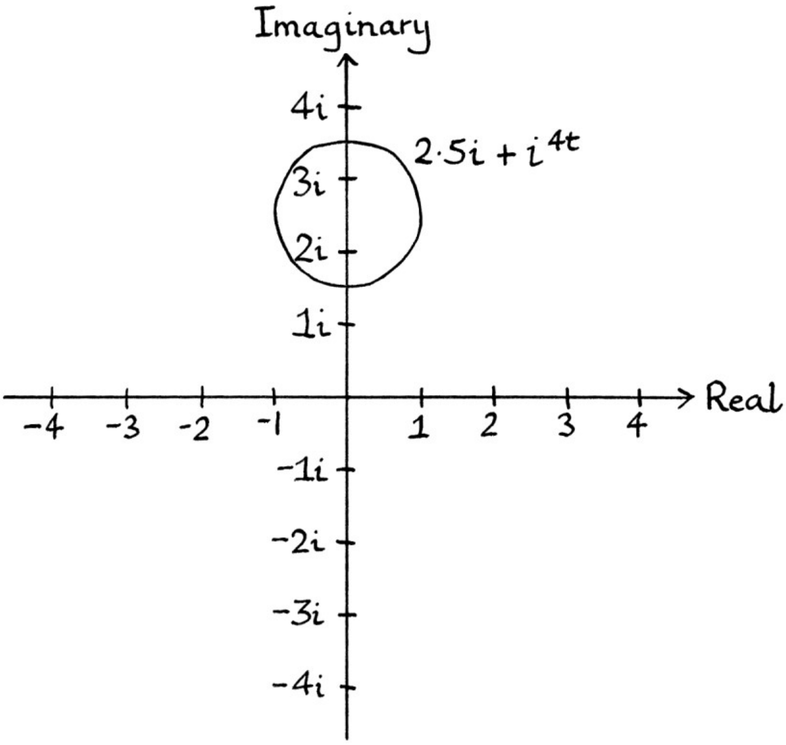
If we want a phase equivalent to a number of radians, we also have to convert it into the quarter-circle angle system. To do this we divide it by 2π to give the portion of a circle represented by that angle, and then we multiply it by 4. Therefore, if we wanted a phase of 3.7 radians, this would be: (3.7 ÷ 2π) * 4 = 2.3555. In the formula, this would appear as:
"$i^{(4t + 2.3555)}$"

A general formula for circles that take into account phase is:
"$z = i^{(4t + \phi)}$"
... where "ϕ" is the phase in *quarter-circle angle units*.

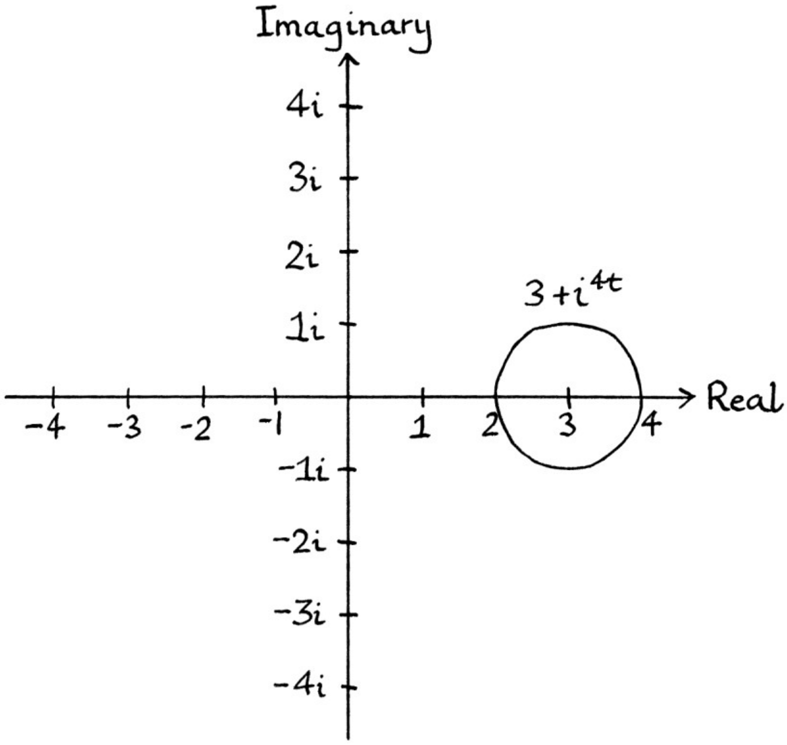
Mean levels

The mean levels for a circle shift it upwards or downwards along the y-axis for the mean level for the derived Sine wave, and to the left or right along the x-axis for the mean level for the derived Cosine wave. As our circle "i^{4t}" is on the Complex plane, we can shift the whole circle up or down by adding an Imaginary number, and we can shift it left or right by adding a Real number.

For example, if we wanted our derived Sine wave to have a mean level of 2.5 units, our formula would be "$2.5i + i^{4t}$". The whole circle will be shifted up the Imaginary axis by 2.5 units, and therefore, the derived Sine wave will be shifted up its y-axis by 2.5 units.



If we want our derived Cosine wave to have a mean level of 3 units, our formula will be "3 + i^{4t}". The whole circle will be shifted to the *right* along the Real axis by 3 units, and therefore, the derived Cosine wave will be shifted up its y-axis by 3 units.



A general formula that takes into account both mean levels is:
"z = $h_c + ih_s + i^{4t}$"
... where:

- "h_c" is the mean level for the Cosine wave (the x-axis position of the centre of the circle).
- "h_s" is the mean level for the Sine wave (the y-axis position of the centre of the circle).

All Attributes Together

If we include all the attributes together, we will have:
"z = $h_c + ih_s + ai^{(4ft + \phi)}$"

Maths with circles

We can perform addition with circles described with variations of "i^{4t}". As "i^{4t}" incorporates both the Sine and Cosine waves at the same time, sometimes this can be easier than adding the waves separately.

One thing to note is that if the result can be expressed as one power of "i", then it is one circle, and the derived waves will be pure waves. If the result cannot be expressed as a single power of "i", the overall shape will not be a circle, and therefore, the derived waves will not be pure waves.

Addition of amplitudes

To add circles with the same frequency and phase, and zero mean level, we just add the values scaling the circles.

If we add the circles "$z = 2i^{4t}$ and "$z = 5i^{4t}$", we end up with "$z = 7i^{4t}$". This result is still a circle, so the derived waves will still be pure waves.

The underlying waves for "$z = 2i^{4t}$" are "y = 2 sin 360t" and "y = 2 cos 360t" (in degrees), or "y = 2 sin 2πt" and "y = 2 cos 2πt" (in radians).

The underlying waves for "$z = 5i^{4t}$" are "y = 5 sin 360t" and "y = 5 cos 360t", in degrees, or "y = 5 sin 2πt" and "y = 5 cos 2πt" in radians.

In degrees, the sum of the two Sine waves is (2 sin 360t) + (5 sin 360t) = 7 sin 360t. The sum of the two Cosine waves is (2 cos 360t) + (5 cos 360t) = 7 cos 360t.

In radians, the sum of the two Sine waves is (2 sin 2πt) + (5 sin 2πt) = 7 sin 2πt. The sum of the two Cosine waves is (2 cos 2πt) + (5 cos 2πt) = 7 cos 2πt.

The summed Sine waves and summed Cosine waves are the derived waves of the resulting circle:
"$z = 7i^{4t}$"

By dealing in circles described by powers of "i", we have *slightly* simplified addition of two waves. For one thing, there is much less typing when waves are treated as circles.

Addition of mean levels

To add circles with the same frequency and phase, and non-zero mean levels, we just add the radiuses (amplitudes) together and the mean levels together.

If we add:
"$z = 5 + 8i + i^{4t}$"
... and:
"$z = 2 + 1i + i^{4t}$"
... we will have:
"$z = 7 + 9i + 2i^{4t}$"

The derived waves of the two *original* circles are:
"$y = 8 + \sin 2\pi t$" and "$y = 5 + \cos 2\pi t$" in radians, or "$y = 8 + \sin 360t$" and "$y = 5 + \cos 360t$" in degrees
... and:
"$y = 1 + \sin 2\pi t$" and "$2 + \cos 2\pi t$" in radians, or "$y = 1 + \sin 360t$" and "$2 + \cos 360t$" in degrees.

The sum of the two derived Sine waves is:
"$y = 9 + 2 \sin 2\pi t$" in radians, or "$y = 9 + 2 \sin 360t$" in degrees

The sum of the two derived Cosine waves is:
"$y = 7 + 2 \cos 2\pi t$", or "$y = 7 + 2 \cos 360t$" in degrees

... and these summed waves are the derived waves of the result, "$z = 7 + 9i + 2i^{4t}$", as they should be.

Addition of phase: method 1

Adding phases with circles described by powers of "i" is not much simpler than adding phases of waves. It is still necessary to do some maths. One thing to remember is that if the frequencies are the same, then the result will still be a circle, and therefore can be represented by a single power of "i".

As an example, we will add the circles described by "$z = i^{(4t + 1.4)}$" and "$z = i^{(4t + 2.2)}$". [Remember that the phases are given in quarter-circle angle units]. There are two ways we can do this. The first way is to reduce the powers of "i" to waves, and calculate the resulting phase in that way:

The underlying waves of the two original circles are, *in radians*:
"y = sin (2πt + 2.1991)" and "y = cos (2πt + 2.1991)"
... and:
"y = sin (2πt + 3.4558)" and "y = cos (2πt + 3.4558)"

[Remember that the power of "i" has a phase based on a system that divides a circle into 4 parts. Therefore, to find the angle in radians, we need to calculate the portion of a circle that that quarter-circle angle unit represents by dividing the value by 4, and then we find out how many radians that is by multiplying it by 2π (or how many degrees that is by multiplying it by 360)].

I explained how to add phases in Chapter 14. The basic idea is to imagine the waves as circles, with one circle rotating around the phase point of the other. To calculate the phase of the sum, we just work out the position of the outer circle's phase point at t = 0. The overall phase will be the angle of the outer object from the origin at t = 0. To calculate the angle, we first calculate its x-axis and y-axis value. These will be (using radians):
cos ((2π * 0) + 2.1991) + cos ((2π * 0) + 3.4558) = −1.5388 units.
sin ((2π * 0) + 2.1991) + sin ((2π * 0) + 3.4558) = 0.5 units.

Therefore, the point is at the coordinates (−1.5388, 0.5). We can calculate the angle as arctan (0.5 ÷ −1.5388) = −0.3142 radians. First, we will convert this to a positive angle to make things clearer. It is −0.3142 + 2π = 5.9690 radians. As we are using arctan, we need to check that this is the angle that we want, out of the two possible angles that would create the gradient. As our point is in the top left quarter of the circle, but this result is in the bottom right quarter of the circle, we actually want the other angle, half way around the circle, which is 5.9690 − π = 2.8274 radians. This can also be written as 0.9π radians.

We then need to work out the distance from the origin of the outer phase point. This will be:
$\sqrt{1.5388^2 + 0.5^2}$ = 1.6180 units.

Therefore, the phase of the resulting circle is 0.9π radians; the amplitude is 1.6180. We need to convert the angle of 0.9π radians into quarter-circle angle units. It is 1.8 quarter-circle angles.

The resulting circle is "z = $1.6180i^{(4t + 1.8)}$"

Addition of phase: method 2

The second method of calculating phase is essentially the same, but quicker. We are adding the circles "$z = i^{(4t + 1.4)}$" and "$z = i^{(4t + 2.2)}$".

We find out the positions of the objects rotating around each circle at t = 0. We will use a calculator that can work with Complex numbers to save time, but we could just as easily do this with Sine and Cosine.

For the first circle, this will be: $i^{((4 * 0) + 1.4)} = i^{1.4} = -0.5878 + 0.8090i$
For the second circle, this will be: $i^{((4 * 0) + 2.2)} = i^{2.2} = -0.9511 - 0.3090i$

The outer phase point at t = 0 will be at the sum of the two points:
$-0.5878 + 0.8090i + -0.9511 - 0.3090i = -1.5388 + 0.5i$

[Note that we could have calculated the total by just adding the powers of "i" on a calculator that can work with Complex numbers.]

The angle of this point is arctan $(0.5 \div -1.5388) = -0.3142$ radians $= +5.9690$ radians. We want the other angle for this gradient: $5.9690 - \pi = 2.8274$ radians. Converting this to quarter-circle angle units, this is $(2.8274 \div 2\pi) * 4 = 1.8$ quarter-circle angle units. The distance of the point from the origin is $\sqrt{1.5388^2 + 0.5^2} = 1.6180$.

Therefore our resulting power of "i" is: $1.6180i^{(4t + 1.8)}$, which is the same result as before.

Addition of different frequencies

Addition of different frequencies for circles described by powers of "i" will end up with shapes that are not circles, and the derived waves will not be pure waves. From this, we can tell that the result will be two or more powers of "i".

For example, $i^{(4 * 2t)}$ added to $i^{(4 * 6t)}$ results in "$i^{(4 * 2t)} + i^{(4 * 6t)}$" or "$i^{8t} + i^{24t}$". As with adding circles and waves of different frequencies normally, we cannot portray the result in a more concise form.

Although the results of such additions might seem slightly pointless, they allow us to find the position of the object rotating around the outer circle at any moment in time just by putting the relevant time into the formula. For example, at t = 3.7 seconds, the outer object will be at:
$i^{(8 * 3.7)} + i^{(24 * 3.7)} = i^{29.6} + i^{88.8} = -0.5 + 1.5388i$.

Calculating this using maths on waves would be more effort. Adding circles with different frequencies in this way remains fairly simple, no matter how many circles we are adding together.

Multiplication

In the previous chapter, we saw how a circle can be described as a Complex number such as "cos (360 * ft) + i sin (360 * ft)". We also saw that if multiplication is performed with two or more such circles, the connection between the circle and its underlying waves breaks down. If we multiply two circles being described with Complex numbers, the result will not match that of multiplying those circles if they are described by their derived waves. This problem also occurs when we perform multiplication with two or more circles described as powers of "i". This might be expected, given that we can say that:
$i^{4t} = \cos 4t + i \sin 4t$
... when "t" is in seconds and Cosine and Sine are working in quarter-circle angle units, or that:
$i^{4t} = \cos 360t + i \sin 360t$
... when "t" is in seconds and Cosine and Sine are working in degrees, or that:
$i^{4t} = \cos 2\pi t + i \sin 2\pi t$
... when "t" is in seconds and Sine and Cosine are working in radians.

If we try to multiply circles using powers of "i", the results will be wrong, in the sense that the result will not represent the result of multiplying the derived waves. We cannot multiply two circles together when using powers of "i".

As we will see in the next chapter, if we are multiplying two exponentials with the same base, we can calculate the result merely by adding the exponents. [The "base" is the part of an exponential that is raised to a power. With "c^x", "c" is the base, and "x" is the exponent.] If we have c^x multiplied by c^y, the result will be $c^{(x + y)}$.

Thinking back to Chapter 16 on the multiplication of waves, if we multiply two pure waves of the same frequency with zero mean levels, we will end up with one pure wave with a non-zero mean level. Given that the multiplication of two exponentials results in an exponential with its exponent as the sum of the two original exponents, it might be clear that no new mean level will ever arise from a multiplication of exponentials. Therefore, without needing to do any maths, we know that multiplication of circles cannot work by multiplying powers of "i".

As an example, we will multiply "$i^{(4 * 2t)}$" by itself. This represents an object rotating around a unit-radius circle at 2 cycles per second. Its derived waves are:
"y = sin (2π * 2t)" and "y = cos (2π * 2t)".

When we multiply "$i^{(4 * 2t)}$" by itself, we end up with:
$i^{(4 * 2t)} * i^{(4 * 2t)}$
$= i^{((4 * 2t) + (4 * 2t))}$
$= i^{(8t + 8t)}$
$= i^{16t}$
$= i^{(4 * 4t)}$

This represents an object rotating around a unit-radius circle at 4 cycles per second. However, if we were multiplying "y = sin (2π * 2t)" by itself, we would end up with a pure wave with a non-zero mean level:
"y = 0.5 + 0.5 sin ((2π * 4t) + 1.5π)"

Likewise, if we multiplied "y = 2 cos (2π * 2t)" by itself, we would end up with a single pure wave with a non-zero mean level:
"y = 0.5 + 0.5 cos (2π * 4t)"

The result of squaring "$i^{(4 * 2t)}$" does not match the results of squaring its derived waves. We could try other examples, and they would just confirm that multiplication of powers of "i" produces results that are unconnected to multiplications of the derived waves.

[We can still multiply circles by numbers and the results will be consistent with multiplying the derived waves by those numbers.]

Conclusion

In this chapter, we saw how it is possible to rotate any point on the Complex plane by any amount anticlockwise, by multiplying its Complex number by a relevant power of "i".

Doing that allows us to identify any point on the Complex plane, by indicating by how much the point at "1 + 0i" would need to be rotated and scaled to get there. We can indicate the rotation and scaling with a power of "i".

We can also describe a circle using the rotational properties of a power of "i", and, therefore, we can also imply the derived Sine and Cosine waves from that circle.

We can incorporate time into the power of "i" to indicate an object rotating around a circle, and thus imply the derived time-related Sine and Cosine waves.

We can perform addition with circles described using powers of "i", and in some cases, this is simpler than addition with the derived waves.

Generally in maths, you will not see people use powers of "i" to describe circles. Instead, they will use 2.71828123... (the number "e") raised to the power of multiples of "i". In other words, they will use "$e^{i\theta}$". This works in essentially the same way, but operates in radians as opposed to quarter-circle angle units. With "$e^{i\theta}$", it is common for people to use it without understanding it at all. If you understand "i" raised to powers, then you will find "e" raised to Imaginary powers straightforward.

People often say that radians are a "natural" way to divide a circle – the use of radians is dictated by mathematical procedures, as opposed to being a socially constructed arbitrary choice. In this chapter, it should also be apparent that quarter-circle angle units are also a "natural" way to divide a circle.

Whether it is useful to describe a circle by using powers of "i" depends on what it is we are trying to achieve. For now, probably the most useful aspect of using powers of "i" is as a stepping stone to understanding Imaginary powers of "e".

Chapter 26: Exponentials

This chapter is a simple explanation of basic maths involving exponentials, roots and logarithms. It is for the benefit of people who do not know much about them, or who have forgotten what they once knew. It will be useful for understanding "e" in later chapters. I have used exponentials before in this book, which means that this chapter will explain some things that I presumed were already known before now. This chapter misses out some important details (such as the exceptions for some of the rules), and includes some details that you might never need to know. If you need to learn about exponentials for a formal education, there are probably better explanations than this one. It is useful to know that some of the rules mentioned in this chapter exist, but you do not need to memorise them. At times in this chapter, I use a larger font to make some symbols clearer.

Exponentials

An exponential is, in essence, an abbreviated way of writing a calculation where a number is multiplied by itself a particular number of times. The most common form of an exponential is the square of a number, which is when a number is multiplied by itself just once. If we wanted to portray the calculation "3 * 3", we could write it in the form "3^2", where "3^2" is called an exponential. If we wanted to portray the multiplication "4 * 4 * 4 * 4 * 4", we could write it in the form "4^5". In the full multiplication, the number 4 appears 5 times. The benefits of shortening a repeated multiplication to an exponential are that exponentials are quicker to write, and they can be used to portray complicated concepts succinctly.

In an exponential, the number that is multiplied by itself is called the "base". The number to which the "base" is raised is called the "exponent". In the exponential "3^4", the number 3 is the base and the number 4 is the exponent. The base becomes multiplied by itself, so that if the multiplication were written out in full, the base would be written the number of times specified by the exponent. For the exponential "3^4", the full multiplication would be 3 * 3 * 3 * 3.

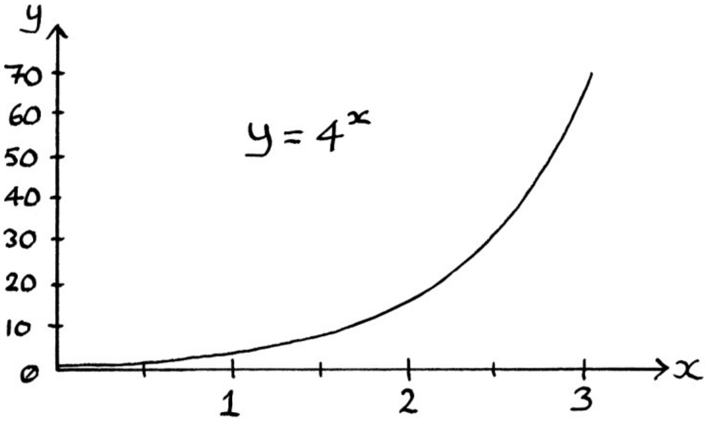
The "exponent" is also sometimes called the "power", which leads to the phrase that one number "is raised to the power" of the other. For example, with the exponential "5^7", the number 5 is "raised to the power" of 7. The terms "exponent" and "power" are often used interchangeably with neither generally being considered better than the other.

To reiterate the meaning of the terms, we can say such things as:
$\text{base}^{\text{exponent}}$
$\text{base}^{\text{power}}$
"the base is raised to the power of the exponent"

If we were typing formulas as text and we could not use a superscript (as in raised-up text), we could use the "^" symbol as a replacement. For example, we could write "2^3" to mean "2^3", which in turn means 2 to the power of 3, or 2 cubed.

Although the result of a simple exponential, such as 10^3, is easy to calculate and understand, some exponentials can be quite complicated, depending on the nature of the base and exponent.

More complicated exponentials can be easier to visualise on a graph. For example, this graph shows the various powers of "4". The graph's formula is "$y = 4^x$":



General rules

Powers of positive integers

Any positive or negative base, whether an integer or not, when raised to the power of a positive *whole* number from 1 upwards, is the same as that base being multiplied by itself that number of times. In other words, "c^x", where "x" is any integer from 1 upwards, means "c * c * c * c * and so on..."

If we have 5^7, it means 5 * 5 * 5 * 5 * 5 * 5 * 5.
If we have 5^2, it means 5 multiplied by 5.
If we have 1.1^3, it means 1.1 * 1.1 * 1.1.
If we have -7^2, it means −7 * −7.

Powers of one

Following on from that, it might be clear that any value, whether negative, positive, an integer or non-integer, when raised to the power of 1, results in itself. It is the same as "c * c * c..., but with only one "c".

Therefore, if we have 5^1, the result is just 5.
If we have 23.0231^1, the result is 23.0231.

Powers of zero

Any value at all, whether negative, positive, an integer, or a non-integer, when raised to the power of zero results in 1.

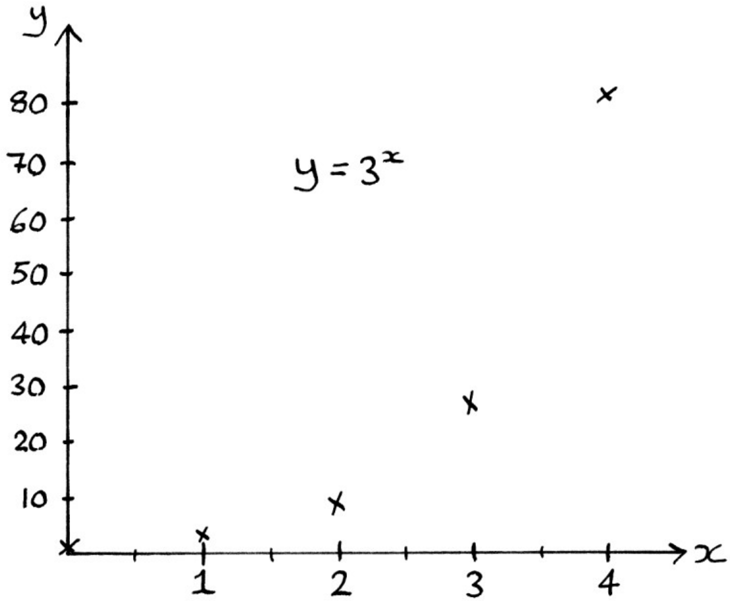
If we have -0.001^0, the result is 1.
If we have 5^0, the result is 1.
If we have 101^0, the result is 1.
If we have 12.3423^0, the result is 1.

Rules for bases from 1 upwards

The following rules apply to positive bases from 1 upwards, whether they are integers or not. In other words, the rules apply to such numbers as 1^x, 2^x, 2.56^x, $1,000,000^x$ and so on, where we will focus on the value of "x".

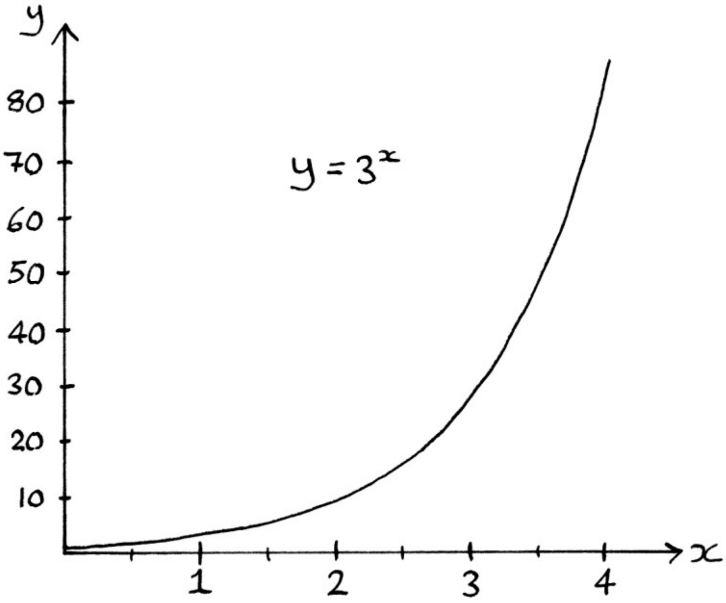
Positive non-integer exponents

The meaning of exponentials becomes harder to visualise when dealing with bases from 1 upwards that have positive exponents over 1 that are not integers – in other words, exponentials such as $4^{1.5}$, $12^{7.07}$ or $3^{2.5}$.
It is much harder to visualise what 3 to the power of 2.5 means, than it is to visualise what 3 to the power of 2 means. We cannot write out "3 * 3 *..." and have half a 3. However, we can figure out what it means by drawing a graph showing 3 raised to the power of every positive integer from 0 upwards. We can draw such a graph by plotting the integer values of "x" in the formula "y = 3^x":

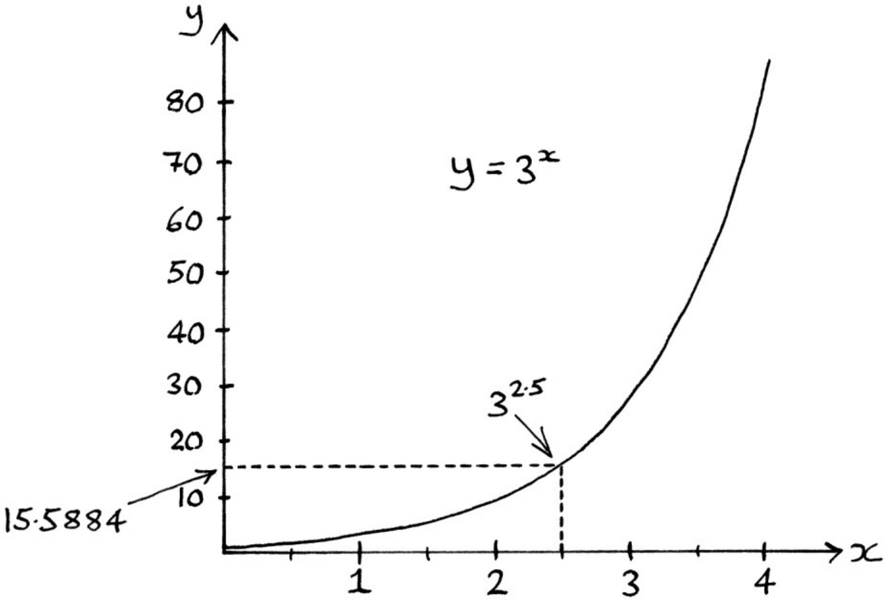


... and then joining those points up.

This is harder than it sounds, but I am trying to explain what the concept means more than I am trying to explain the best way of obtaining the result].



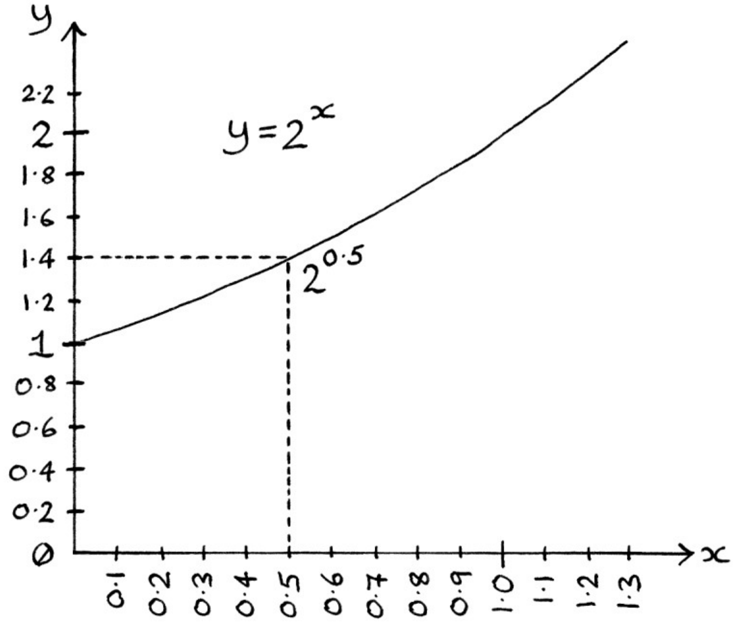
From the joined up graph, we can read off the y-axis value for when “x” is 2.5. If we could read the graph with a reasonable level of accuracy, we would see that $3^{2.5}$ is roughly 15.6. A calculator will give the result as 15.58845727 to 8 decimal places.



Therefore, although it might seem odd to have a number raised to a value that is not an integer, it still makes sense, and it has a valid result. In everyday life, we would use a calculator to do this, instead of drawing a graph.

Positive non-integers between 0 and 1 as exponents

We can solve an exponential with an exponent that is a fraction between 0 and 1 by just examining the entire graph of that base to the power of x. [Of course, we would have the problem of creating such a graph in the first place.] For example, the result of $2^{0.5}$ can be seen on the graph of "$y = 2^x$". It is 1.4142.



However, that does not reveal the interesting properties of such exponents. The interesting properties are revealed if we calculate some examples on a calculator:
$2^{0.5} \approx 1.4142$ = the square root of 2 = $\sqrt{2}$
$2^{0.3333} \approx 1.2599$ = the cube root of 2 = $\sqrt[3]{2}$
$2^{0.25} \approx 1.1892$ = the 4th root of 2 = $\sqrt[4]{2}$
$2^{0.2} \approx 1.1487$ = the 5th root of 2.
$2^{0.125} \approx 1.0905$ = the 8th root of 2.

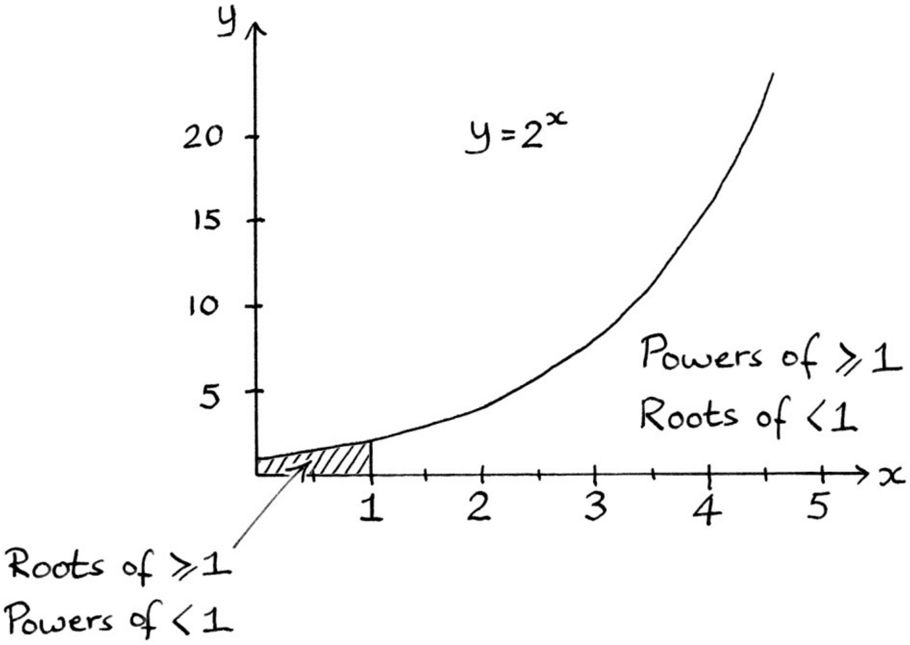
It turns out that for exponents that are fractions between 0 and 1, the exponential is really another way of indicating a root of the base – in other words, a square root, a cube root and so on. In this way, "$c^{0.5}$" means the same thing as "$\sqrt{c}$ ". The exponential "$c^{0.33333...}$" means the same thing as the cube root of "c" or $\sqrt[3]{c}$. The exponential "$c^{0.25}$" is the same as the fourth root of "c".

To convert between an exponent between 0 and 1, and a root, we work out the reciprocal of the exponent, and the type of the root will be that number. For example, if we have the number $3^{0.5}$, the exponent is 0.5. The reciprocal of that is 1 ÷ 0.5 = 2. Therefore, $3^{0.5}$ is the same as the 2nd root of 3, which is another way of saying, the square root of 3.

If we have the number $12^{0.125}$, then its exponent is 0.125. The reciprocal of that is 1 ÷ 0.125 = 8. Therefore, $12^{0.125}$ is the same as the 8th root of 12, which we can write as $\sqrt[8]{12}$.

We can extend the idea to take into account exponents over 1, in which case, we will end up with fractional roots. For example:
3^2 is the 0.5th root of 3, or $\sqrt[0.5]{3}$
6^3 is the 0.3333th root of 6, or $\sqrt[0.3333]{6}$

We can mark the way that exponents and roots match on a graph such as "y = 2^x":



Negative values as exponents

A negative exponent, such as c^{-x}, is another thing that seems confusing at first glance, but is less confusing when we see how it works. If we have the number 3^{-2}, we cannot work it out by writing out the threes as so: 3 * 3 * 3 *... , so we need another method.

It is easiest to think of what such an exponential means by looking at some examples. We can try these on a calculator.

$2^2 = 4$

$2^1 = 2$

$2^0 = 1$

$2^{-1} = 0.5$

$2^{-2} = 0.25$

$2^{-3} = 0.125$

... or:

$3^2 = 9$

$3^1 = 3$

$3^0 = 1$

$3^{-1} \approx 0.3333$

$3^{-2} \approx 0.1111$

$3^{-3} \approx 0.03704$

What is happening is more obvious if we see what the results are equivalent to:

$2^2 = 4$	$= 1 \div 0.25$	
$2^1 = 2$	$= 1 \div 0.5$	
$2^0 = 1$	$= 1 \div 1$	$= 1 \div 2^0$
$2^{-1} = 0.5$	$= 1 \div 2$	$= 1 \div 2^1$
$2^{-2} = 0.25$	$= 1 \div 4$	$= 1 \div 2^2$
$2^{-3} = 0.125$	$= 1 \div 8$	$= 1 \div 2^3$

From this, it might be clear that a negative exponent is the same as the reciprocal of that whole exponential if it had had a *positive* exponent. If we have an exponential with a negative exponent, we can give the exponential a positive exponent and divide 1 by it, and the result will be the same. In other words:

$$c^{-x} = (1 \div c^x)$$

Examples are:
$3^{-3} = 1 \div 3^3 = 1 \div 27 \approx 0.03704$
$100^{-2} = 1 \div 100^2 = 1 \div 10{,}000 = 0.0001$.
$1.34^{-0.25} = 1 \div 1.34^{0.25} \approx 0.9294$.

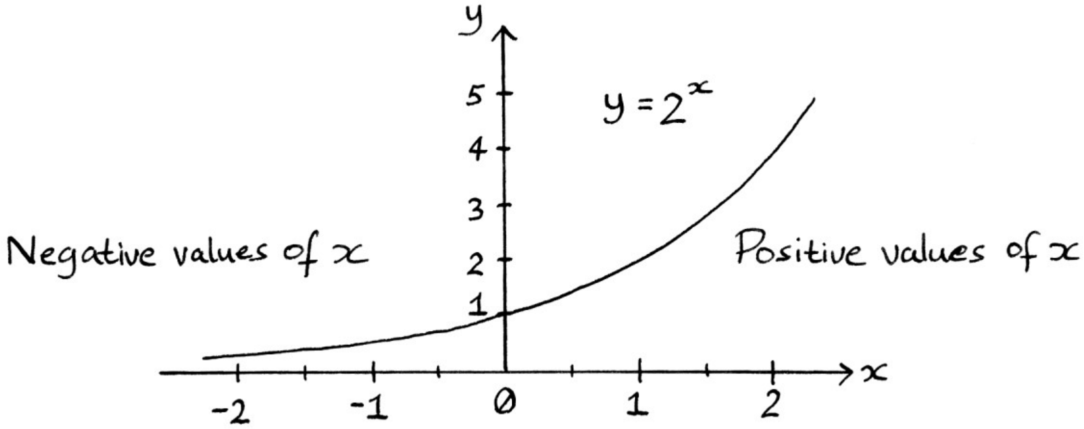
This is straightforward once you see how it works.

The rule works in reverse too:

$$c^x = 1 \div c^{-x}$$

For example, $3^3 = 1 \div 3^{-3} = 1 \div 0.03704 = 27$.

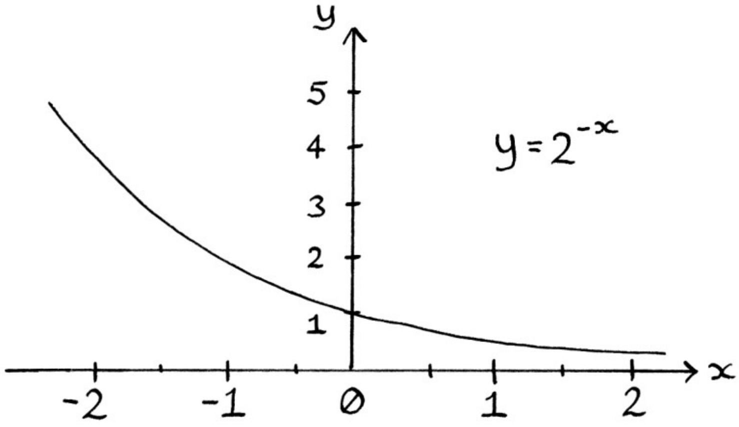
If we were to extend the graph for "$y = 2^x$" so that there were negative values of x, the graph would show negative values as exponents:



If we could accurately read y-axis values off the graph, we would see things such as:

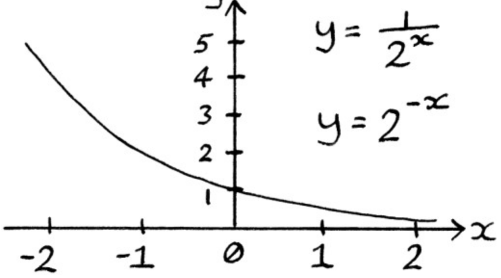
- When x is −1, y is $2^{-1} = 0.5$
- When x is −2, y is $2^{-2} = 0.25$
  ... and so on.

We can also plot the graph for an exponential with a negative exponent. For example, the graph for “$y = 2^{-x}$” looks like this:



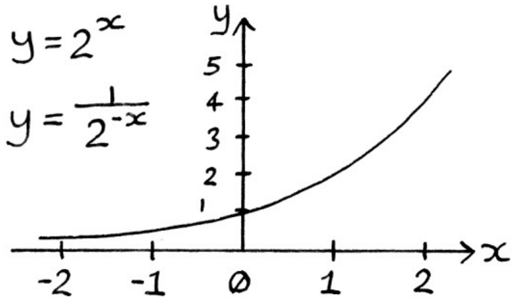
The graph is a mirror image of the graph for “$y = 2^x$”. When “x” on this graph is positive, the equation is finding the result of the base to the power of the negative of the value of “x”. Therefore, “y” becomes ever smaller as “x” increases. The y-axis values tend towards zero as “x” increases in value, but without ever reaching it. When “x” on this graph is negative, the equation is finding the result of the base to the power of the positive value of “x”. Therefore, “y” becomes greater as “x” decreases, and approaches infinity for large negative values of “x”.

It is important to note that the graph of:
“$y = 2^{-x}$”
... is identical to the graph of:
“$y = 1 \div (2^x)$”:



It is also the case that they are the same as the graph for “$y = 0.5^x$”.

The graph of “$y = 2^x$” is identical to the graph of “$y = 1 \div (2^{-x})$”:



Bases between just over 0 and just under 1

Earlier in this chapter, we saw that any base, whether positive or negative, an integer or a non-integer, when raised to the power of a positive *whole* number from 1 upwards, is the same as that base being multiplied by itself that number of times. In other words, “c^x”, where “x” is any integer from 1 up to infinity, means “c * c * c * c and so on.” [Note how this is for when the exponent is an *integer*]

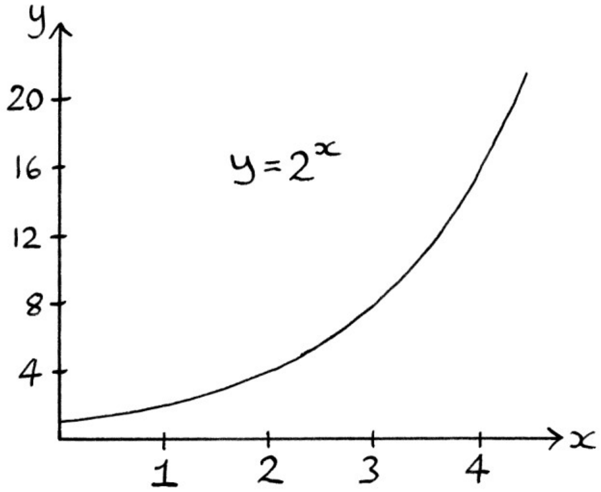
This is still true for bases between just over 0 and just under 1. However, there is a difference between the effects of doing this with numbers from 1 upwards and the effects of doing this with numbers from just over 0 to just under 1. Multiplying the first set of numbers produces larger results for each extra multiplication. Multiplying the second set produces smaller results for each extra multiplication. Therefore, the graphs for each set will be different.

For example, 0.5^3 = 0.5 * 0.5 * 0.5 = 0.125. The result of raising a number between 0 and 1 to a power is a smaller value than the one we started with.

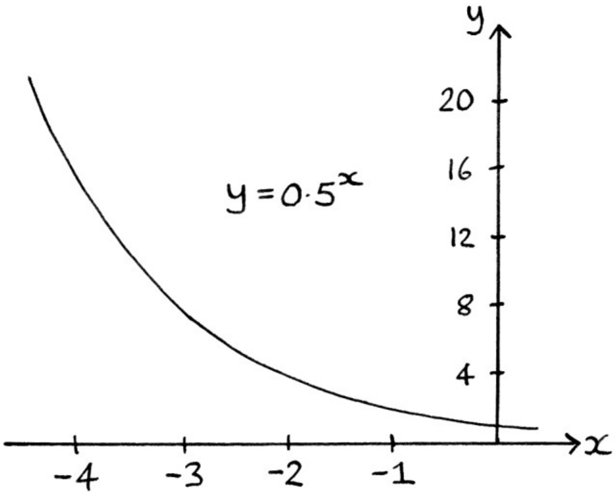
This is easiest to see when the exponents are integers, but in fact, any *positive* base between 0 and 1, when raised to any positive exponent will behave in the same way.

Mirrored exponentials

The graph for "$y = 2^x$" looks like this, as we already know:



As "x" increases in size, the result of "2^x" increases in size rapidly. On the other hand, the graph for "$y = 0.5^x$" looks like this: [which is also the graph for "$y = 2^{-x}$" and "$y = 1 \div (2^x)$"]



As "x" increases, the result of "0.5^x" becomes much smaller. This is easiest to remember if we think about how:
$2^3 = 2 * 2 * 2 = 8$
... and:
$0.5^3 = 0.5 * 0.5 * 0.5 = 0.125$

The two graphs are mirror images of each other. In fact, any graph with a positive base of 1 or higher, raised to a power of "x" will be a mirror image of the reciprocal of that base, all raised to x.

For example:
The graph of "$y = 5^x$" is the mirror image of the graph of "$y = (1 \div 5)^x$", which is "$y = 0.2^x$"
The graph of "$y = 10^x$" is the mirror image of the graph of "$y = (1 \div 10)^x$", which is "$y = 0.1^x$"
The graph of "$y = 100^x$" is the mirror image of the graph of "$y = 0.01^x$"
The graph of "$y = 1.2^x$" is the mirror image of the graph of "$y = 0.8333^x$".

This ultimately means that any base raised to a power is the same as the reciprocal of the-reciprocal-of-that-base raised to the same power. Or, to put it another way: "c^x" is the same as the reciprocal of "$(1 \div c)^x$", or to put this more succinctly:

$$c^x = 1 \div (1 \div c)^x$$

Therefore:
$2.3^5 = 1 \div (1 \div 2.3)^5$
$111^{2.76} = 1 \div (1 \div 111)^{2.76}$

Similarities

As you might have noticed, the graph for a base that is between 0 and 1, when raised to the power of "x", has the same basic shape as a base that is 1 or more, when raised to the power of negative "x".

For example, "$y = 0.5^x$" has the identical shape to "$y = 2^{-x}$".

Conversely, these graphs will both be the mirror images of "$y = 2^x$" and "$y = 0.5^{-x}$".

This shows an interesting property of exponential numbers. As I mentioned earlier, an exponential with a negative exponent is the same as 1 divided by the entire exponential as if the exponential had a *positive* exponent. Therefore, "2^{-2}" is the same as "$1 \div (2^2)$". As these give the same result as "0.5^2", which can also be thought of as "$(1 \div 2)^2$", it means that: $1 \div (2^2) = (1 \div 2)^2$.

From this, we can come up with general formulas:

$$c^{-x} = 1 \div (c^x) = (1 \div c)^x$$

... and:

$$c^x = 1 \div (c^{-x}) = (1 \div c)^{-x}$$

Negative bases

There are two ways to think of exponentials with negative bases:

- As “a negative base raised to a positive exponent”. In other words: $(-c)^x$.
- As the negative of “a positive base raised to a positive exponent”. In other words: $-(c^x)$.

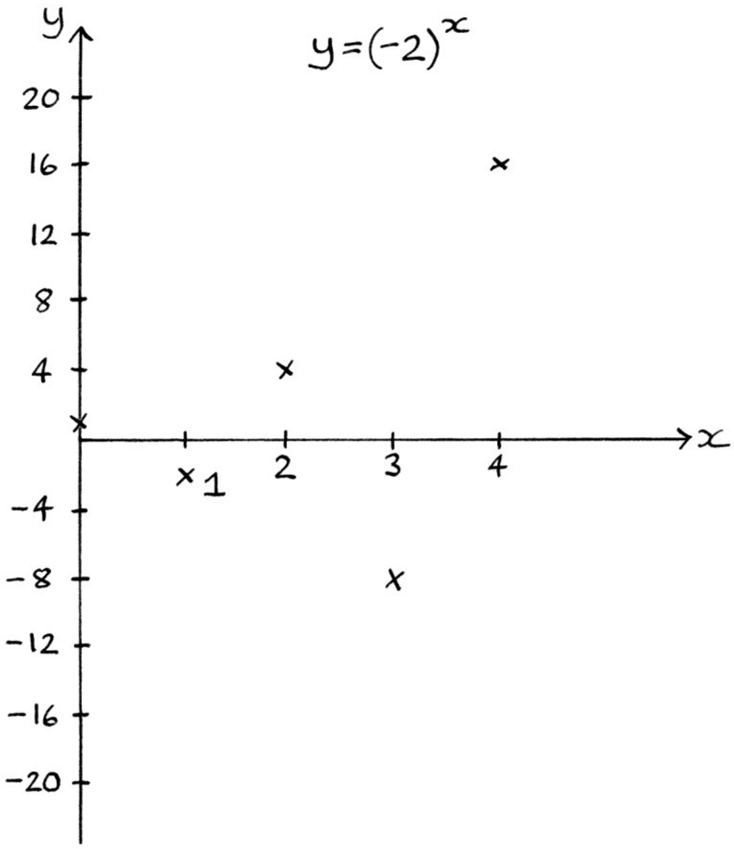
These produce very different results.

Negative bases with positive exponents

Exponentials that have negative bases raised to positive exponents are interesting because the results cannot make a curve on a graph. We can demonstrate this with a few examples:

$(-2)^2 = +4$
$(-2)^3 = -8$
$(-2)^4 = +16$
$(-2)^5 = -32$
$(-2)^6 = +64$
$(-2)^7 = -128$

The graph for “$y = (-2)^x$” has values of “y” that fluctuate from positive to negative for every consecutive *integer* value of “x”. One might think that we could join up the points, thus making an untidy looking chart. However, it would actually be meaningless to join up the dots, as the non-integers between the integers cannot be portrayed in this way.



The big difficulty for formulas such as “y = (−c)^x” is that it is difficult to know what the y-axis value would be when “x” is not an integer. For example, it is hard to know how to calculate the result of something such as “(−2)$^{4.5}$”. For one thing, it is impossible to calculate it from a single graph.

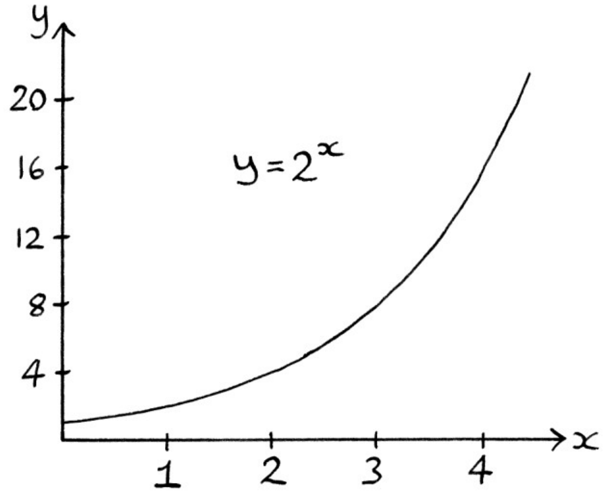
The actual result can be given using Complex numbers, but for now, we will skip why that is. For the exponents of 4, 4.5, 4.55 and 5, we would have these results:
(−2)^4 = 16
(−2)$^{4.5}$ = 22.6274i, which is an Imaginary number.
(−2)$^{4.55}$ = −3.6645 + 23.1370i, which is a Complex number.
(−2)^5 = −32, which is a Real number.

Given that non-integer values of “x” in “y = (−2)^x” have to be portrayed with Complex numbers, we cannot portray the results on a simple graph such as the one above. We will return to these exponentials later in this chapter.

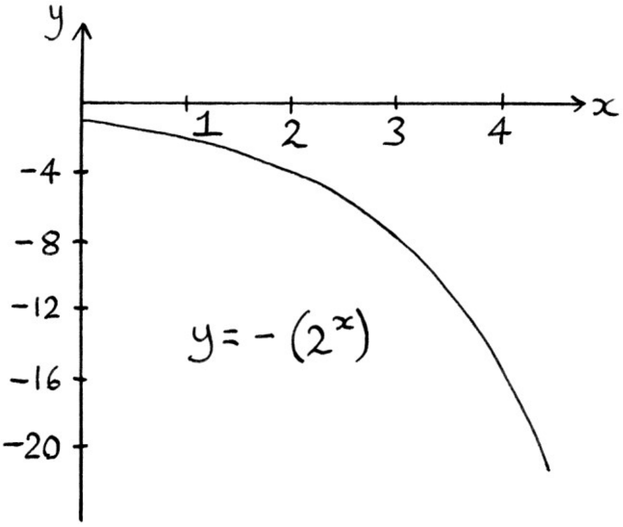
Negatives of positive bases to positive exponents

By “the negative of a ‘positive base to a positive exponent’”, I mean exponentials such as this: $-(c^x)$. In other words, everything about the exponential follows the rules for bases either between 0 and 1, or over 1, and then the result is made negative.

As we know, if the base is over 1, then an example formula such as “$y = 2^x$” looks like this:

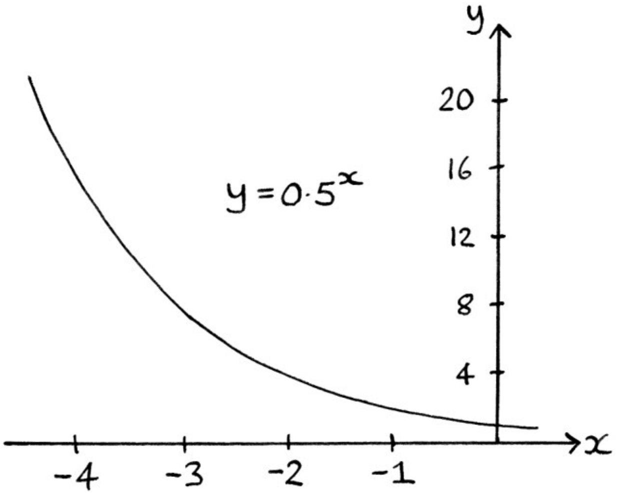


... which means that the negative version: “$y = -(2^x)$” will look like this:

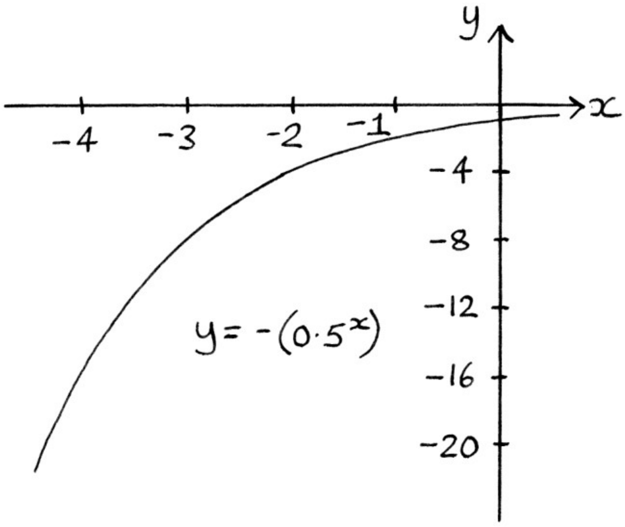


In other words “$-(c^x)$” is the same as “c^x” flipped downwards over the x-axis. Every value of “y” that would have been positive for “c^x” is now negative.

Also, as we know, if the base is between 0 and 1, then an example formula such as "$y = 0.5^x$" looks like this:



... which means that the negative version: "$y = -(0.5^x)$ will look like this:



Summary

As a summary of the various ideas we have seen so far, if we start with the formula "c^x", then:

If "x" is 1, the result is "c".
If "x" is 0, the result is 1.
If "x" is a positive integer, the result is "c" multiplied by itself "x" times.

If "c" is 1 or more, then:

- If "x" is 1 or more and not an integer, the result can be read off a graph.
- If "x" is between 0 and 1, the result is the same as the $(1 \div x)^{th}$ root of "c".
- If "x" is negative, the result is the same as "$1 \div c^x$".

If "c" is between 0 and 1, then:

- Everything above is still true, but the result becomes lower for higher values of "x".
- The result will the same as "$1 \div (c^{-x})$" or "$(1 \div c)^{-x}$"

If "c" is negative, and the calculation is "$(-c)^x$", then:

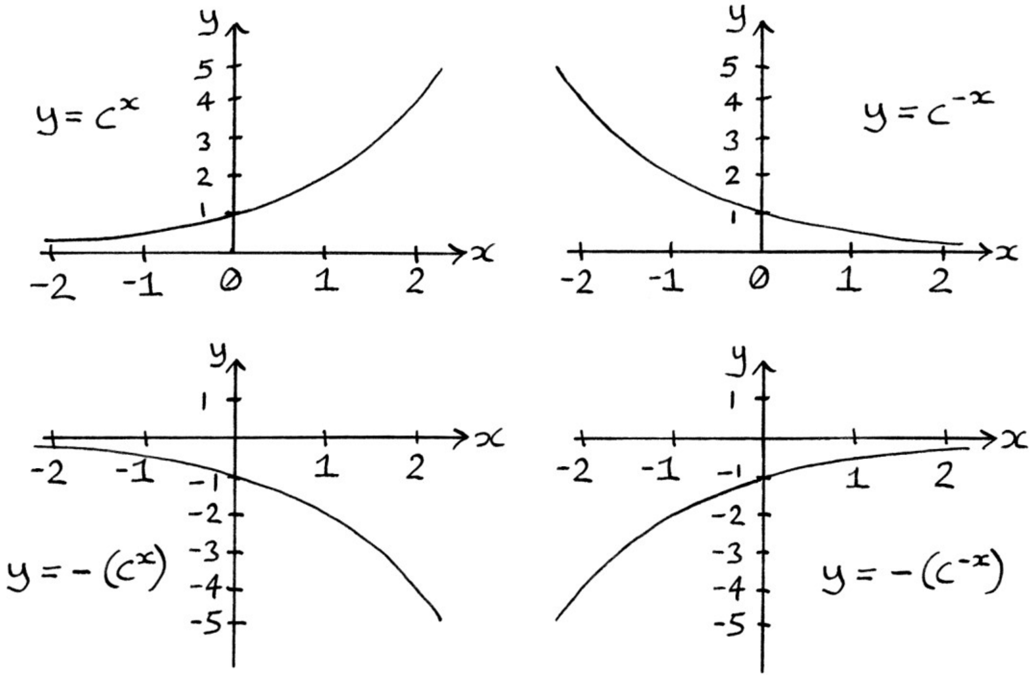
- We can easily calculate results for *integer* values of "x".
- For non-integer values of "x", we have to resort to Complex numbers.

If "c" is negative, and the calculation is "$-(c^x)$", then:

- Everything works as if "c" were positive, but with the negative of the result.

Graph Shapes

There are 4 basic graph shapes for values raised to an exponent. There are several ways of describing each one:



[These presume that "c" is a value greater than 1.]

Simple maths with exponentials

Some maths is straightforward with exponentials.

Multiplication

If we have the numbers 3^5 and 3^7, this is the same as having:
3 * 3 * 3 * 3 * 3
... and:
3 * 3 * 3 * 3 * 3 * 3 * 3

Therefore, if we multiply 3^5 by 3^7, it is the same as multiplying:
3 * 3 * 3 * 3 * 3
... by:
3 * 3 * 3 * 3 * 3 * 3 * 3
... which is:
3 * 3 * 3 * 3 * 3 * 3 * 3 * 3 * 3 * 3 * 3 * 3
... which is the same as:
3^{12}

The rule for multiplication is that any two exponential numbers that share the same base can be multiplied by adding the exponents. A general rule for this is:

$c^x * c^w = c^{(x + w)}$

Division

Division works in a similar way. If we have 5^3 and 5^2 then this is the same as having:
5 * 5 * 5
... and:
5 * 5

If we want to divide 5^3 by 5^2, then this is the same as:
5 * 5 * 5 divided by 5 * 5
...or:
(5 * 5 * 5) ÷ (5 * 5)
... which equals:
5, which is the same as 5^1

To divide one exponential by another with the same base, we just subtract the exponent of one from the other. A general rule for this is:

$c^{x} \div c^{w} = c^{(x-w)}$

The rule works even if we are dividing a smaller number by a larger number. For example, $12^{2} \div 12^{4} = 12^{-2}$. Written out in full, this amounts to $144 \div 20{,}736 = 0.006944 = 12^{-2}$. Using the formula from earlier in this chapter, we can also rewrite this as $(1 \div 12)^{2}$, although doing so does not really make anything clearer.

Sometimes it is easier to perform calculations if the values are treated as exponents than if the values are treated as actual numbers. Of course, converting the numbers to and from exponents takes some effort, so whether it is easier or not depends on the situation.

Powers to powers

If a base is raised to a power, and afterwards, the result of that is raised to another power, it is the same as if the base had been raised to those powers multiplied by each other. In other words:

$(c^{x})^{w} = c^{xw}$

We can test this with the example of: $(2^{3})^{4}$:
This is $(2 * 2 * 2)^{4}$
... which results in:
(2 * 2 * 2) * (2 * 2 * 2) * (2 * 2 * 2) * (2 * 2 * 2)
... which is the same as:
2 * 2 * 2 * 2 * 2 * 2 * 2 * 2 * 2 * 2 * 2 * 2
... which is the same as:
2^{12}

Therefore, we have shown that $(2^{3})^{4} = 2^{3*4} = 2^{12}$.

We might be able to tell that this would be true as the first power results in several copies of the base being multiplied together, so raising that to another power will create even more copies of the base being multiplied together. It is something that might not seem obvious at first, but becomes clearer the more you think about it.

As a more complicated example, we will look at 1.6 raised to the power of 2.5, all raised to the power of 11, or to express this properly:
$(1.6^{2.5})^{11}$
This is the same as:
$1.6^{2.5*11}$
... which equals:
$1.6^{27.5}$

Splitting up exponentials

Given that we can simplify powers of powers, and given that we can combine bases of the same value, we can also split up any given exponential into pieces. Doing this might make particular calculations easier, or it might make the meaning of a number easier to visualise.

As an example, we will look at the number 5^{20}.

We can turn this into a power of a power, and in fact, we can turn it into several different powers of powers:
$5^{20} = (5^{10})^2 = (5^5)^4 = (5^4)^5 = ((5^2)^5)^2 = (5^{0.5})^{40}$

We can turn 5^{20} into a multiplication of two or more exponentials with a base of 5:
$5^{20} = 5^{10} * 5^{10} = 5^{15} * 5^5 = 5^1 * 5^{19}$

Roots

A root, such as a square root, is really the opposite of an exponential.

A square root finds the number that must be multiplied by itself to produce a value. For example:
$\sqrt[2]{9} = 3$
This tells us that 3 * 3 = 9. We can put this as an exponential: $3^2 = 9$.

A cube root finds the number that must appear three times in a multiplication to produce a value. For example:
$\sqrt[3]{8} = 2$
This tells us that 2 * 2 * 2 = 8. We can put this as an exponential: $2^3 = 8$.

Higher roots work in the same way, the 11th root of 48,828,125 is written as so:
$\sqrt[11]{48{,}828{,}125}$
The result is 5. Therefore, $5^{11} = 48{,}828{,}125$.

We can also have a root that has no effect:
$\sqrt[1]{12}$
The result of this is 12. This is a similar idea to how 12^1 is 12.

In a root, there are two parts:

- The number being rooted, which is called the “radicand”. The word “radicand” ultimately comes from the Latin word “radix”, which means “root” in all its various senses. [The word “radish”, as in the *root* vegetable, also comes from the word “radix”.] The suffix “-and” means someone or something being subjected to something. Therefore, the prefix “radic-” and the suffix “-and” can be thought of as meaning “that which is rooted”. Knowing all of this might make the term “radicand” easier to remember.

- The number that shows the extent of the rooting, which is called the “index”. The English noun “index” comes from the Latin noun “index”, one of the meanings of which is a marker or indicator. Therefore, to remember the term “index”, one can think of it as “indicating” the extent of the rooting.

Given all of the above, we can say:

$$\sqrt[index]{radicand}$$

The actual √ symbol is called the "root symbol", the "radical symbol" or the "radix" symbol. I am going to call the calculation as a whole a "root".

Given that roots and exponentials are the opposite of each other, if we have one, we can express it in terms of the other. We can express a root as an exponential, and we can express an exponential as a root. The rule for doing this is:

$$\sqrt[x]{a} = a^{(1 \div x)}$$

... or:

$$a^x = \sqrt[(1 \div x)]{a}$$

In other words, if we have a root, the result will be equal to the radicand (the number being rooted) raised to the power of the reciprocal of the index. [The reciprocal of a number is 1 divided by that number.] If we have an exponential, the result will be the same as the reciprocal of the exponent rooting the base. This is easier to understand with examples:

$\sqrt[2]{7} = 7^{(1 \div 2)} = 7^{0.5}$

$\sqrt[5]{4} = 4^{(1 \div 5)} = 4^{0.2}$

$\sqrt[6.7]{256} = 256^{(1 \div 6.7)}$

$10^2 = \sqrt[(1 \div 2)]{10} = \sqrt[0.5]{10}$

$8^3 = \sqrt[(1 \div 3)]{8}$

Roots and exponentials

If we have a root that is the base of an exponential, the root and exponent can sometimes cancel each other out.

For example:

$(\sqrt{3.213})^2$

... means the square root of 3.213, all squared. The square root and the squaring cancel out to leave just 3.213.

Another example:

$(\sqrt[7.7]{123})^{7.7}$

This is the 7.7th root of 123, all raised to the power of 7.7. The root and the exponent cancel each other out, and we are left with 123.

If the root and exponent are not the same, we can still simplify some calculations.
For example:
$(\sqrt[2]{12})^4$
This is the square root of 12, all raised to the power of 4. In this example, the exponent can cancel out the square root, so we end up with 12^2

As another example:
$(\sqrt[27]{11.1})^3$
This is the 27th root of 11.1, all raised to the power of 3. We can reduce this as so:
$\sqrt[9]{11.1}$
This is just the 9th root of 11.1.

The general rules for any root raised to a power are:

- If we multiply the exponent by a number, we can multiply the index of the root by the same number, and the result will be the same.
- If we divide the index of a root by a number, we can divide the exponent by the same number, and the result will be the same.

These rules allow us to simplify exponentials and roots. We can express them mathematically as so:

$(\sqrt[x]{n})^y = (\sqrt[ax]{n})^{ay}$
... and:
$(\sqrt[x]{n})^y = (\sqrt[x/a]{n})^{y/a}$
... where:

- "x" is the index of the root
- "y" is the exponent of the exponential
- "a" is any number scaling the root and power

One might guess that the rules are true by how roots make things smaller and exponents make things bigger. If we make the index of the root larger, thus making the result smaller, we need to make the exponent of the exponential bigger to keep the result the same.

Knowing all this allows us to simplify some roots and exponentials that might seem difficult otherwise.

For example:

$(\sqrt[6i]{-1})^{12i}$

This is the 6ith root of −1, all raised to the power of 12i. We can divide the root and exponent by 6i, thus keeping the overall meaning the same, and we will have:

$(\sqrt[1]{-1})^{2}$

... which, because the first root of a number is that number, would normally be written as:

$(-1)^{2}$

... which is:

1

Contents of a root

If we have a root such as this:

$\sqrt[4]{16}$

... then it is the same as:

$\sqrt[2]{4}$

For any root, if we halve the index, and square root the radicand, the result will be the same. Similarly, if we divide the index by 3, and cube root the radicand, the result will be the same.

The general rule for this is that the overall result will be the same if we divide the index by a particular number, and root the radicand by that same number at the same time.

A rule for this is:

$$\sqrt[x]{c} = \sqrt[x \div a]{(\sqrt[a]{c})}$$

... where "a" is any number.

We are reducing the amount by which the radicand is being rooted at the same time as reducing the size of the radicand. As before, doing such a thing can sometimes simplify an equation.

As an example, we will look at the 12th root of 512:

$\sqrt[12]{512}$

We can alter the formula so that the index (12) is divided by 3, and the radicand (512) becomes cube-rooted while still in place: [This is shown in a larger font to make it clearer.]

$$\sqrt[12]{512} = \sqrt[(12 \div 3)]{\sqrt[3]{512}} = \sqrt[4]{\sqrt[3]{512}} = \sqrt[4]{8} = 1.6818$$

Scaling up

As well as scaling down the index, we can scale it up too. If we have:

$\sqrt[x]{c}$

... then it is the same as:

$\sqrt[(x * a)]{c^a}$

In other words, if we multiply the index of a root by a particular value, and we raise the radicand to the power of that value at the same time, then the overall result will be the same. We are increasing the amount by which the radicand is rooted, at the same time as increasing the size of the radicand.

As an example, we will look at the 4th root of 65,536:

$\sqrt[4]{65{,}536}$

If we multiply the index (4), by 2, and raise the radicand (65,536) to the power of 2, we will still have the same overall result:

$$\sqrt[4]{65{,}536} = \sqrt[(4 * 2)]{(65{,}536)^2} = \sqrt[8]{4{,}294{,}967{,}296} = 16$$

This is similar to when we had a root that was the base of an exponential. Here we have an exponential that is within the root (it is the radicand). Any radicand can be turned into an exponential, and any base can be turned into a root. When we have a root as the base of an exponential, or an exponential as the radicand of a root, we can scale the index of the root and the exponent of the radicand in the same way and the overall result will be the same.

Simple maths with roots

As a root is really another way of expressing an exponential, the maths that can be done with exponentials can also be done with roots.

Multiplication

One way to multiply two roots with the same radicand is to convert them into exponentials, use the rules for multiplying exponentials with the same base, and then convert the result back to a root. For example, if we have this multiplication:

$$\sqrt[3]{13} * \sqrt[5]{13}$$

... then we know it is the same as:

$$13^{0.3333} * 13^{0.2}$$

[We know this because a root can be converted into an exponential by turning the radicand into the base, and using the reciprocal of the index as the exponent].

Then, as the bases are the same, we can calculate the result of that by adding the exponents:

$13^{0.5333}$

... which is:

$13^{(8 \div 15)}$

We convert the exponential back into a root:

$$\sqrt[\left(\frac{15}{8}\right)]{13}$$

... which is:

$$\sqrt[1.875]{13}$$

The result is the 1.875th root of 13. From this, we can deduce that the rule for multiplying two roots with the same radicand is:

$$\sqrt[a]{c} * \sqrt[b]{c} = \sqrt[(a*b)/(a+b)]{c}$$

In other words, if we have two roots with the same radicand, and indices of “a” and “b”, and we multiply them together, the resulting index will be:
(a * b) ÷ (a + b)

We can test this with an example:
$\sqrt[6]{64} * \sqrt[3]{64}$

The result will have an index of:
(6 * 3) ÷ (6 + 3)
... which is:
18 ÷ 9
... which is:
2

The result is:
$\sqrt[2]{64}$

Division

Division works in a similar way. If we start with indices of “a” and “b”, the resulting index will be: (b * a) ÷ (b − a). The rule is:

$$\sqrt[a]{c} \div \sqrt[b]{c} = \sqrt[(b*a)/(b-a)]{c}$$

For example:
$\sqrt[3]{13} \div \sqrt[5]{13}$

... will produce a root with an index of:
(5 * 3) ÷ (5 – 3)
... which is:
15 ÷ 2
... which is:
7.5

The result is:
$\sqrt[7.5]{13}$

Logarithms

A logarithm finds the value to which a number must be raised to produce a particular result. In other words, a logarithm finds "x" in this equation:

$$c^x = d$$

If we wanted to know the exponent to which the number 7 needs to be raised to produce 49, we would use logarithms. The calculation we would be trying to solve is this:
$7^x = 49$

The "x" in this calculation will be 2. The number 7 must be raised to the power of 2 to produce the result of 49. In other words, the missing exponent is 2. We can write the calculation in full as: $7^2 = 49$.

Using mathematical jargon, we would say, "The base 7 logarithm of 49 is 2." The base of our exponential is 7, so we would be using the "base 7 logarithm". If the base of our exponential were, say, 12, we would use the "base 12 logarithm". If the base of our exponential were, say, 2, we would use the "base 2 logarithm". We have to use the type of logarithm related to the base of the exponential involved.

On a typical calculator, you might see a "log" or "$\log_{10}$" button. This will find the "base 10 logarithm". In other words, it will solve problems such as "$10^x = 1000$" or "$10^x = 45$". It will find the exponent in an exponential where the base is ten. Calculator manufacturers presume that the base 10 logarithm will be used more often than other logarithms, which is why there is often a dedicated button for base 10 logarithms, and which is why the button frequently is not labelled "$\log_{10}$" or "base 10 logarithm", but just "log".

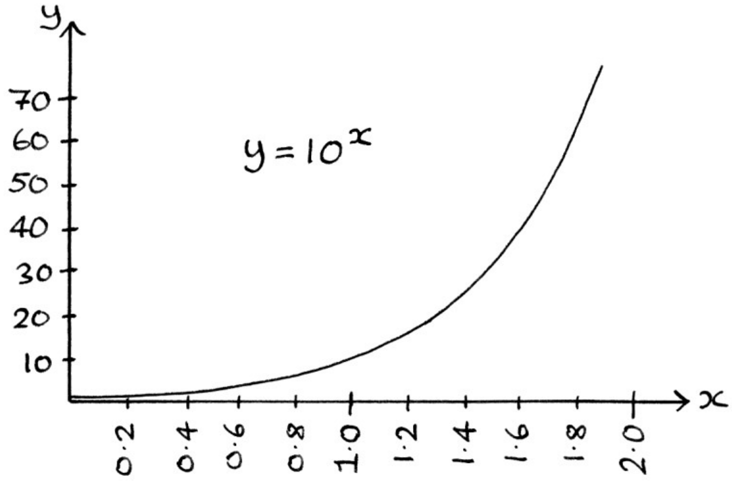
On a typical calculator, you might also see a "$\log_2$" button. This will find the "base 2 logarithm". It will solve problems such as "$2^x = 8$" or "$2^x = 102$". It will find the exponent where the base is 2. Supposing we have the calculation:
$2^x = 1024$
... then we enter 1024 and press the "$\log_2$" button, and the calculator will find "x", which in this case will be 10. This means that the number 2 needs to be raised to the power of 10 to result in 1024. In other words, $2^{10} = 1024$.

Calculators usually have an “ln” button. This is the “natural logarithm” button. This finds the exponent for a calculation where the base is the number “e”, which is 2.71828182... This is really the “log 2.71828182...” button, or the “log e” button. The “ln” button finds “x” in calculations such as:
“$2.71828182^x = 535.4917$”
... or:
“$2.71828182^x = 22{,}026.4658$”

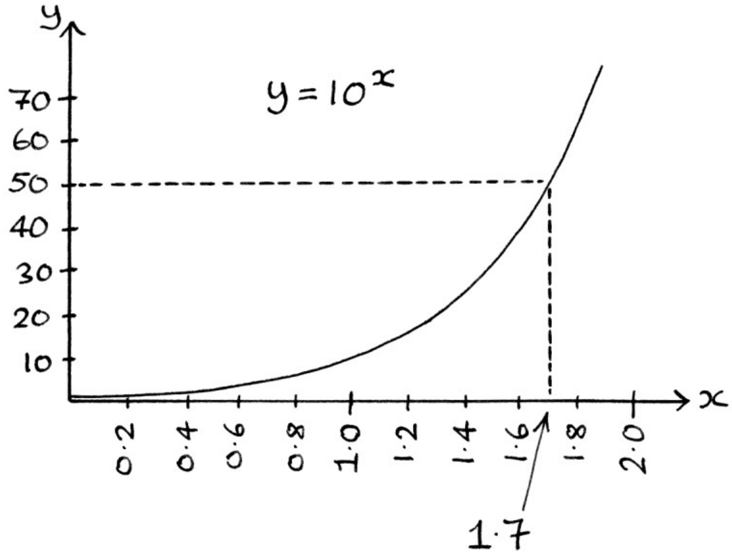
Better calculators can find the logarithms for an exponential with any base. For example, $3.5^x = 22$.

Solving logarithms using a graph

If we have a graph showing “$y = 10^x$”, then, not only can we use the graph to see any power of 10, but we can also use the graph to solve base 10 logarithms. We look up the result on the y-axis, and the x-axis value of the curve at that point will be the exponent of a power of 10 that would produce that result.

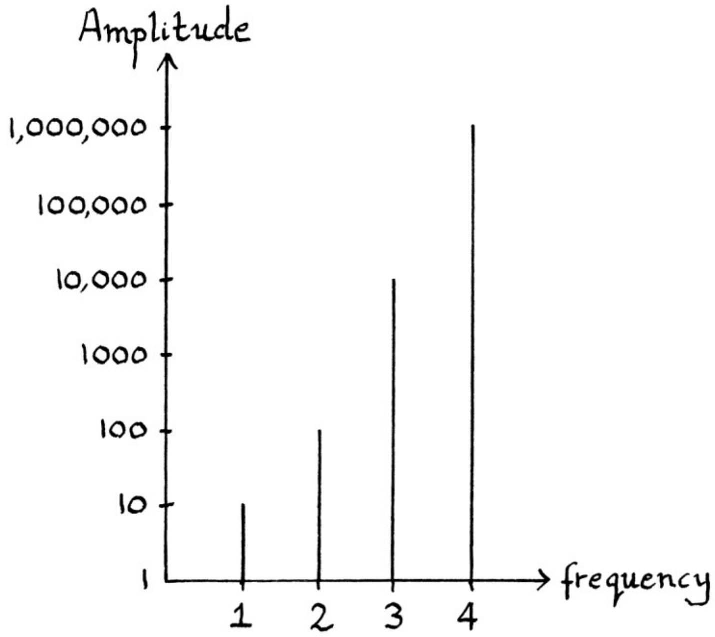


We will solve $10^x = 50$. We will find the place on the curve where $y = 50$, and the x-axis value at that point will be the base 10 logarithm of 50. Reading the graph gives an x-axis value of 1.7, which is an approximate answer because the graph cannot be read to much accuracy. Therefore, $10^{1.7} \approx 50$. In reality, the base 10 logarithm of 50 is 1.6990 to four decimal places, so $10^{1.6990} = 50$.



Logarithmic scales

We looked at logarithmic scales in Chapter 15 on the frequency domain. A logarithmic scale is one where every value is a power of a particular number. In the following graph, the y-axis scale is logarithmic. Every value is a power of 10:



The first entry on the amplitude axis is 10^0, the second entry is 10^1, the third entry is 10^2, the fourth entry is 10^3, and so on. We can summarise the nature of the entries by saying that they are the results of 10^x, where "x" goes upwards from zero.

If we were filling in the details of such a graph, we would use logarithms. More specifically, because every y-axis entry is calculated using 10^x, we would use base 10 logarithms. If we had the value 123 and wanted to find out where it belonged on the y-axis, we would need to find the power of 10 that would be equal to 123. In other words, we would need to solve $10^x = 123$. We could do this with a calculator: $\log_{10} 123 = 2.08991$. In other words, $10^{2.08991} = 123$. Therefore, the entry for 123 will be just after the second entry on the y-axis, so 0.08991 of the way between 100 and 1000 on the scale.

Negative bases again

Earlier in this chapter, we saw how the result of a negative Real number raised to the power of a non-integer Real number resulted in a Complex number. In other words, exponentials of the form "$(-a)^x$" when "x" is not an integer result in Complex numbers. As an example of this, $(-2)^{4.55} = -3.6645 + 23.1370i$. We can calculate such a result with a calculator that can work with Complex numbers. However, the rule for finding the result in other ways is easy to figure out. In this section, we will work out ways of calculating such exponentials through observation, and then we will use our knowledge of exponentials and roots to understand what is happening.

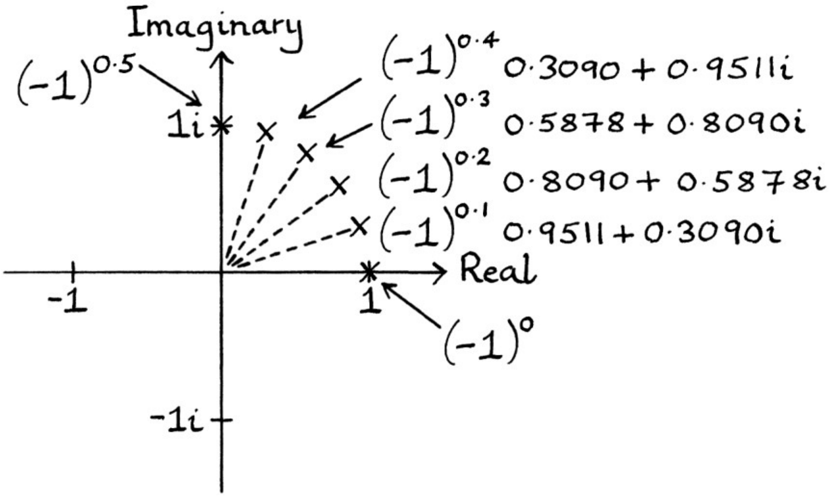
Powers of −1

First, we will look at exponentials with −1 as the base, and try to find a rule for what is happening. The easiest way to find a pattern in such exponentials is to try various exponents of −1 with a calculator, and look at the results.

If we have $(-1)^0$, we would end up with 1, which we can also write as "1 + 0i"
$(-1)^{0.1}$ produces 0.9511 + 0.3090i
$(-1)^{0.2}$ produces 0.8090 + 0.5878i
$(-1)^{0.3}$ produces 0.5878 + 0.8090i
$(-1)^{0.4}$ produces 0.3090 + 0.9511i
$(-1)^{0.5}$ produces 0 + 1i
$(-1)^{0.6}$ produces −0.3090 + 0.9511i
$(-1)^{0.7}$ produces −0.5878 + 0.8090i
$(-1)^{0.8}$ produces −0.8090 + 0.5878i
$(-1)^{0.9}$ produces −0.9511 + 0.3090i
$(-1)^{1.0}$ produces −1 + 0i
$(-1)^{1.1}$ produces −0.9511 − 0.3090i
$(-1)^{1.2}$ produces −0.8090 − 0.5878i
$(-1)^{1.3}$ produces −0.5878 − 0.8090i
$(-1)^{1.4}$ produces −0.3090 − 0.9511i
$(-1)^{1.5}$ produces 0 − 1i
$(-1)^{1.6}$ produces 0.3090 − 0.9511i
$(-1)^{1.7}$ produces 0.5878 − 0.8090i
$(-1)^{1.8}$ produces 0.8090 − 0.5878i
$(-1)^{1.9}$ produces 0.9511 − 0.3090i
$(-1)^{2.0}$ produces 1 + 0i
... and after this point, the results repeat.

If we were to plot the Real part of each of these results, we would end up with a Cosine wave. If we were to plot the Imaginary part of these results, we would end up with a Sine wave. You might be able to tell this from the way the resulting values rise and fall, and how the pattern of Real values is repeated in the Imaginary values a quarter of the way through the list. Given that, you might be able to tell that these results are the coordinates of points around a circle.

If we plot the results on the Complex plane, we can see that every point is exactly one unit away from the origin of the axes. The points are also spaced at angles of 18 degrees. We can confirm all this with Pythagoras's theorem and arctan. [This is a situation in which degrees are more useful than radians, as it would be harder to recognise the pattern with radians.]


Imaginary
$(-1)^{0.5}$
1i
$(-1)^{0.4}$ 0·3090 + 0·9511i
$(-1)^{0.3}$ 0·5878 + 0·8090i
$(-1)^{0.2}$ 0·8090 + 0·5878i
$(-1)^{0.1}$ 0·9511 + 0·3090i
Real
-1
1
$(-1)^{0}$
-1i

The points all lie on the circumference of a unit-radius circle. If we plotted many more points, we would draw out a perfect circle.

We will try to achieve a better understanding of powers of −1. The result of $(-1)^0$ is a point 1 unit from the origin at an angle of 0 degrees. It is "1 + 0i". The result of $(-1)^{0.1}$ is "0.9511 + 0.3090i", which is a point 1 unit away from the origin at an angle of 18 degrees. All the points are 18 degrees apart. The results of the exponentials repeat when we reach $(-1)^{2.0}$, which is "1 + 0i". The result of $(-1)^0$ is the same as $(-1)^{2.0}$.

We can say that for exponentials with "−1" as the base, the exponents are acting as angles in a system where there are 2 angle units in a circle. In other words, the circle is divided into two pieces. We will refer to these angle units as "half-circle angle units". This means that the exponent is really an angle. When we perform a calculation such as "$(-1)^{x}$", we are really indicating the point on a unit-radius circle at an angle of "x" half-circle angle units.

The rule for powers of −1

We saw that “$(-1)^x$” indicates the point on a unit-radius circle at an angle of “x” half-circle angle units. [As “x” is an angle, we could replace it with the symbol “θ” as in “$(-1)^\theta$”, but in this case, it is clearer if we do not.]

We can create a rule that will easily solve any power of −1:
“$(-1)^x = \cos x + i \sin x$”, when “x” is the exponent, and Cosine and Sine are working in *half-circle angle units*. [We know this rule is true because the points draw a circle, and Cosine and Imaginary Sine also draw a circle.]

We can rewrite the formula for Cosine and Sine in degrees:
“$(-1)^x = \cos((x \div 2) * 360) + i \sin((x \div 2) * 360)$”
... which is:
“$(-1)^x = \cos(180x) + i \sin(180x)$”
... where “x” is the exponent, and Cosine and Sine are working in degrees. The formula is like this because one half-circle angle unit is equal to 180 degrees.

The formula for radians will be:
“$(-1)^x = \cos((x \div 2) * 2\pi) + i \sin((x \div 2) * 2\pi)$”
... which is:
“$(-1)^x = \cos(\pi x) + i \sin(\pi x)$”
... where “x” is the exponent, and Cosine and Sine are working in radians. The formula is like this because one half-circle angle unit is equal to π radians.

To test the formulas, we will solve “$(-1)^{3.45}$”. A calculator that can work with Complex numbers will give the result as:
“−0.1564 – 0.9877i”

If we use the degrees formula, we will be calculating:
$\cos(180 * 3.45) + i \sin(180 * 3.45)$ in degrees
$= -0.1564 - 0.9877i$

If we use the radians formula, we will be calculating:
$\cos(\pi * 3.45) + i \sin(\pi * 3.45)$ in radians
$= -0.1564 - 0.9877i$

These results show that our formulas work. [Strictly speaking, the formulas *must* work because each side of each equation is identifying the same point on a unit-radius circle’s edge.]

Powers of −2

Powers of −2 are more complicated than powers of −1, but are straightforward once we figure out the underlying rule. We will use a calculator to work out the results of various powers of −2:

$(-2)^{0.0}$ produces 1 + 0i
$(-2)^{0.1}$ produces 1.01932 + 0.3312i
$(-2)^{0.2}$ produces 0.9293 + 0.6752i
$(-2)^{0.3}$ produces 0.7236 + 0.9960i
$(-2)^{0.4}$ produces 0.4078 + 1.2549i
$(-2)^{0.5}$ produces 0 + 1.4142i
$(-2)^{0.6}$ produces −0.46384 + 1.4415i
$(-2)^{0.7}$ produces −0.9549+ 1.3143i
$(-2)^{0.8}$ produces −1.4086 + 1.02339i
$(-2)^{0.9}$ produces −1.7747 + 0.5766i
$(-2)^{1.0}$ produces −2 + 0i

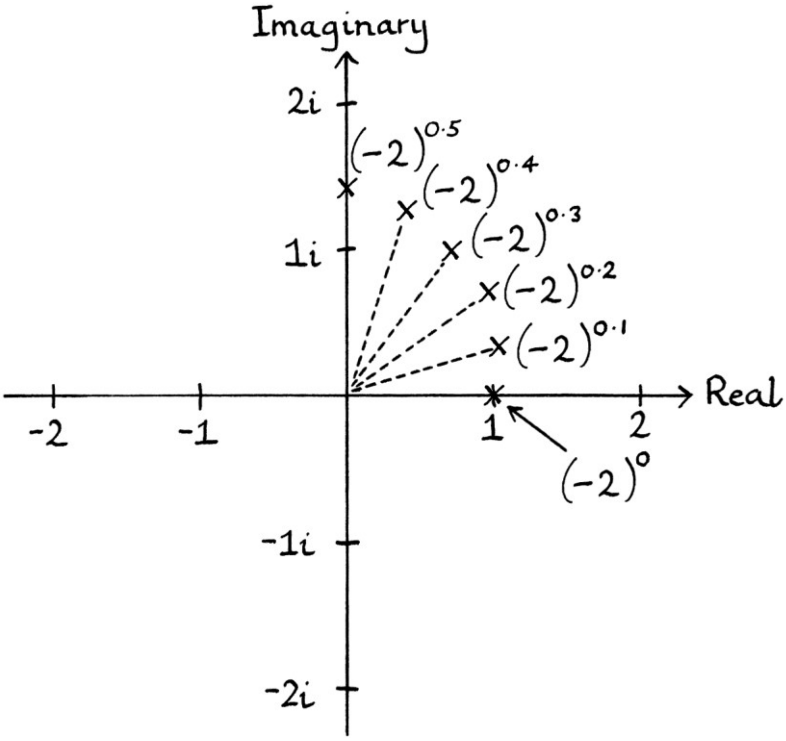
There is still a pattern, but it is less obvious. We can calculate the angle portrayed by the first few Complex numbers:

The angle of "1 + 0i" is 0 degrees.
The angle of "1.01932 + 0.3312i" is 18 degrees.
The angle of "0.9293 + 0.6752i" is 36 degrees.
The angle of "0.7236 + 0.9960i" is 54 degrees.
The angle of "0.4078 + 1.2549i" is 72 degrees.

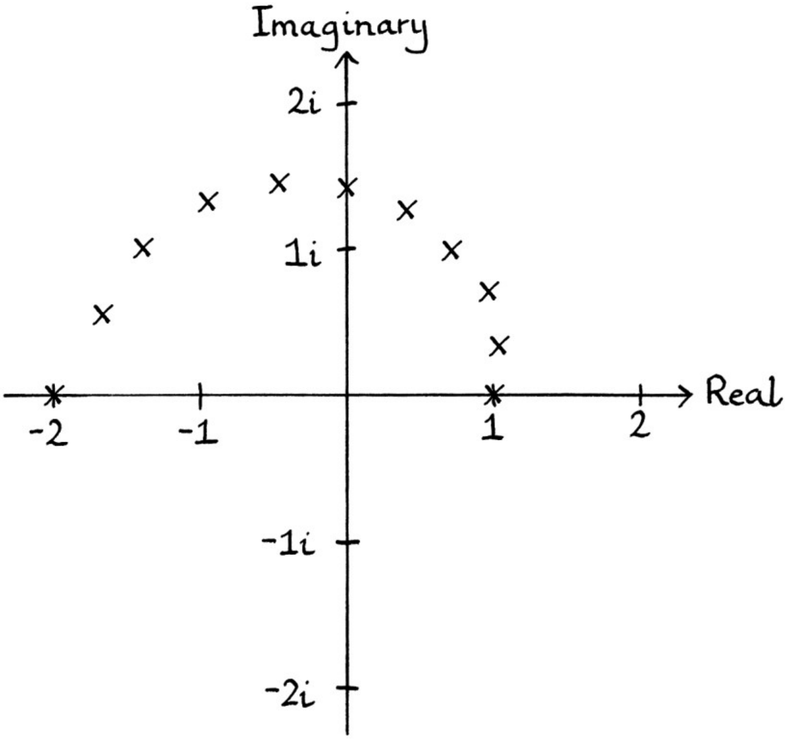
This shows that the angles of each point increase by 18 degrees for every 0.1 increase in the exponent. This is the same as when we were dealing with exponents of −1.

By the way that the Complex numbers become ever larger, we can see that the points are not all the same distance from the origin:

"1 + 0i" is 1 unit from the origin.
"1.01932 + 0.3312i" is 1.071773463 units from the origin (to 8 decimal places).
"0.9293 + 0.6752i" is 1.14869835 units from the origin (to 8 decimal places).
"0.7236 + 0.9960i" is 1.23114441 units from the origin (to 8 decimal places).
"0.4078 + 1.2549i" is 1.31950791 units from the origin (to 8 decimal places).



Each of these numbers is 1.071773463 times larger than the one preceding it. On the Complex plane, the points draw a spiral. The first few points look like this:



We can find the relevance of 1.071773463 by finding its base 2 logarithm on a calculator. [We would know to do this by guessing the idea, after trying many other calculations.] In other words, we want to find the exponent in an exponential where the base is 2, and the result is 1.071773463. We want to solve this equation:
$2^x = 1.071773463$
... which can be rephrased as this:
$x = \log_2 1.071773463$
... and the result is:
$x = 0.1$

Therefore, we know that $2^{0.1} = 1.071773463$.

For every increase of the exponent by 0.1, the point indicated by the exponential rotates anticlockwise by 18 degrees, and moves another 1.071773463 times further away from the origin of the axes. We can also say that for every increase of the exponent by 0.1, the point indicated by the exponential rotates anticlockwise by 18 degrees, and moves another "$2^{0.1}$" times away from the origin of the axes.

We now know enough to create rules to calculate exponentials with "−2"as the base:
"$(-2)^x = 2^x \cos(180x) + i * 2^x \sin(180x)$", where Cosine and Sine are working in degrees.

"$(-2)^x = 2^x \cos(\pi x) + i * 2^x \sin(\pi x)$", where Cosine and Sine are working in radians.

We can test this with an example. A calculator that can work with Complex numbers will give the result of "$(-2)^{6.51}$" as:
$-2.8628 + 91.09424i$

If we use the degrees formula with Cosine and Sine working in degrees, we have:
$2^{6.51} \cos(180 * 6.51) + i * 2^{6.51} \sin(180 * 6.51)$
$= -2.8628 + 91.09424i$

If we use the radians formula with Cosine and Sine working in radians, we have:
$2^{6.51} \cos(\pi * 6.51) + i * 2^{6.51} \sin(\pi * 6.51)$
$= -2.8628 + 91.09424i$

Powers of any negative base

From seeing the formula for powers of −2, we can guess that the formulas for any negative base will be:

“$(-a)^x = a^x \cos(180x) + i * a^x \sin(180x)$”
... where “a” is a Real number, and Cosine and Sine are working in degrees, and:
“$(-a)^x = a^x \cos(\pi x) + i * a^x \sin(\pi x)$”
... where “a” is a Real number, and Cosine and Sine are working in radians.

We can test these with an example. A calculator that can work with Complex numbers will give $(-17.4)^{5.11}$ as:
−2,054,666.07670 – 739,725.3048i

The degrees formula will be:
$17.4^{5.11} \cos(180 * 5.11) + i * 17.4^{5.11} \sin(180 * 5.11)$
... which is:
−2,054,666.07670 – 739,725.3048i

The radians formula will be:
$17.4^{5.11} \cos(5.11\pi) + i * 17.4^{5.11} \sin(5.11\pi)$
... which is also:
−2,054,666.07670 – 739,725.3048i

Using our knowledge of exponentials

Now that we have worked out rules by observing the behaviour of the exponentials, we will work out what is happening by using what we have learnt about exponentials and roots in this chapter.

The exponential $(-1)^x$ can be altered so that the base becomes a root, and the exponent becomes scaled by the same amount as the index of the root. In other words, we will be using this formula from earlier:
$(\sqrt[x]{n})^y = (\sqrt[ax]{n})^{ay}$

As a first step, we will rephrase $(-1)^x$ so that the base is in the form of a root where the index is 1. In other words, the root symbol will be in the formula, but it will have no effect:

$(\sqrt[1]{-1})^x$

We will then scale the index of the root, and the exponent of exponential by 2:

$(\sqrt[2]{-1})^{2x}$

This is the square root of −1, all raised to the power of 2x. The square root of −1 is "i", so the formula is really:

i^{2x}

Therefore, whenever we have been raising −1 to a power, we were really dealing with a variation of raising "i" to a power, and that is why we obtained Complex numbers in the results. As we know from Chapter 25, an exponential such as "i^x" identifies a point on a unit-radius circle at an angle of "x" quarter-circle angle units. We can now see why an exponential of "$(-1)^x$" identifies a point on a unit-radius circle at an angle of "x" *half*-circle angle units.

When it comes to negative bases other than −1, we can also use our knowledge from this chapter to explain the results. Supposing we have this exponential:
$(-2)^x$
... then we can split it up into this:
$2^x * (-1)^x$

In this way we can see that $(-2)^x$ is really a scaling of $(-1)^x$. We can fully understand $(-1)^x$ because we know it is really "i^{2x}", and we know how powers of "i" work. We can now fully understand $(-2)^x$ as we know it is "$2^x * (-1)^x$", and we know how powers of 2 work.

From all this, whenever we see an exponential of the form:
"$(-a)^x$"
... we know that it is really:
"$a^x * (-1)^x$"
... and we know that that, in turn, is really:
"$a^x * i^{2x}$"

This explains all the behaviour we see when we have an exponential with a negative base. The reason that *integer* values of "x" result in Real numbers (and not Complex numbers) is straightforward too. Exponentials with "*i*" as the base treat the exponent as an angle in quarter-circle angle units. If we have "i^x" then:

- When "x" is 0, the result will be "1 + 0i" (at 0 degrees), which is Real.
- When "x" is 1, the result will be "0 + 1i" (at 90 degrees), which is Imaginary.
- When "x" is 2, the result will be "−1 + 0i" (at 180 degrees), which is Real.
- When "x" is 3, the result will be "0 – 1i" (270 degrees), which is Imaginary.
- When "x" is 4, the result will be "1 + 0i" (0 degrees), which is Real.

When "x" is any integer, the result will be at 0, 90, 180, or 270 degrees, which will either be all Real or all Imaginary.

If we have "i^{2x}", as we do in this situation, then:

- When "x" is 0, the result will be "1 + 0i" (at 0 degrees), which is Real.
- When "x" is 1, the result will be "−1 + 0i" (at 180 degrees), which is Real.
- When "x" is 2, the result will be "1 + 0i" (at 0 degrees), which is Real.
- When "x" is 3, the result will be "−1 + 0i" (at 180 degrees), which is Real.
- When "x" is 4, the result will be "1 + 0i" (at 0 degrees), which is Real.

When "x" is any integer, the result will be at 0 degrees or 180 degrees, which means it must be a Real number.

Any scaling of the result of "i^{2x}" by another number will maintain its "Real-ness".

The entirely Real results are really coincidences that hide the underlying nature of the exponentials. That integer exponents of negative-base exponentials result in Real numbers makes it seem as if everything here were happening on a normal exponential graph, but in reality, everything here is happening on a circle.

Summary

After all of the above, we can say that:
$(-a)^x$
$a^x * (-1)^x$
$a^x * i^{2x}$
$a^x \cos(180x) + i * a^x \sin(180x)$ when Cosine and Sine are working in degrees
$a^x \cos(\pi x) + i * a^x \sin(\pi x)$" when Cosine and Sine are working in radians
... are all the same.

Conclusion

As with so much in maths, exponentials, roots and logarithms are easiest to understand if you need to solve problems that use them. If you are taught them just for the sake of education, they are also much harder to remember. It is well worth playing with exponentials, roots and logarithms on a calculator to reinforce how they work. Good explanations of exponentials and other maths are available at Paul Dawkins' website: https://tutorial.math.lamar.edu/

Chapter 27: e

The lower-case Latin letter “e” is the symbol for the number:
2.718281828459045235360287....

The number “e” is one of the most interesting numbers in maths. It is a number that plays a part in several seemingly unrelated aspects of reality. As with “π”, the number “e” is irrational – it cannot be expressed by any fraction consisting of integers, and its digits continue forever. There are no patterns in its digits that help in its calculation. As with “π”, the number “e” is a universal mathematical constant. It is independent of any social or cultural influence.

In maths, “e” appears mainly as the base in an exponential. You are more likely to see it in the form of “e^x”, where “x” is any number, than you are to see it as just “e”.

Compound Interest

The simplest situation in which “e” appears relates to compound interest on bank account balances.

Introduction to compound interest

A bank account that pays compound interest is one that pays interest based on the amount of money in the account, and then, if we keep that interest in the account, pays the next amount of interest based on the new total in the account. In other words, supposing we never spend any money: if we start off with £100 in a bank account and are paid 1% interest a year, after one year, we will receive a payment of £1. Our new balance will be £101. After the second year, we will earn 1% interest on that £101. This works out as £1.01. We now have £102.01 in the account. After the third year, we will earn 1% interest on that £102.01 and earn £1.0201, which, if our bank works in hundredths of pennies, will mean that we have £103.0301 in the account.

If we did not have compound interest, the interest each year would only be based on the original sum of money. In that case, in the first year, we would earn 1% of £100. In the second year, we would also earn 1% of £100. In the third year, we would still earn 1% of £100. After 3 years, we would have earned only £3 instead of the £3.0301 we earned with compound interest.

Monthly compound interest

In the above example, regardless of whether we have compound interest or not, after the first year, we will receive £1 interest.

If the bank staff decided to pay the interest once a *month* instead of once a year, they would split the annual interest payment into 12 parts, and pay one part every month. In this way, if the interest rate were 1%, we would earn 1% ÷ 12 ≈ 0.08333% interest every month. This is the same as receiving 0.0008333 times the total amount in the account each month. If we have £100 invested, then at the end of the first month, we would receive:
£100 * 0.0008333 = £0.08333 = 8.3333 pence.

If we did not have compound interest, having monthly interest payments would achieve the same results as having the interest paid annually in one go – we would earn 8.3333 pence each month and end up with £1 interest at the end of the year. However, if we *did* have compound interest, we would receive interest calculated on the interest we had earned each month, and we would receive more money.

We would earn 0.0008333 times the amount in the bank account each month. At the end of each month, the total would be 1.000833 times as much as it was the month before.

At the end of the first month, we would receive 0.0008333 * £100 = 8.3333 pence interest. The total in the bank account would be 1.0008333 * £100 = £100.08333.

At the end of the second month, we would receive 0.0008333 * £100.08333 = 8.3402 pence interest and have a total of 1.0008333 * £100.08333 = £100.1667.

At the end of the third month, the total would be £100.2502.
At the end of the fourth month: £100.3338
At the end of the fifth month: £100.4174
At the end of the sixth month: £100.5010
At the end of the seventh month: £100.5848

At the end of the eighth month: £100.6686
At the end of the ninth month: £100.7525
At the end of the tenth month: £100.8365
At the end of the eleventh month: £100.9205
At the end of the twelfth month: £101.0046

Therefore, by having the interest paid monthly, over the course of a year, we would end up receiving 0.46 pence more than if the interest were paid just once a year.

Formulas for interest

It is easiest to think of the fraction that a percentage represents instead of thinking of the actual percentage rate. Therefore, in this section, I will give the yearly percentage rates in terms of their fractions. This will be the percentage rate divided by 100. In other words, I will say 0.05 instead of 5%, I will say 0.12 instead of 12%, and so on. I will call this fraction the "annual interest fraction".

We can make a formula to work out the total amount of money in the bank account at the end of twelve months. The total is equal to:

original amount * (1 + (annual interest fraction ÷ 12))^{12}

We can express this more mathematically as:
$y = a * (1 + (f \div 12))^{12}$
... where:

- "y" is the resulting total after one year's interest.
- "a" is the original amount.
- "f" is the annual interest fraction.
- "12" is the number of months in one year.

The annual interest fraction is the annual interest rate divided by 100. In other words, it is the amount that we would multiply by the original total to work out that year's interest if it were going to be paid in one go.

That number is then divided by 12, which is the number of months in one year, and also the number of payments in one year. This results in the fraction for one monthly payment.

We add 1 to this so that we include the original total in the result. If we did not do this, we would just end up with the amount paid instead of the original amount added to the amount paid.

We then multiply that by itself for every month we want to count. If we wanted two months, we would square it. If we wanted three months, we would cube it. As we want the result for a full 12 months of interest payments, we raise it the power of 12.

Next, we multiply that by the original amount in the bank account. The result is the original total summed with the paid interest over the full year.

A more general formula

We can make the formula relate less to years and months, so it becomes more general:

$y = a * (1 + (f \div n))^n$
... where:

- "y" is the total after one overall period of time.
- "a" is the original amount in the bank account.
- "f" is the interest fraction for one overall period.
- "n" is the number of payments made in one overall period.

Note that this formula is now independent of time. If the number of payments for an overall period is 12, then it does not matter if the overall period is years, months, seconds or anything. The result will still be the same.

For our original example of 1% interest each year on a £100 initial investment, with the interest paid monthly, we can fill in the formula as so:

Total after one year = $100 * (1 + (0.01 \div 12))^{12}$
≈ £101.004596

This matches the result we had before.

More intervals

We will look at what would happen if we convinced the bank to pay interest every day instead of every month. This would mean that we would receive 0.01 ÷ 365 ≈ 0.00002740 of the total every day. The total in the bank account after one year would be:
$100 * (1 + (0.01 \div 365))^{365} \approx$ £101.005003.

We have ended up receiving about 0.5 pence more than if the interest were only paid once a year. However, this is only an extra 0.0004 pence more than if the interest were paid every month.

If the bank paid compound interest every hour, the formula would be:
$100 * (1 + (0.01 \div 8760)^{8760} \approx$ £101.005016.

The result is only slightly more than before.

If the bank paid interest every second, the formula would be:
$100 * (1 + (0.01 \div 31{,}536{,}000)^{31{,}536{,}000} \approx$ £101.005017.

The extra amount is negligible. It is clear that after a certain point, increasing the payment intervals does not produce a significantly larger result. We will have diminishing returns on the number of interest payments. There is no way to trick the bank into paying a lot more money.

Finding the limit on interest

If we ignore the original bank balance of £100, and just concentrate on the formula, we can examine the limit more closely. The part of the formula that is multiplied by the original balance is this:
$(1 + (0.01 \div n))^n$
... where "n" is the number of interest payments.

We can put in different values of "n" to see what happens:
If "n" is 1, the result is 1.01
If "n" is 10, the result is 1.01004512021025
If "n" is 100, the result is 1.01004966209288
If "n" is 1000, the result is 1.01005011658
If "n" is 1,000,000, the result is 1.0100501669728

It turns out that we are just getting closer and closer to a number that begins 1.010050...

We will try different annual interest rates over a very large number of interest payments. We will say there are a million interest payments, so we have a good approximation of the limits for each interest rate. The general formula for this will be:
$(1 + (f \div 1{,}000{,}000)^{1{,}000{,}000}$
... where "f" is the interest rate fraction.

We will start with an interest rate of 0.01%, which is an interest fraction of 0.0001. We will then move up to higher and higher interest rates.

Annual Interest at 0.01%: $(1 + (0.0001 \div 1{,}000{,}000))^{1{,}000{,}000} \approx 1.00010001$
Annual Interest at 0.1%: $(1 + (0.001 \div 1{,}000{,}000))^{1{,}000{,}000} \approx 1.00100050$
Annual Interest at 1%: $(1 + (0.01 \div 1{,}000{,}000))^{1{,}000{,}000} \approx 1.01005017$
Annual Interest at 10%: $(1 + (0.1 \div 1{,}000{,}000))^{1{,}000{,}000} \approx 1.05170913$
Annual Interest at 100%: $(1 + (1 \div 1{,}000{,}000))^{1{,}000{,}000} \approx 2.71828047$
Annual Interest at 1000%: $(1 + 10 \div 1{,}000{,}000)^{1{,}000{,}000} \approx 22{,}025.36450783$

We can see that there is a range of results depending on the annual interest rate. These are all the approximate numbers that the scale of the interest payments will converge to for various annual interest rates. If we increased the number of interest payments, the accuracy for each result would be higher.

There is clearly some kind of pattern in these numbers, but exactly what, it is hard to tell. You might be able to figure it out with a calculator. It turns out that every result is actually the number 2.718281828... raised to the power of the interest fraction. In other words:

- When the interest rate is 0.01%, the fraction is 0.0001, and the result in the above table is $2.71828^{0.0001} \approx 1.00010001$
- When the interest rate is 0.1%, the fraction is 0.001, the result is $2.71828^{0.001} \approx 1.00100050$
- When the interest rate is 1%, the fraction is 0.01, the result is $2.71828^{0.01} \approx 1.01005017$
- When the interest rate is 10%, the fraction is 0.1, the result is $2.71828^{0.1} \approx 1.10517092$
- When the interest rate is 100%, the fraction is 1, the result is $2.71828^{1} \approx 2.71828183$
- When the interest rate is 1000%, the fraction is 10, the result is $2.71828^{10} \approx 22{,}026.4658$

The number 2.718281828... is of course the number generally known as “e”.

For any interest fraction, if a single interest payment period is split up into an infinite number of smaller periods and paid with compound interest, the maximum interest achievable will be 2.71828 raised to the power of that fraction.

In other words, if we have £100 in a bank account and an annual interest rate of 1%, it does not matter whether the bank pays the annual interest split up into months, days, minutes, seconds, or microseconds, the most money we will ever have at the end of the year will be:
$100 * e^{0.01} = £101.00501671$

Similarly, if we have £1 in a bank account, and an annual interest rate of 100%, even if the interest were paid at an infinite number of intervals throughout the year, the most money we can ever have at the end of the year will be:
$1 * e^{1} = £2.718281828$

Formulas to approximate e

Given all the above, we have a formula to approximate “e”:
$e \approx (1 + (1 \div n))^{n}$
... where “n” is any very large number. The bigger it is, the more accurate the result will be.

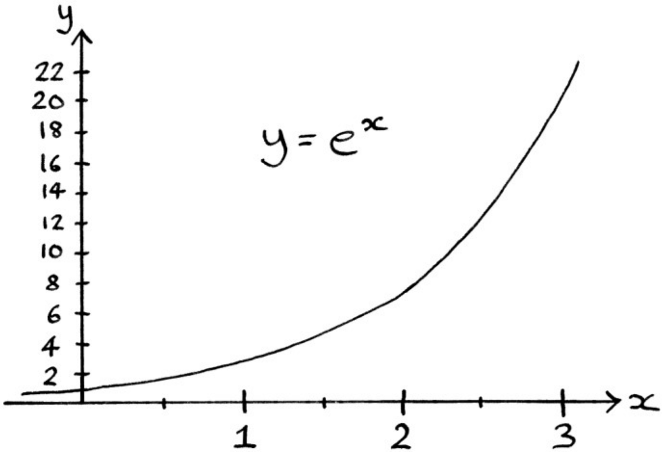
We also have a formula to approximate “e” raised to any power:
$e^{x} \approx (1 + (x \div n))^{n}$
... where “x” is the number to which we want to raise “e”, and “n” is any very large number.

All of the above is the first part of why “e” is an interesting number. From a situation where we might think there is no limit, or maybe a limit with no obvious rules, it turns out that there is a limit that is based around one particular number: “e”.

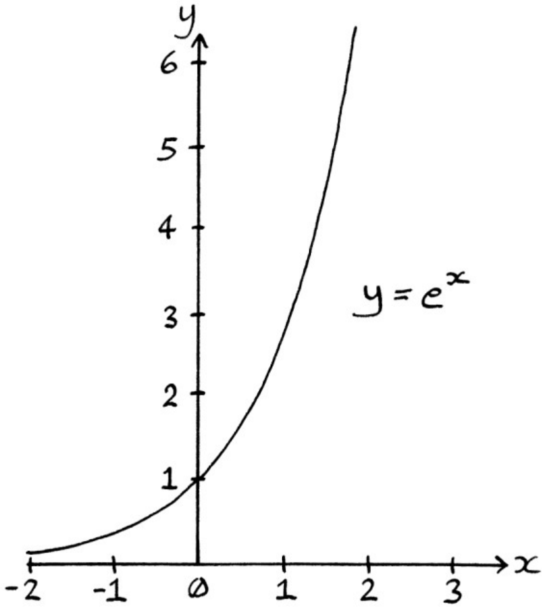
e as the base of an exponential

From the previous section, you might notice that it is not particularly 2.7183 on its own that is the useful number, but 2.7183 raised to a power. In fact, this is true for most of the maths in which "e" is used.

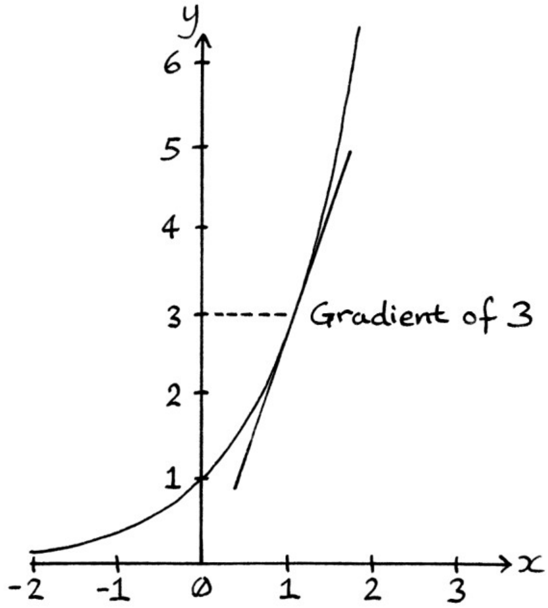
We can plot the graph of "$y = e^x$", and it looks like this (with the x and y-axes drawn to different scales to show a wider range of the formula's characteristics):



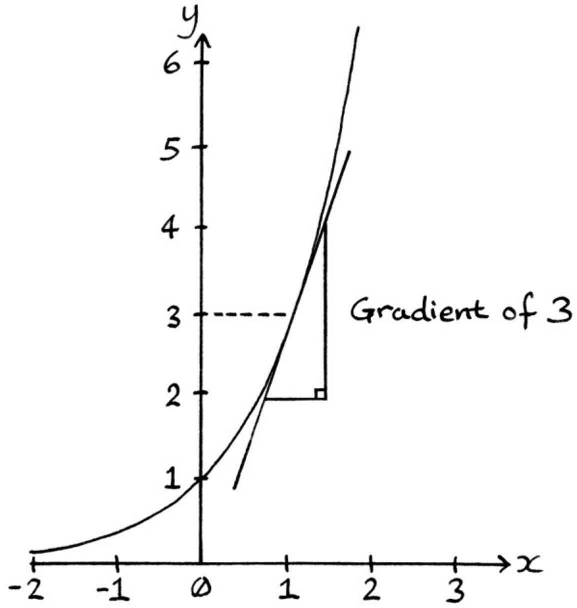
If we draw the x and y-axes to the *same* scale, the graph looks like this:



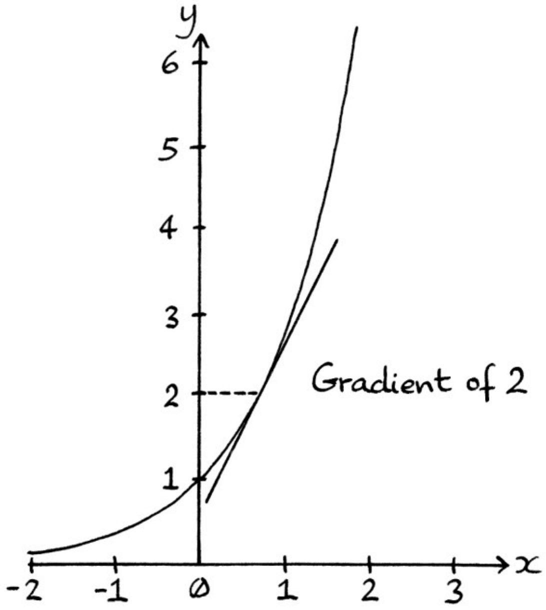
At first glance, there is nothing particularly special about the graph of "$y = e^x$" when the axes are drawn to the same scale. It looks similar to any graph of the type "$y = c^x$". What is interesting about this graph is that the gradient of any point on the curve is equal to the y-axis value at that point. For example, when "y" is 3, the gradient of the curve is also 3. If we were to draw a straight line that was touching the curve at the point where y = 3, the gradient of that line would be 3. [Another way of saying this is that if we were to draw a line that was "tangential" to the curve at that point, its gradient would be 3.]



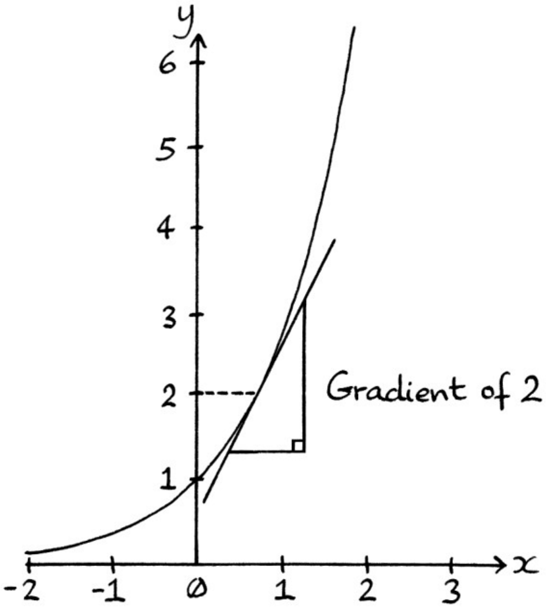
We can check the gradient of the line by drawing a right-angled triangle against it, with part of the line acting as the hypotenuse. Whatever the size of the right-angled triangle, the opposite side divided by the adjacent side will be equal to 3.



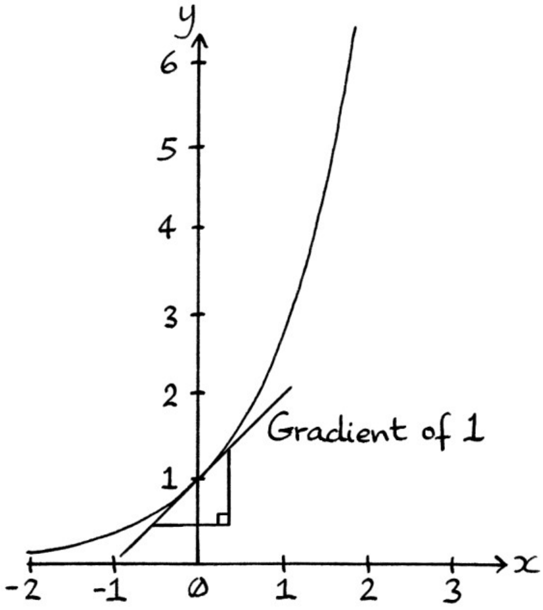
We can draw a straight line that is tangential to the curve when y = 2. That line will have a gradient of 2:



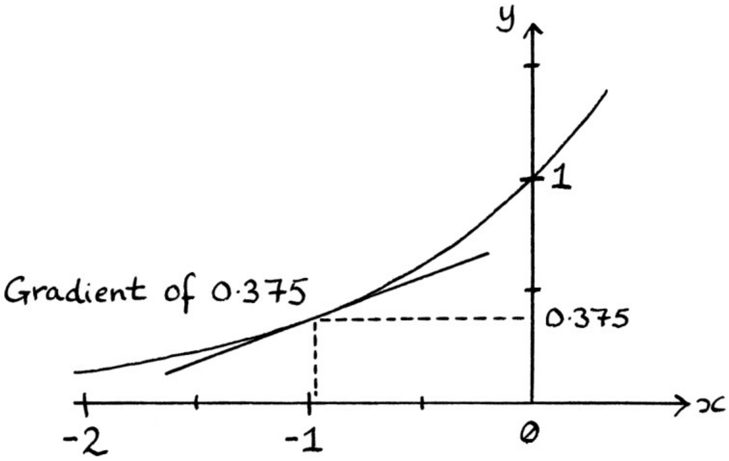
We can check the gradient of this line by placing the hypotenuse of a right-angled triangle against it. However large the triangle is, the opposite side divided by the adjacent side will result in the number 2. To put this another way, the opposite side will be twice the size of the adjacent side.



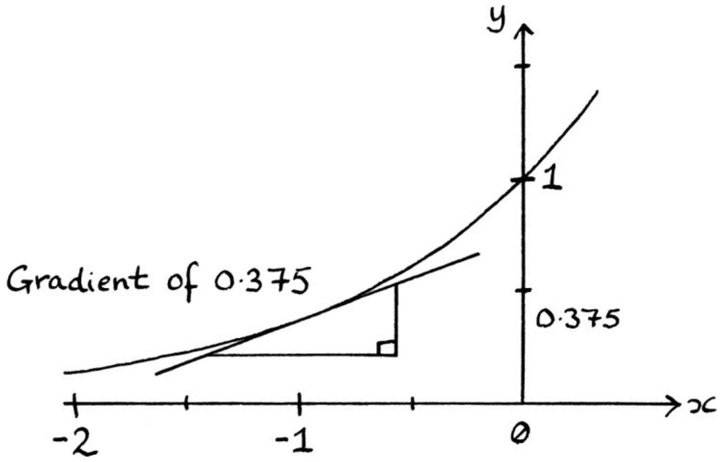
The most obvious example is at y = 1, when the gradient is 1, and the angle is therefore 45 degrees.



The rule still works for values under y = 1. For example, when “y” is 0.375, the gradient is 0.375.



We can check this with a right-angled triangle.



The curve at every point of “y = e^x” has a gradient equal to the y-axis value at that point. Another way of saying this is that the graph of “y = e^x” is also the graph of its own gradients. [If we were talking in terms of calculus, we would say that the derivative of “y = e^x” is “y = e^x”. We could also say that “y = e^x” is infinitely differentiable because however many times it is differentiated, it remains the same. If you do not understand anything about calculus, then it is enough to know that the gradients of “y = e^x” can be portrayed using the formula “y = e^x”.]

The properties of “y = e^x” are unique. The graph of “y = 2.7^x” does not behave in this way, and neither does the graph of “y = 2.8^x”. The value 2.71828... is just right for its graph to have this quality. If we wanted an exponential that showed its own gradient, then we would end up with “y = e^x”.

Notation

The expression “e^x” is sometimes written as “e^x” if the text editor being used cannot do raised-up text. Another way of writing it is as “exp(x)”, where “exp” is shorthand for the “exponential with ‘e’ as the base”. The expression “e^2” would be written as “exp(2)”.

e to the power of ix

As we saw earlier, we can calculate “e” raised to any power by using this formula:
$e^x \approx (1 + (x \div n))^n$
... where “x” is the number to which we want to raise “e”, and “n” is any very large number.

This method works for any value of “x”. Where this idea becomes interesting is when we try values of “x” that are Imaginary numbers. For example, if we used “1i” as the exponent, then we would be trying to calculate “e^{1i}”, or what would normally be written as just “e^i”.

At this level of the explanation, it is probably difficult to visualise what the number “e^i” even means. The first thing to remember when thinking of this is that “e” is the number 2.71828..., so we are looking at 2.71828^i. It is easy to forget that “e” is an actual number and not a variable.

We will use the $(1 + (x \div n))^n$ formula to calculate “e^i”, which to make things more straightforward, we will here write as “e^{1i}”.

We put “1i” into the formula and choose a large number for “n”:
$(1 + (1i \div 10{,}000))^{10{,}000}$
$= (1 + (0.0001i))^{10{,}000}$
$= (1 + 0.0001i)^{10{,}000}$

This is a Complex number raised to the power of 10,000.

We can solve this by multiplying “1 + 0.0001i” by itself 10,000 times.

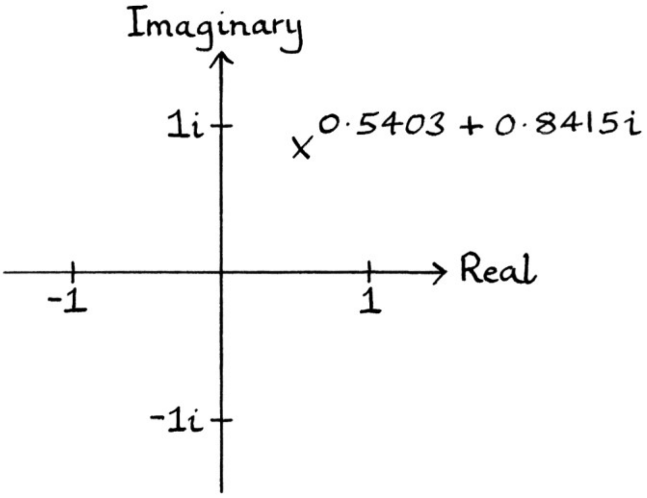
First, we multiply “1 + 0.0001i” by itself:
$(1 + 0.0001i) * (1 + 0.0001i)$
$= 0.0001i + 0.0001i + 1 + (0.0001i)^2$
$= 0.0002i + 1 - 0.00000001$
$= 0.0002i + 0.99999999$
$= 0.99999999 + 0.0002i$

Then, we multiply that result by “1 + 0.0001i”:
$(0.99999999 + 0.0002i) * (1 + 0.0001i) = 0.99999997 + 0.000299999999i$

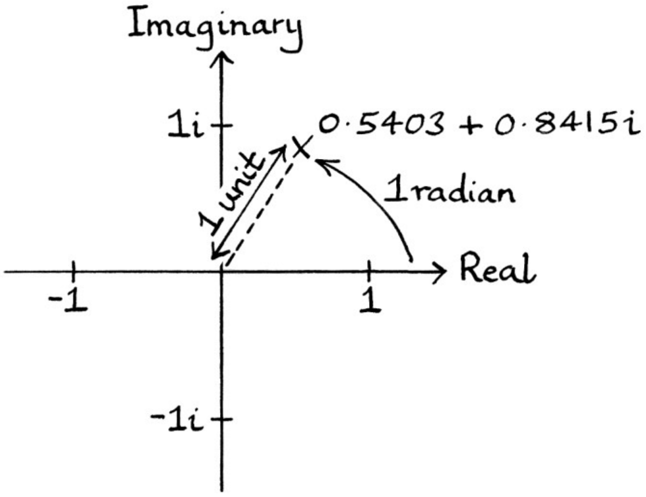
We then multiply that result by “1 + 0.0001i”, and continue in this way until we have achieved “1 + 0.0001i” multiplied by itself 10,000 times. It would be possible to do this on a piece of paper, but it would take a very long time. It is quicker to use a calculator that can work with Complex numbers, or even to write a small computer program to do it.

Eventually, we will end up with “0.5403 + 0.8415i” (to 4 decimal places).

We can plot this point in the Complex plane:

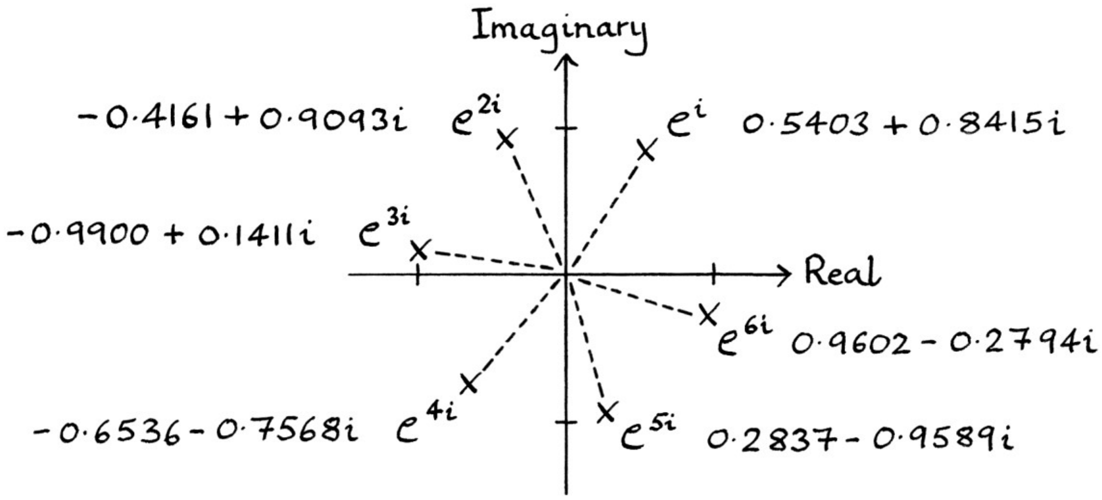


What is interesting here is that the point is at an angle of exactly 1 radian (57.2958 degrees), and it is exactly 1 unit from the origin of the axes.



To obtain a better understanding of what is happening here, we can try other Imaginary values too. For example:
$e^{2i} = -0.4161 + 0.9093i$
$e^{3i} = -0.9900 + 0.1411i$
$e^{4i} = -0.6536 - 0.7568i$
$e^{5i} = 0.2837 - 0.9589i$
$e^{6i} = 0.9602 - 0.2794i$
$e^{7i} = 0.7539 + 0.6570i$
$e^{8i} = -0.1455 + 0.9894i$
$e^{9i} = -0.9111 + 0.4121i$
$e^{10i} = -0.8391 - 0.5440i$

We will plot the first few of these on the Complex plane:



From this set of plotted points, there are three very interesting observations to note:

- Every single point here is exactly one unit from the origin of the axes. The points are all on the edge of the circumference of a unit circle. We can see that "e" to the power of multiples of "i" always produces results that are one unit away from the origin.

- Each point is exactly one radian apart.

- The non-Imaginary part of each exponent is the angle in radians of the point on the unit-radius circle's edge. For example, for "e^{3i}", the point is at 3 radians on the edge of the circle.

To summarise these observations, for any value of “x”, the exponential “e^{ix}” results in a Complex number describing a point on a unit-radius circle at an angle of “x” radians. Given that, we can use the Cosine and Sine of the angle to produce the Real and Imaginary parts of the resulting Complex numbers (when Cosine and Sine are working in radians). For example, “e” raised to the power of “3i” results in a Complex number where the Real part is “cos 3 radians”, and the Imaginary part is “sin 3 radians”. This means that “e^{ix}”, where “ix” is any Imaginary number, results in a Complex number that is the same as “cos x + i sin x”, where “x” is in radians, and Cosine and Sine are working in radians.

From all of this, we can see that “e” raised to an Imaginary power works in a similar way to “i” raised to a power, as explained in Chapter 25. The values “e^{ix}” and “i^{x}” both identify a point on a unit-radius circle based on treating “x” as an angle. Imaginary powers of “e” work in radians; powers of “i” work in quarter-circle angle units.

In addition to this, in a similar way to how a multiplication by “i^{x}” rotates a point anticlockwise by “x” quarter-circle angle units, so does a multiplication by “e^{ix}” rotate a point anticlockwise by “x” radians. [Similarly, in the same way to how a division by “i^{x}” rotates a point *clockwise* by “x” quarter-circle angle units, so does a division by “e^{ix}” rotate a point *clockwise* by “x” radians.]

As the “x” in the “e^{ix}” and “i^{x}” are really angles, I will phrase them as “$e^{i\theta}$” and “i^{θ}” from now on.

More about e to an Imaginary power

In Chapter 25, we saw how a multiplication by “i^{θ}” rotates a point on the Complex plane by “θ” quarter-circle angle units anticlockwise. From knowing that, we can use it to identify the position of a point of interest by saying how much the point at “1 + 0i” would need to be rotated and scaled to get there. [For example, the point at “2.1213 + 2.1213i” can be identified by “$3i^{0.5}$” because, the point at “1 + 0i” would need to be multiplied by “$3i^{0.5}$” to become rotated and scaled to be there. The point at “1 + 0i” would need to be rotated by 0.5 quarter-circle angle units, and its distance from the origin of the axes scaled by 3.]

The same idea is true for “$e^{i\theta}$”, except that it works in radians instead of quarter-circle angle units. We can use a multiplication by “$e^{i\theta}$” to rotate a point on the Complex plane by “θ” radians anticlockwise. Given that, we can use “$e^{i\theta}$” to identify the position of a point of interest by saying how much the point at “1 + 0i” would need to be rotated and scaled to reach that point of interest. [For example, the point at “2.1213 + 2.1213i” can be identified by “$3e^{0.25\pi i}$”. The point at “1 + 0i” needs to be rotated by 0.25π radians, and its distance from the origin of the axes scaled by 3, for it to reach “2.1213 + 2.1213i”.]

“$e^{i\theta}$” and “i^{θ}” work in the same basic way. The difference is that how we are usually introduced to “$e^{i\theta}$” makes it seem more of a way of *identifying* a point on the Complex plane, and how we are usually introduced to “i^{θ}” makes it seem more of a way of *rotating* a point on the Complex plane. In reality, each of them can do both.

There are two main differences between “$e^{i\theta}$” and “i^{θ}”:

- First, “$e^{i\theta}$” rotates or identifies a point on a unit-radius circle when “θ” is an angle in *radians*. This is interesting because it is an example of how radians occur naturally in maths. There is no social or cultural influence in the appearance of radians in the use of “$e^{i\theta}$”. This means that radians are one of the natural ways of dividing a circle. The exponential “i^{θ}” also rotates or identifies a point on a unit-radius circle, but in that case “θ” is an angle in quarter-circle angle units. We can say that quarter-circle angle units are also a natural way of dividing a circle. [Generally, radians are more useful than quarter-circle angle units though.]

- Second, “$e^{i\theta}$” has “i” in the exponent, while “i^{θ}” has “i” in the base.

As we saw in Chapter 25, we can alter the “i^{θ}” way of identifying or rotating a point to use “θ” as an angle in radians. We do this by turning it into “$i^{(2\theta/\pi)}$”. As this formula uses radians, we can say it has exactly the same effect as “$e^{i\theta}$”. They are identical in what they do. We can say that:
“$e^{i\theta} = i^{(2\theta/\pi)}$”
... for any value of “θ”.

[Remember that for “$i^{(2\theta/\pi)}$”, when entering an angle as “θ” in radians, the angle is immediately converted into quarter-circle angle units. Therefore, “2θ/π” is actually an angle in quarter-circle angle units, despite “θ” being an angle in radians.]

We can also alter “$e^{i\theta}$” so that “θ” works in quarter-circle angle units. The exponent incorporates the conversion from quarter-circle angle units into radians, as so:

“$e^{i * 2\pi * (\theta/4)}$”

... which is:

“$e^{(i2\pi\theta)/4}$”

... which is:

“$e^{(i\pi\theta)/2}$”

... which is:

“$e^{0.5i\pi\theta}$”

... which, for no reason than I think the order of symbols looks better, I will write as:

“$e^{0.5\pi i\theta}$”

An angle in quarter-circle angle units is entered as “θ”, and the result will indicate a point on a unit-radius circle’s edge at that angle. We can therefore say:

“$i^{\theta} = e^{0.5\pi i\theta}$”

[Note that when we enter an angle as “θ” in quarter-circle angle units, it is immediately converted into radians. Therefore, “0.5πiθ” is actually an angle in radians, despite “θ” being an angle in quarter-circle angle units.]

Given what we know about “$i^{(2\theta/\pi)}$”, we can say:

“$e^{i\theta} = i^{(2\theta/\pi)} = \cos\theta + i \sin\theta$”

... or

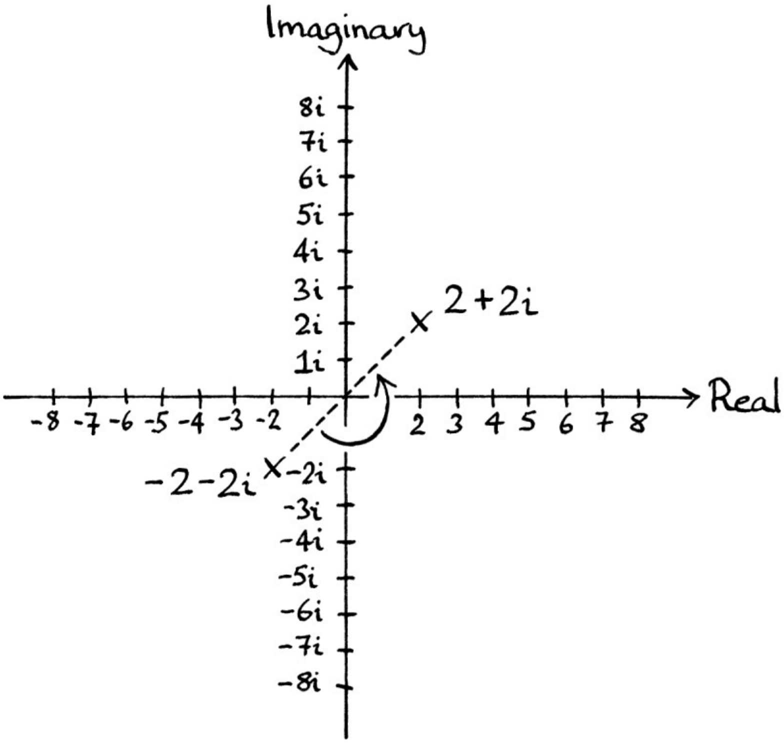
“$e^{i\theta} = \cos\theta + i \sin\theta$”

... where “θ” is an angle in radians, and Cosine and Sine are working in radians. If “$e^{i\theta}$” is equal to “$i^{(2\theta/\pi)}$”, then it should not be surprising that “$e^{i\theta} = \cos\theta + i \sin\theta$”, when “θ” is an angle in radians, and Cosine and Sine are working in radians. Such an idea was explained in Chapter 25.

Rotation

We will look at rotating a point by multiplying it by "$e^{i\theta}$". We will rotate the point at "−2 − 2i" by π radians (180 degrees) anticlockwise. Therefore, we need to solve this calculation:
$e^{i\pi} * (-2 - 2i)$
$= -2e^{i\pi} - 2ie^{i\pi}$

This is a difficult thing to solve in its present form. A good calculator can solve this as "2 + 2i". [We know this is the correct answer because we know that "−2 − 2i" rotated by 180 degrees (π radians) is "2 + 2i"].



One way to solve it by hand is to use our knowledge that "$e^{i\theta} = \cos\theta + i\sin\theta$", and then substitute Cosine and Sine into the formula:
$-2 * (\cos\pi + i\sin\pi) - 2i * (\cos\pi + i\sin\pi)$ [in radians]
$= -2\cos\pi - 2i\sin\pi - 2i\cos\pi - 2ii\sin\pi$
$= -2\cos\pi - 2i\sin\pi - 2i\cos\pi + 2\sin\pi$
$= -2\cos\pi + 2\sin\pi - 2i\sin\pi - 2i\cos\pi$
... which because π radians is 180 degrees, and we can calculate the Cosine and Sine of that in our heads:
$= -2 * -1 + (2 * 0) - (2i * 0) - (2i * -1)$
$= +2 - -2i$
$= 2 + 2i$

Supposing we had done the same calculation with "$i^{(2\theta/\pi)}$", we would have had to solve:
$(-2 -2i) * i^{(2*\pi/\pi)}$
$= (-2 -2i) * i^2$
$= -2i^2 - 2i^3$
$= 2 - (2 * -1i)$
$= 2 - -2i$
$= 2 + 2i$

In this particular case, it is *much* easier to solve the calculation with a power of "i" than it is to solve it with an Imaginary power of "e". When using "e", we had to substitute in Cosine and Sine, while when using "i", we could stay with powers of "i".

Given that "$e^{i\theta}$" is the same as "$i^{(2\theta/\pi)}$", which is the same as "cos θ + i sin θ", when "θ" is an angle in radians, and Cosine and Sine are working in radians, it means that any time we have a difficult calculation with one of these three, we can substitute in one of the others. This will sometimes make things easier.

Identifying a point on the Complex plane

If we have a point that is 3 units from the origin of the axes, and at an angle of 0.5 radians, then there are several ways we can indicate its position:

- We can say it is at "3 cos 0.5 + 3i sin 0.5", when 0.5 is an angle in radians, and Cosine and Sine are working in radians.

- We can say it is at "$3i^{((2*0.5)/\pi)}$" = $3i^{(1/\pi)}$. As we know, this method of identification is really showing the amount the point at "1 + 0i" would need to be multiplied by to become rotated and scaled to reach that point.

- We can say it is at "$3e^{0.5i}$". This method of identification is also showing the amount the point at "1 + 0i" would need to be multiplied by to become rotated and scaled to reach that point.

All of these are exactly the same point. It is easiest to calculate the Complex number that indicates the point's position by using the Cosine and Sine version. The point is at "3 cos 0.5 + 3i sin 0.5" (in radians), which is "2.6327 + 1.4383i".

We could work out the Complex number that is the same as $3i^{(1/\pi)}$, by converting "1/π" quarter-circle angle units into degrees, and then seeing where a point at that angle would be on a 3-unit radius circle. [Remember that "i" raised to a power works in quarter-circle angle units. When we use it to work with "θ" in radians, we are really just adjusting the angle in radians to be in quarter-circle angle units. Therefore, the "1/π" in this example is not in radians, but in quarter-circle angle units.] The angle of "1/π" quarter-circle angle units is ((1/π) / 4) * 360 = 28.6479 degrees. We could mark that point on a circle with a radius of 3 units, and measure the x-axis and y-axis positions. Alternatively, we could use Cosine and Sine: 3 cos 28.6479 (in degrees) and 3i sin 28.6479 (in degrees). This is the same as 3 cos 0.5 and 3i sin 0.5 in radians. We would end up with "2.6327 + 1.4383i".

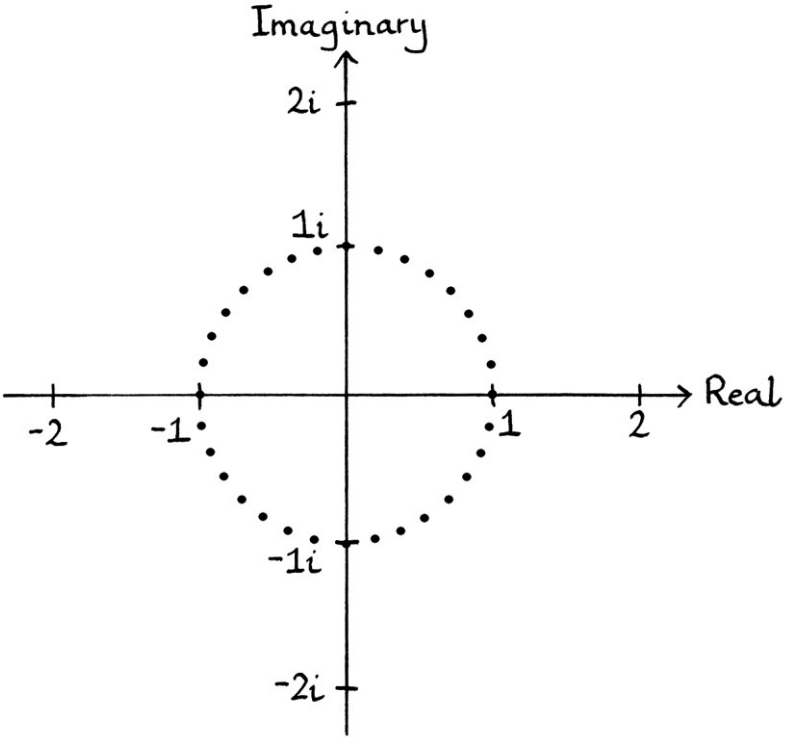
If we use a calculator that can work with Complex numbers to work out any of the three ways of identifying a point, we will end up with "2.6327 + 1.4383i"

From all of this, we can say that:
"3 cos 0.5 + 3i sin 0.5" in radians
"3 cos 28.6479 + 3i sin 28.6479" in degrees
"3 cos (1 ÷ π) + 3i sin (1 ÷ π)" in quarter-circle angle units
"$3i^{(1/\pi)}$"
"$3e^{0.5i}$"
"2.6327 + 1.4383i"
"(2.6327, 1.4383)" [as coordinates]
... all refer to exactly the same point.

Circles

When we looked at “i^{θ}” in Chapter 25, we saw how, if we use a range of angles for “θ”, we can draw a circle. In fact, we can say that “i^{θ}” and its variations *describe* a circle. Everything that we can do with “i^{θ}”, we can also do with “$e^{i\theta}$”.

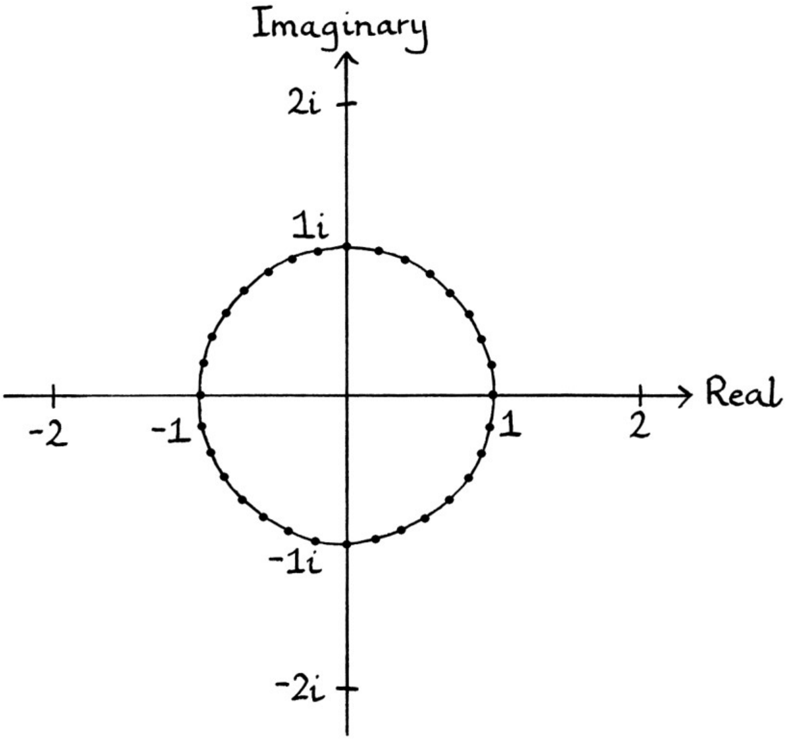
If we repeatedly multiply the point “1 + 0i” by “$e^{0.0625\pi i}$” and draw a dot at each angle, we will end up with this picture:



[0.0625π radians is 11.25 degrees or a thirty-second of a circle].

This is the same picture as in Chapter 25, when we repeatedly multiplied the point “1 + 0i” by “$i^{0.125}$”. [0.125 quarter-circle angle units is the same as 0.0625π radians.]

We can join up the dots and we will have a circle:



If we use smaller intervals, we will draw out a more accurate circle.

An Imaginary power of “e” can be used to identify a point, as well as being used in a multiplication. This means that we can draw a circle by just using “$e^{i\theta}$” with various values of θ, without needing to treat each one as a multiplication.

The formula for “$e^{i\theta}$” with a range of angles from 0 up to 2π radians will draw out a unit-radius circle. This might be obvious given that we could draw out a circle using “cos θ + i sin θ” or “i^{θ}”, and also given that “$e^{i\theta}$” is equal to “cos θ + i sin θ” in radians.

Not only does “$e^{i\theta}$”, with a range of angles from 0 up to 2π radians, draw out a unit-radius circle, but we can say that “$e^{i\theta}$” with a range of angles *is* a unit-radius circle. We can use it as a formula for a circle:
“$z = e^{i\theta}$”
... where:

- “z” indicates the range of points on the Complex plane.
- “e” is the number 2.71828182...
- “i” is the square root of −1.
- “θ” is the range of angles in radians from 0 up to 2π [or higher].

We can scale the radius of this circle by scaling "$e^{i\theta}$". Therefore, "$z = 2e^{i\theta}$" with angles from 0 up to 2π radians will draw a circle with a radius of 2 units. "$z = 0.5e^{i\theta}$" with angles from 0 up to 2π radians will draw a circle with a radius of 0.5 units.

Now that we have a formula for a circle, we can think about the derived waves from that circle. The circle based on "$z = e^{i\theta}$" will have a derived Sine wave and Cosine wave with amplitudes of 1 unit. The circle based on "$z = 2e^{i\theta}$" will have a derived Sine wave and Cosine wave with amplitudes of 2 units.

Time-based waves

Imaginary powers of "e" become more useful when we use them to represent an object rotating around a circle – in other words, when we incorporate time.

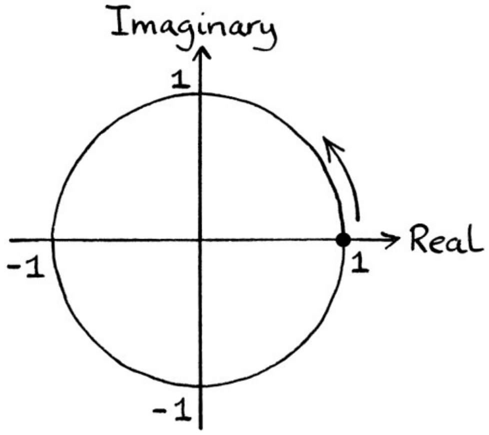
To use time in the "$z = e^{i\theta}$" formula, we need to adjust the time to fit in with the angle system. As "$e^{i\theta}$" treats "θ" as an angle in radians, we need to multiply the time by 2π, so that it can be subjected to being an Imaginary power of "e". Our formula for a basic time-based circle is:
"$z = e^{i(2\pi * t)}$"

The time is multiplied by 2π so it can be treated as an angle, and so that there is a default frequency of one cycle per second. Then it is multiplied by "i", and that all becomes an exponent of "e". We have replaced "θ" in "$e^{i\theta}$" with 2π * t.

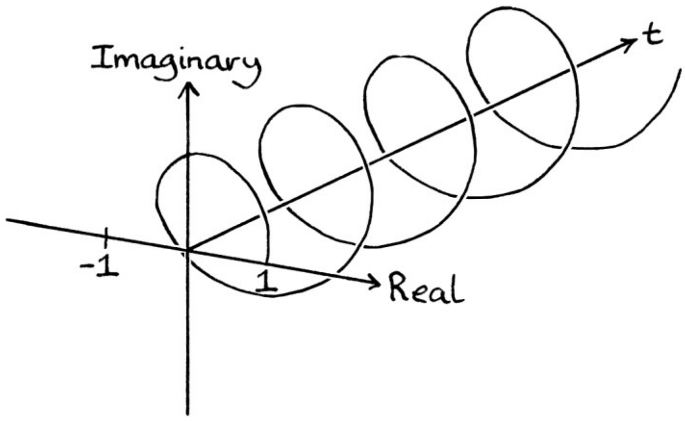
We can rephrase the exponential as:
"$z = e^{(i * 2\pi * t)}$"
... or:
"$z = e^{(i2\pi\ t)}$"
... or, with an order of symbols that is easier to read, as:
"$z = e^{2\pi it}$"

The exponential represents an object rotating around a unit-radius circle at a frequency of 1 cycle per second, with zero phase and zero mean levels. It also implies the derived Sine and Cosine waves, which have the same frequency as each other, amplitudes of 1 unit, zero phases, and zero mean levels.

We can portray this on the Complex plane:



... or we can portray this on the Complex helix chart:



When we alter aspects of the wave to incorporate time, it is important to note that the frequency and phase are replacing the "θ". Therefore, they need to be subjected to being multiplied by "i", and then being an exponent of "e". Therefore, the best way to phrase the formula from the point of view of minimising potential mistakes is:

"$z = e^{i(2\pi t)}$"

Radius and amplitude

We can change the radius of the time circle. For example, a radius of 3 units would be portrayed as so:
“$z = 3e^{2\pi it}$”.
[As always, remember that this is 3 * ($e^{2\pi it}$) and not $(3e)^{2\pi it}$.]

This circle would have derived waves with amplitudes of 3 units.

Frequency

We can change the frequency. A frequency of 5 cycles per second would look like this:
“$z = e^{i * (2\pi * 5t)}$”
... which could also be written as so:
“$z = e^{i(2\pi * 5t)}$”
... or, in this particular case:
“$z = e^{2\pi i * 5t}$”

It is important to remember that it is the non-Imaginary part of the exponent that portrays the frequency. The frequency and the 2π are replacing the “θ” in the original formula. The whole exponent is best thought of as “i * (2π * 5t)” instead of “2πi * 5t”. It pays to make this distinction because of what happens next when we incorporate phase.

Phase

We can alter the phase. The phase has to be given in radians, given how “e” raised to Imaginary powers treats the non-Imaginary part of the exponent as an angle in radians. The phase must be placed in the exponent so that it is grouped with the frequency and 2π, and then multiplied by “i”. We are really swapping the “θ” in “$e^{i\theta}$” with the frequency, phase and 2π. Therefore, a phase of π radians (180 degrees) would look like this:
“$z = e^{i * (2\pi t + \pi)}$”
... or:
“$z = e^{i(2\pi t + \pi)}$”
... which could, if we wanted, be rephrased as:
“$z = e^{(2\pi it + \pi i)}$”

With phase, it is easy to make the mistake of thinking that the exponent should be "$2\pi it + \pi$" (which is wrong), when it should be "$i * (2\pi t + \pi)$". The frequency, phase and the 2π are replacing the "θ" in "$e^{i\theta}$". Therefore, in the time formula, they all become multiplied by "i".

Mean levels

We can also alter the mean levels. The x-axis or Cosine mean level is portrayed with a Real number, and the y-axis or Sine mean level is portrayed with an Imaginary number. This is the same as when we were using "i" raised to a power in Chapter 25. As an example, we will give the circle a horizontal mean level of 1 and a vertical mean level of 2:
"$z = 1 + 2i + e^{2\pi it}$"

A general formula

A general formula for a circle using "e" raised to an Imaginary power is:
"$z = h_c + ih_s + ae^{i(2\pi ft+\phi)}$"
... where:

- "z" indicates the range of points on the Complex plane.
- "h_c" is the horizontal mean level, which will end up as the mean level for the derived Cosine wave.
- "h_s" is the vertical mean level, which will end up as the mean level for the derived Sine wave. This has to be an Imaginary number so that it raises or lowers the circle up or down the Imaginary axis.
- "a" is the radius, which is also the amplitude of the derived waves.
- "e" is the number 2.7182...
- "i" is the square root of −1.
- "2π" is there to scale the time in seconds, so that the time is treated as an angle in radians. This is needed because Imaginary powers of "e" treat the non-Imaginary exponent as an angle in radians.
- "f" is the frequency in cycles per second.
- "t" is the time in seconds.
- "ϕ" is the phase in radians.

Derived waves

If, for example, we have the formula for a circle "$z = 3e^{i(2\pi * 4t)}$", then we know that the derived waves will be "$y = 3 \sin (2\pi * 4t)$" and "$y = 3 \cos (2\pi * 4t)$".

Using terms we learnt in Chapter 23, we can also say:
$\text{Re}\ \{3e^{i(2\pi * 4t)}\} = 3 \cos (2\pi * 4t)$
... and:
$\text{Im}\ \{3e^{i(2\pi * 4t)}\} = 3 \sin (2\pi * 4t)$

The above lines mean that the Real part of the Complex exponential formula:
"$z = 3e^{i(2\pi * 4t)}$"
... is the wave:
"$y = 3 \cos (2\pi * 4t)$"
... while the Imaginary part is the wave:
"$y = 3 \cos (2\pi * 4t)$".

We could also say, after having introduced the formula "$z = 3e^{i(2\pi * 4t)}$", that:
$\text{Re}\ \{z\} = 3 \cos (2\pi * 4t)$
... and:
$\text{Im}\ \{z\} = 3 \sin (2\pi * 4t)$

Why we would want to use exponentials

When we portray a time-based circle using Imaginary powers of "e", there are several advantages:

- By having a formula for a circle, we are really incorporating both derived waves at the same time. This makes everything more concise.
- Certain mathematical calculations on the exponential will also cause the derived waves to undergo those calculations too. For example, if we add two exponentials, we are also adding the derived waves at the same time. If we multiply the exponential by a number, we are also scaling the amplitudes of each of the two derived waves in the same way.
- By adding or subtracting two Imaginary powers of "e" with opposing frequencies, we progress to the next stage of dealing with waves. We will do this in Chapter 28.

The disadvantages of using exponentials are:

- The idea of an exponential representing a circle or waves can be extremely confusing if you have not seen all the steps leading up to the idea. If you have read this book from the start, the idea should be straightforward, even if it seems a bit unwieldy.
- To start with, they require more thought.
- We cannot perform the multiplication of two circles described by exponentials by multiplying the exponentials, and maintain the same relationship with the derived waves. This problem is the same as the one we encountered with "cos θ + i sin θ" in Chapter 24, and "i^{θ}" in Chapter 25.]

Circles and helices

Previously, in this book, I have said that a time-based circle can be thought of as a circle or it can be thought of as a helix – it all depends on what it is we are trying to portray. A helix is useful for portraying the movement of an object rotating around a circle over time, but it is harder to draw, and it is difficult to take measurements from a helix. A circle is a good concise way of portraying the position of an object, but it is harder to distinguish between cycles.

When dealing with time-based Imaginary exponents of "e", it can often be useful to think of them as helices (and helix-like shapes) on the Complex helix chart. The reason for this will become clearer in the next chapter.

The Complex exponential

The exponential "$e^{i\theta}$" and its variations are often referred to as "the Complex exponential". Although there are countless other exponentials that can use Complex numbers, when people refer to "*The* Complex exponential", they are referring to "$e^{i\theta}$" or its variations such as "$e^{2\pi it}$". Using this term is very common, but it can suggest that other Complex exponentials do not exist. A power of "i" is really a Complex exponential, but it is not considered *the* Complex exponential. More confusingly, it is common to refer to "*a* Complex exponential", the meaning of which is dependent on the context. If we are talking about Imaginary powers of "e", then "a Complex exponential" is an Imaginary power of "e". If we are talking generally, then "a Complex exponential" is any exponential that contains Imaginary or Complex numbers. The term "Complex exponential" is sometimes abbreviated to "CE".

The reason why the terms “the Complex exponential” and “a Complex exponential” are used is that they are easier and quicker to write and say, than something such as “exponentials that are Imaginary powers of “e”. The terms could be ambiguous in certain situations, but the usefulness of the terms makes the potential ambiguity worthwhile.

Potential sources of confusion so far

Here are some of the possible sources of confusion surrounding “e” raised to an Imaginary power that you are likely to have, given everything I have said so far in this chapter.

“e” is a number and not an unknown variable

It can be very easy to forget that “e” is a symbol that represents the number 2.718281828459045... Sometimes, when performing calculations, it can be easy to forget that we actually know what “e” is. We can easily calculate something such as “e^2” as it is 2.7183^2. Given the complexity of what we are doing with “$e^{i\theta}$”, it is very easy to forget that “e” is not an unknown variable. We type “e” instead of “2.7182...” to save time and space. Some maths might be easier to understand more thoroughly if we did write 2.7182... instead. This chapter is based around what we can do with variations of “$e^{i\theta}$”, but we could also say it is based around what we can do with variations of “$2.7183^{i\theta}$”, as that is the same thing.

“$e^{i\theta}$” is often misunderstood

Many people who use “e” raised to Imaginary powers do not really understand what it is. They treat it as a magical formula. The idea that “$e^{i\theta}$” is equal to “$\cos \theta + i \sin \theta$” when “$\theta$” is an angle in radians, and Cosine and Sine are working in radians is treated as a miracle, and not an expected fact relating to points on a circle’s edge. The confusion about Imaginary powers of “e” leads to formulas such as “$e^{i\theta}$” being treated, not as calculations, but as static symbols. By this I mean, “e^i” could be replaced by a symbol and it would not change many people’s usage or understanding of it.

Having said all that, it does not particularly matter if people do not understand what Imaginary powers of "e" mean, because one does not need to understand how something works in order to use it.

"$e^{i\theta}$" only works with radians

The formula of "$e^{i\theta}$" indicates the position of a point on a unit-radius circle when "θ" is an angle in *radians*. The formula treats "θ" as an angle in radians regardless of whether we want it to or not. If we supply it with "θ" as an angle in *degrees*, then being an Imaginary exponent of "e" will cause "θ" to be treated as if it were in radians, and any calculations will not work as expected. You will *frequently* see examples in books where an author has incorrectly used Imaginary exponents of "e" with angles in degrees.

The most obvious example that "$e^{i\theta}$" works in radians is when we show the point indicated by "e^{1i}" – this will be a point one unit away from the origin at an angle of 1 radian. This point can also be described as "cos 1 + i sin 1" when Cosine and Sine are working in radians. We can see from this that it cannot be the case that that point could ever be "cos 1 + i sin 1" when Cosine and Sine are working in degrees. It should be clear that "cos 1 + i sin 1" when Cosine and Sine are working in radians is not the same point as "cos 1 + i sin 1" when Cosine and Sine are working in degrees. The point indicated by "cos 1 + i sin 1" when Cosine and Sine are working in degrees is a point that is 1 unit away from the origin and at an angle of 1 degree.

The "θ" in "$e^{i\theta}$" is treated as an angle in radians – it does not matter whether we wish it to be an angle in degrees or any other type of angle unit, it will be only be treated as an angle in radians because that is how "$e^{i\theta}$" works. This is similar to how the "θ" in "i^{θ}" is treated as an angle in quarter-circle angle units.

When we use "$e^{i\theta}$" with time, in the form, "$e^{i(2\pi t)}$", the time has to be multiplied by "2π" because the non-Imaginary part of the exponent is treated as an angle in radians. If we did not multiply the time by "2π", the circle and its derived waves would have frequencies that were "2π" times slower than they should be. For example, the circle "e^{it}" would have a frequency of 1 ÷ 2π cycles per second instead of 1 cycle per second. This idea is the same as when we multiply the time by "2π" in a wave formula such as "y = sin (2π * 3t)", when Sine is operating in radians. Sometimes, you will see people give the exponential as so: "e^{360it}" and believe it to work in degrees. This is wrong and is a sign that they do not understand Imaginary powers of "e". In such an exponential, the "360 * t" would be treated as if it were in radians, and any circles or derived waves would have frequencies that were much faster than intended.

When authors mistakenly think "$e^{i\theta}$" can work in whichever angle system they choose, it is a good example of how one does not need to understand something in order to use it in everyday life. It is also a good example of how "e^{i}" is often treated as a static symbol, with no reference to what it actually is.

e's name

As far as I can tell, 2.7182... was named "e" because of the common convention that mathematical constants be given names from the first half of the Latin alphabet (a, b, c, d, e and so on) and mathematical *variables* be given names from the second half (z, y, x and so on). At the time of its naming, the letters a, b, c, d were already being used for other values. It is sometimes mistakenly said that it is called "e" after the mathematician Euler.

Missing out the Imaginary part

It is easy to miss out the "i" in the exponential accidentally. In other words, it is easy to write "e^{2}" when you mean to say, "e^{2i}". Without the "i" being in the exponent, the sense of the exponential is completely changed. The number "e^{2}" can be easily solved, as it is 2.7183 squared. The number "e^{2i}" is less easily solved, although from having read so far in this chapter, you will know that it will be the Complex number describing the point on a unit-radius circle at an angle of 2 radians. [If you forget that "e" refers to 2.7182..., you might forget that you can solve "e^{2}"]

Including the Imaginary part

When you are completely at ease with Imaginary powers of "e", it can become easy to misread a Real power of "e", such as "e^{5}" as if it were really an Imaginary power of "e", such as "e^{5i}". It pays to be aware of this.

Phase

When we put time, frequency and phase in the exponent of an Imaginary power of "e", it is important to remember that they replace the "θ" in "$e^{i\theta}$". In other words, they appear where the dots are in this: "$e^{i(....)}$". If you are too used to writing the formula with zero phases, then you might forget this rule when you come across a non-zero phase. Ideally, if there is zero phase, the exponential should be written as "$e^{i(2\pi ft)}$" so that when there is a non-zero phase, you remember to write it as "$e^{i(2\pi ft + \phi)}$". If you do not do this, then it can be easy to make the mistake of thinking a formula with phase is "$e^{2\pi ift + \phi}$", which is wrong. You could also think of the correct way as "$e^{2\pi ift + i\phi}$", in which case you could say that the phase has to be given as an Imaginary phase.

Multiplication

Although an Imaginary power of "e" can represent a circle with implied derived Sine and Cosine waves, we cannot multiply two such exponentials and have the result represent the multiplication of the corresponding derived waves. Multiplication breaks the connection of "e" and its derived waves. This problem is identical to the one we had when multiplying circles described with "cos θ + i sin θ" in Chapter 24. This might be expected as "$e^{i\theta}$" is equal to "cos θ + i sin θ" when Cosine and Sine are working in radians. It is also the same problem we had in Chapter 25 with powers of "i".

With Imaginary powers of "e", the problem is easy to see. If we multiply:

$e^{i(2\pi * 3t)}$

... by itself, we just add the exponents as so:

$e^{i(2\pi * 3t) + i(2\pi * 3t)}$

$= e^{6\pi it + 6\pi it}$

$= e^{12\pi it}$

$= e^{i(2\pi * 6t)}$

The result is a pure circle with zero mean levels. We know that if we squared the derived waves of our original circle, we would end up with mean levels for each derived wave. Therefore, our Complex exponential result is not consistent with the squares of the derived waves.

We can perform *addition* of two Imaginary powers of "e", and the result will still be connected to the derived waves. We can also scale the Imaginary power of "e" by a number. We cannot, however, multiply two Imaginary powers of "e" and maintain the connection to the derived waves.

The symbol "j"

You will often see the symbol "j" instead of the symbol "i". They mean exactly the same thing, in that "j" is just another way of expressing the square root of −1. The symbol "j" is used by electrical engineers to avoid confusion with "i" referring to electrical current. If you are used to using "i", then seeing "j" can be potentially temporarily confusing. Therefore, it can be good to see "j" being used for the sake of familiarity. Some formulas that we have seen so far, rewritten with "j", are as so:

"$z = e^{j(2\pi t)}$"
"$z = 1 + 2j + e^{2\pi jt}$"
"$j^{0.5}$"
"$e^{j\theta}$"

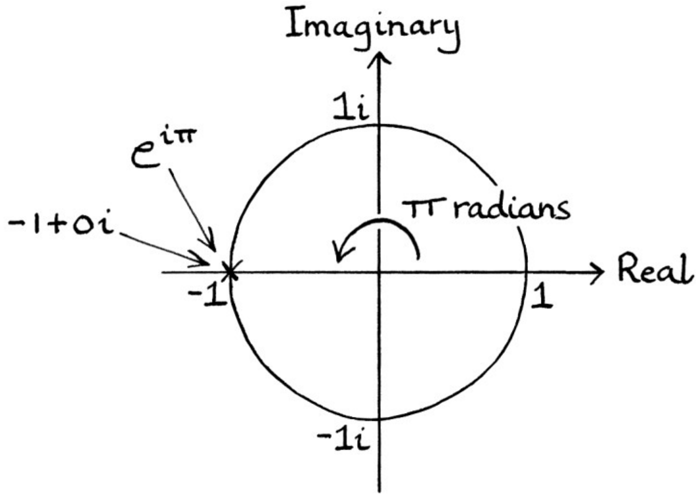
e to the power of iπ = −1

One very common equation that you might see in everyday life is:
$e^{i\pi} = -1$

This is often taken to be an example of incomprehensible maths, but if you have read this far in this book, you will be able to understand it without much difficulty.

The equation is actually short for:
$e^{i\pi} = -1 + 0i$
... so, in other words, it is saying that "$e^{i\pi}$" is the point at "−1 + 0i". When writing Complex numbers, unneeded zeroes are usually ignored, so the Complex number "−1 + 0i" would normally be written as just "−1".

As we know, Imaginary powers of "e" indicate the position of a point on a unit-radius circle where the non-Imaginary part of the exponent is the angle in radians of that point. Therefore, "$e^{i\pi}$" indicates the point on a unit-radius circle at an angle of π radians (which is 180 degrees).



The formula is saying that "$e^{i\pi}$" (which is the point on a unit-radius circle at 180 degrees) can also be identified by the Complex number "−1 + 0i". If you understand powers of "i" (as explained in Chapter 25), and if you understand Imaginary powers of "e" (as explained in this chapter), and if you understand that "−1 + 0i" can be abbreviated to "−1", then the expression "$e^{i\pi} = -1$" should be fairly obvious. It is a very simple statement, but it takes a very long explanation for it to appear as a very simple statement.

We could also say:
$i^2 = -1$
... which, if you think of "i" as a number is obvious because "i" is the square root of −1. If you think of powers of "i" as indicating the positions of points on a unit-radius circle, then this is still straightforward, as it is really "$i^2 = -1 + 0i$". It is saying that the point indicated by "i^2" is at "−1 + 0i".

We could also say:
$i^2 = e^{i\pi}$
... which combines the "powers of 'i'" system and the "Imaginary powers of 'e'" system.

Even when you understand what "$e^{i\pi} = -1$" means, it is still an interesting formula, as it gives a connection between "e", "π", "i" and Real numbers. The formula should not be thought of as complicated but as significant.

The Taylor series

The Taylor series for powers of "e"

Earlier in this chapter, we came up with a formula for calculating powers of "e". The value of "e^x" can be approximated using the formula:
$(1 + (x \div n))^n$
... where "n" is any very large number.

The formula is useful as a basic approximation, but much more accuracy can be achieved by using a Taylor series. The Taylor series for "e^x", looks like this:

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \frac{x^6}{6!} + \cdots$$

... where:

- "!" means "factorial", which means that number multiplied by all the integers lower than it. For example, "5!" means "5 * 4 * 3 * 2 * 1".

[The reason why the Taylor series for "e^x" is like this requires a long explanation, so it is easier just to accept that this Taylor series is correct. Having said that, it would not take particularly long to stumble across this Taylor series by trying out various sums of this type].

Because x ÷ 1! = x, we can also give the formula as so:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \frac{x^6}{6!} + \cdots$$

The more parts of the sum that are completed, the more accurate the result will be.

We will test the Taylor series with a value for which we know the result – we will try "e^1", which is "e" or 2.7182... [Remember that we are calculating "e^1" and not "e^{1i}"]. We put 1 into the formula as so:

$$e^1 = 1 + 1 + \frac{1^2}{2!} + \frac{1^3}{3!} + \frac{1^4}{4!} + \frac{1^5}{5!} + \frac{1^6}{6!} + \cdots$$

Then, we calculate as much of the sum as we want, to gain the accuracy that we need:

$e^1 = 1 + 1 + (1 \div 2) + (1 \div 6) + (1 \div 24) + (1 \div 120) + (1 \div 720)$...

For each step of the sum we end up with a more accurate result. To illustrate this, here is the solved sum for the first few steps:

Step 1: 1
Step 2: 2
Step 3: 2.5
Step 4: 2.66666667
Step 5: 2.70833333
Step 6: 2.71666667
Step 7: 2.71805556
Step 8: 2.71825397

Each extra step adds on a tiny part more, and in this way, the result becomes more and more accurate. If we check on a calculator, we will see that "e^1" is 2.71828183 to 8 decimal places, which is a sign that the sum is progressing in the right direction.

If we had used the $(1 + (x \div n))^n$ formula, with "n" as 10,000, we would have calculated this:

$(1 + (1 \div 10{,}000))^{10{,}000} = (1.0001)^{10{,}000} = 2.71814593$

The Taylor series achieves a more accurate result more quickly. Another benefit of the Taylor series is that if we decide we need more accuracy, we can continue calculating more of the sum. With the other method, we have to start again from scratch with a higher value of "n". On the other hand, "$(1 + (x \div n))^n$" just involves multiplying the same number by itself a very large number of times, while the Taylor series involves separate numbers and different processes.

Another benefit of the Taylor series is that we can avoid rounding errors during the process. If we want 100 digits in our answer, we can perform each stage with slightly more than 100 digits. When the "$(1 + (x \div n))^n$" method is done with a calculator or computer, we cannot be sure if the abilities of the calculator or computer are limiting the accuracy of the result. Generally, the Taylor series is the more useful method.

For interest's sake, here is e^1 calculated on a calculator with "$(1 + (x \div n))^n$" and "n" as 1,000,000:

$(1 + (1 \div 1{,}000{,}000))^{1{,}000{,}000}$
$= (1.000001)^{1{,}000{,}000}$
= 2.71828047

The Taylor series for Imaginary powers of "e"

We can use the Taylor series for "e^x" to calculate "e^{ix}". We could try a value for "ix" in the formula, but instead we will make a generic Taylor series formula for "e^{ix}".

As we saw before, the Taylor series for "e^x" is as so:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \frac{x^6}{6!} + \cdots$$

Therefore, we put in "ix" for every occurrence of "x":

$$e^{ix} = 1 + ix + \frac{(ix)^2}{2!} + \frac{(ix)^3}{3!} + \frac{(ix)^4}{4!} + \frac{(ix)^5}{5!} + \frac{(ix)^6}{6!} + \cdots$$

We will find the result of "e^{2i}", so we will put 2 radians into the formula:

$$e^{2i} = 1 + 2i + \frac{(2i)^2}{2!} + \frac{(2i)^3}{3!} + \frac{(2i)^4}{4!} + \frac{(2i)^5}{5!} + \frac{(2i)^6}{6!} + \cdots$$

The result for each step of the sum is as follows:

After the first step, we have: 1
After the second step, we have: 1 + 2i
Third step: −1 + 2i
Fourth step: −1 + 0.66666667i
Fifth step: −0.33333333 + 0.66666667i
Sixth step: −0.33333333 + 0.93333333i
Seventh step: −0.42222222 + 0.93333333i
Eighth step: −0.42222222 + 0.90793651i

If we used a calculator that can work with Complex numbers, we would find that the "e^{2i}" is −0.41614684 + 0.90929743i. This shows that the Taylor series is not as quick at finding the results of Imaginary exponents of "e" as it is at finding Real exponents of "e". As we know that the point indicated by "e^{2i}" is a point on a unit-radius circle's edge at an angle of 2 radians, we could also have used Cosine and Sine to calculate the result. The same point would be "cos 2 + i sin 2", which is also "−0.41614684 + 0.90929743i".

The Taylor series for "e^{ix}" is not as useful as the Taylor series for "e^{x}", as we can just use Cosine and Sine instead. However, it can be used to confirm that:
"$e^{i\theta} = \cos\theta + i \sin\theta$"
... when "θ" is an angle in radians and Sine and Cosine are working in radians (if we needed yet another way of knowing it).

As we saw in Chapter 2, the Taylor series for the *Sine* of a value in radians is:

$$\sin\theta = \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \frac{\theta^7}{7!} + \frac{\theta^9}{9!} - \frac{\theta^{11}}{11!} \dots$$

... and the Taylor series for the *Cosine* of a value in radians is:

$$\cos\theta = 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \frac{\theta^6}{6!} + \frac{\theta^8}{8!} - \frac{\theta^{10}}{10!} \dots$$

The Taylor series for "$e^{i\theta}$" is:

$$e^{i\theta} = 1 + i\theta + \frac{(i\theta)^2}{2!} + \frac{(i\theta)^3}{3!} + \frac{(i\theta)^4}{4!} + \frac{(i\theta)^5}{5!} + \frac{(i\theta)^6}{6!} + \cdots$$

We know that "i" squared is −1, which means that we can tidy it up slightly:

$$e^{i\theta} = 1 + i\theta + \frac{-\theta^2}{2!} + \frac{-i\theta^3}{3!} + \frac{\theta^4}{4!} + \frac{i\theta^5}{5!} + \frac{-\theta^6}{6!} + \cdots$$

We can tidy up each plus and minus:

$$e^{i\theta} = 1 + i\theta - \frac{\theta^2}{2!} - \frac{i\theta^3}{3!} + \frac{\theta^4}{4!} + \frac{i\theta^5}{5!} - \frac{\theta^6}{6!} - \cdots$$

In the above formula, there are fractions with "i" in them, and there are fractions without "i" in them. We can split them up as so:

$$e^{i\theta} = 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \frac{\theta^6}{6!} - \cdots$$

+

$$i\theta - \frac{i\theta^3}{3!} + \frac{i\theta^5}{5!} -$$

... and from this, we can see that the first part is equal to the Taylor series for Cosine, and the second part is equal to "i" multiplied by the Taylor series for Sine. In other words, if we used the Taylor series to find "$e^{i\theta}$", we would also be finding the Taylor series for "cos θ" added to the Taylor series for "i sin θ". From that we can say that "$e^{i\theta}$ = cos θ + i sin θ" (when Cosine and Sine are working in radians). We already knew this, but the Taylor series reveals it for us in another way.

Without already knowing how circles and waves relate to each other, and without already knowing how "i^{θ}" works, and without having seen the "$(1 + (x \div n))^n$" formula, you might think that the connection was something to do with the Taylor series. If you have read this book from the start, the idea that "$e^{i\theta}$ = cos θ + i sin θ" in radians should have been straightforward and expected long before now. If "$e^{i\theta}$" marks the position of a point on the circumference of a unit-radius circle at an angle of "θ" radians, it should be obvious that that point can also be described with "cos θ + i sin θ" when Cosine and Sine are working in radians.

The trouble with most maths education is that it misses out a lot of information along the way, and just reveals that "$e^{i\theta}$ = cos θ + i sin θ" when thinking about the Taylor series. Therefore, many people find the equivalence amazing because they do not know the background as to why it should be expected. Students generally start using radians very early in their maths education. Being used to radians early on makes the idea that "$e^{i\theta}$" works in radians seem a magical revelation – an angle system that they have used for ages is mysteriously referenced in a seemingly unrelated formula. However, the reason radians are used in maths is primarily due to things such as "$e^{i\theta}$". If "$e^{i\theta}$" worked in a different angle system, we would probably be using a different angle system.

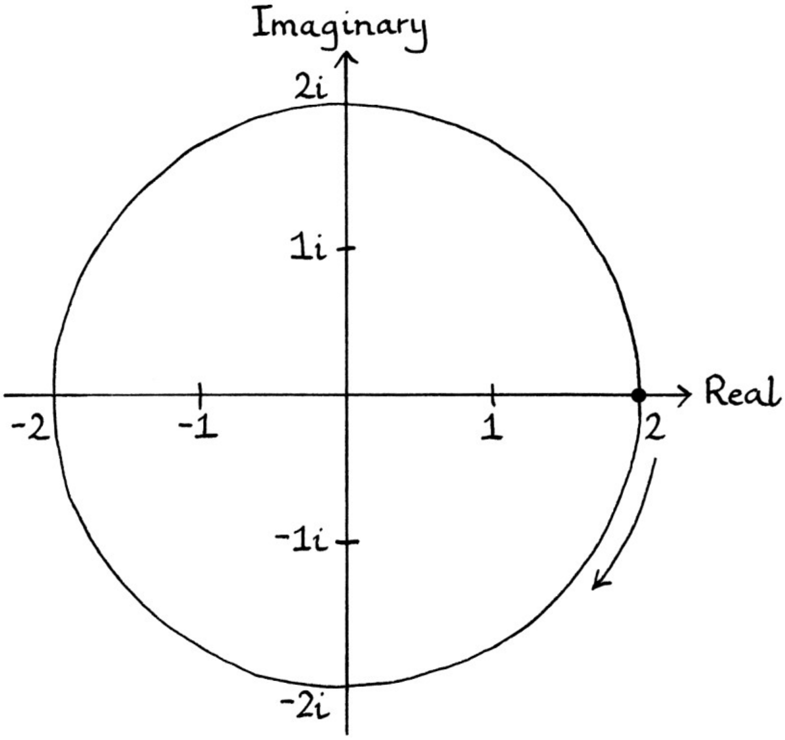
Negative frequencies

We can portray an object rotating clockwise around a circle by making the time-based exponential have a negative frequency. [This is the same as the helix going the wrong way around.]

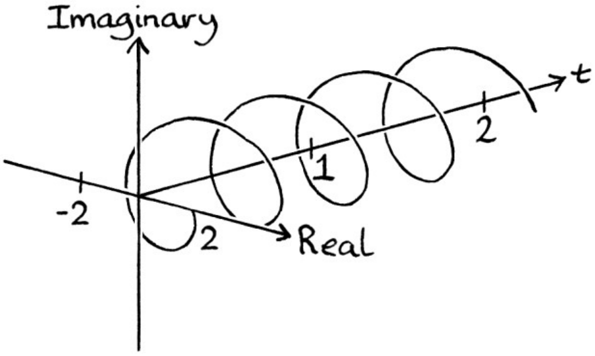
We will look at this circle:
“$z = 2e^{i(2\pi * -2t)}$”

We can also phrase the circle formula as:
“$z = 2e^{(2\pi i * -2t)}$”
... or even:
“$z = 2e^{-4\pi it}$”

This represents an object rotating around a two-unit circle at a frequency of −2 cycles per second. In other words, the object is rotating *clockwise* around the circle at a frequency of 2 cycles per second.



On the Complex helix chart, this looks like this:



We can also say that:
$2e^{(2\pi i * -2t)} = 2 \cos (2\pi * -2 * t) + 2i \sin (2\pi * -2 * t)$
... which is:
$2 \cos (2\pi * -2t) + 2i \sin (2\pi * -2t)$.

Rephrasing the wave part

In Chapter 24, I discussed how a formula such as "cos (360 * −ft) + i sin (360 * −ft)" can be rephrased to have positive frequencies. The trouble with doing this is that a rephrased formula makes it harder to tell that it refers to an object rotating clockwise around a circle, and it makes it harder to know the phase point. There is also the potential for making mistakes if there is a non-zero phase. However, it is common practice for people to phrase the waves in this way. We saw the rules for converting the formulas in Chapter 24:

For degrees, if we have this formula:
$z = \cos ((360 * -ft) + \phi) + i \sin ((360 * -ft) + \phi)$
... then it has the same meaning as this formula:
$z = \cos ((360 * +ft) - \phi) - i \sin ((360 * +ft) - \phi)$

For radians, if we have this formula:
$z = \cos ((2\pi * -ft) + \phi) + i \sin ((2\pi * -ft) + \phi)$
... then it has the same meaning as this formula:
$z = \cos ((2\pi * +ft) - \phi) - i \sin ((2\pi * +ft) - \phi)$

Using the rule for radians, we can rephrase our formula from above, which was:
$2e^{(2\pi i * -2t)} = 2 \cos (2\pi * -2t) + 2i \sin (2\pi * -2t)$
... as:
$2e^{(2\pi i * -2t)} = 2 \cos (2\pi * 2t) - 2i \sin (2\pi * 2t)$

Despite the drawbacks in doing so, it is very common for people to express a negative frequency Imaginary exponential of "e" as being equivalent to positive-frequency waves with a subtraction. [Sometimes, this is done because some people have an aversion to negative frequencies.]

You will see the equivalence so often that it pays to recognise:
"$\cos(2\pi * ft) - i \sin(2\pi * ft)$"
... and its variations as indicating a negative frequency.

You will also see these equivalences:
$e^{-i\theta} = \cos\theta - i \sin\theta$
... or, if we replace "θ" with "x":
$e^{-ix} = \cos x - i \sin x$

Sometimes, these two might be given as examples of negative frequency, but strictly speaking, as they refer to angles and not time, they do not relate to frequency. However, they are still mathematically valid equivalences, and they are more succinct than the time-based formulas.

Conclusion

The number "e", being 2.71828183, is one of the most important numbers to do with waves and circles. When "e" is raised to an Imaginary power, it indicates the position of a point on a unit radius circle at an angle equal to the value of the non-Imaginary part of the exponent in radians. This knowledge can be used to let Imaginary powers of "e" describe circles, which in turn imply the characteristics of a derived Cosine wave and a derived Sine wave. Therefore, Imaginary powers of "e" can be used to describe waves.

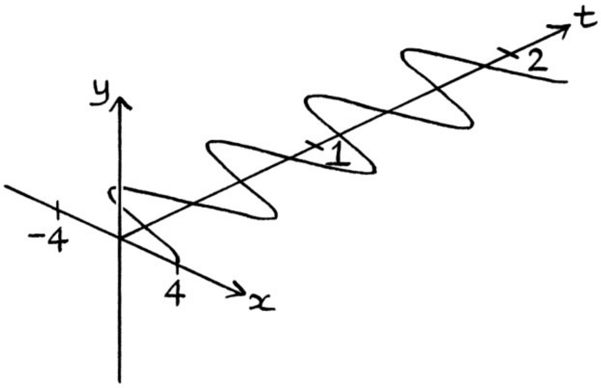
Chapter 28: Waves in terms of exponentials

In more advanced maths using Complex exponentials, waves are given, not in their basic form, but in terms of Complex exponentials. This idea is easier to understand if we think of the Complex exponential as representing a helix in the Complex helix chart (as opposed to thinking of it as just a circle). Instead of giving the formula of a wave in the normal way, we give the formula in terms of the helices that would need to be added or subtracted to create such a wave. As a Complex exponential can be thought of as portraying a helix, we give the formula of a wave in terms of the two Complex exponentials that would need to be added or subtracted to create that wave. This idea will become clearer as we progress through this section.

Adding and subtracting helices

In Chapter 14 on the addition of circles, we saw how adding two identical helices of opposing frequencies results in a two-dimensional Cosine wave lying horizontally in the three-dimensional helix chart. We also saw how subtracting one helix from another, where they have identical characteristics, except for having opposing frequencies, results in a two-dimensional Sine wave sitting vertically in the three-dimensional helix chart.

To refresh your memory, the following picture shows the helix chart containing the result of adding the circles (or helices) made up of:
“y = 2 sin (360 * 2t)” and “y = 2 cos (360 * 2t)”
... and:
“y = 2 sin (360 * −2t)” and “y = 2 cos (360 * −2t)”



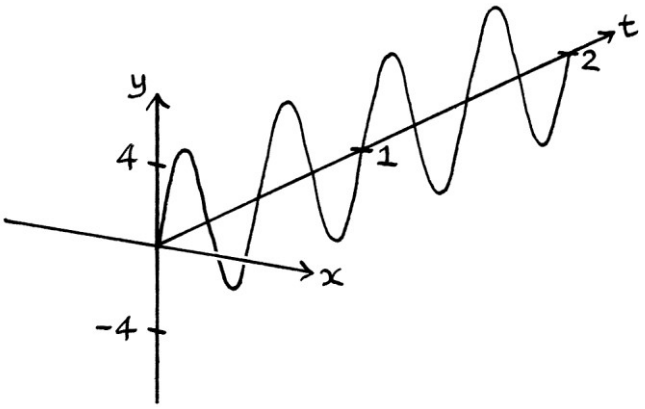
If we viewed the helix chart from underneath, we would see the Cosine wave. The Cosine wave is still a Cosine wave, but it is sitting in the three-dimensional helix chart.

We can see why the addition of two helices (or circles) has this effect by looking at the waves that make up the helices (or circles). The vertically derived signal from the addition will be:
"y = 2 sin (360 * 2t) + 2 sin (360 * −2t)"
... which, because "y = 2 sin (360 * −2t)" is the same as "y = 2 sin ((360 * 2t) + 180)", is:
"y = 2 sin (360 * 2t) + 2 sin ((360 * 2t) + 180)"
... which, because "y = 2 sin ((360 * 2t) + 180)" is the same as "y = −2 sin (360 * 2t)" is:
"y = 2 sin (360 * 2t) − 2 sin (360 * 2t)"
... which results in "y = 0" for all time. This means that the resulting vertically derived signal will be a horizontal line at y = 0 for all time. In other words, if we looked at the coordinates for the resulting helix or circle, all the y-axis values would be zero.

The horizontally derived signal from the addition will be:
"y = 2 cos (360 * 2t) + 2 cos (360 * −2t)"
... which, because "y = 2 cos (360 * −2t)" is the same as "y = 2 cos (360 * 2t)", is:
"y = 2 cos (360 * 2t) + 2 cos (360 * 2t)"
... which results in:
"y = 4 cos (360 * 2t)"
This means that the resulting horizontally derived signal will be the Cosine wave:
"y = 4 cos (360 * 2t)".

To summarise the above, after the addition, all that is left is a Cosine wave with twice the amplitude of the derived Cosine wave with which we started.

When it comes to *subtracting* helices, a slightly different result is achieved. The following picture is the helix chart showing the result of the circle (or helix) made up of:
“y = 2 sin (360 * 2t)” and “y = 2 cos (360 * 2t)”
... minus the circle (or helix) made up of:
“y = 2 sin (360 * −2t)” and “y = 2 cos (360 * −2t)”



The result is a two-dimensional Sine wave sitting vertically within the helix chart. We can see why the subtraction of one helix (or circle) from another results in this by looking at the characteristics of the derived waves. The vertically derived signal from the subtraction will be:
“y = 2 sin (360 * 2t) − 2 sin (360 * −2t)”
... which, because a negative-frequency Sine wave with zero phase is the same as that wave with a positive frequency and a phase of 180 degrees, is:
“y = 2 sin (360 * 2t) − 2 sin ((360 * 2t) + 180)”
... which, because a negative-amplitude Sine wave with a phase of 180 degrees is the same as that wave with a positive amplitude and a phase of zero degrees, is:
“y = 2 sin (360 * 2t) + 2 sin (360 * 2t)”
... which is:
“y = 4 sin (360 * 2t)”

The horizontally derived signal from the subtraction will be:
“y = 2 cos (360 * 2t) − 2 cos (360 * −2t)”
... which is:
“y = 2 cos (360 * 2t) − 2 cos (360 * 2t)”
... which is:
“y = 0” for all time.

Therefore, if we gave the coordinates of the resulting shape on the helix chart, all the x-axis values would be zero for all time. We would only have y-axis values, and they would show a Sine wave.

To summarise all of the above, after the subtraction, all we are left with is a Sine wave with twice the amplitude of the derived Sine wave with which we started.

Adding and subtracting Complex exponentials

A helix can be portrayed with a time-based Complex exponential (an Imaginary power of “e”). Whether we think of such an exponential as portraying a circle or a helix is a matter of choice, as it really portrays both depending on from where we view it.

A two-dimensional wave can be created in the three-dimensional helix chart by adding or subtracting two helices. As we can portray helices with a time-based Imaginary power of “e”, we can create two-dimensional waves in the three-dimensional Complex helix chart by adding or subtracting Imaginary powers of “e”. We can therefore portray pure waves in terms of Imaginary powers of “e”. Note that this is a different situation from a circle or helix having a derived Sine wave and a derived Cosine wave. In this situation, we are literally creating just one pure wave using helices.

When we portray a wave in terms of Complex exponentials, we are using exactly the same ideas as before. One helix added to another with the same but opposing frequency will create a horizontal Cosine wave in the helix chart; one helix subtracted from another helix with the same but opposing frequency will create a vertical Sine wave in the helix chart.

Giving a wave in terms of Imaginary powers of “e” can make some maths simpler because we will still be dealing with exponentials, and exponentials are more suitable for some maths than dealing with Sine and Cosine formulas. It can also make some maths more difficult, and it makes visualising what we are doing harder – there are advantages and disadvantages depending on what we are trying to do.

As an example of what I mean by all of the above, if we want the formula for a Cosine wave, we can give it normally as it would appear on a wave graph as:
“y = cos (2π * ft)”
... or we can give it as it would appear in the Complex helix chart in terms of “$e^{2\pi ift}$” as:

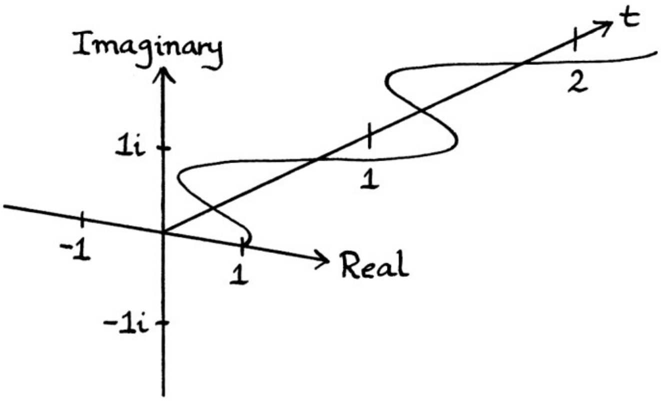
$$z = \frac{1}{2} * (e^{i(2\pi * +ft)} + e^{i(2\pi * -ft)})$$

This formula is one helix added to another helix with the same attributes, but with a negative frequency, then all divided by 2. The division by two is necessary because otherwise the Cosine wave in the Complex helix chart would have twice the desired amplitude. We can also phrase this in a slightly more concise way that moves the negative frequency aspect on to the “i” in the exponent:

$$z = \frac{1}{2} * (e^{i(2\pi ft)} + e^{-i(2\pi ft)})$$

I am giving the formula with a big “half” fraction, instead of turning it into a multiplication of 0.5, to make the formula have the same layout for the Sine wave version, which will we look at shortly.

A Cosine wave in the Complex helix chart looks like this:



If we view the Complex helix chart from underneath, we will see the Cosine wave the correct way around.

For a Sine wave, we can give the formula normally as it would appear on a wave graph as:
"y = sin (2π * ft)"
... or we can give it as it would appear in the Complex helix chart in terms of "$e^{i(2\pi ft)}$" as:

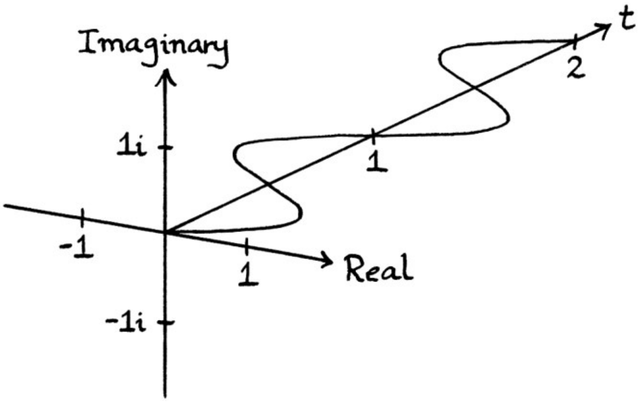
$$z = \frac{1}{2i} * (e^{i(2\pi * +ft)} - e^{i(2\pi * -ft)})$$

We can write this more concisely as:

$$z = \frac{1}{2i} * (e^{i(2\pi ft)} - e^{-i(2\pi ft)})$$

Note that this is not a vertical Sine wave, but a horizontal Sine wave. This formula is one helix subtracted from a second helix with the same, but negative, frequency, all multiplied by "1 ÷ 2i". If you think back to Chapter 23 on Complex numbers, a *division* by "i" rotates a point by −90 degrees. Therefore, the multiplication by "1 ÷ 2i" halves the amplitude of the Sine wave at the same time as rotating every single point by −90 degrees (it rotates them clockwise). The whole wave is rotated from being entirely vertical to being entirely horizontal. In other words, the whole wave goes from being entirely Imaginary to entirely Real.

A Sine wave in the Complex helix chart looks like this:



When we view the Complex helix chart from underneath, we see the Sine wave the correct way around.

The two ways of writing the Sine and Cosine waves look complicated – because they are – but they can simplify many calculations.

Horizontal and vertical waves

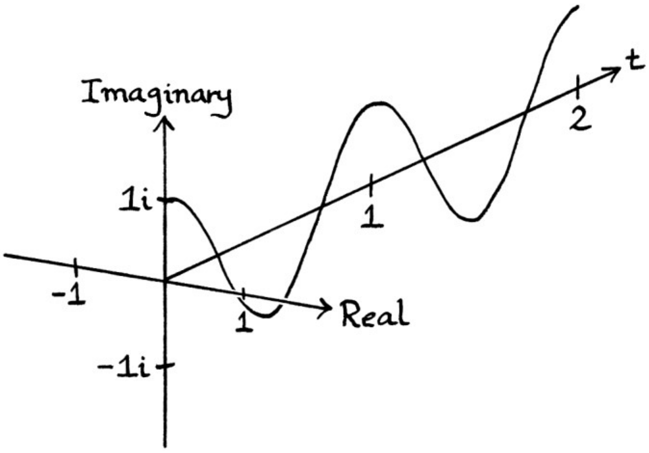
Instead of having the waves sitting horizontally in the Complex helix chart, we could also have them sitting vertically, in which case the formula for the Cosine wave would be:

$$z = \frac{i}{2} * (e^{i(2\pi ft)} + e^{-i(2\pi ft)})$$

... or:

$$z = 0.5i * (e^{i(2\pi ft)} + e^{-i(2\pi ft)})$$

The previous Cosine formula has been multiplied by "i", which rotates every single point by +90 degrees (1 quarter-circle angle unit) anticlockwise. Every single point is now Imaginary. The Cosine wave will look like this in the Complex helix chart:



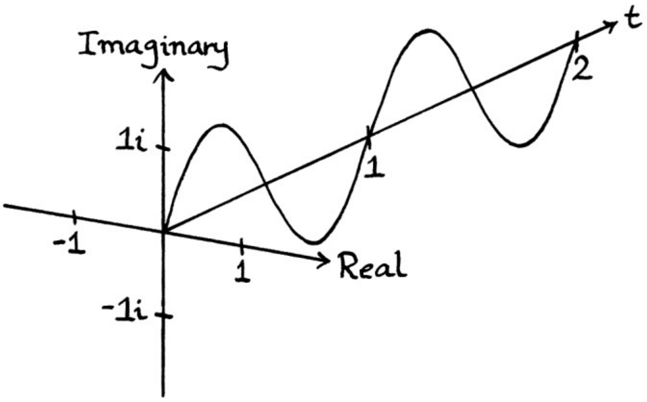
The formula for the Sine wave would be:

$$z = \frac{1}{2} * (e^{i(2\pi ft)} - e^{-i(2\pi ft)})$$

... or:

$$z = 0.5 * (e^{i(2\pi ft)} - e^{-i(2\pi ft)})$$

Whereas the previous horizontal Sine formula rotated every point by −90 degrees, now there is no need to do that. Every single point is now Imaginary. The Sine wave looks like this in the Complex helix chart:



Although we could place the waves vertically or horizontally in the Complex helix chart (or in fact at any angle), the waves are nearly always placed horizontally. There is a good reason for this – it means that every point along their curves is Real. All the Imaginary values of the waves are zero. Doing this means that the waves in the Complex helix chart can be directly equated with waves existing outside of the world of Complex numbers. In other words, we can take a normal wave formula such as "y = sin (2π * 3t)" and have that exact wave portrayed identically in the Complex helix chart. I will explain more about this later in this chapter.

Calculating the formulas

If you did not understand how the addition of two helices worked, or how the subtraction of one helix from another worked, you might want a mathematical way of seeing how the typical formulas for Sine and Cosine in the Complex helix chart are calculated. The formulas are surprisingly easy to calculate mathematically.

First, we take the basic formulas for positive and negative frequencies (with zero phases):

$e^{i(2\pi ft)} = \cos(2\pi ft) + i \sin(2\pi ft)$

$e^{-i(2\pi ft)} = \cos(2\pi ft) - i \sin(2\pi ft)$

If we add these together, we can say that:
$e^{i(2\pi ft)} + e^{-i(2\pi ft)}$ = cos (2πft) + i sin (2πft) + cos (2πft) − i sin (2πft)
... so:
$e^{i(2\pi ft)} + e^{-i(2\pi ft)}$ = 2 cos (2πft)
... and switching each part of the equation around, we have:
2 cos (2πft) = $e^{i(2\pi ft)} + e^{-i(2\pi ft)}$
... so:
cos (2πft) = 0.5 * ($e^{i(2\pi ft)} + e^{-i(2\pi ft)}$)
... which is the Cosine wave as a horizontal wave given in terms of Imaginary powers of "e".

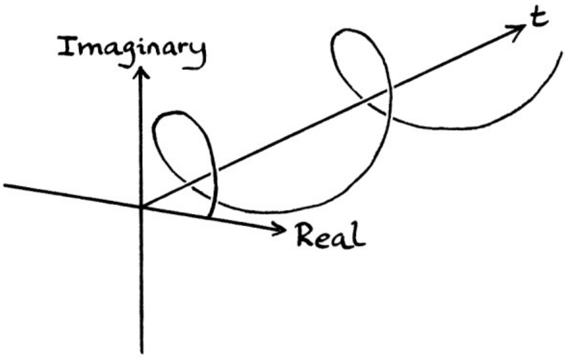
If we subtract the negative-frequency formula from the positive-frequency formula, we can say that:
$e^{i(2\pi ft)} - e^{-i(2\pi ft)}$ = cos (2πft) + i sin (2πft) − (cos (2πft) − i sin (2πft))
... so:
$e^{i(2\pi ft)} - e^{-i(2\pi ft)}$ = cos (2πft) + i sin (2πft) − cos (2πft) + i sin (2πft)
... so:
$e^{i(2\pi ft)} - e^{-i(2\pi ft)}$ = 2 * (i sin (2πft))
... which is:
$e^{i(2\pi ft)} - e^{-i(2\pi ft)}$ = 2i sin (2πft)
... which we can swap around to be:
2i sin (2πft) = $e^{i(2\pi ft)} - e^{-i(2\pi ft)}$
... which we can rephrase to be:
sin (2πft) = ($e^{i(2\pi ft)} - e^{-i(2\pi ft)}$) ÷ 2i
... which is:
sin (2πft) = (1 ÷ 2i) * ($e^{i(2\pi ft)} - e^{-i(2\pi ft)}$)
... which is the Sine wave as a horizontal wave given in terms of Imaginary powers of "e".

The Complex helix chart

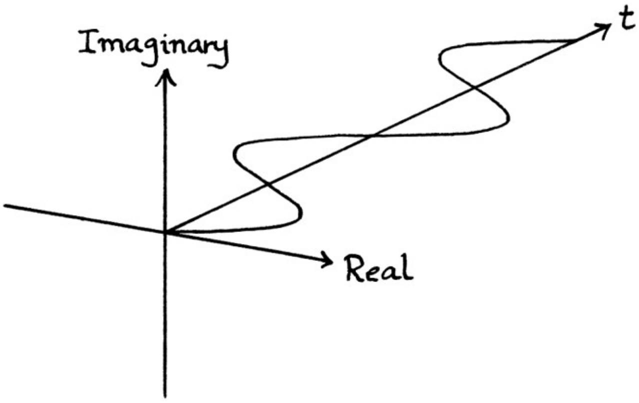
The first thing to notice from the formulas is that we are really creating a Cosine wave with the sum of two helices, and we are creating a Sine wave by subtracting one helix from another, and then rotating it. Although one can think of Imaginary powers of "e" as representing either circles or helices, in this case, it is better to think of them as helices.

Previously in this book, Sine and Cosine waves have been flat lines on flat graphs. The waves were portrayed on two-dimensional axes – we had the y-axis and we had the time axis. On a helix chart, we have three dimensions: x, y and the time axis. On a Complex helix chart, which is essentially the same thing, we have three dimensions too, and they are the Real axis, the Imaginary axis and the time axis. By portraying the Sine and Cosine waves in terms of Imaginary powers of "e", we are really placing the flat Sine and Cosine waves into the three-dimensional Complex helix chart. We are saying that a three-dimensional helix added to, or subtracted from, another three-dimensional helix results in a two-dimensional wave. Our two-dimensional wave is sitting in the three-dimensional helix chart.

This is a helix in the Complex helix chart:

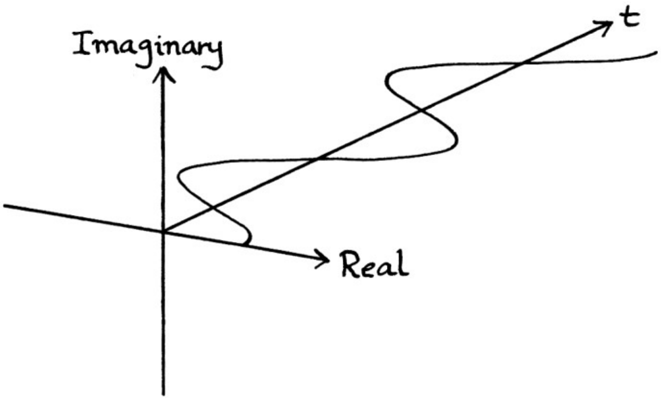


This is the horizontal Sine wave in the Complex helix chart:



It is a two-dimensional Sine wave in the three-dimensional Complex helix chart. It has been created by subtracting one helix from another, where the helices have equal but opposing frequencies, then halving the result, and rotating it by −90 degrees, so that it is horizontal. Note that this Sine wave is not the helix viewed from the side – that is a different concept that refers to the *derived* Sine wave from the circle or helix. This Sine wave is sitting in the Complex helix chart itself.

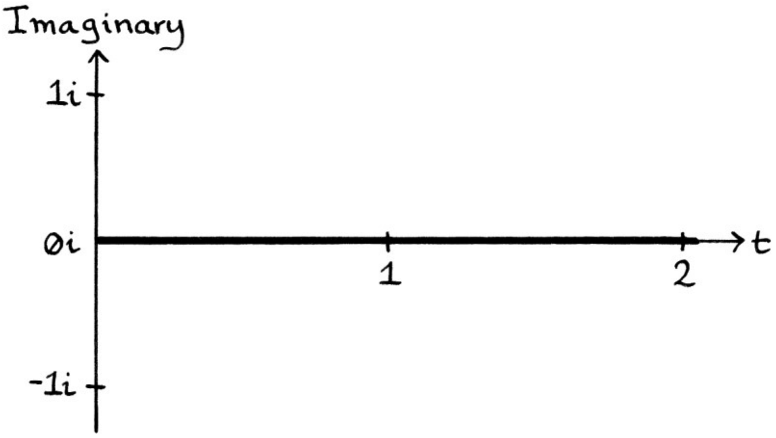
This is a Cosine wave in the Complex helix chart:



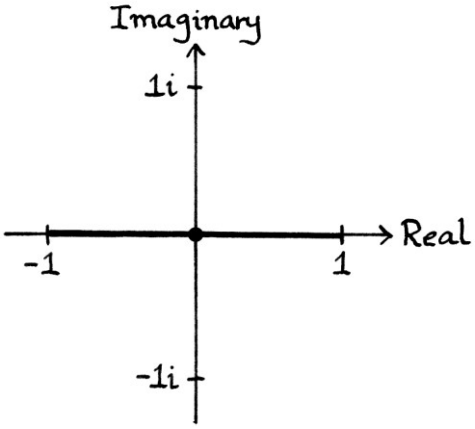
It is a two-dimensional Cosine wave in the three-dimensional Complex helix chart. It has been created by adding two helices of equal but opposing frequencies, and then halving the result.

A Sine or Cosine wave sitting in the Complex helix chart is still two-dimensional, in that it has no width, but it is a different concept to a Sine wave or Cosine wave on the typical “y-axis, time-axis” graph. It is a two-dimensional line in a three-dimensional grid. This distinction is important. The waves in the Complex helix chart are not derived waves from a circle or helix. Instead, they are duplicates of the derived waves – duplicates that have been created in the Complex helix chart.

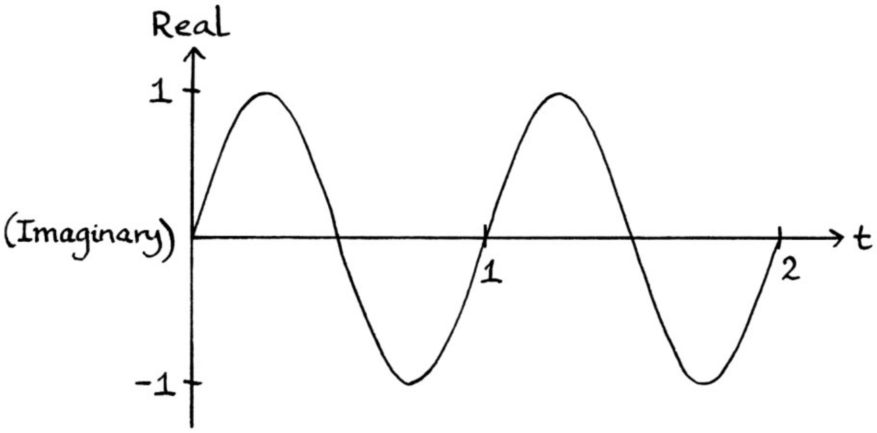
Previously, when we had a proper helix, we could see the Sine wave by looking at the helix side on, we could see the Cosine wave by looking at the helix from underneath, and we could see the circle by looking at the helix from the end. The waves were not literally in the helix chart, but we could see their two-dimensional shapes by looking at the sides of a helix. Now that we are putting the waves into the Complex helix chart, things become different. Supposing we have the above Sine wave in the Complex helix chart, if we view it from the side, with the time-axis pointing to the right, we will just see a straight line:



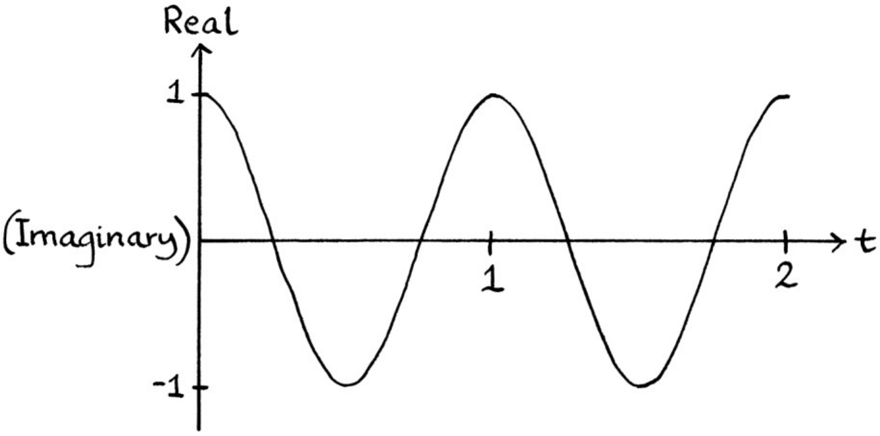
If we view the Complex helix chart from the end with the time axis pointing away from us, we will just see a horizontal line, the length of which will be twice the amplitude of the wave:



If we view the Complex helix chart from underneath, with the Imaginary axis pointing away from us, we will see the Sine wave:



All of this is true for when the Cosine wave is sitting flat in the Complex helix chart. The Complex helix chart has to be viewed from underneath, with the Imaginary axis pointing away from us, to see it as the Cosine wave it is:



Plotting points

The idea of a Sine wave or Cosine wave sitting in the Complex helix chart might seem to be a complicated idea to start with, but we can make it easier to visualise by plotting some points along the graph. We will fill in points in this formula for a Sine wave with a frequency of 1 cycle per second:

$$z = \frac{1}{2i} * (e^{i(2\pi t)} - e^{-i(2\pi t)})$$

[To make things easier, we will use a calculator that can work with Complex numbers.]

At t = 0 seconds, the point indicated by the formula will be at the coordinates: (0, 0) and 0 seconds down the time axis. It will have the Complex number "0 + 0i" at 0 seconds.

At t = 0.05 seconds, the point is at the coordinates (0.3090, 0) and 0.05 seconds down the time axis. The Complex number for this is 0.3090 + 0i. This point, and in fact, all the points will be entirely Real. All the Imaginary values will be zero.

At t = 0.1, the point is at the coordinates (0.5878, 0) and 0.1 seconds down the time axis.

At t = 0.15, the coordinates are (0.8090, 0) and 0.15 seconds down the time axis.

At t = 0.2, the coordinates are (0.9511, 0) and 0.2 seconds down the time axis.

At t = 0.25, the coordinates are (1, 0) and 0.25 seconds down the time axis. This is the first peak of the wave. The Complex number for this point is "1 + 0i".

If we continued, we would see that the Sine wave consists entirely of Real values (x-axis values) in the Complex helix chart. The Imaginary part (the y-axis value) will always be zero. If the formula for the Sine wave did not have the multiplication by $\frac{1}{2i}$ it would be entirely Imaginary. That multiplication makes it entirely Real.

We will do the same thing with the formula for a Cosine wave with a frequency of 1 cycle per second. The formula is as so:

$$z = \frac{1}{2} * (e^{i(2\pi t)} + e^{-i(2\pi t)})$$

At t = 0 seconds, the coordinates are (1, 0), and 0 seconds down the time axis.

At t = 0.05 seconds, the coordinates are (0.9511, 0) and 0.05 seconds down the time axis.

At t = 0.1 seconds, the coordinates are (0.8090, 0) and 0.1 seconds down the time axis.

At t = 0.15 seconds, the coordinates are (0.5878, 0) and 0.15 seconds down the time axis.

At t = 0.2 seconds, the coordinates are (0.3090, 0) and 0.2 seconds down the time axis.

At t = 0.25, the coordinates are (0, 0) and 0.25 seconds down the time axis. This is the first place where the wave crosses the time axis (or y = 0, or x = 0). The Complex number for this point is “0 + 0i”, which would normally be written as just “0”.

If we continued, we would see that the Cosine wave consists entirely of Real values (x-axis values). The Imaginary values (y-axis values) are always zero.

From all of this, we can see that the Sine wave and Cosine wave constructed by adding helices are flat, but still in the Complex helix chart. As we saw earlier in this chapter, it would also be possible to create Sine waves and Cosine waves that consisted entirely of Imaginary points in the Complex helix chart. It would also be possible to create Sine and Cosine waves that were at an angle in the Complex helix chart and consisted of both Real and Imaginary components. [We created a Sine wave at an angle of 45 degrees in the helix chart in Chapter 17 on the multiplication of circles. We saw an entirely Imaginary Sine wave in the helix chart in Chapter 14 on the addition of circles, although we did not know the term Imaginary back then, so it was just a “vertical” Sine wave. The Sine wave in the Complex helix chart in these examples would be vertical if it had not been rotated by −90 degrees by the division by “i”.]

The name of the wave in the Complex helix chart

Given that these Sine and Cosine waves appear in the Complex helix chart, it might be better to distinguish them from normal Sine and Cosine waves by calling them something simple such as “the Sine wave in the helix chart”, or “the helix-chart Sine” or similar. In practice, the distinction of the waves being in the helix chart does not seem to be emphasised as much as it could be. You might see the term “Complex sinusoid”, which depending on the person using it might mean this sort of wave, or it might mean a helix in the Complex helix chart. This term is not to be confused with “complex waveform”, which just means a normal signal made up of the sum of normal waves. There is so much risk in being ambiguous in using the term “Complex sinusoid” that, in my view, it is better to avoid using the term. [I also think the term “complex waveform” should never be used due to how it has nothing to do with Complex numbers. Other people might disagree though.]

Equivalences

When the waves are sat horizontally in the Complex helix chart, all their points can be referred to using just Real numbers. The Real values of a wave such as this are identical to the y-axis values of a normal wave, not sat in the Complex helix chart. This means that we can say:

$$\sin(2\pi ft) = \frac{1}{2i} * (e^{i(2\pi ft)} - e^{-i(2\pi ft)})$$

The above formula states that a Sine wave is *exactly* mathematically equivalent to one Complex Exponential subtracted from another in this way, rotated clockwise by 90 degrees to make it all Real, and then halved. We could put the same values in either side of the equivalence and we would produce exactly the same results.

We can also say:

$$\cos(2\pi ft) = \frac{1}{2} * (e^{i(2\pi ft)} + e^{-i(2\pi ft)})$$

The above formula states that a Cosine wave is exactly mathematically equivalent to one Complex Exponential added to another, and then halved. Again, we could put the same values in either side of the equivalence and we would end up with exactly the same result.

These formulas are two of the most significant formulas in this book. The equivalences mean that if we have a Sine wave formula or a Cosine wave formula, we can portray its meaning *exactly* using Imaginary powers of "e". We can switch from non-Complex formulas to Complex formulas and back because they are mathematically the same thing. This allows us to use the maths of exponentials to deal with waves. This, in turn, can often simplify calculations. [It can also often make them more complicated too.]

When these equivalences are phrased in this way in maths books, the idea of helices is usually forgotten about, and instead everything is thought of mathematically. Not thinking about, or not knowing about, the helices is a reason that a lot of people struggle to comprehend what the equivalences really mean, or why they are as they are. The equivalences would seem very complicated without the background of waves, Complex numbers, "i", Imaginary powers of "e", and helices that we have looked at throughout this book. However, they should be reasonably straightforward if you have read this book so far.

There are really two ways to think about these equivalences:

- We can think of them entirely mathematically, with no regard to helices. In this way, the equivalences are saying that one way of describing a wave produces *exactly* the same results as another, more complicated, way of describing a wave.
- Or, we can think of them more "visually" by saying that the characteristics of a wave can be portrayed by the addition or subtraction of two helices.

Personally, I would say that the formulas on each side of the equivalences are *mathematically* the same, but *conceptually* different. In other words, when it comes to entering values into the formulas and obtaining results, there is no difference between say:

$$\sin(2\pi f t)$$

... and

$$\frac{1}{2i} * (e^{i(2\pi f t)} - e^{-i(2\pi f t)})$$

When it comes to maths, they are exactly the same thing. However, they are really two different concepts. The first is a two-dimensional wave in a two-dimensional world; the second is a two-dimensional wave in a three-dimensional world. The first is a two-dimensional wave; the second is a combination of three-dimensional helices.

Sums of waves

Given the above equivalences, we can create signals by adding waves, and then give the results either in terms of normal Sine and Cosine formulas, or in terms of the exponentials that are equivalent to those waves.

For example, if we add:
"y = sin (2πt)"
... and:
"y = sin (2π * 3t)"
... we would end up with:
"y = sin (2πt) + sin (2π * 3t)"

We could also give the sum in terms of Imaginary powers of "e", in which case we would have:

$$z = \frac{1}{2i} * (e^{i(2\pi t)} - e^{-i(2\pi t)})$$

... added to:

$$z = \frac{1}{2i} * (e^{i(2\pi * 3t)} - e^{-i(2\pi * 3t)})$$

... and we would end up with:

$$z = \frac{1}{2i} * \left(e^{2\pi i t} - e^{-2\pi i t}\right) + \frac{1}{2i} * (e^{2\pi i * 3t)} - e^{-2\pi i * 3t})$$

Putting in any value of "t" into the exponentials will produce an identical result to putting that value of "t" into the sum of waves. They are two ways of expressing exactly the same thing. Obviously, in this example, treating the waves as exponentials unnecessarily complicates things – it is much harder to know what is being described by the exponentials. However, this example demonstrates what is possible.

Positive and negative frequencies

In the above example, the resulting signal has the formula:

$$z = \frac{1}{2i} * \left(e^{2\pi it} - e^{-2\pi it}\right) + \frac{1}{2i} * (e^{2\pi i*3t)} - e^{-2\pi i*3t})$$

As we know, the second exponential of each pair refers to an object rotating clockwise around a unit-radius circle – they refer to negative frequencies. This means that there are positive and negative frequencies in this form of the signal. When the resulting signal is phrased normally, there are only positive frequencies:

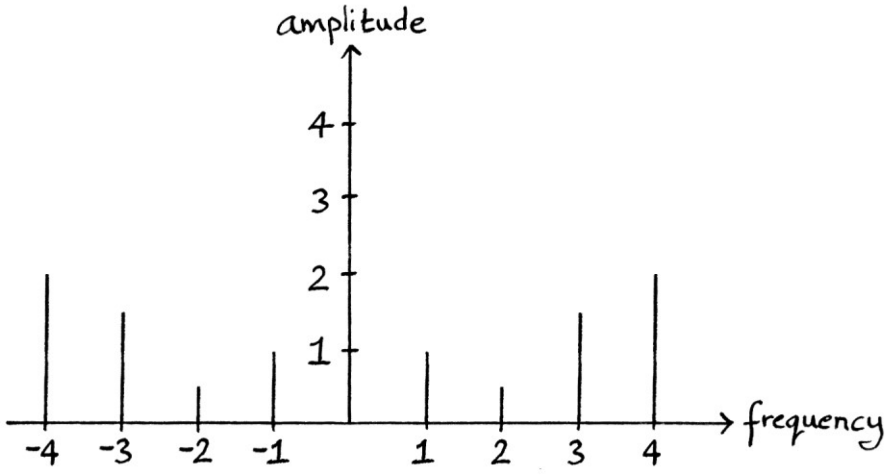
“y = sin (2πt) + sin (2π * 3t)”

A time-based Sine wave can be thought of as indicating the y-axis position of an object rotating around a circle. If an object rotates anticlockwise, the frequency is positive; if it rotates clockwise, the frequency is negative. By phrasing a Sine wave or a signal in terms of Imaginary exponents of “e”, we are saying it consists of objects rotating in different directions around circles. This conflicts with our normal understanding of Sine or Cosine. However, it is important to note that this situation is different because the exponential form of a wave *is a portrayal of the wave within the Complex helix chart*. Although the exponential form of a wave is mathematically the same as the normal form, it is still really a portrayal of the wave using another way of thinking.

The best way of thinking about exponential forms of waves is that they are not really Sine waves or Cosine waves, but *portrayals* of Sine waves or Cosine waves in the realm of the Complex helix chart. The frequencies we see are the frequencies of this portrayal. It is best to think of both the positive and negative frequencies of this portrayal as being illusionary – they refer to the *portrayal* of the wave, and not to the actual wave that is being portrayed. [This is similar to how we could paint a picture of a person with oil paints, but the person who has been painted is not made of paint in real life]. As it is, the positive frequency of the portrayal will be the same as the frequency of the actual wave, but in my view, it is still better to think of both frequencies as being illusionary. However we think of the situation, the negative frequency aspect only exists in this portrayal and does not reflect the frequency of the actual wave.

Putting Sine and Cosine in terms of Imaginary powers of "e" means that any portrayal of a wave will have both positive and negative frequencies. What is more, any *sum* of waves will also consist of waves with positive and negative frequencies. Strictly speaking, we should say that it is the portrayal of the sum that has positive and negative frequencies. This means that when using Fourier series analysis on a signal *while dealing with it in the Complex Exponential form*, we would need to take into account both the positive and negative frequencies of the portrayal of each constituent wave. In other words, the list of test frequencies would be positive and negative integer multiples of the frequency of the signal being analysed. In the list of discovered constituent waves, for every positive frequency, there would be a corresponding negative frequency with the same absolute value. For example, the constituent waves might have frequencies such as: −3, −2, −0.5, +0.5, +2, +3. We would not be analysing a signal in terms of Sine or Cosine waves, but in terms of Sine or Cosine waves being portrayed *in the Complex helix chart.*

This is one of the reasons that some frequency domain graphs have both positive and negative frequencies that are mirrored along the frequency axis. In such cases, the frequency domain graphs are not showing the frequencies of the original waves, but the frequencies of the Complex Exponential equivalents of the original waves.



Given that the negative frequencies really only exist in the Complex helix chart portrayal of a wave, some people would say that these negative frequencies are illusionary. In a received radio signal, for example, there will be no negative frequencies in the constituent waves *in reality*. There are not really any objects rotating the wrong way around a circle. However, if we portray the constituent waves as being the addition or subtraction of two helices of opposing frequencies, there will be apparent negative frequencies due to the way we are portraying the waves. A Sine wave or a Cosine wave only has one frequency, but when we create a portrayal of a Sine wave or a Cosine wave in the realm of the Complex helix chart, the *portrayal* has two frequencies that are the same but of opposing directions. As I said before, personally, I would go one step further than saying that the negative frequencies are illusionary – I would say that both the negative and positive frequencies are illusionary – they exist in the portrayal. I think this is a better way of thinking about the situation, and a better way of maintaining the separation between waves in the Complex helix chart and the actual waves.

Using the analogy of a bucket containing unmixable liquids, if there were some way of examining a bucket's contents that required holding two mirrors over the top of it, one would see twice the contents of the bucket – there would not literally be twice the contents in existence, but such a method of examining it would create that illusion.

Characteristics of the waves

We can alter the formulas for Sine and Cosine in the Complex helix chart to change the amplitude, frequency, phase and mean levels.

We will look at the Sine wave first. For an amplitude of 1 unit, a frequency of 1 cycle per second, zero phase, and zero mean level, the normal Sine wave would have the formula "y = sin 2πt". The formula for that Sine wave in terms of Imaginary exponents of "e" would be:

$$z = \frac{1}{2i} * (e^{i(2\pi t)} - e^{-i(2\pi t)})$$

Amplitude

If we wanted an amplitude of 2 units, we would scale the whole formula with a multiplication by 2:

$$z = \frac{2}{2i} * (e^{i(2\pi t)} - e^{-i(2\pi t)})$$

... which we could also phrase as:

$$z = \frac{1}{2i} * (2e^{i(2\pi t)} - 2e^{-i(2\pi t)})$$

Frequency

If we wanted to have a frequency of 3 cycles per second, we would multiply the time *in both exponentials* by 3:

$$z = \frac{1}{2i} * (e^{i(2\pi * 3t)} - e^{-i(2\pi * 3t)})$$

Phase

When it comes to phase, we have to remember that the exponents are really "i * (...)" and "−i * (...)" where the dots are replaced by the frequency, phase, time and the 2π. If we have no phase then not remembering this does not matter, but when it comes to non-zero phases, this is important.

When the exponents are kept in the form of "i * (...)" and "−i * (...)", the phases applied to each exponential are always the same. [It is not the case that the second phase is the negative of the first one.] If the exponents are in the form where the multiplication by "i" or "−i" has been carried out, then the second phase will be the negative of the first one, and both phases will be Imaginary.

If we wanted a phase of 0.125π radians (22.5 degrees), we would have:

$$z = \frac{1}{2i} * (e^{i(2\pi t + 0.125\pi)} - e^{-i(2\pi t + 0.125\pi)})$$

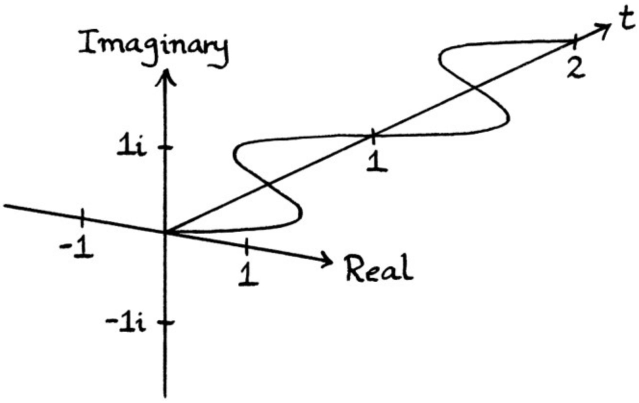
If we perform the multiplications by "i" and "−i" in the exponents, a phase of 0.125π radians would look like this:

$$z = \frac{1}{2i} * (e^{(2\pi i t + 0.125 i\pi)} - e^{(-2\pi i t - 0.125 i\pi)})$$

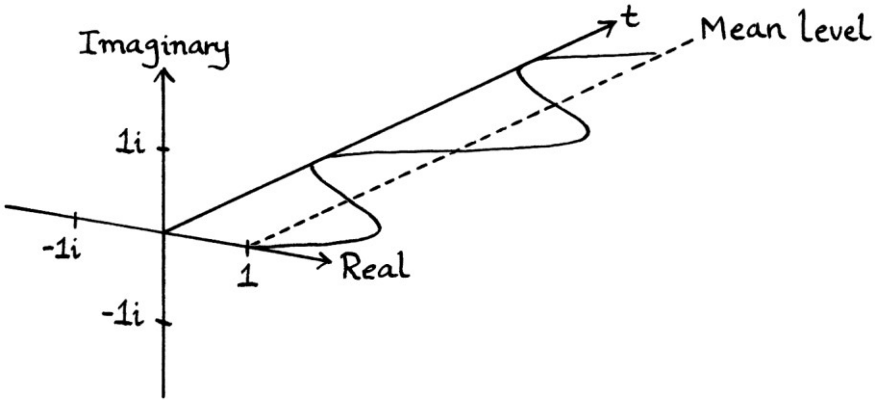
Mean levels

For a normal Sine wave, we only have one mean level. However, this Sine wave is being portrayed in the Complex helix chart. Therefore, it can have two mean levels. The Real mean level will move it along the Real axis; the Imaginary mean level will move it up or down the Imaginary axis. [Note that you are unlikely to see an Imaginary mean level used in practice.]

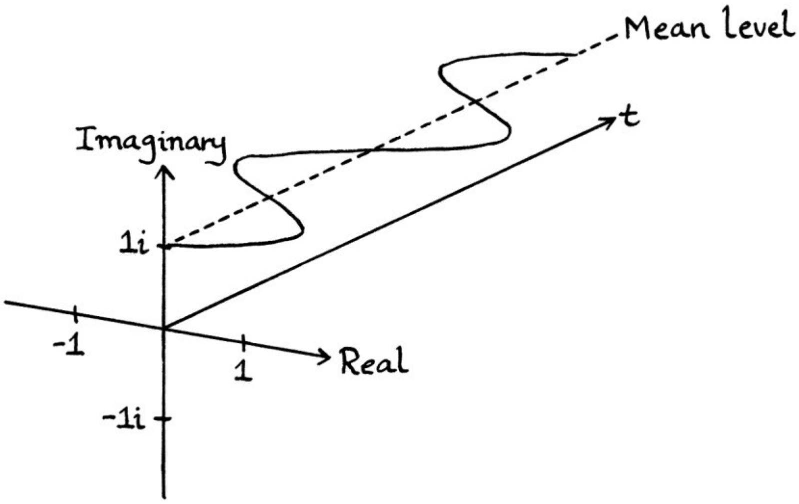
Our Sine wave with zero mean levels looks like this:



If we give it a positive Real mean level, it will become shifted towards the positive end of the Real axis:



If we give it a positive Imaginary mean level, it will become shifted up the Imaginary axis:



As an example of mean levels in use, a Real mean level of 2 units and an Imaginary mean level of −3 units would be portrayed in this way in the formula:

$$z = 2 \; - \; 3i \; + \; \frac{1}{2i} * (e^{i(2\pi t)} \; - \; e^{-i(2\pi t)})$$

A general formula

A general formula for a Sine wave is:

$$z = h_r + ih_i + \frac{1}{2i} * (Ae^{i(2\pi ft \, + \, \phi)} - Ae^{-i(2\pi ft \, + \, \phi)})$$

... or, if we tidy up the exponent:

$$z = h_r + ih_i + \frac{1}{2i} * (Ae^{(2\pi ift \, + \, \phi i)} - Ae^{(-2\pi ift \, - \, \phi i)})$$

... where:
“h_r” is the mean level for the Real axis.
“h_i” is the mean level for the Imaginary axis.
“A” is the amplitude.
“f” is the frequency in cycles per second.
“ϕ” is the phase in radians.

An easy mistake to make when it comes to formulas involving phase is to forget that the exponents for frequency and phase are all subjected to a multiplication by "i".

A Cosine wave is altered in much the same way. As with a Sine wave, a Cosine wave in this situation can have two mean levels. A general formula for a Cosine wave is:

$$z = h_r + ih_i + \frac{1}{2} * (Ae^{i(2\pi ft + \phi)} + Ae^{-i(2\pi ft + \phi)})$$

... which we could also write in this way if we tidy up the exponents:

$$z = h_r + ih_i + \frac{1}{2} * (Ae^{(2\pi ift + \phi i)} + Ae^{(-2\pi ift - \phi i)})$$

Equivalent formulas

Note that the above formulas are ones that allow the two-dimensional waves to sit anywhere in the three-dimensional Complex planes. Generally, in maths, we would only be interested in the two-dimensional waves that sit flat in the Real axis, or in other words, the waves that are entirely Real – the waves that are mathematically equivalent to the standard non-exponential way of describing waves. For this reason, the Imaginary mean level will never appear in Complex exponentials that are equivalents to normally-described waves. Therefore, the Complex exponential formula that is equivalent to a non-Complex Sine wave formula will be:

$$z = h_r + \frac{1}{2i} * (Ae^{i(2\pi ft + \phi)} - Ae^{-i(2\pi ft + \phi)})$$

... where the mean level, "h_r", is a Real number.

The Complex exponential formula that is equivalent to a non-Complex Cosine wave formula will be:

$$z = h_r + \frac{1}{2} * (Ae^{i(2\pi ft + \phi)} + Ae^{-i(2\pi ft + \phi)})$$

... where the mean level, "h_r", is a Real number.

An example

We can test the formulas by entering a value for a time for which we know the result. If we have this normal Sine wave formula:
“y = 1 + 3 sin ((2π * 2.5t) + 0.2π)”
... then we know that at 0.5 seconds, “y” will be 3.4271 units.

When the wave is portrayed in the Complex helix chart, that same point should be at “3.4271 + 0i” and 0.5 seconds down the time axis. We will fill the details into the Complex formula. As the Sine wave is flat in the Complex helix chart, a *Real* mean level will slide the wave up and down from the point of view of the wave – it will slide it and down the Real axis. We do not want an Imaginary mean level in our formula. Our formula is:

$$z = 1 + \frac{1}{2i} * (3e^{i(2\pi*2.5*0.5\ +\ 0.2\pi)}\ - 3e^{-i(2\pi*2.5*0.5\ +\ 0.2\pi)})$$

This can be slightly simplified by sorting out the multiplications by “i” in the exponents:

$$z = 1 + \frac{1}{2i} * (3e^{(2\pi i*2.5*0.5\ +\ 0.2\pi i)}\ - 3e^{(-2\pi i*2.5*0.5\ -\ 0.2\pi i)})$$

... which is:

$$z = 1 + \frac{1}{2i} * (3e^{(2.5\pi i\ +\ 0.2\pi i)}\ - 3e^{(-2.5\pi i\ -\ 0.2\pi i)})$$

... which is:

$$z = 1 + \frac{1}{2i} * (3e^{2.7\pi i} - 3e^{-2.7\pi i})$$

... which is:

$$z = 1 + \frac{3e^{2.7\pi i}}{2i} - \frac{3e^{-2.7\pi i}}{2i}$$

We can sort out the division by “2” as so:

$$z = 1 + \frac{1.5e^{2.7\pi i}}{i} - \frac{1.5e^{-2.7\pi i}}{i}$$

In Chapter 23, we saw a way of solving divisions by “i”. The rule was:

$$\frac{a}{i} = -ia$$

... where “a” is any number, whether a Real number or a Complex number.

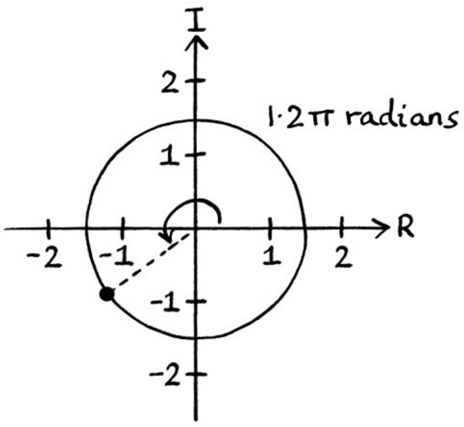
Therefore, we can rephrase the formula as:
$z = 1 + -i(1.5e^{2.7\pi i}) - -i(1.5e^{-2.7\pi i})$
... which is:
$z = 1 - 1.5ie^{2.7\pi i} + 1.5ie^{-2.7\pi i}$

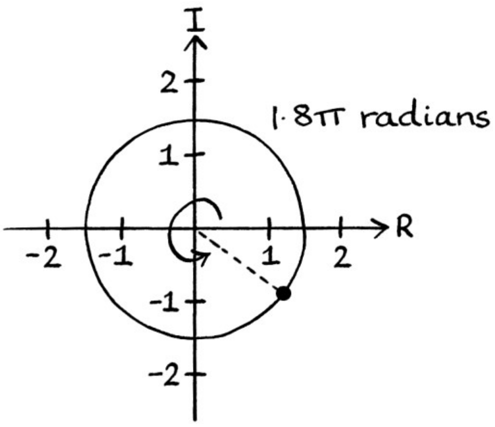
We know that a multiplication by “i” rotates a point by 90 degrees, which is 0.5π radians. This means that for both exponentials, we can remove the multiplication by “i” at the same time as adding 0.5π radians to the exponents, and it will be as if nothing has changed: [Note that the second exponential has a *negative* exponent.]
$z = 1 - 1.5e^{3.2\pi i} + 1.5e^{-2.2\pi i}$

As there are 2π radians in a circle, we can change the first exponent from 3.2πi to 1.2πi as so:
$z = 1 - 1.5e^{1.2\pi i} + 1.5e^{-0.2\pi i}$

As a point on a circle’s edge marked with a negative angle can also be marked with a positive angle that is that value subtracted from 360 degrees (or 2π radians), we can rephrase the second exponential to have a positive exponent. This gives us:
$z = 1 - 1.5e^{1.2\pi i} + 1.5e^{1.8\pi i}$

We can actually solve the exponential parts of this by drawing two circles, and seeing where the points marked by $1.5e^{1.2\pi i}$ and $1.5e^{1.8\pi i}$ are. The first point is on a circle with a radius of 1.5 units, and at an angle of 1.2π radians; the second point is on a circle with a radius of 1.5 units, and at an angle of 1.8π radians.





Supposing we could read the circles very accurately, we would see that the first point is at "−1.2135 – 0.8817i", and the second point is at "1.2135 – 0.8817i". [The circles are mirror images of each other]. We could also use Cosine and Sine to calculate the position of these points, and so use a basic calculator to find where they are. The first point is at: "1.5 cos 1.2π + i * 1.5 sin 1.2π" and the second point is at: "1.5 cos 1.8π + i * 1.5 sin 1.8π".

Our main calculation becomes:
z = 1 – (−1.2135 – 0.8817i) + (1.2135 – 0.8817i)
z = 1 + 1.2135 + 0.8817i + 1.2135 – 0.8817i
z = 1 + 1.2135 + 1.2135
z = 3.427

If we had used a calculator that can work with Complex numbers to solve the initial calculation, we would have ended up with 3.42705098 (to 8 decimal places).

This confirms that the exponential version of the wave formula works correctly. Clearly, it is much quicker with a calculator, but this example shows how much can be achieved without one.

Rotating the waves

Given that the waves are two-dimensional, yet in a three-dimensional helix chart, we can alter the formulas to rotate them too. To do this, we rotate every single point of the wave. Whether there is any purpose in doing this is another matter, and you are unlikely to see this being done in a typical maths book. [One reason for *not* doing it is that no normally-described wave can ever be equivalent to one that is not entirely Real in the three-dimensional Complex plane.]

The formula for rotating the Sine wave by "r" radians, while keeping its centre at the position of the mean levels will be:

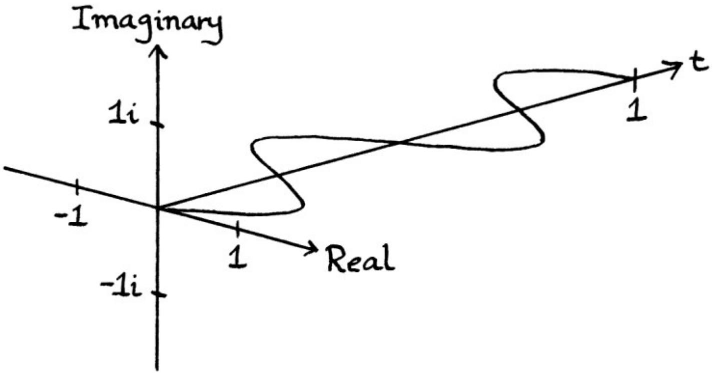
$$z = h_r + ih_i + \frac{e^{ir}}{2i} * (ae^{i(2\pi ft\,+\,\phi)} - ae^{-i(2\pi ft\,+\,\phi)})$$

This multiplies the whole formula, save for the mean levels, by an Imaginary power of "e", thus rotating the wave around its centre by a number of radians anticlockwise. [We could also have rotated the wave by using powers of "i".]

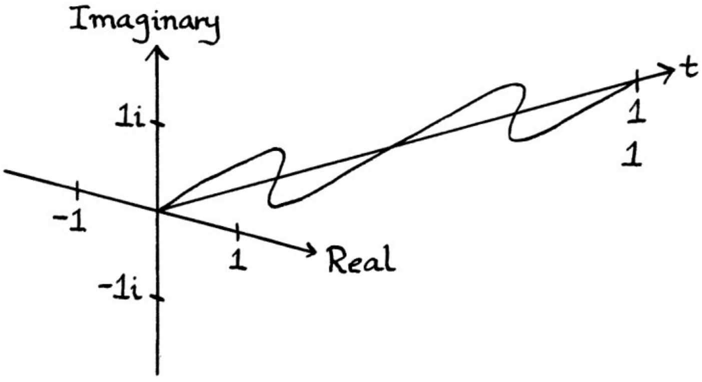
We can test the new formula by calculating one point of a formula. We will use the following formula, which is a Sine wave with an amplitude of 1 unit, a frequency of 2 cycles per second, a phase of zero radians, mean levels of zero units, and a rotational amount of 45 degrees (0.25π radians) anticlockwise:

$$z = \frac{e^{0.25\pi i}}{2i} * (e^{i(2\pi*2t)} - e^{-i(2\pi*2t)})$$

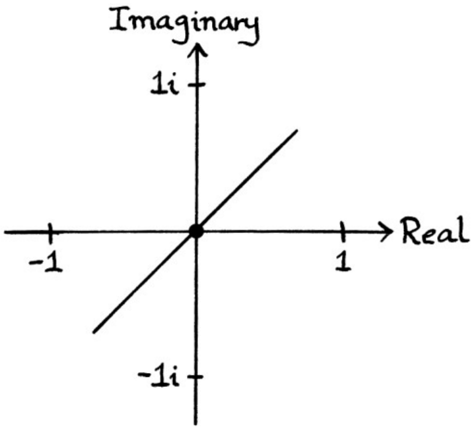
The curve of this wave will be at 45 degrees (0.25π radians) within the Complex helix chart. Without being rotated, it would have looked like this:



Now that it has been rotated anticlockwise by 45 degrees (0.25π radians), it will look like this: [It is difficult to draw this in a clear way]



Viewed end on, the Complex helix chart looks like this:



We can simplify the formula by carrying out the main multiplication:

$$z = \frac{e^{0.25\pi i}}{2i} * (e^{i(2\pi * 2t)} - e^{-i(2\pi * 2t)})$$

... becomes:

$$z = \frac{e^{0.25\pi i} * e^{i(2\pi * 2t)}}{2i} - \frac{e^{0.25\pi i} * e^{-i(2\pi * 2t)}}{2i}$$

[Sidenote: we can perform multiplication with exponentials in this situation and the results will still be correct, as I will explain later in this chapter.]

We know that "$c^x * c^y = c^{x+y}$", which means that we can rephrase the formula as:

$$z = \frac{e^{0.25\pi i + i(2\pi * 2t)}}{2i} - \frac{e^{0.25\pi i + -i(2\pi * 2t)}}{2i}$$

...which is:

$$z = \frac{e^{0.25\pi i + 4\pi i t}}{2i} - \frac{e^{0.25\pi i + -4\pi i t}}{2i}$$

... which, after rearranging the exponents, is:

$$z = \frac{e^{4\pi i t + 0.25\pi i}}{2i} - \frac{e^{-4\pi i t + 0.25\pi i}}{2i}$$

We know that "a ÷ i = −ia", and that "a ÷ 2 = 0.5a" (obviously), so this becomes:

$$z = -0.5ie^{4\pi it + 0.25\pi i} - -0.5ie^{-4\pi it + 0.25\pi i}$$

... so:

$$z = -0.5ie^{4\pi it + 0.25\pi i} + 0.5ie^{-4\pi it + 0.25\pi i}$$

We know that a multiplication by "i" rotates by 1 quarter-circle angle unit (90 degrees, or 0.5π radians), so we can remove the "i" from the second exponential and add 0.5π radians on to the exponent. A multiplication by "−i" rotates by −1 quarter-circle angle units (−0.5π radians). Therefore, we can remove the "−i" from the first exponential and subtract 0.5π radians from the exponent.

$$z = 0.5e^{4\pi it\ -\ 0.25\pi i} + 0.5e^{-4\pi it\ +\ 0.75\pi i}$$

The negative phase in the first exponential can be changed to be a positive phase [−0.25π radians is the same angle as 2π − 0.25π = 1.75π radians]:

$$z = 0.5e^{4\pi it\ +\ 1.75\pi i} + 0.5e^{-4\pi it\ +\ 0.75\pi i}$$

... and this is a simplified version of the formula to create the Sine wave at an angle of 45 degrees. This is one object rotating clockwise around a circle added to an object rotating anticlockwise around a circle.

We can test the formula by checking a known result. As the wave has a frequency of 2 cycles per second, we know that at 0.125 seconds, the wave will be at its first peak, which should be at the point 0.7071 + 0.7071i, and 0.125 seconds down the time axis. We will put 0.125 seconds into the formula:

$$z = 0.5e^{4\pi i * 0.125\ +\ 1.75\pi i} + 0.5e^{-4\pi i * 0.125\ +\ 0.75\pi i}$$

... which is:

$$z = 0.5e^{0.5\pi i\ +\ 1.75\pi i} + 0.5e^{-0.5\pi i\ +\ 0.75\pi i}$$

... which is:

$$z = 0.5e^{2.25\pi i} + 0.5e^{0.25\pi i}$$

... which, because 2.25π radians is the same angle as 0.25π radians, is:

$$z = 0.5e^{0.25\pi i} + 0.5e^{0.25\pi i}$$

... and, because these exponentials are the same, they can be added together as so:

$$z = e^{0.25\pi i}$$

Given how Imaginary powers of "e" work, we know that this is a point on a unit-radius circle at an angle of 0.25π radians (45 degrees). Therefore, we know that the point's Complex number is "0.7071 + 0.7071i", which confirms that the Sine wave in the original formula has been rotated by +45 degrees.

The formula for rotating the *Cosine* wave by "r" radians, while keeping its centre at the position of the mean levels is:

$$z = h_r + ih_i + \frac{e^{ir}}{2} * (ae^{i(2\pi ft\,+\,\phi)} + ae^{-i(2\pi ft\,+\,\phi)})$$

More Sine and Cosine formulas

Given that a Sine wave with a 90-degree (0.5π radian) phase is the same as a Cosine wave with zero phase, we can say that:

$$\frac{1}{2} * (ae^{i(2\pi ft\,+\,\phi)} + ae^{-i(2\pi ft\,+\,\phi)}) = \frac{1}{2i} * (ae^{i(2\pi ft\,+\,0.5\pi\,+\,\phi)} - ae^{-i(2\pi ft\,+\,0.5\pi\,+\,\phi)})$$

In other words, the standard Cosine exponential formula is the same as the standard Sine exponential formula if the Sine formula has 0.5π radians (90 degrees) added to both the phases.

This means that we can also portray a Cosine wave using this formula:

$$z = h_r + ih_i + \frac{1}{2i} * (ae^{i(2\pi ft\,+\,0.5\pi\,+\,\phi)} - ae^{-i(2\pi ft\,+\,0.5\pi\,+\,\phi)})$$

... which is the Sine formula with 0.5π radians added to the phases, and we can also portray a Sine wave with this formula:

$$z = h_r + ih_i + \frac{1}{2} * (ae^{i(2\pi ft\,-\,0.5\pi\,+\,\phi)} + ae^{-i(2\pi ft\,-\,0.5\pi\,+\,\phi)})$$

... which is the Cosine formula with 0.5π radians subtracted from the phases.

[Again, it pays to remember that normally, we would not require the Imaginary mean level for the Sine wave or the Cosine wave. This is because we would want the waves to be entirely Real so that they were mathematically identical to their non-Complex-exponential equivalents.]

Potential sources of confusion

For obvious reasons, it is very easy to be confused by seeing Sine and Cosine waves given in terms of Imaginary exponents of "e". It is very easy to find people who do not really understand the idea, which is not surprising as it takes a long time to explain it.

The easiest way to avoid being confused by giving Sine and Cosine in terms of Imaginary exponents of "e" is to think of them as being *portrayals* of waves and not as being the actual waves. It is best to think of the positive and negative frequencies appearing in the formulas as being the frequencies of the portrayal, and not the frequencies of the actual wave. The frequencies of the portrayal should be thought of as illusionary.

Although the exponential forms of a wave are mathematically equivalent to the normal forms of a wave, they actually refer to different concepts. A wave given in the normal form is two-dimensional in a two-dimensional chart; a wave given in terms of Imaginary powers of "e" is two-dimensional in a three-dimensional chart.

Definitions

You might occasionally hear people say that the Sine and Cosine functions are *defined* as being the addition or subtraction of two Imaginary exponents of "e". Although one could define Sine and Cosine in this way, it would be an incredibly complicated definition for two concepts that can be otherwise defined very easily. It would not be a helpful definition for anyone trying to understand what Sine and Cosine are. It can also be a sign that someone does not understand the basics of waves, which is not uncommon, even for people who understand more complicated aspects of waves. It is better to define the Sine function as giving the y-axis value of a point on a unit-radius circle's edge at a particular angle, and the Cosine function as giving the x-axis value of a point on a unit-radius circle's edge at a particular angle.

A similar point is that some people say that all Sine waves and Cosine waves are really made up of both positive and negative frequencies. What they should really be saying is that all Sine waves and Cosine waves are made up of both positive and negative frequencies *when we are portraying them in terms of Complex exponentials*. If we are not portraying them with Complex exponentials, then a Sine wave or a Cosine wave only has one frequency that might be positive or negative. [It would usually be thought of as being positive.]

Lack of helices

In explanations of the equivalences of pure waves and Complex Exponentials, it is rare that you will see pictures of helices, and it is rare that the addition or subtraction of helices will even be mentioned. Most books treat the equivalences as being an algebraic phenomenon, without regard to the simplest way of understanding why the equivalences work.

i and j

As I have said before, it pays to be acquainted with the letter “j” being used in formulas instead of the letter “i”. Different authors might use one or the other. Whether you choose to use “i” or “j” is often an arbitrary choice. It is easy to become too used to seeing either “i” or “j” in formulas, and then being temporarily confused when you see the other one. An example of the letter “j” being used in a formula from this chapter is as follows: [Everything is the same, but “i” is replaced with “j”.]

$$\sin(2\pi ft) = \frac{1}{2j} * (e^{j(2\pi ft)} - e^{-j(2\pi ft)})$$

Maths on exponentials

It is possible to perform addition and multiplication on waves when they are written in terms of Complex exponentials. The results will be the same as the calculations performed on the non-Complex exponential form of the waves. Note that converting a wave into its Complex exponential form often does not make calculations any easier or simpler.

When we were portraying circles with exponentials (whether with powers of “i” or Imaginary powers of “e”), we could use addition, but when we performed multiplication, the connection between the results and the derived waves broke down. We do not have this problem when we multiply waves that have been encapsulated in the forms of exponentials. The situation is different now, as we are not dealing in the *derived* waves of a circle or helix, but in waves that are made from adding or subtracting circles or helices.

Addition

We will add these two waves, but phrased in the form of Imaginary powers of “e”:
“y = 1+ 1.23 sin (2π * 5t)”
... and:
“y = −3 + 2.56 sin ((2π * 5t) + 1.5π)”

These two waves are mathematically identical to those of:

$$z = 1 + \frac{1}{2i} * (1.23e^{i(2\pi \,*\, 5t)} - 1.23e^{-i(2\pi \,*\, 5t)})$$

... and:

$$z = -3 + \frac{1}{2i} * (2.56e^{i(2\pi \,*\, 5t \,+\, 1.5\pi)} - 2.56e^{-i(2\pi \,*\, 5t \,+\, 1.5\pi)})$$

When we add these exponential waves, we end up with:

$$1 + \frac{1}{2i} * (1.23e^{i(2\pi \,*\, 5t)} - 1.23e^{-i(2\pi \,*\, 5t)})$$
$$+$$
$$-3 + \frac{1}{2i} * (2.56e^{i(2\pi \,*\, 5t \,+\, 1.5\pi)} - 2.56e^{-i(2\pi \,*\, 5t \,+\, 1.5\pi)})$$

This is a lot more effort to solve than if we tried to add the waves normally, so we will leave this addition as it is. This is a good sign that Complex exponentials can sometimes complicate things more than is needed.

Multiplication

We will multiply this wave by itself:
“y = 1.2 sin ((2π * 3t)”.

The Complex exponential that is mathematically identical to this wave is:

$$z = \frac{1}{2i} * (1.2e^{i(2\pi\, *\, 3t)} - 1.2e^{-i(2\pi\, *\, 3t)})$$

Therefore, we will be calculating this sum:

$$\frac{1}{2i} * \left(1.2e^{i(2\pi\, *\, 3t)} - 1.2e^{-i(2\pi * 3t)}\right) * \frac{1}{2i} * (1.2e^{i(2\pi\, *\, 3t)} - 1.2e^{-i(2\pi\, *\, 3t)})$$

Because $\frac{1}{2i} * \frac{1}{2i}$ is $\frac{1}{-4}$ or $\frac{-1}{4}$ or -0.25, we can write it as:

$$-0.25 * \left(1.2e^{i(2\pi\, *\, 3t)} - 1.2e^{-i(2\pi\, *\, 3t)}\right) * (1.2e^{i(2\pi\, *\, 3t)} - 1.2e^{-i(2\pi * 3t)})$$

It is simpler to do the maths if we sort out the exponents as so:

$$-0.25 * \left(1.2e^{6\pi it} - 1.2e^{-6\pi it}\right) * (1.2e^{6\pi it} - 1.2e^{-6\pi it})$$

Given that (a + b) * (a + b) = aa + ab + ba + bb = aa + 2ab + bb, we can solve this as follows: [written over several lines to make it easier to read]

−0.25 * (
$\left(1.2e^{6\pi it}\right) * (1.2e^{6\pi it})$
+
$2 * \left(1.2e^{6\pi it}\right) * (-1.2e^{-6\pi it})$
+
$\left(-1.2e^{-6\pi it}\right) * (-1.2e^{-6\pi it})$
)

To solve a multiplication of exponentials with the same base, we add the exponents. Therefore, we end up with this:

-0.25 * (
$\left(1.44e^{12\pi it}\right)$
+
$2 * \left(-1.44e^{0\pi it}\right)$
+
$\left(1.44e^{-12\pi it}\right)$
)

... which, written on one line is:

$$-0.25 * \left(\left(1.44e^{12\pi it}\right) + 2 * \left(-1.44e^{0\pi it}\right) + \left(1.44e^{-12\pi it}\right)\right)$$

... which, because any exponential with a zero exponent is 1, is:

$$-0.25 * \left(\left(1.44e^{12\pi it}\right) + 2 * (-1.44) + \left(1.44e^{-12\pi it}\right)\right)$$

... which is:

$$-0.25 * \left(\left(1.44e^{12\pi it}\right) - 2.88 + \left(1.44e^{-12\pi it}\right)\right)$$

... which is:

$$-0.36e^{12\pi it} + 0.72 - 0.36e^{-12\pi it}$$

... which, rearranged, is:

$$0.72\ - 0.36e^{12\pi it} - 0.36e^{-12\pi it}$$

The above formula is a mean level and a wave encoded as the sum of two exponentials. We can reframe the exponentials in the standard form of a Sine wave. We start doing this by grouping the exponentials together as so:

$$0.72 + \left(-0.36e^{12\pi it} - 0.36e^{-12\pi it}\right)$$

Next, we divide the whole exponential section by 2i, and multiply each individual exponential by 2i to keep the ultimate meaning the same:

$$0.72 + \frac{1}{2i}(-2i * 0.36e^{12\pi it} - 2i * 0.36e^{-12\pi it})$$

... which is:

$$0.72 + \frac{1}{2i}(-i * 0.72e^{12\pi it} - i * 0.72e^{-12\pi it})$$

We can rephrase the exponents to make the frequencies clearer:

$$0.72 + \frac{1}{2i}(-i * 0.72e^{i(2\pi * 6t)} - i * 0.72e^{-i(2\pi * 6t)})$$

For the next two steps, we need to remember that multiplying a point on the Complex plane by an Imaginary power of “e” with positive “i” in the exponent will rotate that point anticlockwise by a number of radians equal to the non-Imaginary part of the exponent. For example, a point multiplied by “e^{i2}” will become rotated by 2 radians anticlockwise. Conversely, an Imaginary power of “e” with a *negative* “i” in the exponent, when multiplied by a point, will rotate that point clockwise by a number of radians equal to the non-Imaginary part of the exponent. For example, if we multiply a point by “e^{-i4}”, it will become rotated *clockwise* by 4 radians. We also know that a multiplication by “i” rotates a point by 1 quarter-circle angle unit (90 degrees or 0.5π radians) anticlockwise, and a multiplication by “−i” rotates a point by 1 quarter-circle angle unit clockwise.

For the first exponential in the above formula, we have an Imaginary power of “e” with a positive “i” in the exponent being multiplied by “−i”. The exponential can be thought of as rotating anticlockwise by “2π * 6t” radians, and the “−i” can be thought of as rotating *clockwise* by 0.5π radians. Therefore, we can remove the multiplication by “−i” (thus rotating everything *anticlockwise* by 0.5π radians), and at the same time, subtract 0.5π radians from the non-Imaginary part of the exponent of the exponential (thus rotating everything *clockwise* by 0.5π radians). This will leave the overall meaning of the multiplication and the exponential the same. [For more information about doing such things, see the section “Seemingly complicated maths” in the next chapter.] We end up with:

$$0.72 + \frac{1}{2i}(0.72e^{i((2\pi * 6t)-0.5\pi)} - i * 0.72e^{-i(2\pi * 6t)})$$

For the second exponential in the formula, we have an Imaginary power of "e" with a *negative* "i" in the exponent. This exponential can be thought of as rotating *clockwise* by "2π * 6t" radians. We also have a multiplication by "−i" before the exponential, but we will want to keep the negative sign in the whole formula (to make the whole formula match that for a Sine wave). Therefore, we will just focus on the "i" part of the multiplication. This "i" on its own causes an anticlockwise rotation. We can remove the "i" (thus rotating everything clockwise by 0.5π radians), and at the same time subtract 0.5π radians from the non-Imaginary part of the exponent of the exponential (thus rotating everything anticlockwise by 0.5π radians). Again, this will leave the overall meaning of the multiplication and the exponential the same:

$$0.72 \, + \frac{1}{2i}(0.72e^{i((2\pi \, * \, 6t)-0.5\pi)} - 0.72e^{-i((2\pi \, * \, 6t)-0.5\pi)})$$

If we want to, we can make the phases positive by adding 2π to each one:

$$0.72 \, + \frac{1}{2i}(0.72e^{i((2\pi \, * \, 6t)+1.5\pi)} - 0.72e^{-i((2\pi \, * \, 6t)+1.5\pi)})$$

This subtraction now represents a Sine wave with a mean level of 0.72 units, an amplitude of 0.72 units, a frequency of 6 cycles per second, and a phase of 1.5π radians (270 degrees). The non-exponential formula of the Sine wave is:
"y = 0.72 + 0.72 sin ((2π * 6t) + 1.5π)"
... which is what we would have ended up with, had we not used Complex exponentials. This confirms that everything is correct.

A process that would have been extremely simple if done normally has become much more complicated when using Complex exponentials. However, we can see that the relationship between the exponentials and the waves stays consistent throughout the maths. There is no mathematical difference between a wave described normally, and that wave described as the sum of Complex exponentials. Therefore, any maths performed on a wave described in one way will produce the same results as if it had been described in the other way.

Whereas the connection between a Complex exponential and its derived waves breaks down when we multiply two circles together, multiplication of waves in the form of additions or subtractions of Complex exponentials remains consistent no matter what we do with them. This is because a normal wave formula is mathematically identical to its Complex exponential form.

Waves in terms of powers of "i"

For interest's sake, we will convert waves into powers of "i" in the same way that we can convert waves into Imaginary powers of "e". There is probably not much use in knowing how to do this, and you are unlikely to see this done. A power of "i" represents a circle or a helix, so we just need to add or subtract two exponentials with opposing frequencies, and scale the result (and rotate it for Sine).

Sine

The general formula for a Sine wave in terms of "i" raised to powers is as so:

$$z = h_r + ih_i + \frac{1}{2i} * (ai^{(4ft\,+\,\phi)} - ai^{((4\,*\,-ft)\,-\,\phi)})$$

... where:

- "h_r" is the mean level for the Real axis.
- "h_i" is the mean level for the Imaginary axis.
- "a" is the amplitude.
- "f" is the frequency in cycles per second.
- "ϕ" is the phase in *quarter-circle angle units*.

[Note that the Imaginary mean level would need to be zero if we were describing a Sine or Cosine wave in a way that was mathematically the same as one described normally. This is because we would need the wave to be entirely Real.]

We can also phrase the negative frequency exponent more simply, and have the formula as:

$$z = h_r + ih_i + \frac{1}{2i} * (ai^{(4ft\,+\,\phi)} - ai^{(-4ft\,-\,\phi)})$$

The formula is one helix minus a second helix, all divided by 2, and rotated by 90 degrees clockwise. This is exactly the same idea as when we were using Imaginary exponents of "e".

We have a division by 2i. A division by "i" is the same as a rotation by −90 degrees, which is 1 quarter-circle angle unit clockwise. Therefore, we can remove the division by "i", and subtract 1 from the phases in the exponents. We can also change the division by 2 into a multiplication by 0.5 to make the formula tidier:

$$z = h_r + ih_i + 0.5ai^{(4ft + \phi - 1)} - 0.5ai^{(-4ft - \phi - 1)}$$

Remember that the phases are in quarter-circle angle units. To keep the formula simple, we will leave them as quarter-circle angle units.

As an example of the formula in use, we will convert this wave into its exponential form:
1 + 2.5 sin ((2π * 3t) + 0.25π)
... where Sine is working in radians, and the phase is in radians.

First, we convert this into a quarter-circle angle unit wave, and we have:
1 + 2.5 sin ((4 * 3t) + 0.5)
... where Sine is working in quarter-circle angle units, and the phase is in quarter-circle angle units.

We fill in the relevant parts of our exponential formula. We ignore the Imaginary mean level as it is not relevant here. We will have:

$$z = 1 + (0.5 * 2.5)i^{((4 * 3t) + 0.5 - 1)} - (0.5 * 2.5)i^{((-4 * 3t) - 0.5 - 1)}$$
... which is:
$$z = 1 + 1.25i^{(12t - 0.5)} - 1.25i^{(-12t - 1.5)}$$

We can test if everything is correct by putting in a time and seeing if the results match. We will put in the time as 0.0127 seconds:

$1 + 1.25\mathrm{i}^{((12 * 0.0127) - 0.5)} - 1.25\mathrm{i}^{((-12 * 0.0127) - 1.5)}$
$= 1 + 1.25\mathrm{i}^{(0.1524 - 0.5)} - 1.25\mathrm{i}^{(-0.1524 - 1.5)}$
$= 1 + 1.25\mathrm{i}^{(-0.3476)} - 1.25\mathrm{i}^{(-1.6524)}$
= 1 + 1.06825 – 0.6491i – (−1.06825 – 0.6491i)
= 1 + 1.06825 – 0.6491i + 1.06825 + 0.6491i)
= 1 + 1.06825 + 1.06825
= 3.1365

If we had put the same time into our original wave formula, we would have had:
1 + 2.5 sin ((2π * 3 * 0.0127) + 0.25π)
= 1 + 2.1365
= 3.1365
... which is exactly the same result.

Cosine

The general formula for a Cosine wave is as so:

$$z = h_r + ih_i + \frac{1}{2} * (ai^{(4ft\,+\,\phi)} + ai^{((4\,*\,-ft)\,-\,\phi)})$$

... where:

- "h_r" is the mean level for the Real axis.
- "h_i" is the mean level for the Imaginary axis.
- "a" is the amplitude.
- "f" is the frequency in cycles per second.
- "ф" is the phase in *quarter-circle angle units*.

[Again, the Imaginary mean level is not needed if we want to have a wave that is mathematical identical to one described without exponentials. In such a case, we want the wave to be entirely Real.]

We can make the negative frequency exponent simpler:

$$z = h_r + ih_i + \frac{1}{2} * (ai^{(4ft\,+\,\phi)} + ai^{(-4ft\,-\,\phi)})$$

We can convert the big half into a multiplication by 0.5:

$$z = h_r + ih_i + \ 0.5ai^{(4ft\,+\,\phi)} + 0.5ai^{(-4ft\,-\,\phi)}$$

Supposing we had the Cosine wave:
"y = 7 cos ((2π * 11t) + 1.3π)"
... where Cosine is working in radians, and the phase is in radians, then we can convert it into an exponential form with "i" as the base.

First, we convert the Cosine wave from one based on radians to one based on quarter-circle angle units. To convert the phase, we divide 1.3π by 2π to see how much of a circle that represents: 1.3π ÷ 2π = 0.65. Then we multiply that by 4 to see how many quarter-circle angle units that is: 0.65 * 4 = 2.6. We end up with:
"y = 7 cos ((4 * 11t) + 2.6)
... where Cosine is working in quarter-circle angle units, and 2.6 is an angle in quarter-circle angle units.

We then fill in the exponential formula:

$$z = (0.5 * 7)i^{((4\,*\,11t)\,+\,2.6)} + (0.5 * 7)i^{((-4\,*\,11t)\,-\,2.6)}$$

... which is:

$$z = 3.5i^{((4\,*\,11t)\,+\,2.6)} + 3.5i^{((-4\,*\,11t)\,-\,2.6)}$$

We can test this by entering a time. We will try a time of 0.1278 seconds:

$$z = 3.5i^{((4\,*\,11\,*\,0.1278)\,+\,2.6)} + 3.5i^{((-4\,*\,11\,*\,0.1278)\,-\,2.6)}$$

... which is:

$$z = 3.5i^{(5.6232\,+\,2.6)} + 3.5i^{(-5.6232\,-\,2.6)}$$

... which is:

$$z = 3.5i^{(8.2232)} + 3.5i^{(-8.2232)}$$

... which is:

z = 6.5742

If we had put a time of 0.1278 seconds into our original non-Complex wave formula, we would have had exactly the same result:

7 cos ((2π * 11 * 0.1278) + 1.3π)
= 6.5742

Conclusion

This chapter has contained the most complicated concepts in this book so far. They are complicated ideas, but the steps to reach this point are straightforward. The main idea in this chapter is that a pure wave formula can be rewritten in terms of the addition or subtraction of two Imaginary powers of “e”.

Chapter 29: e and i in more depth

In this chapter, we will look at “e” and “i” in more depth, and we will learn more about Complex numbers. There is much more to know about all of these than I will explain in this book.

Other bases with Imaginary exponents

To understand Imaginary exponents of “e” most thoroughly, it is necessary to explore Imaginary powers of other numbers. Doing this will show that some characteristics of “$e^{i\theta}$” are not unique or particularly special.

As we know, “$e^{i\theta}$” will mark the position of a point on the circumference of a unit-radius circle at an angle of “θ” radians. For example, “e^{1i}” marks the position of a point on the circumference at 1 radian. [Normally, “e^{1i}” would be written as “e^{i}”]. We can rewrite “e^{1i}” as “2.7183^{1i}” to emphasise that “e” is not an unknown variable.

If we calculate “$2.7183^{i\theta}$” for values of “θ” of 1, 2, 3 and 4, we will end up with these Complex numbers:
$2.7183^{1i} = 0.5403 + 0.8415i$
$2.7183^{2i} = -0.4161 + 0.9093i$
$2.7183^{3i} = -0.9900 + 0.1411i$
$2.7183^{4i} = -0.6536 - 0.7568i$

If we plotted these points on the Complex plane, we would see that they are spaced at angles of 1 radian, and are all 1 unit away from the origin. This is to be expected, given what we know about “$2.7183^{i\theta}$”.

Now we will calculate “$2.6^{i\theta}$” for values of “θ” of 1, 2, 3, and 4. For now, we will calculate the position of these points using a calculator that can work with Complex numbers. The points are:
$2.6^{1i} = 0.5772 + 0.8166i$
$2.6^{2i} = -0.3337 + 0.9427i$
$2.6^{3i} = -0.9624 + 0.2716i$
$2.6^{4i} = -0.7773 - 0.6291i$

If we plotted these points on the Complex plane, we would see that these points are still all one unit away from the origin. The points that we plotted are not spaced at angles of 1 radian, but at a slightly smaller angle. We can use arctan to calculate this angle: arctan (0.8166 ÷ 0.5772) = 0.9555 radians or 54.7468 degrees. [For reference, 1 radian is 57.2958 degrees.]

We will now calculate "$2.8^{i\theta}$" for values of "θ" of 1, 2, 3, and 4. These points are:
$2.8^{1i} = 0.5151 + 0.8571i$
$2.8^{2i} = -0.4693 + 0.8831i$
$2.8^{3i} = -0.9986 + 0.05271i$
$2.8^{4i} = -0.5596 - 0.8288i$

If we plotted these points on the Complex plane, we would see that these are still all one unit from the origin of the axes. These points are spaced at slightly more than 1 radian apart from each other. They are spaced at angles of arctan(0.8571 ÷ 0.5151) = 1.02962 radians, which is 58.9928 degrees.

From these examples, we can see that "$e^{i\theta}$" is far from unique in drawing a circle. In fact, all positive Real numbers greater than 1 could be swapped with "e", and we would still have a unit-radius circle. For example, $10^{1i} = -0.6682 + 0.7440i$, which is 1 unit away from the origin of the axes. The most obvious thing that stands out about "$e^{i\theta}$" is that it treats "θ" as an angle in radians, while other exponentials treat "θ" as being some other type of angle unit. Another unique aspect of "$e^{i\theta}$" is that its non-Imaginary equivalent, "e^{θ}" [which would be best given as "e^{x}" because the exponent is no longer an angle], produces results that are always the gradient of the curve at that point.

From all of this, we can see that if we were happy working in an angle system other than radians, we could use *any* base raised to an Imaginary power (as long as that base is a positive Real number greater than 1). We would still be able to describe circles, and we would still be able to use those circles to think about waves.

How to calculate bases for angle units

If we used a base other than “e”, we would need a way of knowing what angle division that base worked in. Conversely, if we wanted to use a different angle system to radians, we would need to know what base to use.

The path to understanding how to calculate these is to examine equivalent Imaginary and non-Imaginary exponential pairs. Although an exponential such as “c^x” is in a different realm to an exponential to “c^{ix}”, they are related to each and can be used to give clues about the behaviour of each other. We can think of “c^x” and “c^{ix}” as being “exponential pairs”.

If we take the exponential “$e^{i\theta}$”, we know that when “θ” is 2π, the object rotating around a circle represented by the exponential will have completed exactly one revolution. One way of saying this is that “$e^{i2\pi} = 1 + 0i$”. We can say that this is the maximum rotation for an object rotating around a circle. Any more rotation will cause it to repeat its path.

The non-Imaginary pair of “$e^{i2\pi}$” is “$e^{2\pi}$”. This number is 535.49165552 to 8 decimal places. We can use this number as the basis for calculating bases and Imaginary exponents for angle systems other than radians. The idea we will be using is:
If “$c^x = 535.4917$”, then “c^{ix}” will be the point at exactly one revolution.
... where:
“c” is the base that we will be needing for a particular angle system
“x” is the number of divisions in a circle for that particular angle system.

We can also say that:
If “$c^x = 535.4917$”, then “$c^{ix} = 1 + 0i$”

When thinking of graphs, we are saying that the point on the “$y = c^x$” graph when $y = 535.4917$, is analogous to the point on the unit-radius circle at “$1 + 0i$”.

Quarter-circle angle units

We will calculate what base we would need to use Imaginary exponents that work in the quarter-circle angle system. There are 4 quarter-circle angle units in one circle. Therefore, we know that:
"c^{4i} = 1 + 0i"
... where "c" is the unknown base that we want to find.

By switching to the non-Imaginary pair of "c^{4i}", we can say that:
"c^4 = 535.4917"
... where "c" is the unknown base that we want to find.

Using our knowledge of exponentials, we can rephrase the equation as:
c = $\sqrt[4]{535.4917}$
[We have taken the fourth root of each side]
... which means that:
c = 4.81047738 (to 8 decimal places).

Therefore, we can say that:
4.8105^4 = 535.4917

Using the Imaginary version of the exponential, we can say:
4.8105^{4i} = 1 + 0i

Therefore, we can say that "$4.8105^{i\theta}$" will indicate the point on a unit-radius circle at an angle of "θ" quarter-circle angle units. We can test this using a calculator that can work with Complex numbers. We will find the point on a unit-radius circle at an angle of 0.5 quarter-circle angle units:
$4.8105^{0.5i}$ = 0.7071 + 0.7071i
... and, by now, we should know that "0.7071 + 0.7071i" is the point one unit away from the origin at 45 degrees. The angle of 45 degrees is the same as 0.5 quarter-circle angle units.

We can also say that:
"$4.8105^{i\theta}$ = cos θ + i sin θ" when "θ" is an angle in quarter-circle angle units, and Cosine and Sine are working in quarter-circle angle units. [Being pedantic, it does not actually matter whether "θ" is an angle in quarter-circle angle units or not because it will be treated as one anyway because of it being an Imaginary exponent of 4.8105.]

When we were dealing with Real powers of "i" in Chapter 25, we saw that we could identify any point on a unit-radius circle with "i^{θ}", where "θ" was an angle in quarter-circle angle units. Now, we have a way of identifying any point on a unit-radius circle with "$4.8105^{i\theta}$", where "θ" is an angle in quarter-circle angle units. This allows us to say that:
"$i^{\theta} = 4.8105^{i\theta}$"
... where "θ" is an angle in quarter-circle angle units.

This is really the main connection between Real powers of "i" and Imaginary powers of "e".

If we wanted, we could also say:
"$i^{\theta} = 4.8105^{i\theta} = \cos\theta + i\sin\theta$"
... where "θ" is an angle in quarter-circle angle units, and Cosine and Sine are working in quarter-circle angle units.

If we multiply a point by "$4.8105^{i\theta}$", it will rotate it by "θ" quarter-circle angle units. For example, to rotate the point at "0 + 1i" by 90 degrees, which is 1 quarter-circle angle unit, we multiply "0 + 1i" by "4.8105^{1i}". If we have a calculator that can work with Complex numbers, we will see that the result of this is: "−1 + 0i", which is the expected result.

Degrees

We can find the base, that when raised to an Imaginary power, works in degrees. There are 360 degrees in a circle, so we know that:
"$c^{360i} = 1 + 0i$"
... where "c" is the unknown base that we want to find.

By switching to the non-Imaginary equivalent of "c^{360i}", we can say that:
"$c^{360} = 535.4917$"
... where "c" is the unknown base that we want to find.

We can rephrase that as:
$c = \sqrt[360]{535.4917}$
... which, in English, means "c" is the 360th root of 535.4917. This produces:
c = 1.01760649 (to 8 decimal places).

We can now indicate any point on a unit-radius circle with the formula “$1.01761^{i\theta}$” where “θ” is an angle in degrees. We can check this by using the point at 45 degrees, for which we already know the answer. We calculate:
1.01761^{45i}
... which is:
0.7071 + 0.7071i
... and we know that this point is 1 unit from the origin of the axes at an angle of 45 degrees.

Using the degrees formula, we can say:
“$1.01761^{i\theta} = \cos\theta + i \sin\theta$”, when “θ” is an angle in degrees, and Cosine and Sine are working in degrees.

In Chapter 25, we saw how we could use powers of “i” with degrees. The exponential “$i^{(\theta/90)}$” indicates the point one unit away from the origin at an angle of “θ” degrees. [Remember that the exponent as a whole is in quarter-circle angle units, and the division by 90 turns a value in degrees into quarter-circle angle units.] We can now say that:
“$i^{(\theta/90)} = 1.01761^{i\theta}$”
... where “θ” is an angle in degrees.

We can also use a multiplication of “$1.01761^{i\theta}$” to rotate a point by “θ” degrees.

Whole-circle angle units

A circle that has been “divided” into 1 portion uses whole-circle angle units. There is one whole-circle angle unit in a circle. To find the base that works with whole-circle angle units, we need to solve this:
“$c^{1i} = 1 + 0i$”
... where “c” is the unknown base that we want to find.

By switching to the non-Imaginary equivalent of “c^{1i}”, we can say that:
“$c^{1} = 535.4917$”
... where “c” is the unknown base that we want to find.

As any number to the power of 1 remains the same, we know that “c” in this case must be 535.4917. In other words, $535.4917^{1} = 535.4917$.

Therefore, we can say that:
$535.4917^{1i} = 1 + 0i$

What is more, we can now say that "$535.4917^{i\theta}$" gives the position of a point on a unit-radius circle at an angle of "θ" whole-circle angle units.

We can also say that:
"$535.4917^{i\theta} = \cos\theta + i \sin\theta$", when "θ" is an angle in whole-circle angle units, and Cosine and Sine are working in whole-circle angle units.

We can also say that a multiplication by "$535.4917^{i\theta}$" rotates a point on the Complex plane by "θ" whole-circle angle units.

Unknown divisions

Supposing we had a particular base, and wanted to know the angle system that that base worked in, we would need to solve this equation:
$c^x = 535.4917$
... where "c" is the base, which we know, and "x" is the number of divisions that that base works with, which we want to find out.

We can use logarithms to solve this:
$x = \log_c (535.4917)$

Supposing we had a base of 10, then we would need to solve this:
$x = \log_{10} (535.4917)$
... which is:
$x = 2.72875271$

This means that if we use a base of 10 in an exponential with an Imaginary exponent, the exponential will treat the Real part of the exponent as being an angle in a system that divides a circle into 2.7288 parts.

Therefore, we can say:
$10^{2.7288i} = 1 + 0i$
... and:
"$10^{i\theta} = \cos\theta + i \sin\theta$", where "θ" is an angle in a system that divides a circle into 2.7288 parts, and Cosine and Sine are working in that system.

Table of bases and exponents

There are countless ways to divide up a circle, and therefore, there are countless exponentials that can indicate the position of points around a unit-radius circle. As we have seen, the rules for calculating the bases and angle systems are as so:

We can calculate the base for an angle system using the formula:
“$c = \sqrt[x]{535.4917}$”
... where:

- “c” is the base that we want to find.
- “x” is the number of portions into which the circle has been divided, which we know.
- 535.4917 is “$e^{2\pi}$”.

We can calculate the angle system for a particular base by using the formula:
“$x = \log_c (535.4917)$”
... where:

- “x” is the number of portions into which the circle has been divided, which is the value we want to find.
- “c” is the base, which we know.
- 535.4917 is “$e^{2\pi}$”.

On the next page is a table of a few examples of bases and angle systems. A value taken from first column, when raised to the power of the corresponding value from the second column, will result in 535.4917. The same exponential with the exponent multiplied by “i” will indicate a point on a unit-radius circle at an angle that is equivalent to 360 degrees.

Base for a particular circle division	**Number of divisions in the circle**	
535.491655524765	1	
23.140692632779	2	This base is “Gelfond's constant”.
10.089090581019	2.7182...	This circle is divided into “e” divisions.
10	2.728752707684	
8.120527396670	3	
7.389056098931	3.1415... (π)	This base is e^2 and the circle is divided into π pieces. This base is more obvious if you remember that $(e^2)^\pi = e^{2\pi}$.
4.810477380965	4	This is the base for quarter-circle angle units.
4.304530324517	4.304530324517	The number of divisions in this circle is the same as the base. In other words, $4.3045^{4.3045i}$ indicates a point on a unit-radius circle’s edge at the equivalent of 360 degrees, in a system that divides a circle into 4.3045 pieces.
3.513585624286	5	
3.1415... (π)	5.488792932594	
2.718281828459 (e)	2π (6.28...)	This is $e^{2\pi}$.
2	9.064720283654	
1.874456087585	10	
1.369107777062	20	
1.064847773329	100	
1.017606491206	360	This is the base for working in degrees.
1.006302965923	1000	
1	∞	This circle would need to be divided into an infinite number of pieces.

Of the numbers in the table, one that stands out is 4.30453032 (given to 8 decimal places). With this number:
$4.30453032^{4.30453032} = 535.49165552$
... and:
$4.30453032^{4.30453032i}$ indicates the position of a point at the equivalent of 360 degrees on a unit-radius circle.

We can also say that:
$4.30453032^{4.30453032i} = 1 + 0i$
... which can also be written as:
$4.30453032^{4.30453032i} = 1$.

This is an interesting equation as it means that 4.30453032 is the solution to:
$x^{xi} = 1$
... and:
$x^x = e^{2\pi}$

The step from i to e

I have just explained how to calculate any base or exponent for an Imaginary exponential. The trouble with my explanation so far is that it glosses over the important details. Here is a more thorough explanation that repeats some of what I have just said. We will start from scratch as if the discoveries from earlier in this chapter had not been made. This explanation also leads on to how we can solve seemingly complicated calculations involving "i".

In Chapter 25, we used powers of "i" to identify points on a unit-radius circle. In Chapter 27, we saw that Imaginary powers of "e" worked in an identical way (but in a different angle system). The connection between the two was made by *observing* how Imaginary powers of "e" worked, rather than through using maths. In this section, we will use maths to show how they are connected.

In Chapter 25, we could alter the way the powers of "i" formula worked, so that it could operate with "θ" in degrees, radians or other angle units. We had four formulas based on powers of "i" (although there are obviously countless possible angle units):
We can use "i^{θ}", where "θ" is treated as an angle in quarter-circle angle units.
We can use "$i^{4\theta}$", where "θ" is treated as an angle in whole-circle angle units.
We can use "$i^{\theta/90}$", where "θ" is treated as an angle in degrees.

We can use "$i^{2\theta/\pi}$", where "θ" is treated as an angle in radians.
For each of these, putting "θ" into the formula immediately converts the angle unit into quarter-circle angle units. In other words, "θ" is the angle in a particular type of angle units, but the exponent as a whole is treated as being in quarter-circle angle units.

For each of these, we can also say that they are equivalent to "cos θ + i sin θ", when "θ" is an angle in the relevant angle system, and Cosine and Sine are working in that angle system. For example, "$i^{\theta/90}$ = cos θ + i sin θ", when "θ" is an angle in degrees, and Cosine and Sine are working in degrees.

To see how these can be connected to Imaginary powers of "e", we need to find a way of changing each of these formulas so that instead of them consisting of "i" raised to a power, they become a Real number raised to a multiple of "i". We want to change them from being "$i^{\theta?}$" to "$c^{i\theta}$"
...where:

- "θ" represents an angle in a particular system of dividing a circle.
- "?" represents some value multiplying or dividing "θ"
- "c" is the new base for the exponential. The value represented by "c" will be a Real number.

Doing this will result in the meaning of the exponential as a whole become more abstract.

Quarter-circle angle units

We will start with quarter-circle angle units, which use the simplest of the formulas: "i^{θ}".

We want to find the value "c" in the following:
$c^{i\theta} = i^{\theta}$

In other words, we want to find the base, that, when raised to an Imaginary power, treats "θ" as an angle in quarter-circle angle units. To do this, we will make both halves refer to an actual point around the unit circle. We will use the point at "i^1", which is a quarter of the way around the circle. This is an arbitrarily chosen point, but it makes the maths simpler.

Therefore, we have:
$c^{1i} = i^1$
... where "c" is the unknown base. This would normally be written as:
$c^i = i$
Using our knowledge of exponentials, this can be rewritten as:
$c = \sqrt[i]{i}$
[We took the ith root of each side]

In English, this says that "c" is equal to the ith root of "i".

We could resort to using a calculator that can work with Complex numbers, but for now, we will say we do not know how to solve $\sqrt[i]{i}$.

We can say that "$(\sqrt[i]{i})^{i\theta}$" will indicate the position of a point on a unit-radius circle where "θ" is an angle in quarter-circle angle units. We can also say that:
$(\sqrt[i]{i})^{i\theta} = \cos\theta + i \sin\theta$
... where "θ" is an angle in quarter-circle angle units, and Cosine and Sine are working in quarter-circle angle units.

We can use this base to indicate points on a unit-radius circle. For example, the point at 2 quarter-circle angle units on a unit-radius circle can be represented with: $(\sqrt[i]{i})^{2i}$.

Without needing a calculator, we can use our knowledge of exponentials to solve this. The ith root of something, when raised to the power of "i" will cause the root and the power to cancel out. Therefore, we are left with:
$(i)^2$
... which, as we know, is −1. The point's position is at −1 + 0i.

As another example, the point at 4 quarter-circle angle units can be represented with: $(\sqrt[i]{i})^{4i}$. Again, the ith root and the power of "i" cancel out, and we are left with:
$(i)^4$
... which is
$(-1)^2$
... which is +1. The point's position is at 1 + 0i.

Although we can use "$(\sqrt[i]{i})^{i\theta}$" to identify points, we still cannot make it into a nice Real number raised to an Imaginary power, as we do not yet know enough to solve the ith root of "i". We will come back to this later.

Whole-circle angle units

For whole circle-angle units, the formula using powers of “i” was“$i^{4\theta}$”. To turn this into a formula that has a Real base and an Imaginary exponent, we need to solve this:
$c^{i\theta} = i^{4\theta}$

As a step towards solving this, we will use the value of “θ” that represents the equivalent of 180 degrees – in other words, 0.5 whole-circle angle units. Again, the choice of this number is completely arbitrary. We end up with this:
$c^{0.5i} = i^{4*0.5}$
$c^{0.5i} = i^{2}$
$c^{0.5i} = -1$
$c = \sqrt[0.5i]{-1}$

If we treat the radicand −1 as being $(-1)^1$, which means the same thing, we can double the index of the root, while halving this exponent, and the overall result will be the same. Therefore, $\sqrt[0.5i]{-1}$ is the same as $\sqrt[0.25i]{(-1)^{0.5}}$. The exponential $(-1)^{0.5}$ is the square root of −1, which is “i”. Therefore, we can say:
$c = \sqrt[0.25i]{i}$

We can identify the position of any point around the unit circle by giving $\sqrt[0.25i]{i}$ raised to the power of “iθ”, where “θ” is an angle in whole-circle angle units:
$(\sqrt[0.25i]{i})^{i\theta}$

We can also say that $(\sqrt[0.25i]{i})^{i\theta} = \cos\theta + i \sin\theta$” where Cosine and Sine are working in whole-circle angle units, and “θ” is being treated as an angle in whole-circle angle units.

As an example of this working, the point at 0.5 whole-circle angle units can be represented with:
$(\sqrt[0.25i]{i})^{i0.5}$

This ends up as −1, which means the point is at “−1 + 0i”.

Degrees

To rebase the formula for degrees, which is “$i^{\theta/90}$”, we need to solve this:
$c^{i\theta} = i^{\theta/90}$

To solve this, we will use the value of “θ” that represents the equivalent of 180 degrees, which obviously, is 180 degrees. Therefore, we end up with this:
$c^{180i} = i^{180/90}$

This reduces to:
$c^{180i} = i^2$
$c^{180i} = -1$
$c = \sqrt[180i]{-1}$
$c = \sqrt[90i]{i}$

We can use this as a base to the power of “iθ” where “θ” is an angle in degrees. The formula of $(\sqrt[90i]{i})^{i\theta}$ gives the position of a point on a unit-radius circle at an angle of “θ” degrees. That point’s position can be also given by “cos θ + i sin θ” when “θ” is in degrees, and Cosine and Sine are working in degrees. Or, to put this another way:
$(\sqrt[90i]{i})^{i\theta} = \cos\theta + i \sin\theta$
... when Cosine and Sine are working in degrees.

As an example of this working, the point at 270 degrees can be represented with:
$(\sqrt[90i]{i})^{i270}$

We can reduce the root and power by thinking of them as multiples of 90i, and we end up with:
$(\sqrt[1]{i})^3$
$= (i)^3$
$= -i$.

Therefore, the point is at 0 − 1i.

Radians

To rebase the formula “$i^{2\theta/\pi}$”, we need to solve this:
$c^{i\theta} = i^{2\theta/\pi}$

To solve this, we will use the value of “θ” that represents the equivalent of 180 degrees, which is the angle of π. We end up with this:
$c^{i\pi} = i^{(2*\pi)/\pi}$

This can be simplified as:
$c^{i\pi} = i^{(2\pi)/\pi}$
$c^{i\pi} = i^2$
$c^{i\pi} = -1$
$c = \sqrt[i\pi]{-1}$
$c = \sqrt[0.5i\pi]{i}$

We can use this as a base to the power of “iθ” where “θ” is an angle in radians. The formula $(\sqrt[0.5i\pi]{i})^{i\theta}$ gives the position of a point on a unit-radius circle at an angle of “θ” radians. As always, that point can also be identified using “cos θ + i sin θ” when “θ” is in radians, and Cosine and Sine are working in radians:
$(\sqrt[0.5i\pi]{i})^{i\theta}$ = cos θ + i sin θ, when Cosine and Sine are working in radians.

As an example of this in use, the point on a unit-radius circle at an angle of π radians is situated at:
$(\sqrt[0.5i\pi]{i})^{i\pi}$
... which is:
$(\sqrt[0.5]{i})$
... which is:
i^2
... which is:
-1
... which is the point at −1 + 0i.

Changing the bases

Our original exponentials with "i" as a *base* were:
For a circle divided into 4 parts: "i^{θ}".
For a circle as one part: "$i^{4\theta}$".
For a circle in degrees: "$i^{\theta/90}$".
For a circle in radians: "$i^{2\theta/\pi}$".

Our new exponentials with "i" as an *exponent* are:
For a circle divided into 4 parts: "$(\sqrt[i]{i})^{i\theta}$".
For a circle as one part: "$(\sqrt[0.25i]{i})^{i\theta}$".
For a circle in degrees: "$(\sqrt[90i]{i})^{i\theta}$".
For a circle in radians: "$(\sqrt[0.5i\pi]{i})^{i\theta}$".

With no actual evidence, we will presume that these new bases can all be converted into Real numbers, which is something that would make them a lot easier to use. We could just use a calculator that can work with Complex numbers, but we need to find a way that will show the connection in a more obvious way, so that we can see what is happening.

One fact about all of these exponentials is that they can all be used to refer to the same points around the circle. Therefore, they can all be equivalent if the appropriate values are fed into them that refer to the same point.

For example, if we put the equivalent of an angle of 180 degrees into each exponential formula, they will all result in −1 + 0i, which is −1:
$(\sqrt[i]{i})^{i2} = -1$
$(\sqrt[0.25i]{i})^{0.5i} = -1$
$(\sqrt[90i]{i})^{i180} = -1$
$(\sqrt[0.5i\pi]{i})^{i\pi} = -1$
... which means that all of *these* formulas are the same as each other. However, this does not help solve anything because all of these formulas were derived from the same original formula. Using these will, at best, mean we end up with a conclusion such as "1 = 1".

If we put the equivalent of an angle of 360 degrees into each formula, all the formulas will result in 1 + 0i, which is 1:

$(\sqrt[i]{i})^{i4} = 1 + 0i$

$(\sqrt[0.25i]{i})^{i} = 1 + 0i$

$(\sqrt[90i]{i})^{i360} = 1 + 0i$

$(\sqrt[0.5i\pi]{i})^{i2\pi} = 1 + 0i$

... which means that all of these are the same as each other too. The significant thing about this idea is that in each case we have an exponential that is equal to the full circle. For this to be so, and presuming Imaginary exponents follow the rules of non-Imaginary exponentials, it means that the smaller the base, the larger the exponent must be, and the larger the base, the smaller the exponent must be. We can put the exponentials in order of the size of their exponent, from smallest to largest, as so:

$(\sqrt[0.25i]{i})^{i} = 1 + 0i$

$(\sqrt[i]{i})^{i4} = 1 + 0i$

$(\sqrt[0.5i\pi]{i})^{i2\pi} = 1 + 0i$

$(\sqrt[90i]{i})^{i360} = 1 + 0i$

For all of these exponentials to result in “1 + 0i”, it must be the case that ($\sqrt[0.25i]{i}$) is a larger number than ($\sqrt[i]{i}$), which in turn is a larger number than ($\sqrt[0.5i\pi]{i}$), which is a larger number than ($\sqrt[90i]{i}$). We can also tell this by how the index of each root in the list is higher than the one before it, so the roots as a whole must be lower. There must be an inverse relationship between the bases and the exponents for them all to equal the same amount.

Given *that*, and given that we are presuming the bases can all be expressed with Real numbers, if we remove the “i” from the exponents, the formulas would all be equal to the same Real number. We will call this number “S” for “the Special Number”. In other words:

$(\sqrt[0.25i]{i})^{1} = S$

$(\sqrt[i]{i})^{4} = S$

$(\sqrt[0.5i\pi]{i})^{2\pi} = S$

$(\sqrt[90i]{i})^{360} = S$

... where “S” is the same Real number in each case.

This means that the number "S" in the Real world is an equivalent to the full circle in the Complex plane. It is the value that links an exponential with a Real base and Real exponent to a circle represented by a Real base and an Imaginary exponent.

Given that "S" represents the non-Complex equivalent to a full circle, and that "S" is calculated with exponential numbers, it will be the case that $\sqrt{S}$ will be the non-Complex equivalent of half a circle (180 degrees), $\sqrt[4]{S}$ will be quarter of a circle (90 degrees), $\sqrt[8]{S}$ will be an eighth of a circle (45 degrees), and so on.

This means that any exponential with a Real base and a Real exponent, that is equal to, say, $\sqrt{S}$, will, if the exponent is multiplied by "i", be equal to a point on the Complex plane that marks the position of a point on a unit-radius circle at 180 degrees.

In other words:
If c^x = S, then c^{ix} identifies the position of a point on a unit-radius circle at an angle of 360 degrees.
If c^x = S, then $c^{0.5ix}$ (in other words, $\sqrt{c^{ix}}$) identifies the position of a point on a unit-radius circle at an angle of 180 degrees.
If c^x = S, then $c^{0.75ix}$ identifies the position of a point on a unit-radius circle at an angle of 270 degrees.

Finding S

If we can find "S", we will be able to calculate the bases for all our new exponentials, and for any other exponentials we care to have, and with a minimum of effort.

We know that:
c^x = S
...where "c" is a base, and "x" is the number of divisions in a circle. Therefore, if we knew what "S" is, we would be able to find the base for any system of dividing up a circle using this equation:
c = $\sqrt[x]{S}$
... where "x" represents the desired divisions in a circle, and "c" is the base that will produce an exponential, with an Imaginary exponent, that will operate according to that angle system.

We know that:
$(\sqrt[0.25i]{i})^1 = S$
$(\sqrt[i]{i})^4 = S$
$(\sqrt[0.5i\pi]{i})^{2\pi} = S$
$(\sqrt[90i]{i})^{360} = S$
[Note how these all have non-Imaginary exponents]

Therefore, we know that "S" is exactly $\sqrt[0.25i]{i}$. However, this is no use to us as we cannot find the 0.25ith root of "i" using the information that we have.

We also know that "S" is $(\sqrt[0.5i\pi]{i})^{2\pi}$. This comes from the formula for a circle divided up into 2π portions: $(\sqrt[0.5i\pi]{i})^{2\pi i}$. It is obvious that the base for this will be the number "e", but unfortunately, I do not know a way of working this out without actually knowing it beforehand. If we guess in advance that "e" is the base, it is possible to test that it is correct. However, if we do not know that "e" is the base, there does not seem to be any way (that I can think of) of knowing what the base is. Anyway, if we use "e" as the base, everything else will fall into place.

In the second from last paragraph, I said that we could not, yet, solve a calculation such as $\sqrt[0.25i]{i}$. The reason for this is that all the methods for solving such a calculation involve prior knowledge of "$e^{i\theta}$". Therefore, it would have been cheating to use those methods when we had not shown that "e" was the base. A calculator could have found the result, but it would have used prior knowledge of "$e^{i\theta}$".

In Chapter 27, when calculating Imaginary powers of "e" using $(1 + (x \div n))^n$, we saw how "$e^{i\theta}$" identifies points around a unit-radius circle that are at an angle of "θ" radians from the origin. Given that, the missing base for an exponential that divides a circle into 2π portions must be "e". We can now say that $\sqrt[0.5i\pi]{i}$ must be equal to "e".

Once we know that $\sqrt[0.5i\pi]{i} = e$, everything else can be easily solved. We can now calculate "S". The number "S" will be the result of "$e^{2\pi}$", which is 535.491655524765 [to 12 decimal places].

Note that there is nothing necessarily special about "$e^{2\pi}$" in this situation. The breakthrough in finding the value of "S" was through finding another way to calculate a base. In this case, we knew a way of calculating "$e^{2\pi i}$", and so that let us calculate "S". If there were some other formula for calculating the base for circles divided up into other numbers of pieces (e.g. 360 pieces, 4 pieces, 112 pieces and so on), then that would have worked just as well. The number 535.4917 could be said to be equal to countless other exponentials, and it is a matter of choice that we say it is "$e^{2\pi}$" as opposed to anything else. [Although, on the other hand, "e" and "π" are probably the only number pair for which we already have names and symbols, and they have an advantage in that their values are usually programmed into calculators.]

It would be interesting to know if there are other ways of calculating "S" that do not require "e" and "π".

Now that we know S

From knowing "S", we can calculate the bases for our other circle formulas using the equation:
c = $\sqrt[x]{S}$.

Quarter-circle angle units

A circle that has been divided into 4 portions will have the base:
$\sqrt[4]{535.4917}$ = 4.810477380965 [to 12 decimal places].

Therefore, $4.8105^{i\theta}$ gives the position of a point on the unit-radius circle at an angle of "θ", where "θ" is an angle in quarter-circle angle units.

We can also say that "$4.8105^{i\theta}$ = cos θ + i sin θ", when Cosine and Sine are operating in the quarter-circle angle unit system.

Given that our provisional base for quarter-circle angle units was $\sqrt[i]{i}$, we now know that:
$\sqrt[i]{i}$ = 4.81047738

By finding "S", we have become able to solve what looked like a very complicated calculation. We can say that:
$\sqrt[i]{i}$ = 4.8105 = $\sqrt[4]{535.4917}$ = $\sqrt[4]{e^{2\pi}}$

Looking back at our first system of indicating points around a unit-radius circle using "i^{θ}", we can say that:
"i^{θ} = $4.8105^{i\theta}$

Whole-circle angle units

A circle that has been “divided” into 1 portion will have 535.4917 as its base.

Therefore, $535.4917^{i\theta}$ gives the position of a point on the unit-radius circle at an angle of “θ”, where “θ” is an angle in whole-circle angle units.

We can also say that $535.4917^{i\theta} = \cos\theta + i\sin\theta$, when Cosine and Sine are operating in whole-circle angle units.

We can also say that:
$\sqrt[0.25i]{i} = 535.4917 = e^{2\pi} = 4.8105^4$

Degrees

A circle that has been divided into 360 portions will have the base:
$\sqrt[360]{535.4917} = 1.017606491206$ [to 12 decimal places].

Therefore, $1.0176^{i\theta}$ gives the position of a point on a unit-radius circle at an angle of “θ”, where “θ” is an angle in degrees.

We can also say that $1.0176^{i\theta} = \cos\theta + i\sin\theta$, when Cosine and Sine are working in degrees.

We can also say that:
$\sqrt[90i]{i} = 1.0176$
... and:
$1.0176^{360} = 535.4917^1 = e^{2\pi} = 4.8105^4$

Radians

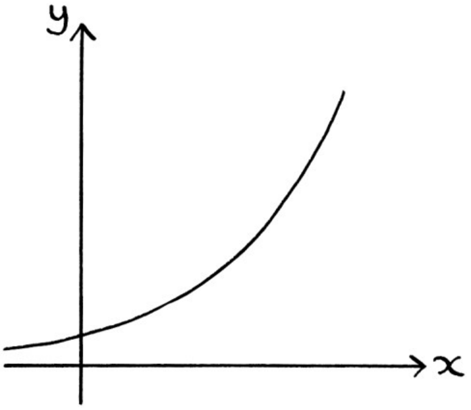
As we already know, a circle that has been divided into 2π portions will have as its base: $\sqrt[2\pi]{535.4917} = 2.718281828459$ or “e”.

Therefore, $e^{i\theta}$ gives the position of a point on the unit-radius circle at an angle of “θ”, where “θ” is an angle in radians.

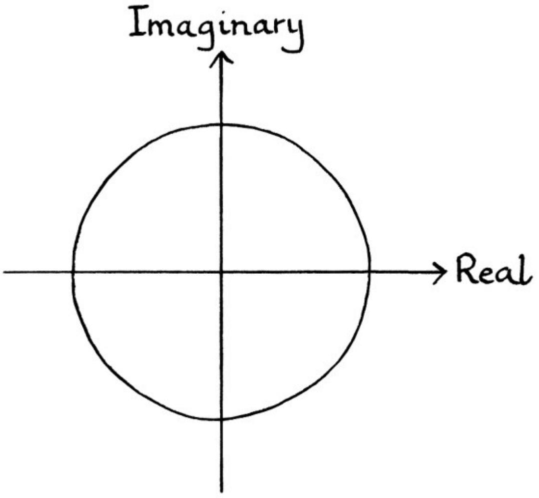
We can also say that $e^{i\theta} = \cos\theta + i\sin\theta$, when Cosine and Sine are working in radians. We can also say that $(\sqrt[0.5i\pi]{i}) = e$.

Increasing powers of "i"

From what we have learnt, it should now be clear that if we have a Real number to a power of "x", it will indicate a series of points on an exponential curve on x and y-axes:



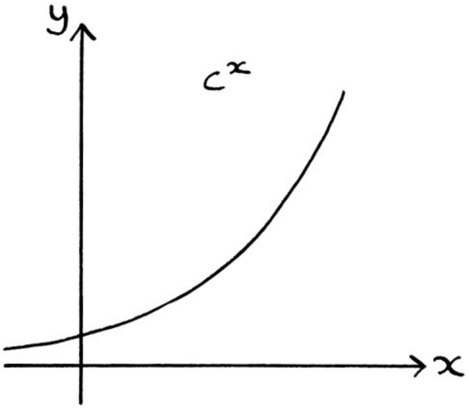
... and if we have a Real number to a power of "ix" (or in other words, iθ), it will indicate a series of points on a unit-radius circle on the Complex plane:



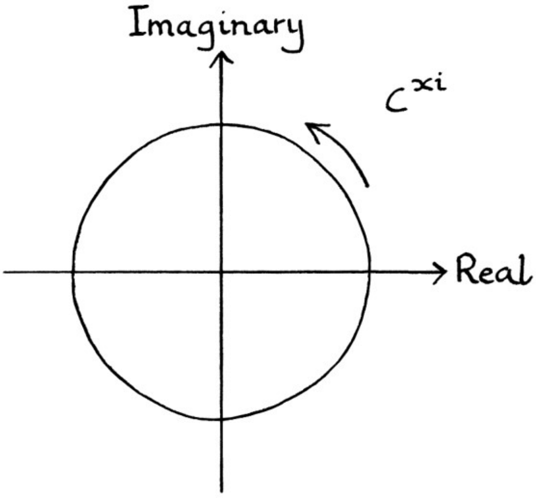
The 4 stage cycle of "i" powers

Where this idea becomes more interesting is when we repeatedly raise an existing exponential to the power of "i". As an example, we will use a generic formula: "c^x" where "c" is any Real number.

If we were to plot the points "$y = c^x$" over a range of "x", we would end up with a curve on x and y-axes as in this picture. As "x" increases, so does the result of "c^x".

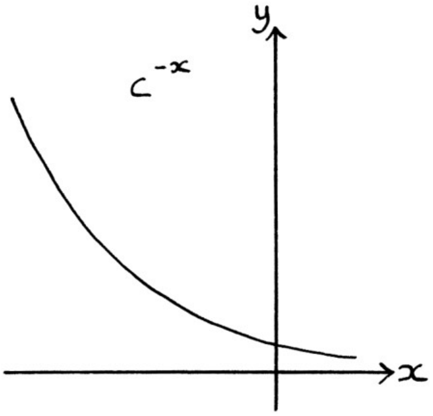


If we now raise this entire exponential to the power of "i", we will have: $(c^x)^i = c^{xi}$. In this formula, "x" is actually an angle in a system of angles related to the base "c". If we were to plot the points of "c^{xi}" for a range of "x", we would draw a unit-radius circle. As "x" increases, "c^{xi}" indicates a point on a unit-radius circle at ever-increasing angles, which is the same thing as saying that as "x" increases, "c^{xi}" indicates points that are further anticlockwise around the circle:

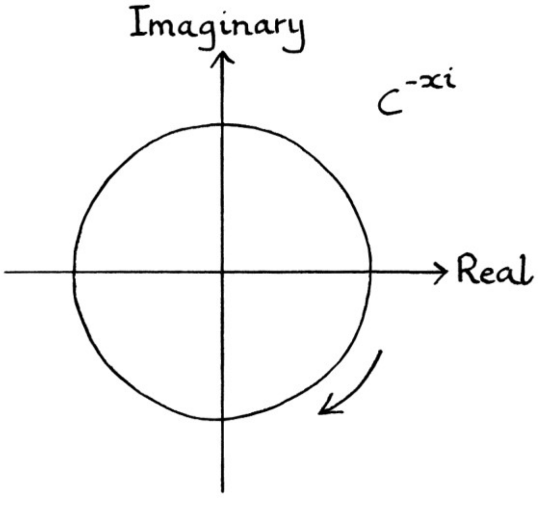


So far, we already knew all of this.

Now, if we raise the entire exponential "c^{xi}" to the power of "i", we will have $(c^{xi})^{i} = c^{xii} = c^{(x * -1)} = c^{-x}$. We have gone from a circle on the Complex plane, back to a curve on x and y-axes. What is more, this time the curve is "c^{-x}" which is a mirrored version of the exponential curve with which we started. As "x" increases, the result of "c^{-x}" decreases. The curve looks like this:

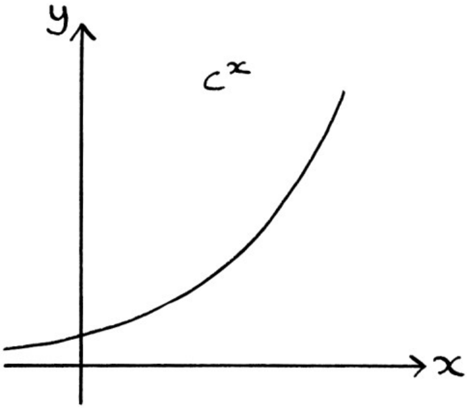


Now, if we raise the entire exponential, "c^{-x}" to the power of "i", we will have:
$(c^{-x})^{i} = c^{-xi}$. We have gone from the backwards exponential curve to a circle on the Complex plane again. However, this time, as "x" increases, "c^{-xi}" indicates a point on the unit-radius circle at ever-*decreasing* angles. In other words, as "x" increases, "c^{-xi}" indicates points that are further *clockwise* around the circle.

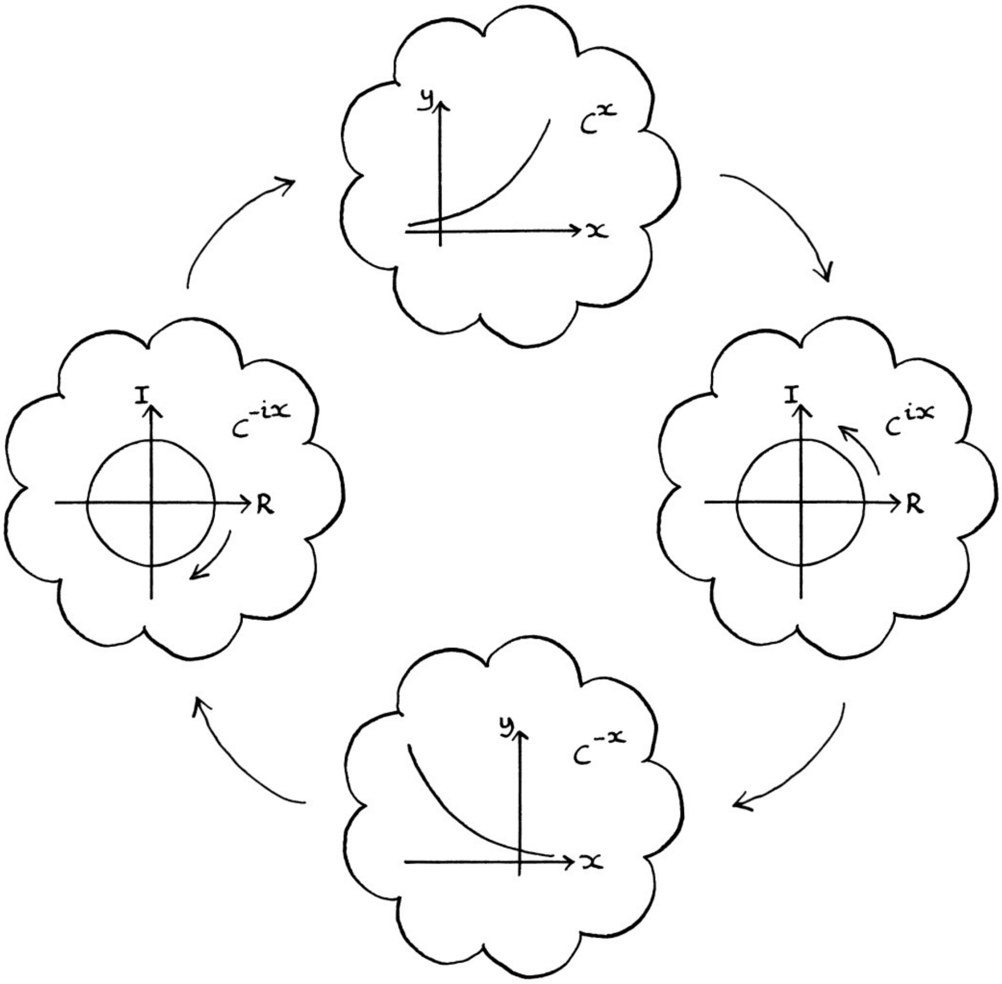


And, if we raise "c^{-xi}" to the power of "i", we have:
$(c^{-xi})^{i} = c^{-xii} = c^{(-x * -1)} = c^{x}$

We have ended up with the curve on x and y-axes that we started with:



To summarise this series of events, every time we raise something to the power of “i” we switch from a curve on x and y-axes to a circle on the Complex plane, or we switch from a circle on the Complex plane to a curve on x and y-axes. There is also a progression, or cycle, of four stages that goes: curve, circle, backwards curve, backwards circle. At whichever stage we start, raising the exponential to the power of “i” will move it on to the next stage:



An example with “e”

To show the cycle working in practice, we will look at what happens to “e^x”. [Note that this is “e^x” and not “e^{ix}”]

“e^x” is a curve on x and y-axes. When raised to the power of “i”, it becomes:
$(e^x)^i = e^{xi} = e^{ix}$
... which is a circle on the Complex plane.

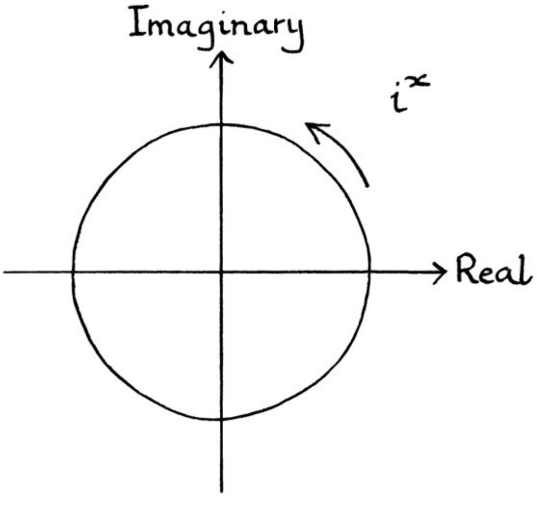
When that is raised to the power of “i”, we have:
$(e^{ix})^i = e^{xii} = e^{x * -1} = e^{-x}$
... which is a “backwards” curve on x and y-axes.

When that is raised to the power of “i”, we have:
$(e^{-x})^i = e^{-xi} = e^{-ix}$, which is a “backwards” circle on the Complex plane. The higher “x” is, the further *clockwise* around the circle, the result will be.

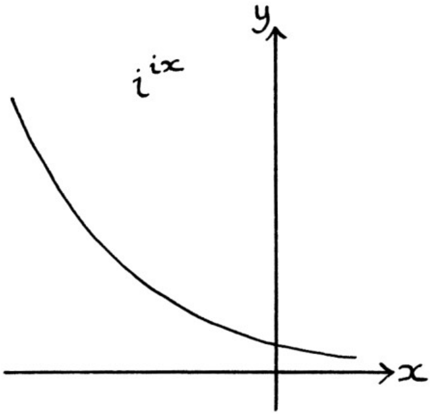
When that is raised to the power of “i”, we have:
$(e^{-ix})^i = e^{-ixi} = e^{-xii} = e^{(-x * -1)} = e^x$
... which is what we started with.

An example with “i”

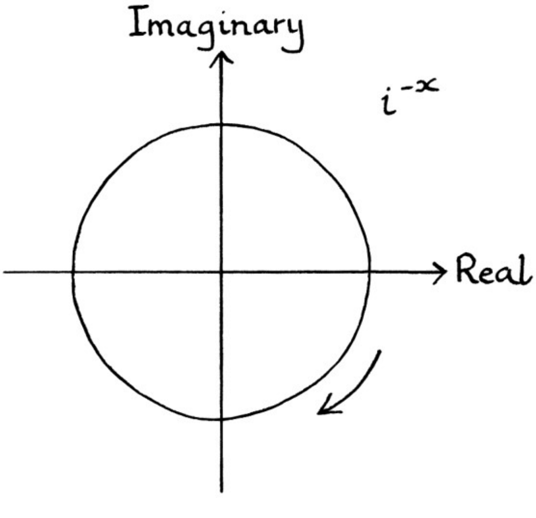
The formula “i^x” is already referring to a circle on the Complex plane. Therefore, it is at stage 2 in the cycle. The higher “x” is in “i^x”, the further anticlockwise around the circle the result will be.



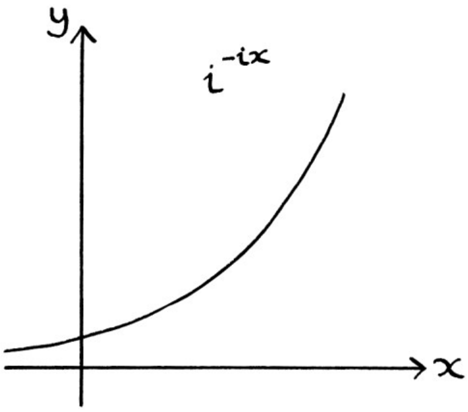
When "i^x" is raised to the power of "i", we have:
$(i^x)^i = i^{ix}$
... which refers to a backwards curve on x and y-axes. This means that for any value of "x", the result of "i^{ix}" will always be a Real number. The higher "x" is, the smaller the result will be.



When that is raised to the power of "i", we have:
$(i^{ix})^i = i^{iix} = i^{-x}$
... which refers to a backwards circle on the Complex plane. This means that for any value of "x", the result of "i^{-x}" will be a Complex number. The higher "x" is, the further *clockwise* around the circle the result will be.



When “i^{-x}” is raised to a power of “i”, we have:
$(i^{-x})^i = i^{-ix}$
... which refers to a forwards curve on x and y-axes. The higher “x” is, the higher the result will be.



When “i^{-x}” is raised to a power of “i”, we have:
$(i^{-ix})^i = i^{-iix} = i^x$
... which is what we started with, and refers to a forwards circle on the Complex plane.

“i” to the power of “i”

Where the four-stage cycle becomes useful is in visualising the solving of seemingly difficult calculations such as “i^i” or “ $\sqrt[i]{i}$ ”. There are already methods to calculate such things using the knowledge that “$e^{i\theta} = \cos\theta + i\sin\theta$” [in radians], but they take some thought. We can visualise what we are doing more easily using the knowledge of the four-stage cycle.

We know that “i^x” represents a circle on the Complex plane. Therefore, if we raised that to the power of “i”, we know that the resulting formula, “i^{ix}”, would be the next stage in the four-stage cycle. Therefore, it would be a backwards curve on x and y-axes. In that case, any results would always be Real numbers. In other words, for any Real value of “x”, “i^{ix}” would be a Real number. Therefore, if “x” is 1, we would have “i^{1i}”, which is “i^i”, and we would know that the solution must be a Real number.

Actually calculating the value of “i^i” is easy. We know that “i^x” divides a circle up into 4 pieces. Using what we learnt from earlier in this chapter, this means that it is identical to “4.8105^{ix}”. Therefore, “i^{ix}” is identical to “4.8105^{iix}”, which ends up as “4.8105^{-x}”. This confirms that “i^{ix}” is a backwards exponential curve on x and y-axes. The value “i^i” is equal to “i^{1i}”, which means it is the same as 4.8105^{-1}, which is 1 ÷ 4.8105 = 0.20787958 [to 8 decimal places, and calculated from the full value of 4.810477380965...]

Usually in explanations of calculating “i^i”, people are surprised that “i^i” should result in a Real number. This is because they do not realise that raising something to the power of “i” changes the nature of what the exponential is about. Previously, one might have tried to plot the result of “i^i” on the Complex plane – in other words, by marking its position at “0.2079 + 0i”, and in doing that, it seems very confusing that something should end up rotated and scaled to that point on the Real axis. However, as we now know, the value 0.2079 in this case does not belong on the Complex plane – it belongs on the backwards exponential curve on x and y-axes. It represents the point on the curve of “4.8105^{-x}” where “x” is 1.

ith Roots of i

An ith root of “i” such as “ $\sqrt[i]{i}$ ” seems completely incomprehensible, but it is actually reasonably easy to solve. [We actually found out what “ $\sqrt[i]{i}$ “ is when we found the base for a circle with 4 divisions earlier in this chapter, but we did not really *solve* it – instead we just knew what its equivalence was].

We will solve a generic root: the ixth root of i: $\sqrt[ix]{i}$. Using our knowledge of exponentials, we can rephrase this as:
“$i^{(1 \div ix)}$”
... which we will write as:
“$i^{(1 / ix)}$”.

This is now “i” raised to a multiple of “i”. From our knowledge of the four-stage cycle, we know that “i” raised to a multiple of “i” will have a result that is on the backwards exponential curve on the x and y-axes. Therefore, the result will be a Real number.

We know that “i^x” is equivalent to “4.8105^{ix}”.

Therefore, “$i^{(1/ ix)}$” will be $4.8105^{(i * 1/ix)} = 4.8105^{(i/ix)} = 4.8105^{(1/x).}$

Therefore, $\sqrt[i]{i} = \sqrt[1i]{i} = i^{(1/1i)} = 4.8105^{(1/1)} = 4.8105$.

As another example, if we wanted to calculate “ $\sqrt[3i]{i}$ ”, it would be equivalent to “$i^{(1/3i)}$” = $4.8105^{(1/3)}$ = 1.68809179 [to 8 decimal places, and calculated from the full value of 4.810477380965...]

Exponentials with negative numbers

In Chapter 26, we saw how the results of powers of negative numbers were Complex numbers. The formula for solving these was:
$(-a)^x = a^x \cos(180x) + i * a^x \sin(180x)$
... when Cosine and Sine are working in degrees, or:
$(-a)^x = a^x \cos(\pi x) + i * a^x \sin(\pi x)$”
... when Cosine and Sine are working in radians

The formula $(-a)^x$ is easiest to understand if we know that it is the same as:
$a^x * (-1)^x$
... or:
$a^x * i^{2x}$

Imaginary powers of “e”

Given that “$(-1)^x$” identifies points around a unit radius circle, we can portray it in terms of Imaginary powers of “e”. The exponential $(-1)^x$ works in half-circle angle units; the exponential e^{ix} works in radians. Therefore, we convert half-circle angle units into radians, which we do by multiplying the exponent of “e” by π. [To convert any half-circle angle unit into radians, we find out the portion of a circle that that angle represents by dividing by 2, and then we multiply that by the number of radians in a circle (2π)]. We end up with:

$(-1)^x = e^{\pi x i}$

We can test this with $(-1)^{1.24}$, which should be the same as “$e^{1.24\pi i}$”, which should be “−0.7290 – 0.6845i”. With any of the countless methods we could use to solve “$e^{1.24\pi i}$”, we will end up with “−0.7290 – 0.6845i”.

Given that $(-2)^x$ indicates a point on a circle, but at ever-increasing radiuses, we can also put the formula into Imaginary powers of “e”. It will be:
$(-2)^x = 2^x * e^{\pi x i}$

We can also give a general formula for negative bases:
$(-a)^x = a^x * e^{\pi x i}$
... where “a” is the base and “x” is the exponent.

Raised to the power of “i”

Given that a negative base raised to a Real exponent results in a Complex number on the Complex plane, if we raise a negative-base exponential to the power of “i”, it will shift the result into the next stage of the four stage cycle. Therefore, the result will be on the backwards exponential curve, and be a Real number.

We can test this with an example. We will solve:
$(-1)^{2.2i}$

This is the same as:
$i^{(2 * 2.2i)}$
... which is:
$i^{4.4i}$

[Note that this is not the same as “$i^{0.4i}$”. This result will not be on a circle, but on a backwards exponential curve. On a circle, “$i^{4.4}$” is the same as “$i^{0.4}$”, but this has “i” in the exponent, so it is a different concept.]

There are two ways we can go from here. Earlier in this chapter, when we were converting Real exponents of “i” to Imaginary exponents of “e”, we learnt that “i^x” is equivalent to “4.81047738^{ix}”. Therefore, “$i^{4.4i}$” will be equivalent to:
“$4.8105^{4.4ii}$”
... which is:
“$4.8105^{-4.4}$”
... which is:
0.0009963.

We could also have used our knowledge from earlier in this chapter that “i^i” is 0.20787958 (to 8 decimal places). First, we turn “$i^{4.4i}$” into an exponential as the base of a second exponential:
$(i^i)^{4.4}$
... then because “i^i” is 0.2079, we can phrase this as:
$0.2079^{4.4}$
... which is also:
0.0009963.

Seemingly complicated maths

Now that we know more about Imaginary numbers, we can solve calculations that would have seemed incomprehensible earlier.

We have several formulas and ideas that can help us.

From Chapter 23, we know that:

- A multiplication by "i" rotates a point by +90 degrees.
- A division by "i" rotates a point by −90 degrees.
- A multiplication by "−i" rotates a point by −90 degrees.
- A division by "−i" rotates a point by +90 degrees.

From those, and our knowledge of powers of "i", we can say that:

- A multiplication by "i^{θ}" rotates a point by "+θ" quarter-circle angle units.
- A division by ""i^{θ}" rotates a point by "−θ" quarter-circle angle units.
- A multiplication by "$-i^{\theta}$" rotates a point by "−θ" quarter-circle angle units.
- A division by "$-i^{\theta}$" rotates a point by "+θ" quarter-circle angle units.

As we know that Imaginary powers of "e" work in a similar way to powers of "i", we can say that:

- A multiplication by "$e^{i\theta}$" rotates a point by "+θ" radians.
- A division by "$e^{i\theta}$" rotates a point by "−θ" radians.
- A multiplication by "$-e^{i\theta}$" rotates a point by "−θ" radians.
- A division by "$-e^{i\theta}$" rotates a point by "+θ" radians.

Related to those, we have:

- A multiplication by "$e^{-i\theta}$" rotates a point by "−θ" radians.
- A division by "$e^{-i\theta}$" rotates a point by "+θ" radians.
- A multiplication by "$-e^{-i\theta}$" rotates a point by "+θ" radians.
- A division by "$-e^{-i\theta}$" rotates a point by "−θ" radians.

As we know other bases that are alternatives to using "e", we can also say the same things for:
"$4.8105^{i\theta}$" where "θ" is an angle in quarter-circle angle units.
"$535.4917^{i\theta}$" where "θ" is an angle in whole-circle angle units.
"$1.0176^{i\theta}$" where "θ" is an angle in degrees.

As we know those bases in terms of roots of “i” and −1, we can also say the same things for each of these:
“$(\sqrt[i]{i})^{i\theta}$” where “θ” is an angle in quarter-circle angle units.
“$(\sqrt[0.25i]{i})^{i\theta}$” where “θ” is an angle in whole-circle angle units.
“$(\sqrt[90i]{i})^{i\theta}$” where “θ” is an angle in degrees.
“$(\sqrt[0.5i\pi]{i})^{i\theta}$” where “θ” is an angle in radians.

From all of these, we can instantly know the results of some calculations, such as how:

$$\frac{0 + 1i}{(\sqrt[90i]{i})^{45i}}$$

... results in the point at “0 + 1i” being rotated by −45 degrees, as does:

$(\sqrt[90i]{i})^{-45i} * (0 + 1i)$

... as does this:

$-i * (\sqrt[90i]{i})^{45i} * (0 + 1i)$

The result of all three is “0.7071 + 0.7071i”, because that is “0 + 1i” rotated by −45 degrees. [Remember that most people would write “0 + 1i” as just “i”. I am keeping Complex numbers written in full to make things clearer.]

We also know that an Imaginary power of “e” with a positive “i” rotates a point by a number of radians anticlockwise. Given that, whenever we have such an “e” exponential multiplied by “i” (such as “$i * e^{i\theta}$”), we can remove the multiplication by “i” and add 1 quarter-circle angle unit (0.5π radians) on to the *angle* in the exponent, and it will still refer to the same place [*but only if the exponent had a positive “i” in it.*] We can say that:

- $i * e^{0.6\pi i}$, which would best be written as: $i * e^{i(0.6\pi)}$
- $e^{1.1\pi i}$, which would best be written as: $e^{i(1.1\pi)}$

... are the same. They both refer to the point “−0.9511 – 0.3090i”.

[Note how I emphasise that the addition is to *the angle*, and not to the exponent as a whole. To do this alteration correctly, we must remember that the *angle* given in an exponent is the non-Imaginary part. Ideally, we would think of such an exponential as "$e^{i(...)}$", with the dots representing the angle in radians. Not thinking of this distinction will cause you to make mistakes. To restate this, if the exponent is, say, "1.2πi", then the angle is "1.2π" and not "1.2πi". If we take 0.5π radians off *the angle*, we end up with "0.7π" *as the angle*, and "0.7πi" *as the exponent*. That is why it is better to separate the "i" from the rest of the exponent as in "i(1.2π)". Often in such situations, the angle will be tangled up with the "i", so they need to be untangled before any additions or subtractions.]

If we have an Imaginary power of "e" with a negative "i" in the exponent ("−i"), then it rotates a point by a number of radians *clockwise*. Given that, whenever we have such an "e" exponential multiplied by "i" (such as "$i * e^{-i\theta}$"), we can remove the multiplication by "i" and *subtract* 1 quarter-circle angle unit (0.5π radians) from the *angle* in the exponent, and it will still refer to the same place. We can say that:

- $i * e^{-0.6\pi i}$, which would best be written as: $i * e^{-i(0.6\pi)}$
- $e^{-0.1\pi i}$, which would best be written as: $e^{-i(0.1\pi)}$

... are the same. They both refer to the point "0.9511 – 0.3090i".

We know that a division by "i" is the same as a multiplication by "−i", which both result in a rotation of −90 degrees. The rule, given mathematically, is as follows:

$$\frac{a}{i} = -ia$$

... where "a" is any number, whether Real or Complex.

This means that we can know that "$e^{2i} \div i$" is the same as "$-i * e^{2i}$".

A multiplication by "−i" rotates a point by −90 degrees (−0.5π radians). Therefore, in the above result ("$-i * e^{2i}$"), we can remove the multiplication by "−i", and subtract 0.5π radians from the angle in the exponent and we will have the same result:

$e^{i(2 - 0.5\pi)}$

Therefore, we can say that:

"$e^{2i} \div i$"

"$-i * e^{2i}$"

"$e^{2i - 0.5\pi i}$"

... are all the same.

Conclusion

This chapter was intended to increase your understanding of Complex numbers. You might not need to know much of this chapter, but knowing more than you need will make maths easier in general.

www.timwarriner.com

Chapter 30: Calculus

Calculus

In this chapter, I will give a very basic introduction to calculus. The explanation will be sufficient for understanding the calculus used in this book, but it will probably not be enough for you to pass maths exams.

Calculus is the overall name for two different processes: "differentiation" and "integration". These terms sound complicated, but they refer to reasonably simple ideas. Defined most simply, "differentiation" is the process that finds the gradient of a curve or line, while "integration" is the process that that finds the curve or line that has a particular gradient.

We can express these definitions slightly more thoroughly: if we are given a curve or a line, then:

- Differentiation finds the curve or line for which every y-axis value is the gradient of the corresponding place on our curve or line.

- Integration finds the curve or line for which every gradient is equal to the y-axis value of the corresponding place on our curve or line.

We could also say:

- Differentiation finds the *formula* that describes all the gradients of a particular curve or line.

- Integration finds the *formula* of a curve or a line for which all its y-axis values are the gradients of a given formula.

The process of integration is essentially the opposite of the process of differentiation. The definitions will become clearer as this chapter progresses.

Calculus is often thought of as being difficult, and I think this is for the following reasons:

- It is a new way of thinking about ideas compared with the maths that precedes it in a typical education. The maths leading up to calculus can be fairly intuitive, while calculus itself needs more thought. If you have understood most of this book so far, then you will understand at least the basics of calculus.

- The notation is unnecessarily obtuse. Most lessons in calculus start with the notation, which makes learning it much harder.

- It is usually taught for the sake of knowing it, and not for any particular reason. It is harder to learn something if you do not have a reason to learn it. If you do not need to know any calculus, it can seem a dull subject.

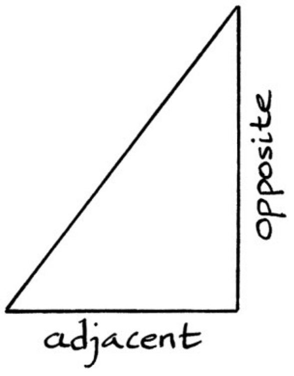
- It uses unhelpful vocabulary for naming concepts.

- As you learn more about the subject, you become confronted with examples that are more and more obscure.

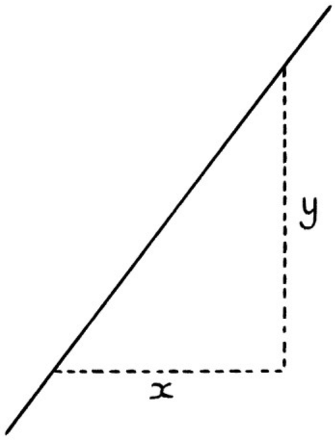
Basic calculus is not difficult. In this book, we will only need to know the most basic aspects of calculus.

Gradients

As explained in Chapter 2, a gradient is a measure of the steepness of a line. If we have a right-angled triangle, the gradient of the hypotenuse is the opposite side divided by the adjacent side:

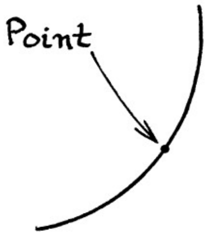


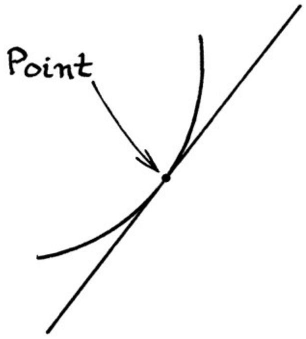
If we have a straight line, the gradient between any two points is the change in the y-axis values divided by the change in the x-axis values. [This is the same as saying the rise divided by the tread].

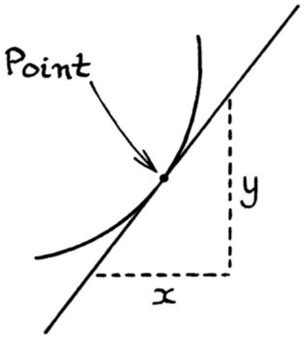


Such an idea is identical to imagining a right-angled triangle placed next to the line and then dividing the opposite side by the adjacent side.

When it comes to curved lines, the gradient of a particular point can be calculated by placing a straight line tangentially against that point [by which I mean touching that point], and then calculating the gradient of the straight line.



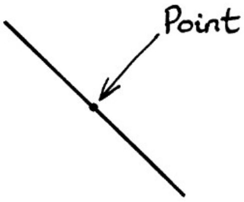




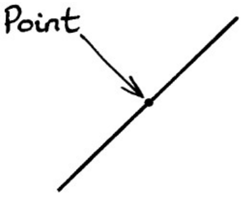
The difficult part in doing this is making sure that the straight line is at the correct angle to the curve. This can really only be achieved by making the straight line cross the curve at points either side and equidistant of the point of interest. The more we zoom into the curve, the more accurately this can be achieved, but ultimately, there is no limit to how far we can zoom into a curve. If the straight line touches two points either side of the point of interest, its gradient is really the average gradient between those two points.

Things become conceptually more complicated when the straight line touches just one point. If we look at such a straight line, we know its gradient. If we ignore the straight line, and look at the point itself, we are really finding the gradient of a single infinitely small point. There is no change in y-axis or x-axis at *one* point. However, we can still say that this point has a gradient *in the context of the points either side of it*. If we took the point out of its context, and removed it from the rest of the curve, the point itself could not be said to have a gradient. That point could be part of any curve or line, and without knowing more about the curve or line, we cannot say it has a gradient.

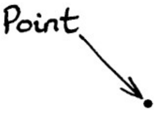
For example, the marked point in this line has a particular gradient:



The same point as part of a different line has a different gradient:



The point on its own cannot have a gradient:



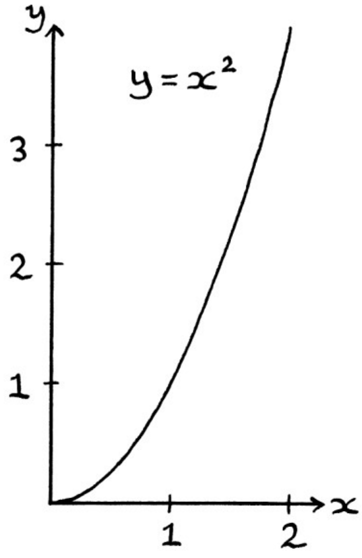
A point that sits on its own cannot be said to have a gradient. We need to know how a line approaches or leaves that point to know what the gradient would be at that point. However, while the point is part of the line, we can say that it is possible to calculate the gradient of a single point. This is what we might call the "instantaneous gradient". Strictly speaking, we cannot calculate the gradient of a single point by using the change in the y-axis divided by the change in the x-axis. However, we can calculate it by using calculus.

The gradient of a line shows the rate of change of that line. If we had a graph showing the distance in metres travelled by a vehicle over time, then the gradient calculated over any one second would show the number of metres travelled per second. It would show the "rate of travel", which we would normally call the "speed". If we had a graph showing the number of litres of rainwater collected in a container over time, then the gradient calculated over any one second would show the rate that the volume of water increased in litres per second. That gradients show the rate of change is what makes them a useful concept in some aspects of real life.

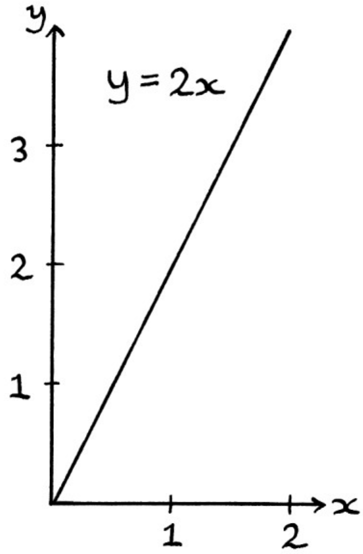
Differentiation

If we have the formula for a line or curve on a graph, it is often possible to calculate a second formula that shows the gradient of every single point on that line or curve. For example, if we have the formula "$y = x^2$", then the gradient of every single point along the curve can be calculated with the formula "$y = 2x$".

This is the graph of "$y = x^2$":



This is the graph of “y = 2x”:



We can calculate the gradient of “y = x^2” when “x” is 1 by looking at the “y = 2x” graph when “x” is 1 there, or by putting 1 into the “y = 2x” formula. The gradient is: 2 * 1 = 2.

When “x” is 5, the gradient of the “y = x^2” curve is 2 * 5 = 10.

When “x” is 0.5, the gradient of the “y = x^2” curve is 2 * 0.5 = 1.

The formula “y = 2x” gives the gradients of every point on the “y = x^2” graph. This means that for any point on the “y = x^2” graph, we can double the x-axis value of that point and the result will be the gradient. This allows us to find the *exact* gradient of a *single* point on our “y = x^2” curve. We are no longer finding the average gradient around a particular point, but the exact gradient of a single point.

Terminology

The formula “y = 2x” is called the “derivative” of “y = x^2”. In everyday English, the word “derivative” means “that which has been derived”. In maths, this definition is more specific, and means “the formula that shows the gradients of the original formula.” We could also refer to the derivative as “the gradient formula”. The process of calculating the derivative of a formula is called “differentiation”. In everyday English, one sense of the verb “to differentiate” is “to make different”. In maths, “to differentiate” means to calculate the derivative formula. Given how language changes, there was probably a time when the mathematical meaning of “differentiate” referred more to how a formula was altered to become a second formula than to how it was altered to become specifically a derivative formula. Similarly, there was probably a time when the word "derivative" in maths referred more to how one formula had been created from another, than to how specifically the gradient formula had been created from another formula. As it is, the current mathematical meanings of "differentiation" and "derivative" are much more specific than the actual English words would suggest. The words are not particularly good ones – they are long and in no way descriptive. They are one of the reasons that calculus can be confusing. Better terms than "derivative" and "differentiation" would be "gradient formula" and "finding the gradient formula", but, as with so much in maths, no one likes to change anything.

As an example of the terms in use, we might say something such as, "Let's use differentiation to calculate the derivative."

The term "calculus" refers to the whole subject of differentiation and derivatives, and also to integration, which we will look at later in this chapter. We might say, "We use calculus, specifically differentiation, to find the derivative."

Examples of derivatives

As we have seen, the derivative of "$y = x^2$" is "$y = 2x$". Any point on the graph of "$y = x^2$" has a gradient that can be calculated with the formula "$y = 2x$". We can also say that the graph of "$y = 2x$" shows the gradients of every possible x-axis value of the graph of "$y = x^2$".

The derivative of "$y = x^3$" is "$y = 3x^2$". The formula "$y = 3x^2$" can be used to find the gradient of any point on the curve "$y = x^3$". For example, the point on the curve at $x = 1.5$, has the gradient of $3 * (1.5^2) = 6.75$.

The derivative of "$y = x^4$" is "$y = 4x^3$". The formula "$y = 4x^3$" can be used to find the gradient of any point on the curve "$y = x^4$".

The derivative of "$y = x^5$" is "$y = 5x^4$".
The derivative of "$y = x^6$" is "$y = 6x^5$".
The derivative of "$y = x^7$" is "$y = 7x^6$".

There is a pattern in all of these derivatives. If we have "x" raised to a power, then the derivative will be that power multiplied by "x raised to that-power-minus-1".

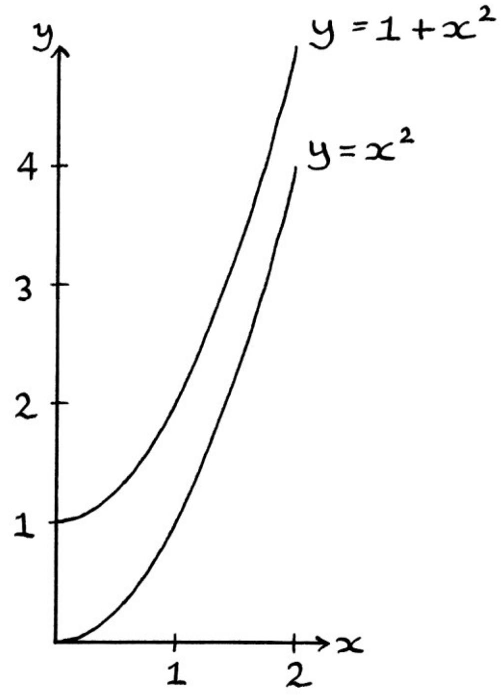
In other words if we have:
“$y = x^{b}$”
... then the derivative will be:
“$y = b * x^{(b-1)}$”

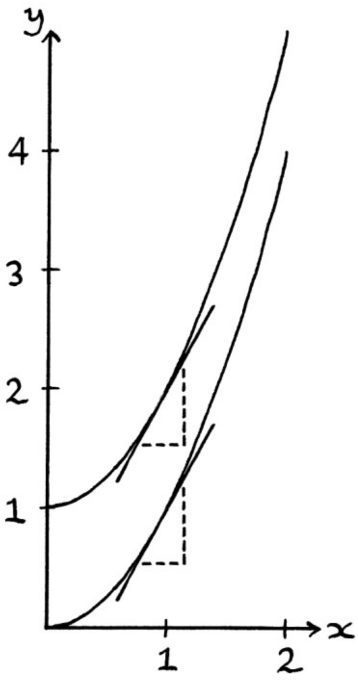
Lines shifted up or down the y-axis

One useful thing to know about gradients is that they relate to the steepness of a line, which means that it does not matter how high or low that line is on the y-axis.

The following two graphs each show:
“$y = x^2$”
... next to the curve of:
“$y = 1 + x^2$”

Although these are different formulas, the curves have identical shapes. The gradients of each curve at corresponding x-axis points are identical:





A formula that is any fixed number added to “x^2” will still have the same gradients as “x^2”. Therefore, all of the following formulas will have the same gradients in the same places:

“$y = x^2$”
“$y = 1 + x^2$”
“$y = 2 + x^2$”
“$y = 3 + x^2$”
“$y = -1 + x^2$”
“$y = 1.6588 + x^2$”
“$y = -1{,}000{,}004 + x^2$”
“$y = 0.000000001 + x^2$”

This means that the derivatives of each of these formulas will be the same. The derivative of each of these formulas is “$y = 2x$”. The position of the curve up or down the y-axis is irrelevant to the gradient. How much we consistently add to every value of the curve is irrelevant to the gradient. We can adjust our rule to take this into account. If we have the formula:
“$y = C + x^b$”
... then the derivative will be:
“$y = b * x^{(b-1)}$”
[We ignore the “C”.]

Scaled formulas

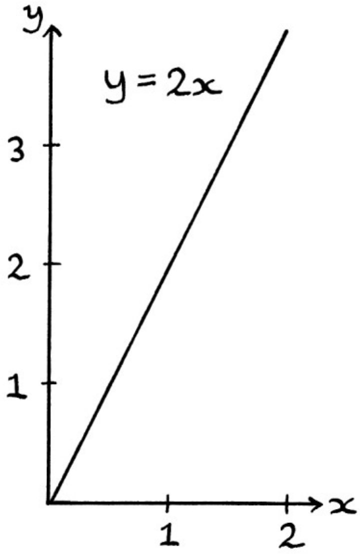
The derivative of:
“$y = 3x^2$”
... which, to clarify, means:
“$y = 3 * (x^2)$”
... is:
“$y = 3 * (2x)$”
... which is:
“$y = 6x$”.

This is, in essence, a similar rule to the first one that we saw. If we have this formula:
“$y = C + ax^b$”
... then its derivative will be:
“$y = (b * a) * x^{(b-1)}$”

Simple formulas

If we have the formula:
“y = 2x”
... then the derivative formula is:
“y = 2”

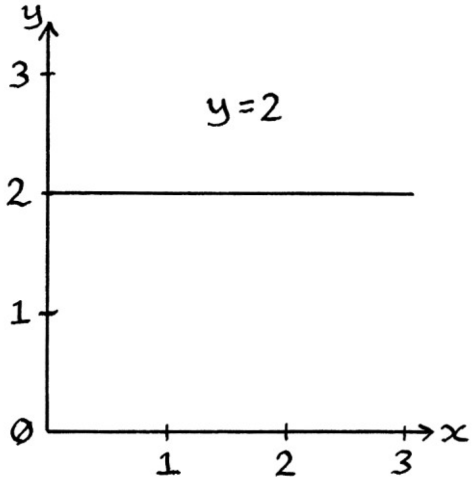
We can know this intuitively because “y = 2x” is a straight line with a constant gradient of 2:



As the derivative formula shows the gradient of this line for all values of “x”, it must be “y = 2”. Whatever the value of “x”, the gradient is 2.

We can also think of:
“y = 2x”
... as being:
“y = 2x^1”
... so the derivative formula is:
“y = (1 * 2) * x^0
... and because any value to the power of 0 is 1, this gives us:
“y = (1 * 2) * 1
... which is:
“y = 2”

The graph of the derivative formula looks like this:



More complicated formulas

If we have a list of items being added together, we find the derivatives of each item in turn, and then we add the results together.

For example, the derivative of:
"$y = 11x^4 + x^3 + x^2 + 5x + 7$"
... is:
"$y = 44x^3 + 3x^2 + 2x + 5$"

We just go through the original formula piece by piece:

- The "$11x^4$" part becomes "$44x^3$" because we multiply the 11 by 4, and then reduce the exponent by 1.

- The "x^3" part becomes "$3x^2$".

- The "x^2" part becomes "2x".

- The "5x" part becomes "5".

- We ignore the "7" part, because it acts solely to raise the whole curve up by 7 units, and so has no effect on the gradient of the whole curve.

If we have a list of items consisting of subtractions, or additions and subtractions, we do the same thing – we find the derivative of each part in turn. For example, the derivative of:
“$y = -7x^2 + 5x^3 - 2x^8$”
... is:
“$y = -14x + 15x^2 - 16x^7$”

It is very useful to remember that a series of additions or subtractions can be split up into its pieces, with each piece converted individually. This can simplify many calculations.

Powers of “e”

There are certain functions that behave in special ways when it comes to finding their derivatives. The ones that are relevant to this book are Sine and Cosine, which we will look at later in this chapter. One function that stands out among the others is “$y = e^x$”. The derivative of the formula:
“$y = e^x$”
... is:
“$y = e^x$”

In other words, the formula and its derivative are the same as each other. This is because every point on the curve of “$y = e^x$” is at a y-axis value that is also its own gradient. [Note that the previous formulas in this chapter were generally “x” as a base, while this has “x” as an exponent.]

We could calculate “e” in a very long-winded way by searching for the exponential curve that is its own gradient. We would start with something such as “$y = 2^x$”, and then keep trying values slightly higher and lower until we achieved the accuracy that we wanted.

To reinforce a basic fact about derivatives, the derivative of the formula:
“$y = 1 + e^x$”
... is also:
“$y = e^x$”.

This should be expected as the derivative shows the gradients, and the gradients are unaffected by the actual y-axis values of the curve. The number 1 just raises the whole of the “e^x” curve upwards by 1 unit, and does not affect the gradients.

Differentiation in general

The reasons that differentiation (finding the derivative formula) works as a process are not particularly hard to understand, but would take too long to explain in this chapter.

When you learn differentiation at school or from books, you will spend most of the time looking at more complicated formulas than the exponentials seen here. Such formulas might involve multiplications, divisions and logarithms for example. Typical teaching of differentiation includes how to find the various maximums and minimums of curves among other things. These are all aspects of calculus that are easier to learn if you have a need, and therefore a desire, to know about them. They can be very useful for some real-life matters, but are not relevant to what we are learning in this book.

Symbols

One of the confusing aspects of basic differentiation is the symbol used to express that a formula is a derivative formula. Previously, in this chapter, I have explained derivatives by saying, “The derivative of ... is ...” In maths, such a statement is usually made in a different way. For example, if we have the function:

“y = x^2”

... then it would be common to say:

$$\frac{dy}{dx} = 2x$$

In its most basic interpretation, the division $\frac{dy}{dx}$ means “the derivative of the formula we are discussing”. In this way, the two letters “dy” divided by the two letters “dx” can be thought of as just a symbol meaning “the derivative”. This interpretation is fine for very basic understanding, but it is better to understand the idea more fully.

Outside of calculus, when we calculate the gradient of a line over any particular region, we take the change in the y-axis and divide it by the change in the x-axis. If we say that our first y-axis reading is y_1, our second y-axis reading is y_2, our first x-axis reading is x_1, and our second x-axis reading is x_2, then we could give the formula to calculate a gradient as:

$$\frac{y_2 - y_1}{x_2 - x_1}$$

If we are calculating the gradient of a curve, it pays to make the readings as close together as possible to achieve the most accuracy. Therefore, we could rephrase the above formula in a slightly un-mathematical way as:

$$\frac{a\ tiny\ change\ in\ y}{a\ tiny\ change\ in\ x}$$

... or:

$$\frac{a\ tiny\ difference\ in\ y}{a\ tiny\ difference\ in\ x}$$

We will now look at the derivative “symbol” again. It is:

$$\frac{dy}{dx}$$

This is really a gradient calculation. Instead of saying “a tiny difference in y”, it says “dy”. Instead of saying “a tiny difference in x”, it says, “dx”. We can think of each letter “d” as being shorthand for “a tiny difference in”. Instead of referring to the gradient at one particular point of the curve, this is referring to the whole formula of the curve. Given that we are dealing with a formula that gives “y” in terms of calculations with “x”, the division $\frac{dy}{dx}$ is really shorthand for “all the gradients of the formula that was just mentioned involving “y” and “x” can be expressed using...”

The $\frac{dy}{dx}$ division becomes altered if we are using letters other than “y” and “x”. For example, if we had a formula that involved time such as:

“y = t^2”

... then we would say that:

$$\frac{dy}{dt} = 2t$$

This ultimately means “the derivative of the formula just mentioned, which involves ‘y’ and ‘t’, is 2t”. We could also think of it as saying “the tiniest change in ‘y’ divided by the tiniest change in ‘t’ for the formula as a whole is 2t.”

If we had the formula:

“a = t^2”

... then we might say:

$$\frac{da}{dt} = 2t$$

In everyday differentiation, the $\frac{dy}{dx}$ division is treated as if it were a static symbol, although the letters ‘y’ and ‘x’ become replaced if appropriate. If you ever need to say the division out loud, it is pronounced as “dee why, dee exe” if the letters are “dy” and “dx”.

Another form of the $\frac{dy}{dx}$ division joins it with the original formula as so:

$$\frac{d\,x^2}{dx} = 2x$$

This is saying that the derivative of “x^2” is 2x, without referencing “y”.

The $\frac{dy}{dx}$ division is one of the worst “symbols” in maths. It looks complicated, its meaning is unintuitive, and it is incompatible with modern typed text. An alternative is to use an apostrophe as in the following example.

If we have the formula:
“$y = x^2$”
... then:
“$y' = 2x$”

In this case, y' means the derivative of the function involving y. This notation can also be used to show the derivative of a derivative:

If we have the formula:
$y = x^3$
... then:
$y' = 3x^2$
... and:
$y'' = 6x$
... and:
$y''' = 6$

The derivative of a derivative is called “the second derivative”. The derivative of a second derivative is called “the third derivative” and so on. The apostrophe is easier to use than $\frac{dy}{dx}$ but unlike $\frac{dy}{dx}$ it might need clarification that it refers to derivatives, and not some other process. Its use can be confusing if apostrophes and inverted commas are being used non-mathematically in surrounding text.

Outside of a maths lesson or something involving a lot of calculus, it is often easier just to write the words “the derivative of ... is ...”, than to use symbols. It is easier to read, and it is less confusing to people who have forgotten or never knew calculus. Someone can always search for what “derivative” means, but it is much harder to search for what $\frac{dy}{dx}$ means.

Integration

Integration is the process that finds the “integral” of a formula. An integral is the opposite of a derivative. In other words, an integral formula is the opposite of a derivative formula. If we start with a given formula, then, as we know, the derivative formula shows the gradients at every point of the curve of that formula. The integral formula, on the other hand, shows the curve for which the given formula has the gradients. In other words, if formula B is the derivative formula of formula A, then formula A is the integral formula of formula B.

We know that the derivative of:
“$y = x^2$”
... is:
“$y = 2x$”.

The integral of:
“$y = 2x$”
... is:
“$y = x^2 + C$”.
[I will explain what “C” is in a moment.]

If we find the derivative of:
“$y = x^2 + C$”
... we end up with:
“$y = 2x$”.

You might think that the integral should be “$y = x^2$”, but that is just *one* of the possible formulas for which “$y = 2x$” is the derivative. There are countless formulas for which “$y = 2x$” is the derivative. These include:

“$y = x^2 + 1$”
“$y = x^2 + 2$”
“$y = x^2 + 3$”
“$y = x^2 + 4$”
“$y = x^2 - 1.000001$”
“$y = x^2 + 0.000000001$”
“$y = x^2 + 7{,}999{,}998$”
... and so on.

All of the above formulas have the same curve, but are raised up or down the y-axis by the addition of a number. As gradients relate to the rate of change of a curve, they are independent of how high or low the curve is on the y-axis. Therefore, there are really an infinite number of formulas that have the same gradients along their curves, and therefore, there are really an infinite number of formulas that have the same derivative. Consequently, there are really an infinite number of formulas that are the integral of a particular formula. Instead of writing out an infinite number of integrals as the result, we acknowledge their existence by putting in the addition of the letter "C". The letter "C", in this case, stands for "constant", and is a symbol that means "any possible number at all". Therefore:
"$y = x^2 + C$"
... really means "x^2" moved up or down the y-axis by any amount.

It is important to remember to add the "C" to the integral formula. Otherwise, the resulting formula would suggest that there were only one result. Normally, in maths books, the formula would be written as "$y = x^2 + C$", with the "+ C" afterwards, but we could instead add the "C" beforehand to be consistent with how we add mean level on to a wave: "$y = C + x^2$". Sometimes, it is clearer with the "C" beforehand. The addition of "C" has exactly the same effect as the mean level in a wave formula, in that it moves the curve or line upwards or downwards. However, the "C" cannot be called "mean level" here because, unless we were specifically finding the integral of a pure wave, it would not be the average level of the curve.

We know that the derivative of:
"$y = x^3$"
... is:
"$y = 3x^2$"

The integral of:
"$y = 3x^2$"
... is:
"$y = x^3 + C$"

Again, there are an infinite number of formulas for which "$y = 3x^2$" is the derivative, such as:
"$y = x^3$"
"$y = x^3 + 11.45$"
"$y = x^3 + 700$"
"$y = x^3 + 4$"
"$y = x^3 – 22.9$"
... and so on.

Therefore, there is not just one integral. We have to acknowledge this fact by including the addition of "C" in the result. If we do not, then we are suggesting that there is only one result. By including the "C", it also helps us to visualise the situation better, and gives us a more thorough understanding of what we are doing.

If we are given the formula:
"$y = x^3 + C$"
... which contains the symbol "C" to mean "any value", then the derivative formula will be:
"$y = 3x^2$"

This should be expected. The "C" becomes ignored because the height of a curve up or down the y-axis is not relevant to the gradients of a curve.

The integral of:
"$y = 4x^3$"
... is:
"$y = x^4 + C$"

The integral of:
"$y = 5x^4$".
... is:
"$y = x^5 + C$"

The integral of:
"$y = 6x^5$".
... is:
"$y = x^6 + C$"

The integral of:
"$y = 7x^6$".
... is:
"$y = x^7 + C$"

The rule for calculating the integral of a formula such as:
“$y = ax^b$”
... is that we divide “a” by “b + 1”, and then increase the exponent “b” by 1. We then add “C” to the whole formula. In other words, we divide anything scaling the base of the exponential by the exponent plus 1, then we add one to the exponent, and then we add “C” to the whole formula.

We end up with:

$$y = \frac{a}{b+1} * x^{b+1} + C$$

... which can be clearer with the “C” written at the beginning:

$$y = C + \frac{a}{b+1} * x^{b+1}$$

The process is easier to do than to explain. We can use the rule to find the integral of “y = 5”. We just have to realise that 5 is the same as $5 * x^0$. Therefore, the integral is:
“$y = 5x^1 + C$”
... which we would normally write as:
“y = 5x + C”

Working backwards, we can now say that the derivative of the formula:
“y = 5x + C”
... is:
“y = 5”

When we have a list of items being added or subtracted, we find the integral of each item in turn, and keep the additions or subtractions the same. This is similar to when we were finding the derivative of a list of items.

The integral of:
“$y = 3x^2 + 7x – 4$”
... is:
“$y = x^3 + 3.5x^2 – 4x + C$”

Despite there being several parts to this formula, we only add “C” once. As “C” represents “any number”, adding more than one “C” is unnecessary. Just one “C” on its own is the same as “C + C” or “C + C + C” and so on.

The derivative of:
“$y = x^3 + 3.5x^2 - 4x + C$”
... is:
“$y = 3x^2 + 7x - 4$”

Terminology

Outside of the context of calculus, the English word “integral” is the adjectival form of the word “integer”. We might say “an integral number” to mean an integer. The word “integral” is also an adjective that means “necessary to make something complete”. We might say that wheels are an integral part of a vehicle. The Latin adjective “integer” means whole, complete, or unblemished. This is related to the Latin verb “integrare”, which means “to restore to a former condition”, “to make new” or “to make whole again”. The English word “integral” *in the context of calculus* can be thought of as meaning “that which has been restored to its former condition” or “that which has been made whole again”. In other words, if we have formula A, we derive the derivative formula B. We then *restore* it back to formula A by finding the integral. The process of integration turns our derivative formula back into the original formula with which we started – it restores it. Knowing all of this might make it easier to remember the term “integral” and stop it seeming a completely irrelevant word for the process it describes.

Anti-derivatives

An integral is the formula for which a given formula is the derivative. To put this more pedantically, an integral is the *general basis* for all the formulas for which a given formula is the derivative. The addition of “C”, which refers to every possible value that could be added to the rest of the formula, means that an integral is not really a specific formula, but a guide to all the possible formulas. This idea leads on to a new term: “anti-derivative”. An anti-derivative is *one* of the possible formulas for which a given formula is the derivative. For example:
“$y = x^3 + 2$”
... is an anti-derivative of:
“$y = 3x^2$”

“$y = x^3 + 7$”
... is also an anti-derivative of:
“$y = 3x^2$”

“$y = x^3$”
... is also an anti-derivative of:
“$y = 3x^2$”

The difference between an anti-derivative and an integral is that an integral is a general formula that encompasses all the possible curves for which a particular formula is the derivative. An anti-derivative is just one of the specific formulas for which a particular formula is the derivative. For any given formula, there is just one integral, but an infinite number of anti-derivatives, all with the same shape, but different values added to them.

For example, for the formula “$y = 2x$”:

- The integral is “$y = x^2 + C$”
- One of the anti-derivatives is “$y = x^2 + 1$”
- Another of the anti-derivatives is “$y = x^2 + 2.7$”
- Yet another of the anti-derivatives is “$y = x^2 + 1034$”
  ... and so on.

The concepts of integrals and anti-derivatives are so similar that many people outside of maths books do not bother making the distinction, and will just call both concepts “integrals”.

It pays to remember that the integral for a particular formula will always be referred to as **the** integral (as there can be only one). An anti-derivative will always be referred to as **an** anti-derivative (as it will be one among countless others). An integral formula will always have the addition of “C”. An anti-derivative will not generally have the addition of “C” unless the “C” is being used for another purpose as an unknown variable in a calculation.

A sensible question that one might ask on first learning these terms is, “Why do two almost identical concepts have such different names?” Another question one might ask is “Why do these concepts still have such names nowadays, when everyone agrees that the names are confusing?” I do not know the answers to either of these questions. If I were naming them, I would call them something such as “the general anti-derivative” and “a specific anti-derivative”. The term “anti-derivative” is a more intuitive term than “integral”, as it hints that it might be the inverse of “derivative”.

The process of finding an anti-derivative is called “anti-differentiation”. Despite the process being the same as the process of integration, it has been given a completely different name.

Powers of “e”

We saw in the section on differentiation that the derivative of:
“$y = e^x$”
... is also:
“$y = e^x$”.

From what we have just learnt, we can now say that:
The integral of “$y = e^x$” is “$y = e^x + C$”
One of the anti-derivatives is “$y = e^x$”
Another of the anti-derivatives is “$y = e^x + 1$”
Yet another of the anti-derivatives is “$y = e^x + 2$”
Yet another of the anti-derivatives is “$y = e^x + 78.95$”
Yet another of the anti-derivatives is “$y = e^x - 12$”
... and so on.

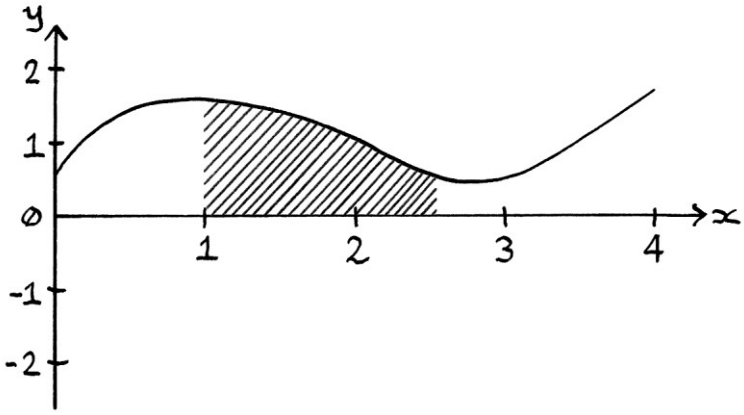
The curves of all the anti-derivatives of a function will all be parallel to each other. They will only differ in how far up or down the y-axis they are.

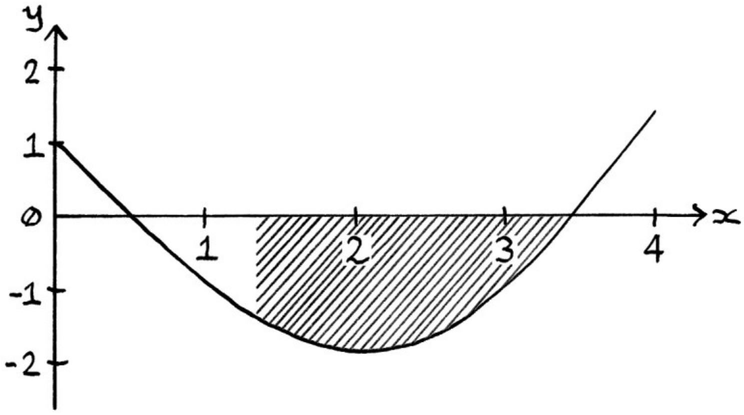
Integration in general

As with differentiation, most of the academic subject of integration concerns itself with formulas that are more complicated than those shown here.

Integration to calculate areas

The most useful aspect of integration is that we can use it to find the exact area between a section of a curve (or line) and the x-axis if we have the formula for the curve or line. [The reasons why this is so are not particularly complicated, but would take too long to explain to fit into this one chapter.] The area above the x-axis is treated as being positive, and the area beneath the x-axis is treated as being negative. Therefore, if a curve or line exists both above and below the x-axis, the area below the x-axis is subtracted from the area above.





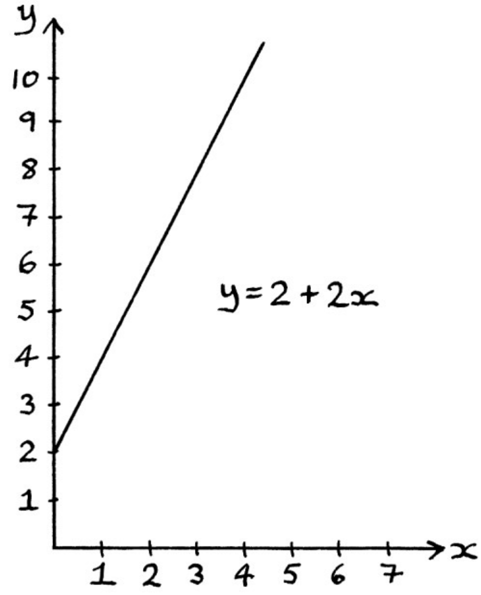
To find the area between a curve and the x-axis:

- We calculate the integral formula of the curve.
- We put in the starting x-axis value of the area we want into the integral formula, and calculate the result.
- We put in the ending x-axis value of the area we want into the integral formula, and calculate the result.
- We then subtract the first result from the second result.

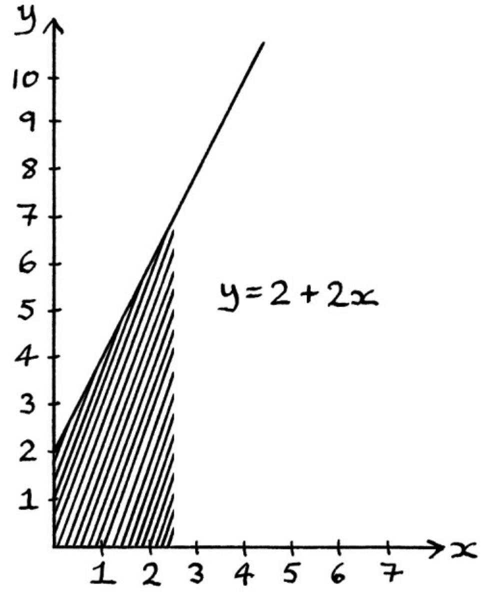
We will explore this idea with some simple examples.

Example 1

We will look at the graph of “y = 2 + 2x”. This is just a diagonal line that starts at y = 2:



We will find the area beneath the line from x = 0 up to x = 2.5:



We can do this with basic maths in two different ways. The area can be calculated as a rectangle added to a triangle, in which case, it is:
(2.5 * 2) + ((2.5 * 5) ÷ 2) = 11.25 square units.

We could also calculate the area as the average height multiplied by the x-axis distance we are measuring:
$((2 + 7) \div 2) * 2.5 = 11.25$ square units.

Now we will use integration to do the same thing. The integral of the formula is:
"$y = 2x + x^2 + C$".

First, we set "x" to 0 and solve the equation as so:
"$y = (2 * 0) + (0^2) + C$".
... which is:
"$y = 0 + 0 + C$"
... which is:
"$y = C$"

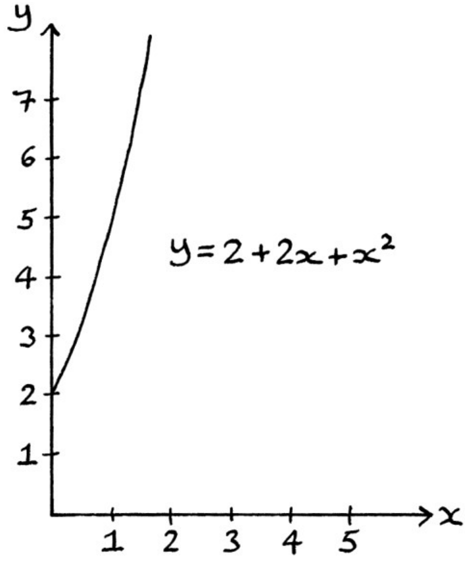
Next, we set "x" to 2.5, and solve the equation as so:
"$y = (2 * 2.5) + (2.5^2) + C$"
... which is:
"$y = 5 + 6.25 + C$"
... which is:
"$y = 11.25 + C$". This is our second value.

We subtract the first result from the second result:
$(11.25 + C) - C$
$= 11.25 + C - C$
$= 11.25$

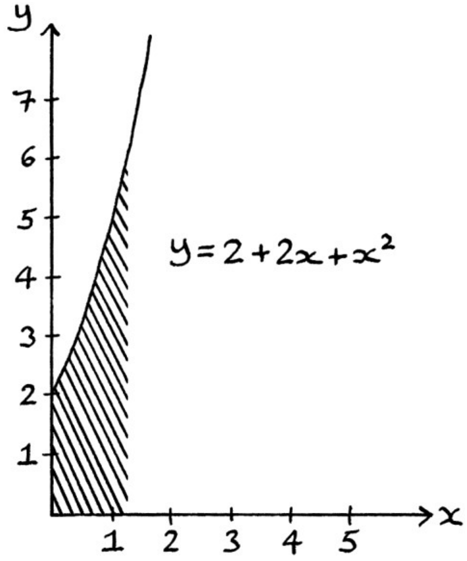
Therefore, the area from $x = 0$ up to $x = 2.5$ is 11.25 square units. This matches our earlier calculations, but this time, we have achieved the result with calculus.

Example 2

We will try a more complicated formula: “$y = 2 + 2x + x^2$”. It looks like this:



We will find the area between the curve and the x-axis between x = 0 and x = 1.25:



To do this, we will find the integral. It is:
“$y = 2x + x^2 + 0.3333x^3 + C$”

First, we solve the formula with 0 replacing “x”:
“y = (2 * 0) + (0^2) + (0.3333 * 0^3) + C”
... which is:
“y = 0 + 0 + 0 + C”
... which is:
“y = C”

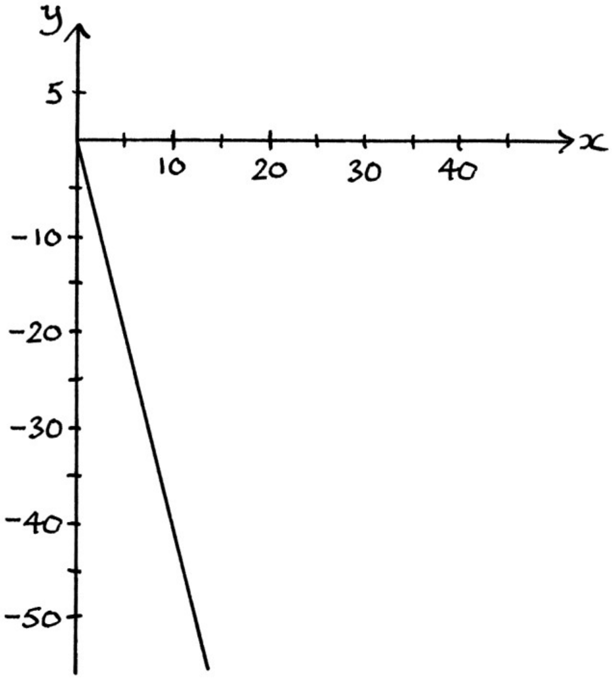
Next, we solve the formula with “x” as 1.25:
“y = 2.5 + 1.5625 + 0.6510 + C”
... which is:
“y = 4.7135 + C”

We then subtract the first result from the second result:
4.7135 + C – C
= 4.7135 square units.

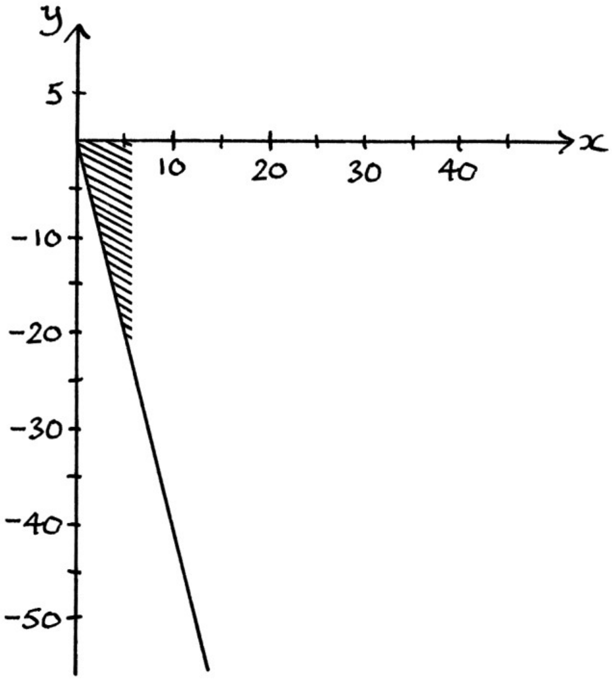
Therefore, 4.7315 is the number of square units between the curve and the x-axis. In this example, it is much harder to know that this is correct from looking at the graph. If we had used any other method to calculate the area, we would have needed to perform much more work (perhaps by calculating the areas of countless little rectangles), and, at best, we would have ended up with an approximation of the area. Calculus gives us the *exact* area.

Example 3

We will now look at a graph that consists entirely of negative y-axis values. We will look at the graph of:
“y = −4x”.



Using integration to find the area between a curve or line and the x-axis works whether the line is above or below the x-axis. It even works if the line crosses the x-axis. We will find the area between the line and the x-axis from x = 0 up to x = 6:



We can calculate this area using normal maths – it will be $(-24 * 6) \div 2 = -72$ square units, but we will use integration to find the area anyway. [Note that normally, we would not think of an area as being negative, but in calculus, it pays to do so. A negative area is one that is beneath the x-axis; a positive area is one that is above the x-axis. Some curves, such as a Sine wave, move both above and below the x-axis, so it is useful to be able to distinguish between positive and negative areas.]

The integral of:
“$y = -4x$”
... is:
“$y = -2x^2 + C$”

We enter the value 0 into the formula:
“$y = -2 * (0^2) + C$”
... which is:
“$y = C$”

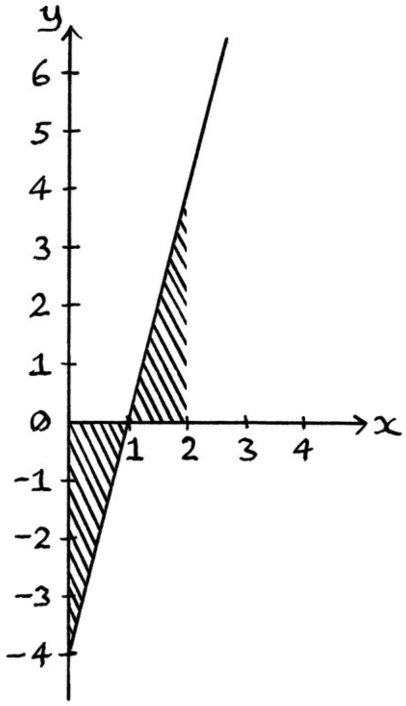
Next, we enter the value 6 into this formula, and we have:
“$y = (-2 * 36) + C$”
... which is:
“$y = -72 + C$”.

We subtract the first result from the second result:
$-72 + C - C$
$= -72$ square units.

Therefore, the area between the line and the x-axis is −72 square units.

Example 4

We will now look at the formula "y = −4 + 4x". This line moves from under the x-axis to above it. This means that the area between the line and the x-axis has both positive and negative parts. We will find the area from x = 0 to x = 2.



We can calculate the area with normal maths. The negative area is:
(−4 * 1) ÷ 2 = −2 square units.

The positive area is:
(2 * 4) ÷ 2 = +2 square units.

We add these together to obtain a total area of 0 square units.

Now we will use integration to find the result. The integral of:
"y = −4 + 4x"
... is:
"y = −4x + 2x^2 + C"

We put 0 into the formula:
“$y = (-4 * 0) + (2 * 0^2) + C$”
... which is:
“$y = 0 + 0 + C$”
... which is:
“$y = C$”

We then put 2 into the formula:
“$y = (-4 * 2) + (2 * 2^2) + C$”
... which is:
“$y = -8 + 8 + C$”
...which is:
“$y = 0 + C$”
... which is:
“$y = C$”

We then subtract the first value from the second value:
$C - C$
= 0 square units.

Therefore, the total area between the line and the x-axis over the length of $x = 0$ to $x = 2$ is 0 square units.

More complicated examples

In the examples so far, the first value has always ended up as just “C”. This is due to the nature of the examples, and how we have always started at $x = 0$. We do not have to start at $x = 0$, and we can, if we want, find the area between any two points.

We will look at the formula “$y = -4 + 4x$” again. We will find the area from $x = 1$ to $x = 3$. The integral is, as before, “$y = -4x + 2x^2 + C$”.

First, we put 1 into the formula:
“$y = (-4 * 1) + (2 * 1^2) + C$”
... which is:
“$y = -4 + 2 + C$”
... which is:
“$y = -2 + C$”

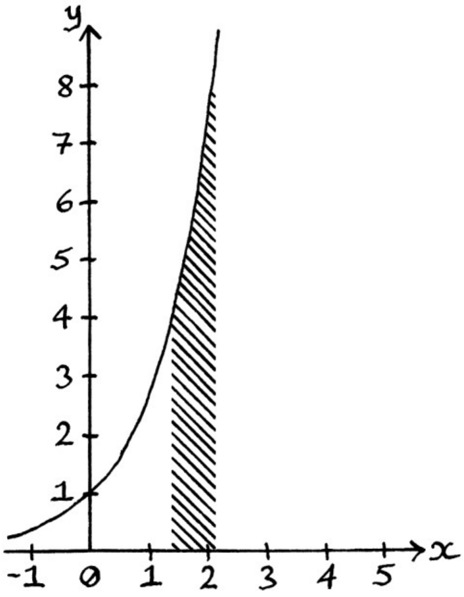
Next, we put 3 into the formula:
"$y = (-4 * 3) + (2 * 3^2) + C$"
... which is:
"$y = -12 + 18 + C$"
... which is:
"$y = 6 + C$"

We then subtract the first result from the second result:
$6 + C - (-2 + C)$
$= 6 + C + 2 - C$
= 8 square units.

Therefore, the area between the line and the x-axis from x = 1 to x = 3 is 8 square units.

Example 5

We will find the area between the curve of:
"$y = e^x$"
... and the x-axis between x = 1.3 and x = 2.1.



The integral of the formula is:
"$y = e^x + C$"

First, we put 1.3 into the integral formula:
“$y = e^{1.3} + C$”
... which is:
“$y = 3.6693 + C$”

Next, we put 2.1 into the integral formula:
“$y = e^{2.1} + C$”
... which is:
“$y = 8.1662 + C$”

We then subtract the first result from the second result:
$8.1662 + C – (3.6693 + C)$
$= 8.1662 + C −3.6693 − C)$
$= 4.4969$

This means that the area from $x = 1.3$ up to $x = 2.1$ is 4.4969 square units.

Example 6

We will find the area between the curve and the x-axis for the formula:
“$y = 4x^3 + 2x + 7 + e^x$”
... between $x = 0$ and $x = 2$.

The integral of the formula is:
“$y = x^4 + x^2 + 7x + e^x + C$”

First, we put 0 into the integral formula:
“$y = 0 + 0 + 0 + 1 + C$”
... which is:
“$y = 1 + C$”

Next, we put 2 into the integral formula:
“$y = 2^4 + 2^2 + (7 * 2) + e^2 + C$”
... which is:
“$y = 16 + 4 + 14 + e^2 + C$”
... which, because “e” is a known number, is:
“$y = 16 + 4 + 14 + 2.71828183^2 + C$”
... which is:
“$y = 41.3891 + C$”

We subtract the first result from the second result, and we have:
41.3891 + C – (1 + C)
= 41.3891 + C – 1 – C
= 40.3891 square units.

Example 7

We will use the same formula as in the previous example, and find the area between x = 1 and x = 5. The integral is, as before:
"$y = x^4 + x^2 + 7x + e^x + C$"

We put 1 into the integral formula:
"y = 1 + 1 + 7 + e + C"
... which is:
"y = 1 + 1 + 7 + 2.7183 + C"
... which is:
"y = 11.7183 + C"

We put 5 into the integral formula:
"$y = 5^4 + 5^2 + (7 * 5) + e^5 + C$"
... which is:
"$y = 625 + 25 + 35 + e^5 + C$"
... which is:
"$y = 625 + 25 + 35 + 2.718281835^5 + C$"
... which is:
"y = 625 + 25 + 35 + 148.4132 + C"
... which is:
"y = 833.4132 + C"

We then subtract the first result from the second result:
833.4132 + C – (11.7183 + C)
= 833.4132 + C – 11.7183 – C
= 821.6949

Therefore, the total area between the curve and the x-axis between x = 1 and x = 5 is 821.6949 square units.

Integration terminology and notation

So far in this chapter, we have used the term "integral" or "integral formula" to refer to the integral formula whether we are using it to find the curve for which a second formula is the gradient, or whether we are using it to find the area between a curve and the x-axis. If we need to make a distinction, we can use the terms "indefinite integral" and "definite integral".

A "definite integral" is an integral formula that is being used to find the area between a curve and the x-axis. It is called "definite" because there are "defined" boundaries that mark the start and end of the section for which we want to find the area. The term "definite integral" makes more sense in relation to the symbol for the idea, as we shall see later in this section.

An "indefinite integral" is a general integral formula that is not being used to calculate an area.

The distinction between the two terms is very minor and relates just to how an integral formula is being used. If an integral formula is being used to calculate an area, it is called a "definite integral"; if it is not, it is called an "indefinite integral". Outside of maths exams, the distinction is almost irrelevant, as the same integral formula will be used in each case.

Notation

The subject of calculus as a whole becomes more difficult as formulas that are more complicated are introduced. In this chapter, I have intentionally left out complicated formulas, as I want to explain the basic idea of calculus simply. For someone who is being introduced to the subject of integration, the most complicated aspect is the notation. Now that we have seen the basics of integration without notation, the notation will be easier to understand.

If we wanted to express the idea that the integral formula of:
"$y = 2x$"
... is:
"$y = x^2 + C$"
... then we would write the following:

$$\int 2x\,dx = x^2 + C$$

The ∫ symbol is the letter “S” written in a fancy elongated way. It is short for the word “sum”, although knowing this does not particularly help in understanding its meaning. The letters “dx” have a similar meaning to the letters “dx” in the derivative formula. They essentially mean “a tiny difference in x”. When we looked at derivative formulas, the division $\frac{dy}{dx}$ referred to the basis of calculating a gradient at a specific point. The “dy ÷ dx” symbol means a tiny difference in “y” divided by a tiny difference in “x”. It is essentially a symbolic way of indicating the gradient, but it is saying that the gradient of every specific point on the curve is represented by the following formula. Integration is the opposite of differentiation. The integral equation is *essentially* saying that the sum of tiny bits of “x” is equal to “x^2 + C”. However, the reason it makes the statement in this way is due to the basis of why and how integration works in the first place. Why integration works is reasonably straightforward, but would make this chapter much longer than it needs to be. For the sake of this chapter, it is easiest to ignore any underlying meaning of the ∫ symbol and the presence of “dx”, and treat it all as just an accepted way of expressing integrals. Examples of the ∫ symbol in use are:

$$\int e^x \, dx = e^x + C$$

In the above formula, we are saying that the integral of:
“y = e^x”
... is:
“y = e^x + C”
... but without explicitly mentioning “y”.

The next formula is the same as the previous one, except that we are dealing with “t” as the variable and not “x”. Therefore, the “dx” becomes replaced with “dt”.

$$\int e^t \, dt = e^t + C$$

In the next formula:

$$\int \sin\theta \;\, d\theta = C + sin\,(\theta - 0.5\pi)$$

... we are saying that the integral of:
“y = sin θ” [when Sine is working in radians, and “θ” is an angle in radians]
... is:
“y = C + sin (θ – 0.5π)”

... which is something that we will see later in this chapter. As we are working with "θ", we use "dθ" instead of "dx".

The basic rule for using the ∫ symbol is that we express integrals as so:

$$\int "something\ involving\ x"\ dx = "the\ integral\ of\ that\ formula"$$

Integrals to calculate areas

If we wanted to express that we were using integration to calculate the area between the line of:
"y = 2x"
... and the x-axis between x = 0 and x = 1, then we would write the following:

$$\int_0^1 2x\,dx = 1$$

In this case, the ∫ symbol has the start and end x-axis values written at its top and bottom. The above formula essentially means:

"The area between the line of "y = 2x" and the x-axis between x = 0 and x = 1, as calculated using integration, is 1 square unit."

When calculating areas, instead of giving the result as a formula, we give the result as the actual resulting area.

As another example:

$$\int_0^1 e^x\,dx = 1.7813$$

This means that the area between the curve of "y = e^x" and the x-axis from x = 0 to x = 1, as calculated with integration, is 1.7813 square units.

An almost identical example is as so:

$$\int_0^1 e^t\,dt = 1.7813$$

This means that the area between the curve of "y = e^t" and the t-axis from t = 0 to t = 1, as calculated with integration, is 1.7813 square units. The difference here is that we are using "t" instead of "x", so we include "dt" instead of "dx". There is also a conceptual difference in that if "t" stands for the time in seconds, then our area is made up of y-axis units multiplied by t-axis seconds. In this way, the concept of area is slightly more obscure than if we were just multiplying the same type of unit.

As another example:

$$\int_{3}^{5} e^x \, dx = 128.3276$$

This means that the area between the curve of "y = e^x" and the x-axis from x = 3 to x = 5, as calculated with integration, is 128.3276 square units.

Another example:

$$\int_{-2}^{2} 7x \, dx = 0$$

This means that the area between the line of "y = 7x" and the x-axis from x = −2 to x = +2, as calculated with integration, is zero square units.

As another example:

$$\int_{0}^{2\pi} \sin\theta \, d\theta = 0$$

... where Sine is working in radians, and "θ" is an angle in radians.

This formula is saying that the area between the curve of "y = sin θ" and the θ-axis from θ = 0 to θ = 2π, as calculated with integration, is zero square units. [I will explain more about calculus involving waves later in this chapter. For now, if we imagine one cycle of an angle-based wave, with the second half being the negative of the first half, we can see why the area for one full cycle would be zero.]

Definite and indefinite integrals

When we use an integral to calculate an area, it is called a "definite integral" because there is a defined range. The following is a definite integral:

$$\int_{-6}^{4} 2x\,dx = -20$$

A definite integral uses the ∫ symbol with values at the top and bottom to indicate the range in which we are interested. A definite integral calculates an area.

When we calculate an integral otherwise, it is called an "indefinite integral" because we are interested in the entire curve, and not in a particular defined area beneath it. The following is an indefinite integral:

$$\int 2x\,dx = x^2 + C$$

An indefinite integral uses the ∫ symbol without any values at the top or bottom. It results in a formula.

Definite integrals and indefinite integrals are the two types of integral. Therefore, it would be valid to use the term "integral" when discussing either of them. Using the word "definite" or "indefinite" helps to distinguish between the two.

Summary of calculus terms

Here is a list of the main terms to do with calculus that we have seen so far:

Derivative

A derivative is a formula that shows all the gradients of a second formula. Another way of saying this is that a derivative is a function for which the y-axis of every point is equal to the gradient of the point at the corresponding x-axis value of a given function.

Differentiation

Differentiation is the process of calculating the derivative formula of a given formula.

Integral

An integral is the *general* formula for which a given formula is the derivative. It is the opposite of a derivative. An integral must contain an addition of the constant "C", to indicate that there are really countless possible formulas for which a given formula is the derivative.

Note that some people also use the term "integral" to mean "integer" – for example, "integral number", "integral result" and so on. This can be mildly confusing if they use the term while discussing some aspect of calculus.

Anti-derivative

An anti-derivative is one specific formula for which a given formula is the derivative. There are an infinite number of anti-derivatives for any one formula, all with the same curve, but differing by how high or low they are on the y-axis. There is only one integral for each formula.

Anti-differentiation

Anti-differentiation is the name of the process of finding an anti-derivative. It ultimately means exactly the same thing as integration, but anti-differentiation finds one specific formula, while integration finds a general formula.

Indefinite integral

An indefinite integral is a term that distinguishes the type of integral, with the focus on how the integral is being used. An indefinite integral is just the general integral formula.

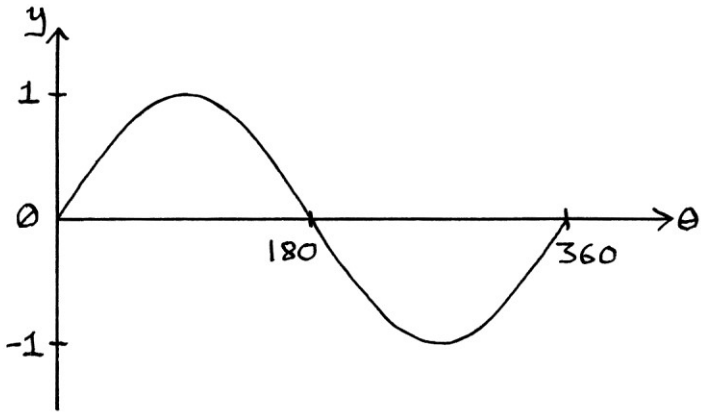
Definite integral

A definite integral is an integral that is being used to calculate the area between a curve and the x-axis between two defined x-axis points.

Angle-based waves

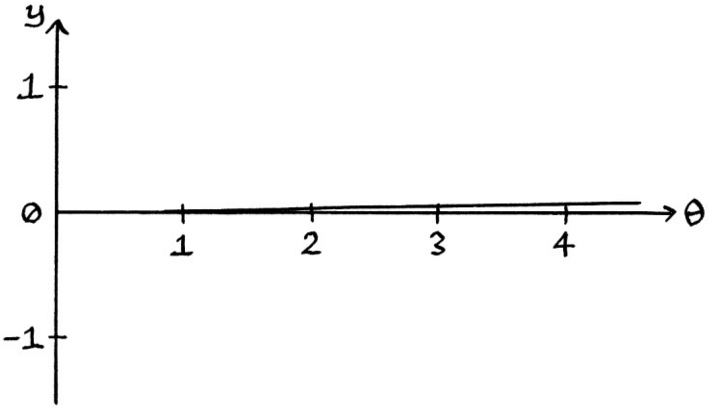
We will now look at the derivatives and integrals of waves. Calculus with waves is often more straightforward than other calculus. To introduce the idea, we will examine what it actually means to calculate the *gradient* of a wave. Outside of calculus, if we were calculating the gradient of a point on the graph of a curve such as “y = x^2”, we would need the units for both “x” and “y” to have the same dimensions. We then divide the tiniest change in “y” by the difference in “x” over that change. If the units are the same, we end up with the gradient. If the units are different, then obviously we do not. For example, if the y-axis units are in centimetres, but the x-axis units are in metres, then the calculation of the gradient would be wrong. [We could still calculate the gradient from the formula with calculus, but not from the graph.]

A “y = sin θ” wave graph shows the y-axis heights of a unit-radius circle’s edge at evenly spaced angles from the origin. Previously in this book, it has not mattered which angle system we used to measured those angles. A “y = sin θ” wave graph with the angles measured in degrees, and Sine operating in degrees, looks like this:



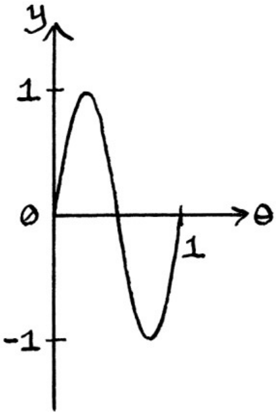
The trouble with this graph is that the units for the θ-axis are not to the same scale as the units for the y-axis. On the page, there are 360 degrees in about the same space as three y-axis units. Therefore, we cannot measure any gradients directly from the graph. If we want to calculate the gradient from the graph, we need one unit of the y-axis to take up the same amount of space as one unit of the θ-axis. We could redraw the graph with the θ-axis units taking up the same amount of space as the y-axis units. This would mean that the graph of the curve would be 360 times longer than it is high. Such a picture cannot be drawn clearly on a normal piece of paper – either the y-axis units have to be drawn at such a small scale that the curve appears as a nearly straight line, or the θ-axis units will not fit on to the page.

Here are the first 4 degrees of the curve:



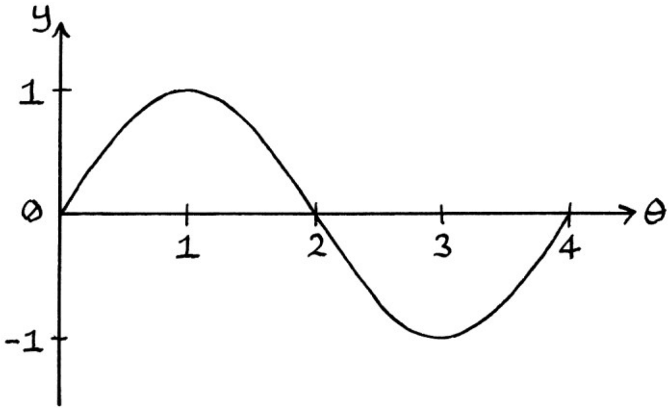
Despite not being able to show the *whole* graph in an easy form, we can imagine what it looks like.

There are angle systems other than degrees, so we will see what happens when we use them instead. First, we will draw a Sine wave graph with the θ-axis showing whole-circle angle units. In other words, every angle is the portion of a circle at which the y-axis values occur. [For example, quarter of the way around the circle is 0.25 whole-circle angle units.] In the following picture, the θ-axis is drawn from 0 whole-circle angle units up to 1 whole-circle angle unit. The graph has been drawn so that one y-axis unit has the same dimensions as one θ-axis unit:

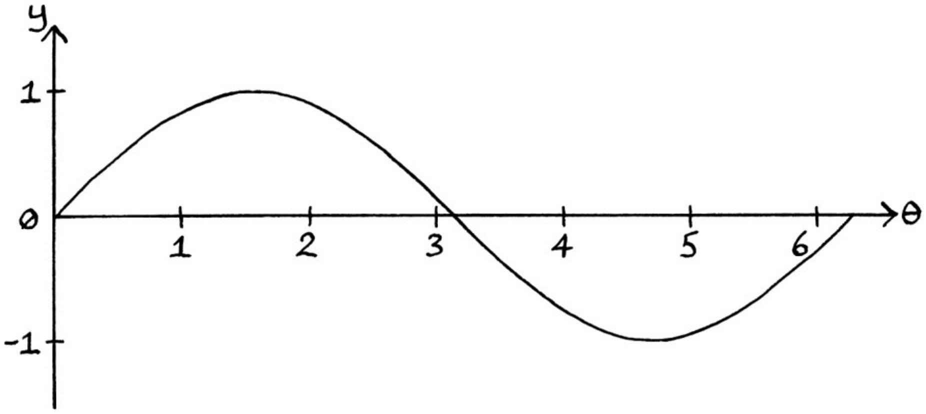


Although this is still a Sine wave graph that represents the y-axis values of points around a unit-radius circle at evenly spaced angles, the gradients of the wave curve are clearly steeper than those of the wave that used degrees.

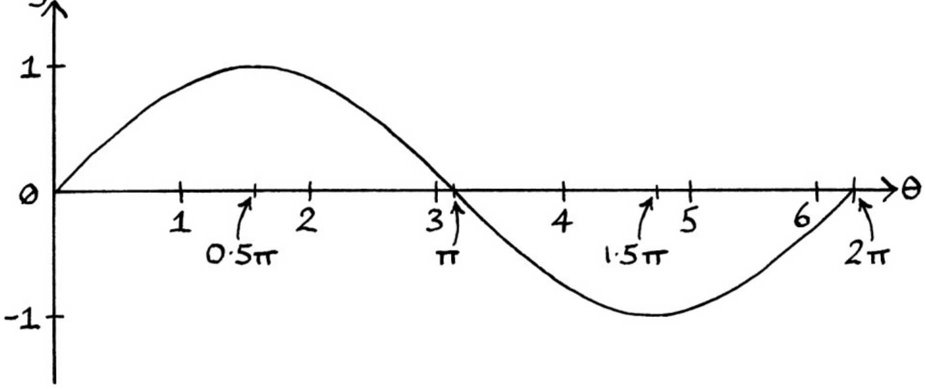
We can redraw a Sine wave graph with the θ-axis measured in quarter-circle angle units. As there are 4 quarter-circle angle units in a circle, one cycle of the wave will be 4 times longer than one unit of the y-axis:



We will now draw the Sine wave with radians:



Note that previously in this book, I have labelled the θ-axis for radians in terms of "π", such as 0.5π, π, 1.5π, and so on, while in the above graph, I have labelled the θ-axis with normal numbering. The same graph with the significant points along the θ-axis marked as multiples of "π" looks like this:

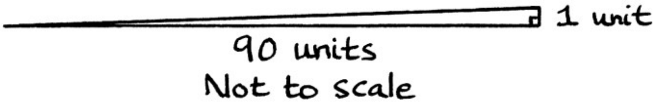


As we can see from the above graphs, each different type of angle results in a Sine wave graph of a different length. Given that, it should be clear that the *gradients* of a wave curve will depend on which angle system we are using.

Gradients

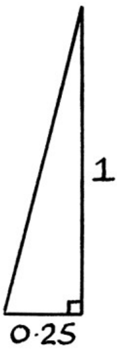
For interest's sake, we will calculate the average gradient from the start of each wave (the equivalent of 0 degrees) to the first peak (the equivalent of 90 degrees). For a degrees-based wave, this will be from θ = 0 degrees (where y = 0) to θ = 90 degrees (where y = 1).

The gradient is the change in the y-axis divided by the change in the x-axis (the θ-axis). We can portray this on a right-angled triangle to make it clearer, in which case, we calculate the opposite side divided by the adjacent side. [The following triangle is not drawn to the correct proportions, or else, either it would look like a flat line, or it would not fit on the page.]



This gradient of the hypotenuse is:
= 1 ÷ 90
= 0.01111

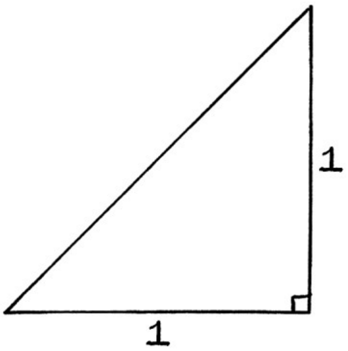
Now we will calculate the average gradient between the equivalent angles but for whole-circle angle units. Therefore, we will calculate the average gradient between 0 whole-circle angle units and 0.25 whole-circle angle units. At 0 whole-circle angle units, the y-axis value of the curve is zero units, and at 0.25 whole circle angle units, the y-axis value of the curve is 1. We can portray the situation with a right-angled triangle, which in this case, we can draw to the correct scale:



The gradient is calculated as so:
= 1 ÷ 0.25
= 4.

Therefore, although we have a “y = sin θ” wave showing the same information, we have a different gradient for the same portion of a circle. This is because we are dividing by a different value when we calculate the gradient.

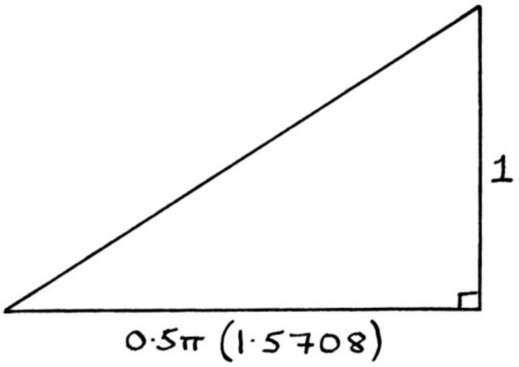
Now we will calculate the average gradient for quarter-circle angle units for the same part of the wave. This is the section from θ = 0 up to θ = 1 quarter-circle angle units. The change in the y-axis is 1 unit; the change in the θ-axis is also 1 unit. We can portray this as a right-angled triangle (which is drawn to the correct proportions):



We calculate the gradient with:
= 1 ÷ 1
= 1.

Again, because we are using a different type of angle, we end up with a different gradient.

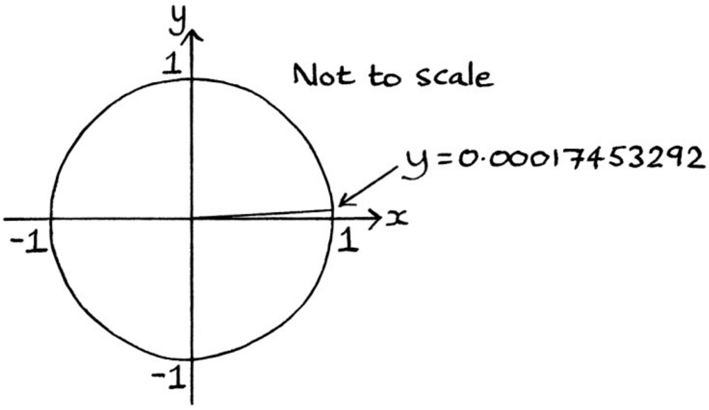
We will now calculate the gradient for the same section of a wave in radians. This will be from 0 radians up to 0.5π radians (1.5708 units). The triangle that represents the gradient we are finding is as follows, and is drawn to the correct proportions.



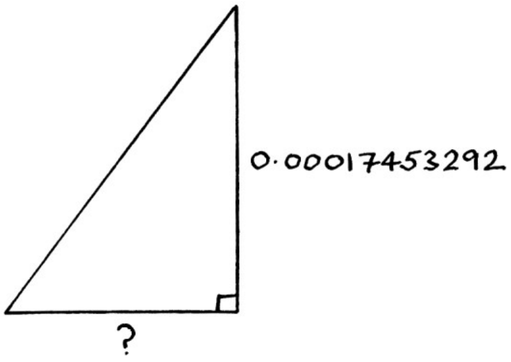
The gradient of the hypotenuse of this triangle, and the average gradient of the curve from θ = 0 radians up to θ = 0.5π radians is:
= 1 ÷ 0.5π
= 1 ÷ 1.5708
= 0.6366

From all of the above, we can see that the gradient of any part of a wave is dependent on the angle system that we are using to describe the angles.

Although it might appear that any angle system is equally good in every situation, when it comes to calculus, radians turn out to be the most useful. To demonstrate why this is, we will look at the average gradient for the very first part of each curve. We will measure the gradient of the equivalent of the first 0.01 degrees, but with different angle units. The y-axis value of the point on a unit-radius circle at 0.01 degrees is 0.00017453292 units. It does not matter which angle system we use, the y-axis value of the point on a unit-radius circle at any equivalent to that angle will be 0.00017453292 units.



Given that, we know that our triangle that represents the gradient calculation will always have an opposite side of 0.00017453292 units. The adjacent side will vary depending on the angle system we are using:



When we are using a wave described with degrees, the length of the adjacent side will be 0.01 degrees. The gradient will be:
0.00017453292 ÷ 0.01
= 0.017453292

For whole-circle angle units, the equivalent of 0.01 degrees is: 0.000027777778 whole-circle angle units. The gradient is:
0.00017453292 ÷ 0.000027777778
= 6.28318528

For quarter-circle angle units, the equivalent of 0.01 degrees is: 0.00011111111 quarter-circle angle units. The gradient is:
0.00017453292 ÷ 0.00011111111
= 1.57079632

For radians, the equivalent of 0.01 degrees is: 0.000174532935. The gradient is:
0.00017453292 ÷ 0.000174532935
= 0.99999999

There are patterns in these calculated gradients:

- For degrees, the gradient is: 0.017453292, which is roughly 2π ÷ 360.
- For whole-circle angle units, the gradient is: 6.28318528, which is roughly 2π.
- For quarter-circle angle units, the gradient is: 1.57079632, which is roughly 2π ÷ 4 = 0.5π.
- For radians, the gradient is: 0.99999999, which is roughly 1.

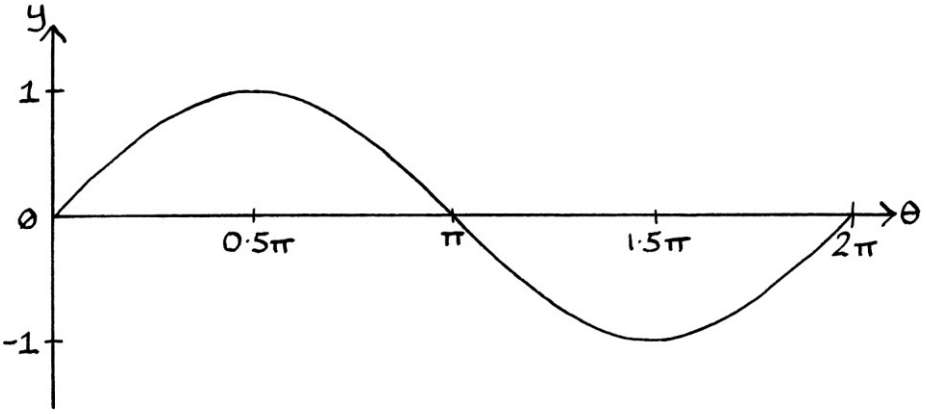
This means that the starting point of a curve of a Sine wave, no matter what the angle system, has a gradient related to 2π. [For radians, the angle system itself has a relationship with 2π.] The reason for this relates to how the circumference of a unit-radius circle is 2π units long. A "y = sin θ" wave shows the y-axis values of points around a circle at evenly spaced angles from the origin. Another way of saying the same thing is that a Sine wave shows the y-axis values of points around a circle's *circumference* at evenly spaced divisions of its circumference. As a unit-radius circle has a circumference of 2π units, any way in which we decide to divide the circumference of a circle will have an unavoidable relationship with the number 2π.

Radians

We could decide to use any angle unit, but when it comes to calculus, radians are by far the most appropriate. The main reason for this is that the gradients of a radians-based "y = sin θ" wave are described by that same wave with a phase 0.5π radians (quarter of a circle) more. To put this slightly more concisely:

For any place on the wave, the gradient of "sin θ" will be "sin θ + 0.5π" *when "θ" is an angle in radians and Sine is working in radians.*

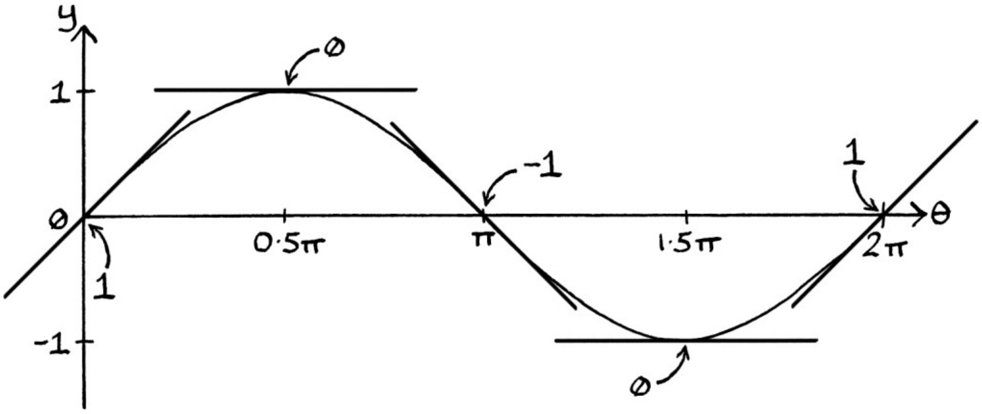
We can display this with pictures. We will start with the graph of "y = sin θ" in radians, with the axes drawn to the same scale as each other. [I am labelling the θ-axis with multiples of "π" to make it clearer where the significant parts of the wave are. This makes it less obvious that the θ-axis is drawn to the same scale as the y-axis.]



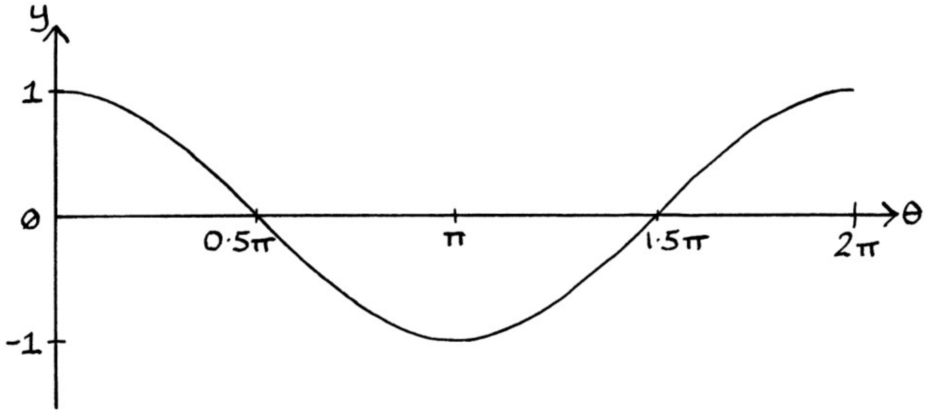
We can make some observations about the gradients of points on this curve:

When θ = 0, the gradient of the wave's curve is 1.
When θ = 0.5π radians, the curve is at its peak, and the gradient is 0.
When θ = π radians, the gradient is −1.
When θ = 1.5π radians, the gradient is 0.
When θ = 2π radians, the gradient is 1 again.

The same picture with the gradients labelled is as so:



The graph that indicates all these gradients is "y = sin θ + 0.5π" [which we could also describe as "y = cos θ"]:



When "θ" is 0 on this "y = sin θ + 0.5π" graph, "y" is 1. The value of 1 is the gradient of the "y = sin θ" graph when θ is 0.

When "θ" on the "y = sin θ + 0.5π" graph is 0.5π, "y" is 0. The value of 0 is the gradient of the "y = sin θ" graph when θ = 0.5π.

In fact, the “y = sin θ + 0.5π” graph lists all of the gradients of the “y = sin θ” graph. Every y-axis value on “y = sin θ + 0.5π” is the gradient of the curve at the corresponding θ-axis point of “y = sin θ”. It is the derivative graph. We can say that “y = sin θ + 0.5π” is the derivative of “y = sin θ”. To put this more mathematically, we would say:

If “y = sin θ”, when “θ” is an angle in radians and Sine is operating in radians, then:

$$\frac{dy}{d\theta} = \sin\ (\theta + 0.5\pi)$$

Generally, we would refer to “y = sin θ + 0.5π” as “y = cos θ”. Therefore, in radians, “y = cos θ” is the derivative formula for “y = sin θ”. We would say:

$$\frac{dy}{d\theta} = \cos\ \theta$$

[For the sake of clarity, we will continue to call the formula “sin (θ + 0.5π)” instead of “cos θ”.]

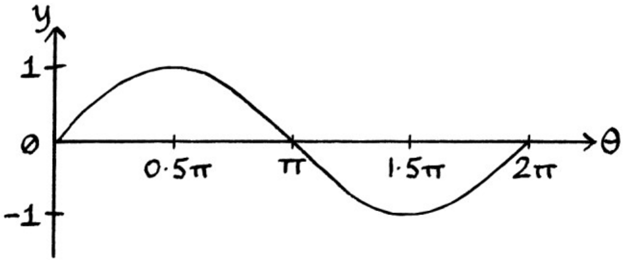
Derivatives

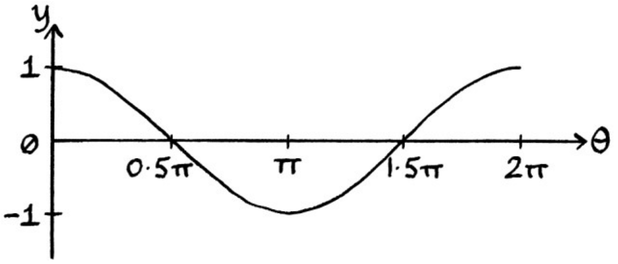
There is a repeating set of derivatives for radian-based Sine waves. If we start with:
“y = sin θ”, then its derivative formula is:
“y = sin (θ + 0.5π)”, and the derivative of that is:
“y = sin (θ + π)”, and the derivative of that is:
“y = sin (θ + 1.5π)”, and the derivative of that is:
“y = sin (θ + 2π)”, which is the same as “y = sin θ”, which we started with.

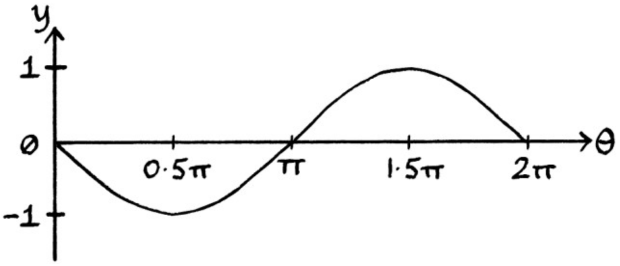
This shows that there are four stages of wave derivatives. Each derivative is the same wave with 0.5π radians added on to the phase.

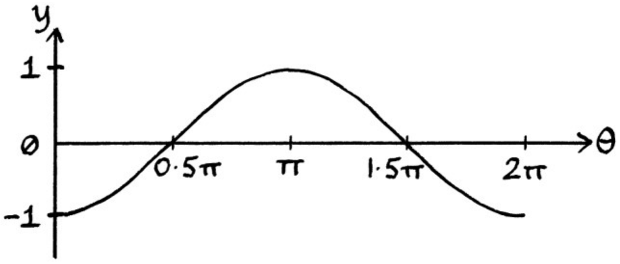
[It is important to understand that adding 0.5π radians (or quarter of a circle) to the phase only works if the waves are entirely in radians.]

The four stages of derivative waves look like this, with each wave being the derivative of the one before it:









Each derivative wave is the previous wave slid to the left by 0.5π radians (quarter of a circle). The next derivative would be the wave that we started with.

Frequently in maths books, the progression of four derivatives is given in terms of Sine and Cosine with zero phases. Doing this makes the formulas simpler but less intuitive. The four derivatives are given as so:

- "y = sin θ"
- "y = cos θ"
- "y = −sin θ"
- "y = −cos θ"

These are exactly the same waves, but expressed in a different way. This is the most common way of seeing the progression of waves, but it is slightly harder to remember and understand why they are as they are.

To make the more common progression of derivatives easier to visualise, we can give their equivalents alongside them:

- "$y = \sin\theta$"
- "$y = \cos\theta$", which is "$y = \sin(\theta + 0.5\pi)$".
- "$y = -\sin\theta$", which is "$y = \sin(\theta + \pi)$" or an upside down "$y = \sin\theta$".
- "$y = -\cos\theta$", which is "$y = \sin(\theta + 1.5\pi)$" or an upside down "$y = \cos\theta$", or an upside down "$y = \sin(\theta + 0.5\pi)$".

The progression of derivatives works for any phase. If we have a wave in the form of "$y = \sin\theta + \phi$", we just add 0.5π radians to the phase to produce the formula for the derivative. For example, the derivative formula for:
"$y = \sin(\theta + 2)$"
... is:
"$y = \sin(\theta + 2 + 0.5\pi)$"

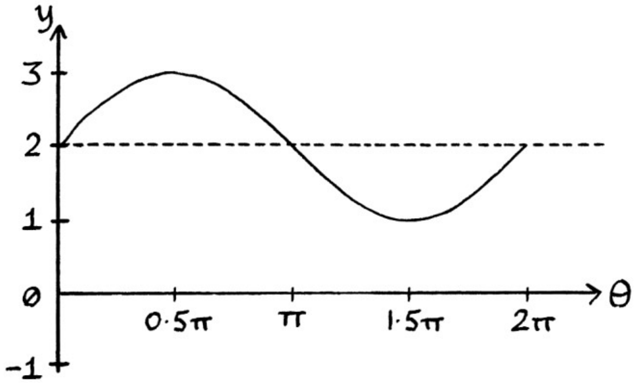
Always remember that all of the above only works in radians. It is not the case, for example, that the derivative of "$y = \sin\theta$", when "θ" is an angle in degrees and Sine is working in degrees, is "$y = \sin\theta + 90$". In a sense, radians are the "natural" circle division for calculus.

Mean levels

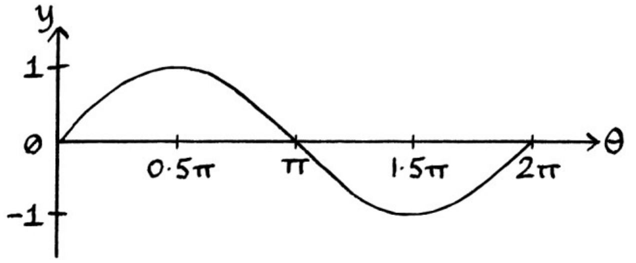
Calculus with waves can make the idea of constants added to formulas easier to comprehend. This is because a constant added to a wave is really just a mean level by another name.

For any wave of the type "$y = \sin\theta$" in radians, where there is a mean level, then its derivative wave will be that same wave with zero mean level and 0.5π added on to the phase. This is because the height of a curve up or down the y-axis is irrelevant to its gradients.

For example, this wave:



... has the same gradients as this wave:



All of the following waves have exactly the same gradients:
“y = 1 + sin θ”
“y = 2 + sin θ”
“y = 0.00001 + sin θ”
“y = −9999.99 + sin θ”

The derivative formula for each of these is “y = sin (θ + 0.5π)”. The mean level is irrelevant to the gradient.

When it comes to integration, the addition of “C” in the integral has exactly the same meaning as an unknown mean level. In fact, it can be easier to understand the relevance of “C” when thinking of waves, than it is when thinking of other formulas.

Integrals

For waves of the form "y = sin θ" *in radians*, the rule to calculate the integral formula is to subtract 0.5π radians from the phase, and then add the constant "C" as the mean level. As always, when performing integration, we have to bear in mind that, as gradients are independent of the actual y-axis values of a curve, there are countless waves that have a particular derivative. Any waves that share the same shape, but have different mean levels will all have the same derivative. As we know from the last section:
"y = sin (θ + 0.5π)"
... is the derivative of:
"y = sin θ"
... and it is also the derivative of "y = 1 + sin θ", "y = 2 + sin θ" and so on.

Therefore, when we find the integral of a wave, we have to acknowledge that there are countless anti-derivatives all with the same curve but with different mean levels. We do this by including the constant "C".

The integral of:
"y = sin θ"
... is:
"y = sin (θ − 0.5π) + C"
... which, to be consistent with a typical wave formula, I will write as:
"y = C + sin (θ − 0.5π)"

The integral of:
"y = sin (θ + 0.5π)"
... is:
"y = C + sin θ"

The integral of:
"y = sin (θ + 2.7π)"
... is:
"y = C + sin (θ + 2.2π)"

Note that is common to see the "C" added on to the end of the formula, and not the start. When it comes to waves, I add it on to the start, first to make it clear that it is not part of everything being Sined, and second, to be consistent with how I have been writing wave formulas with mean levels. It does not matter if you add the "C" to the start or the end of the formula as long as you remember to add it, and as long as what you have written is clear and unambiguous.

The progression of integrals is as follows. Note that I am finding the integral of the formulas without the "C" constant. I will explain why later in this chapter:

- The integral of: "y = sin θ" is "y = C + sin (θ – 0.5π)"
- The integral of: "y = sin (θ – 0.5π)" is "y = C + sin (θ – π)"
- The integral of: "y = sin (θ – π)" is "y = C + sin (θ – 1.5π)"
- The integral of: "y = sin (θ – 1.5π)" is "y = C + sin (θ – 2π)" or "y = C + sin θ"

Many maths books will give a Sine wave with a phase of 0.5π, π or 1.5π in terms of Sine or Cosine waves with zero phases and maybe a negative amplitude. In this way of thinking, the progression is as so:

- The integral of: "y = sin θ" is "y = C − cos θ"
- The integral of: "y = −cos θ" is "y = C − sin θ"
- The integral of: "y = −sin θ" is "y = C + cos θ"
- The integral of: "y = cos θ" is "y = C + sin θ"

Angle systems other than radians

We do not *have* to use radians when performing calculus, but the advantages of radians mean that it makes very little sense to use any other system. The derivatives and integrals of waves based on radians are simple. The derivatives and integrals of waves based on degrees, for example, are much less intuitive. A very common mistake is to forget that the rules for calculus on radian-based waves do not work with other angle systems. To remember why this is so, just imagine a degrees-based wave graph with the y-axis and θ-axis drawn to the same scale and think about the gradients.

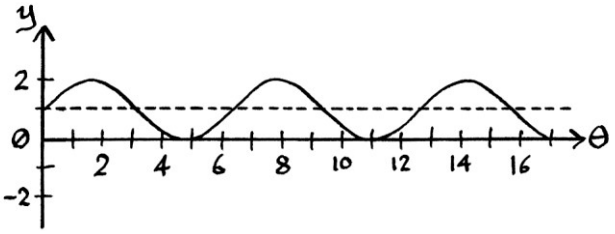
Integrals of waves with mean levels

All of the above integral examples with waves had zero mean levels. An important thing to realise is that, although we can say that the integral of:
"y = sin θ"
... is:
"y = C + sin (θ – 0.5π)"
... we cannot know the integral of:
"y = C + sin (θ – 0.5π)"

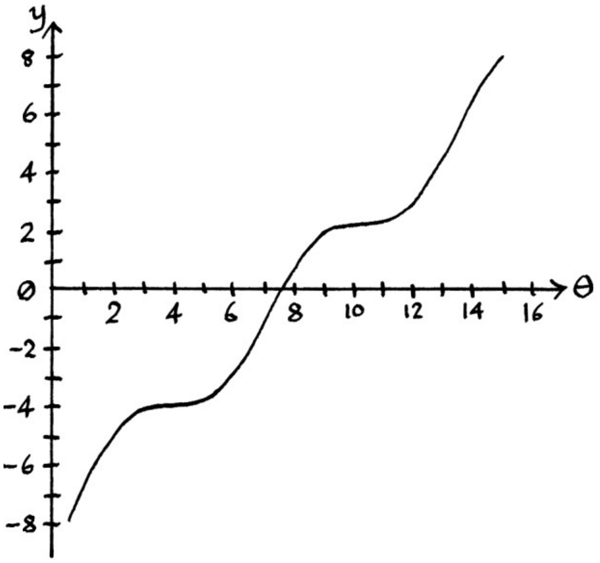
This is because we do not know, and cannot know, what “C” is. “C” is really a symbol that refers to *every* possible value. Finding an integral involves finding the general formula that has gradients equal to the y-axis values of the curve. For a wave such as “y = C + sin (θ – 0.5π)”, we cannot know what those y-axis values are because “C” does not refer to just one value.

Another reason that I gave the above examples zero mean levels is because the results of calculations with non-zero mean levels (where we *do* know the mean level) are more complicated. The integral of a Sine wave with a non-zero mean level is not a pure wave (although it is wave-like), and more importantly, it is *at an angle* on the graph. To illustrate this, we will look at this wave:
“y = 1 + sin θ” [in radians].

The wave, when drawn with the y-axis and θ-axis units to the same size, looks like this:



One of the anti-derivatives looks like this:



[All of its anti-derivatives will have this shape, but will be higher or lower on the y-axis.]

We can see that this is a wiggly line at an angle. Although it is hard to tell, the line has a wave-like quality. The derivative of this line is the wave that we started with.

This anti-derivative is not a pure wave any more. Instead, it is a slightly distorted wave at an angle. The instantaneous gradients of this signal change in a repetitive pattern along its length, but they never become negative. The signal continues forever at this angle and with this shape. The original wave had a positive mean level that raised all the y-axis values to be zero or higher – the original wave's y-axis values never became negative. Therefore, the integral and the anti-derivatives have gradients that never become negative.

The wave-like shape resembles a pure wave that has been rotated from horizontal to have an average gradient equal to the mean level of the original wave. The wave has become distorted as it has been rotated. The signal repeats its shape in the same amount of time as the original wave's frequency. However, because the signal is at an angle, the repetitions as seen along the centre of the signal are further apart – if we rotated the signal to be horizontal, its frequency would be slower than the original wave.

If the original wave had had a mean level of −1 units, and so consisted of only negative values, the sloped integral signal would have moved from the top left to the bottom right of the graph.

Why a wave with a mean level ends up at an angle is easy to understand if we remember that any formula with an addition or subtraction can be split into its parts, and each part integrated or differentiated in turn. If we start with the formula:
"y = 1 + sin θ" [in radians]
... we can split it into the sum of two parts – the mean level and the wave:
"y = 1"
... and:
"y = sin θ"

We then find the integrals of each part, and we end up with:
"y = C + 1θ" [because our x-axis is the θ-axis]
... and:
"y = C + sin (θ – 0.5π)"

Therefore, the full integral is:
"y = C + 1θ + sin (θ – 0.5π)"

One of the anti-derivatives is:
"y = 1θ + sin (θ – 0.5π)"

The graph of "y = 1θ" on its own is a diagonal line where the y-axis value of any point is always equal to the θ-axis value. It is a line at 45 degrees. The formula "y = sin (θ – 0.5π)" is a normal Sine wave with a phase of +1.5π radians (270 degrees). When we add the two parts together, we end up with the wave added to a 45-degree line. This makes a wave-like shape at an angle of 45 degrees.

Time-based waves

We can use calculus on time-based waves. We will continue to use time-based waves that are ultimately based on radians, which means they will have the "2π" frequency correction in the formula. This "2π" and any other frequency in the formula will mean that the progression of wave derivatives does not work quite as simply as before. The derivative will still have a phase of 0.5π radians more, but the amplitude will be different.

A time-based wave is ultimately an angle-based wave with a scaled angle. Any angle-based wave with a value scaling the angle will have a derivative with a different amplitude.

As we know, the derivative of:
"y = sin θ"
... when Sine is working in radians, is:
"y = sin (θ + 0.5π)"

If we swap angles ("θ") with time ("t"), we can make the observation that the derivative of:
"y = sin t"
... when Sine is working in radians, is:
"y = sin (t + 0.5π)"

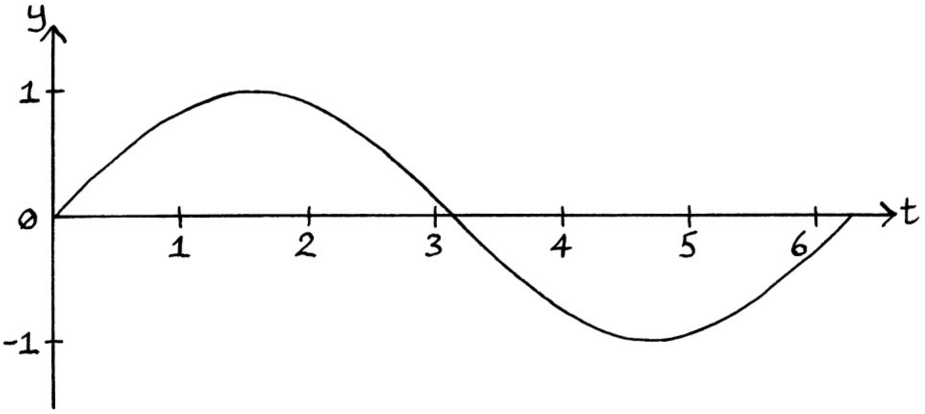
This is because the Sine function does not care about the nature of the values being Sined. Regardless of what we want the value being presented to Sine to be, the Sine function will treat it as an angle. Therefore, if we give the time to the Sine function, the time will be treated as if it were an angle (in this case, in radians). In the above example, the wave would repeat once every 2π seconds, because Sine treats the time as an angle in radians, and a radian-based wave would repeat once every 2π radians.

Generally, when we use time-based waves in radians, we multiply the time by 2π. This has the effect of speeding up the wave so that, by default, it repeats once every second. This means that we can further multiply the time by a frequency value and the cycles will repeat at that frequency. The reasons are explained in more depth in Chapter 4. In radians, a basic time-based wave formula is:
"y = sin (2π * t)"

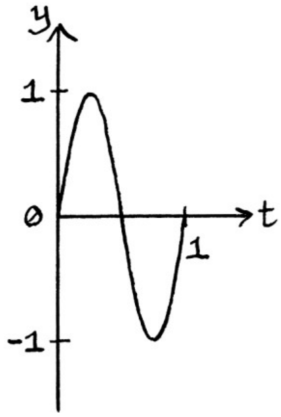
The multiplication by 2π essentially acts as a frequency correction. It speeds the wave up by 2π times (6.2832 times). This means that any wave that would have repeated once every 6.2832 of the units being Sined will now repeat 6.2832 times for every unit being Sined. The wave's peaks and dips occur 6.2832 times more often in the same space. This means that the gradients of the peaks and dips of the curve are steeper than before. This is because they must rise and fall 6.2832 times in the space that they would have risen and fallen just once before.

It is important to note that the derivative of:
"y = sin (2π * t)"
... is *not*:
"y = sin ((2π * t) + 0.5π)"

The graph of "y = sin (t)", in radians, with the y-axis and t-axis drawn to the same scale, looks like this:



The graph of “y = sin (2π * t)” on the other hand, when the axes are drawn to the same scale, looks like this:



The whole curve that used to fit into 2π units (6.2832 units) on the time axis is now squeezed into just 1 unit on the time axis. This means that the gradients of the curve are now steeper. It also means that the area between the first half of the cycle and the t-axis is now less, as is the area between the second half of the cycle and the t-axis. [Note that normally, the y-axis and t-axis are not drawn to the same scale, so such a graph usually appears longer than it does here.]

The derivative signal shows the gradients at any particular t-axis value for the curve, and our 2π time-based wave has a curve with steeper rises and falls (steeper gradients) than before. This means that the derivative wave will have a higher amplitude. [The y-axis values at a particular point on the derivative wave show the instantaneous gradient at the corresponding point of the original wave. Therefore, if the gradient is steeper, those y-axis values will be higher, and the overall amplitude will be larger.]

A general differentiation rule

As the waves discussed in this book are more likely to be relatively simple, with a simple phase, and have the value treated as an angle ("θ" or "t") just scaled by a fixed value, we can use a simplified rule to calculate the derivative formulas. It is worth noting that this rule will not work for more complicated formulas:

If we have this wave in radians:

$$y = h + A\, sin\, ((x * \theta) + \ \phi)$$

... where "x" is scaling the value being treated as an angle in radians, then its derivative will be:

$$y = (x * A)\ sin\, ((x * \theta) + \ \phi + 0.5\pi)$$

In other words, to obtain the derivative:

- We remove the mean level because mean levels are irrelevant to gradients.
- We multiply the amplitude by whatever is scaling the value being treated as an angle.
- We add quarter of a circle (0.5π radians) to the phase of the wave.

The rule, re-written to incorporate "t" instead of "θ", is as follows:

If we have this wave in radians:

$$y = h + A\, sin\, ((x * t) + \ \phi)$$

... where "x" is scaling the value being treated as an angle in radians, then its derivative will be:

$$y = (x * A)\ sin\, ((x * t) + \ \phi + 0.5\pi)$$

Simple examples

The derivative of:
“$y = \sin (2\pi * t)$”
... where “t” is being scaled by 2π, is:
“$y = 2\pi \sin ((2\pi * t) + 0.5\pi)$”

The derivative of:
“$y = \sin (2\pi t + 0.5\pi)$”
... is:
“$y = 2\pi \sin (2\pi t + \pi)$”

The derivative of:
“$y = \sin (2\pi t + \pi)$”
... is:
“$y = 2\pi \sin (2\pi t + 1.5\pi)$”

The derivative of:
“$y = \sin (2\pi t + 1.5\pi)$”
... is:
“$y = 2\pi \sin (2\pi t + 2\pi)$”
... which is:
“$y = 2\pi \sin (2\pi t)$”

The derivative of:
“$y = \sin (2\pi t + 0.1\pi)$”
... is:
“$y = 2\pi \sin (2\pi t + 0.6\pi)$”

The derivative of:
“$y = \sin (2\pi * 5t)$”
... where “t” is being scaled by “$2\pi * 5$”, is:
“$y = (2\pi * 5) \sin ((2\pi * 5t) + 0.5\pi)$”
... which is:
“$y = 10\pi \sin ((2\pi * 5t) + 0.5\pi)$”

The derivative of:
“y = 3 + 2 sin ((2π * 4t) + 0.25π)”
... where “t” is being scaled by “2π * 4”, is:
“y = (2π * 4) * 2 sin ((2π * 4t) + 0.75π)”
... which is:
“y = 2π * 4 * 2 sin ((2π * 4t) + 0.75π)”
... which is:
“y = 16π sin ((2π * 4t) + 0.75π)”

The derivative of that result would be:
“y = (2π * 4) * 16π * sin ((2π * 4t) + 1.25π)”
... which is:
“y = 128π^2 sin ((2π * 4t) + 1.25π)”
... which is:
“y = 1,263.3094 sin ((2π * 4t) + 1.25π)”

The next derivative would be:
“y = (2π * 4) * 1,263.3094) sin ((2π * 4t) + 1.75π)”
... which is:
“y = 31,750.4273 sin ((2π * 4t) + 1.75π)”

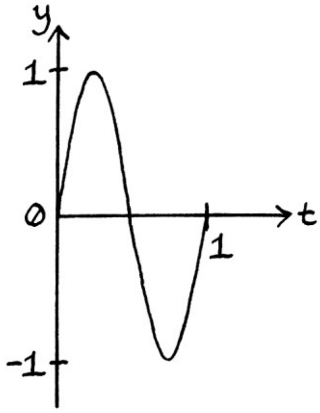
The next derivative is:
“y = (2π * 4) * 31,750.4273) sin ((2π * 4t) + 2.25π)”
... which is:
“y = 797,975.2738 sin ((2π * 4t) + 2.25π)”
... which, because there are 2π radians in a circle, is the same as:
“y = 797,975.2738 sin ((2π * 4t) + 0.25π)”

This is the same *phase* as we started with, but it is not the same wave. When the value being Sined is scaled by a frequency, or a number acting as a frequency, and we repeatedly differentiate it, we might end up with the original *phase*, but we will never end up with the original *wave*.

Integrals

Calculating the integral of a wave, where the item being Sined is scaled by a value, also requires taking into account that value.

If we start with this wave:
“y = sin (2π * 1t)”

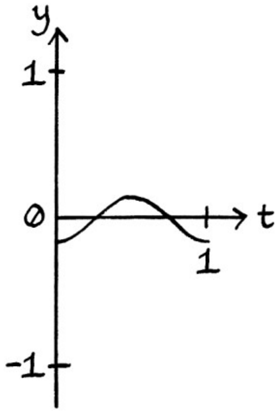


... then all of its anti-derivatives will have the shape of:

$$y = \frac{1}{2\pi} sin\ ((2\pi * 1t) - 0.5\pi)$$

... but will vary in their mean level.

They will have the following shape, but different mean levels:



We will look at a simplified rule for the integration of time-based waves. This will work with waves where "θ" or "t" is being scaled by an unvarying value, and where the phase is a fixed value. It will not work for more complicated situations where, for example, we might have "θ^2", "t^2" or a phase of "ϕt".

If we have this wave in radians:

$$y = A\,sin\;((x * \theta) + \phi)$$

... where "x" is scaling "θ", then its integral will be:

$$C\; + \frac{A}{x}\,sin\;((x * \theta) + \phi - 0.5\pi)$$

The same rule with "t" instead of "θ" is essentially identical in meaning. If we have this wave in radians:

$$y = A\,sin\;((x * t) + \phi)$$

... where "x" is scaling "t", then its integral will be:

$$C\; + \frac{A}{x}\,sin\;((x * t) + \phi - 0.5\pi)$$

In other words:

- We add on the constant "C" to indicate that there are countless answers all with different mean levels.
- We divide the amplitude by the value scaling the angle or the time.
- We subtract 0.5π from the phase.

We can write the integration rules more mathematically as so:

$$\int A\,sin\;((x *\; \theta) + \phi)\;d\theta = C + \frac{A}{x}\,sin\;((x * \theta) + \phi - 0.5\pi)$$

... or in terms of "t":

$$\int A\,sin\;((x * t) + \phi)\;dt = C + \frac{A}{x}\,sin\;((x * t) + \phi - 0.5\pi)$$

As an example, if we have this wave:
"y = sin (2πt)"
... where "t" is being scaled by "2π", then, first, we can make things easier by writing its amplitude as 1:
"y = 1 sin (2πt)"

Its integral will be:

$$y = C \ + \ \frac{1}{2\pi} \, sin \, (2\pi t - 0.5\pi)$$

... which is:

$$y = C \ + \ 0.1592 \, sin \, (2\pi t - 0.5\pi)$$

[Note that some mathematicians or calculus calculators might give a result in a much more complicated form that means exactly the same thing. That result might be more succinct from a mathematical point of view, but it will be less connected to the general layout of a wave formula that we have been using in this book. It pays to be aware of this.]

If we have this wave:
"y = 3 sin ((2π * 4t) + 1.5π)"
... where "t" is being scaled by "2π * 4", then the integral will be:

$$y = C \ + \ \frac{3}{2\pi * 4} \, sin \, ((2\pi * 4t) + 1.5\pi - 0.5\pi)$$

... which is:

$$y = C \ + \ \frac{3}{8\pi} \, sin \, ((2\pi * 4t) + \pi)$$

... which is:

$$y = C \ + \ 0.1194 \, sin \, ((2\pi * 4t) + \pi)$$

The integral of:
"y = 5 sin ((2π * 1.5t) + 0.125π)"
... where "t" is being scaled by "2π * 1.5", is:

$$y = C + \frac{5}{2\pi * 1.5} sin\,((2\pi * 1.5t) - 0.375\pi)$$

This ends up as:
"y = C + 0.5305 sin ((2π * 1.5t) – 0.375π)"
... which, if we give it a positive phase, is:
"y = C + 0.5305 sin ((2π * 1.5t) + 1.625π)"

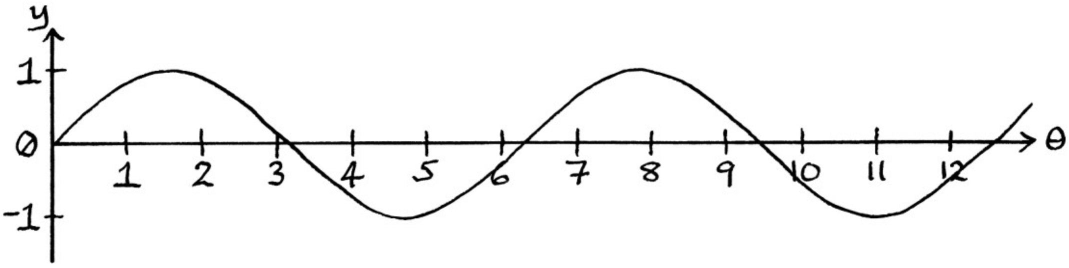
Many textbooks and maths teachers will try to give results with zero phases, so do not be surprised to see results in terms of Cosine with zero phase, or as Sine or Cosine with negative amplitudes and zero phases. It can take some thought to realise that such results mean exactly the same thing.

Note that if the wave already has a mean level before we integrate it, then its integral will not be a pure wave, but instead a pure wave added to a sloping line. We will ignore such things, as they will not be relevant to this book.

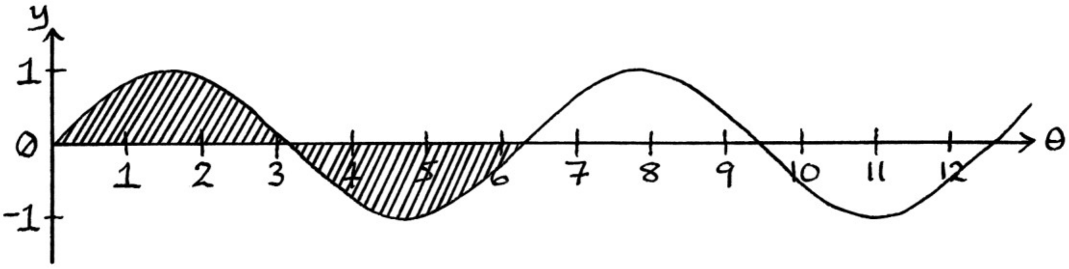
Integration to calculate areas of waves

As with other formulas, we can use integration to calculate the area between the curve of a wave and the x-axis (or the θ-axis or the t-axis).

As an example, we will look at the wave "y = sin θ" in radians:



Without needing to do any maths, there are several observations we can make about the area between the curve and the θ-axis. First, the area over one cycle will be zero. This is because the part of the curve above the θ-axis in one cycle is matched by an identical part beneath the θ-axis. Whatever the area of the positive half of one cycle, it is equal to the area of the negative half. When added together, they will add up to zero.



This idea is similar to the concept of mean level. Over one cycle of this wave, the average y-axis value will be zero. This is because the sum of the second half of y-axis values will be the negative of the sum of the first half of y-axis values. Mean level and area are closely connected.

Another observation we can make is that the area for the whole, or a portion, of one particular cycle will be the same as that of any other cycle. This is because all the cycles are the same.

If the wave continues forever, another observation we can make is that the total area above the θ-axis and the total area below the θ-axis will be equal. In other words, there is zero total area between the curve and the θ-axis. This is related to how the mean level of an eternally long wave is zero.

Areas of parts of angle-based waves

We will use the integral of a wave to calculate some areas. We will use the wave:
"y = sin θ"
... where Sine is working in radians.

Its integral is:
"y = C + sin (θ – 0.5π)"

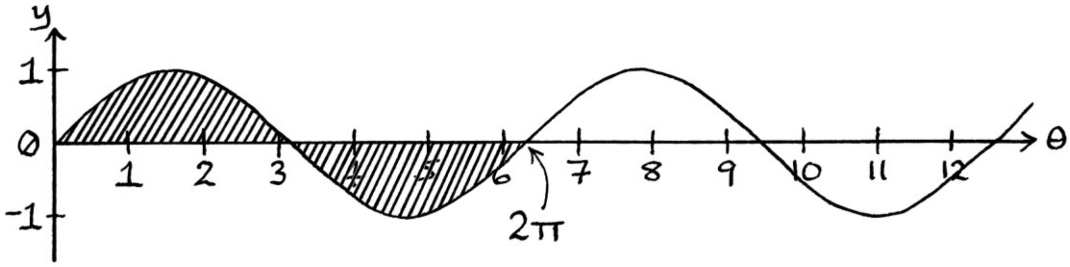
We can also express this as:

$$\int \sin\theta \ d\theta = C \ + \ sin\ (\theta - \ 0.5\pi)$$

We will calculate the area from 0 radians up to 2π radians. Therefore, we will be solving this:

$$\int_0^{2\pi} \sin\theta \, d\theta$$

On the graph, we are calculating this area:



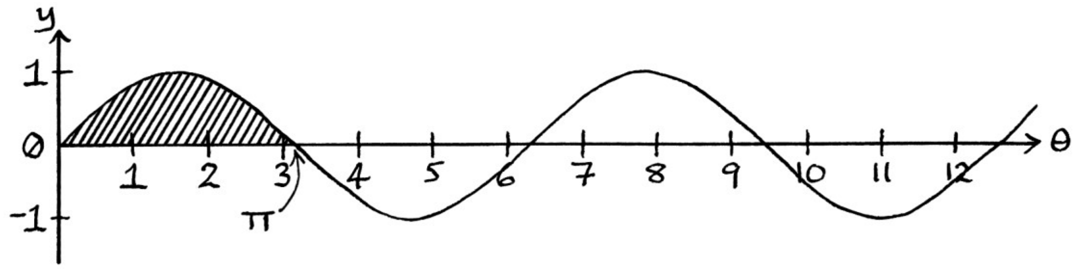
First, we put 2π into the integral formula, and we will have this:
"y = C + sin (2π – 0.5π)"
... which is:
"y = C + sin (1.5π)"
... which is:
"y = C + −1"
... or:
"y = C – 1"

Then, we put zero into the integral formula, and we have this:
“y = C + sin (0 – 0.5π)”
... which is:
“y = C + sin (1.5π)”
... which is:
“y = C + −1”
... or:
“y = C – 1”

Then, we subtract the second result from the first result:
C – 1 – (C – 1)
... which is:
C – 1 – C + 1
... which is:
0.

Therefore, the total area over one cycle is zero square units, which is something that we could tell without needing to use integration.

Now, we will find the area of the first half of the first cycle. This is the area from 0 radians to π radians.



As this is the same wave, we will use the same integral as before:
“y = C + sin (θ – 0.5π)”.

The calculation we will be solving is this:

$$\int_{0}^{\pi} \sin\theta \, d\theta$$

First, we substitute “π” for “θ” in the indefinite integral formula:
“y = C + sin (π – 0.5π)”
... which is:
“y = C + sin (0.5π)”
... which is:
“y = C + 1”

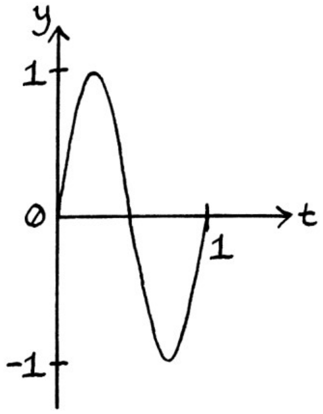
Then, we substitute 0 for “θ” in the indefinite integral formula:
“y = C + sin (0 – 0.5π)”
... which is:
“y = C + sin (1.5π)”
... which is:
“y = C + −1”
... which is:
“y = C – 1”

Then, we subtract the second calculation from the first calculation:
C + 1 – (C – 1)
= C + 1 – C + 1
= 2

Therefore, the area from θ = 0 to θ = π is exactly 2 square units. Although it might be surprising to end up with a result that is exactly 2 square units, the reasons are based around how we are dealing with “2π” and “π” in the graph and the calculations. [2π ÷ π = 2]

Areas of parts of time-based waves

We will now use the integral of a time-based wave to calculate some areas. We will focus on “y = sin (2π * 1t)”



The integral of this is:

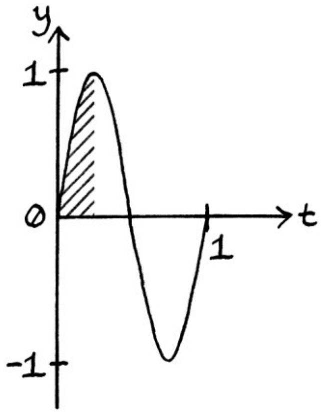
$$y = C + \frac{1}{2\pi} \, sin \left((2\pi * 1t) - 0.5\pi \right)$$

We will calculate the area from 0 seconds up to 0.25 seconds. We will be calculating this:

$$\int_{0}^{0.25} \sin\ (2\pi * 1t)\, dt$$

Note that because the formula involves time in seconds, the 0.25 and the 0 in the integral symbol refer to times in seconds, and not angles in radians. [Although strictly speaking, the results would be identical if they were angles, as the Sine function does not care about the nature of the values being Sined.]

The area we will be calculating is this:



First, we put 0.25 into the integral formula, and we have this:

$$y = C + \frac{1}{2\pi} \, sin \left((2\pi * 1 * 0.25) - 0.5\pi \right)$$

... which is:

$$y = C + \frac{1}{2\pi} \, sin \left(0.5\pi - 0.5\pi \right)$$

... which is:

$$y = C + \frac{1}{2\pi} \, sin \left(0 \right)$$

... which is:

$$y = C$$

Next, we put zero into the integral formula, and we have this:

$$y = C\ +\ \frac{1}{2\pi}\ sin\ ((2\pi * 1 * 0) - 0.5\pi)$$

... which is:

$$y = C\ +\ \frac{1}{2\pi}\ sin\ (0 - 0.5\pi)$$

... which is:

$$y = C\ +\ \frac{1}{2\pi}\ sin\ (-0.5\pi)$$

... which is:

$$y = C\ +\ \frac{1}{2\pi} * -1$$

... which is:

$$y = C\ +\ -0.1592$$

... which is:

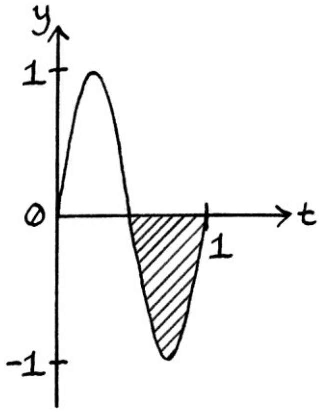
$$y = C\ -\ 0.1592$$

Then, we subtract the second result from the first result:
C – (C – 0.1592)
... which is:
C – C + 0.1592
... which is:
0.1592

Therefore, the area between the curve and the t-axis between t = 0 and t = 0.25 is 0.1592 square units. [As we dealing with both time and general units, the units of area have a slightly more complicated meaning than if both axes were just general units.]

Another example

As another example, we will use the same wave but calculate the area from t = 0.5 to t = 1.



We will be solving the following (where Sine is working in radians):

$$\int_{0.5}^{1} \sin(2\pi * 1t)\, dt$$

The integral formula is, as before:

$$y = C\ + \frac{1}{2\pi}\, sin\, ((2\pi * 1t) - 0.5\pi)$$

First, we put 1 into the integral formula, and we have this:

$$y = C\ + \frac{1}{2\pi}\, sin\, ((2\pi * 1 * 1) - 0.5\pi)$$

... which is:

$$y = C\ + \frac{1}{2\pi}\, sin\, ((2\pi - 0.5\pi)$$

... which is:

$$y = C\ +\ \frac{1}{2\pi}\ sin\ (1.5\pi)$$

... which is:

$$y = C\ +\ \frac{1}{2\pi} * -1$$

... which is:

$$y = C\ -\ \frac{1}{2\pi}$$

... which is:

$$y = C\ -\ 0.1592$$

Then, we put 0.5 into the integral formula to produce this:

$$y = C\ +\ \frac{1}{2\pi}\ sin\ ((2\pi * 1 * 0.5) - 0.5\pi)$$

... which is:

$$y = C\ +\ \frac{1}{2\pi}\ sin\ (\pi - 0.5\pi)$$

... which is:

$$y = C\ +\ \frac{1}{2\pi}\ sin\ (0.5\pi)$$

... which is:

$$y = C\ +\ 0.1592$$

We subtract the second result from the first result:
C – 0.1592 – (C + 0.1592)
= C – 0.1592 – C – 0.1592
= −0.3183 square units.

Therefore, the area between the curve and the t-axis between t = 0.5 and t = 1 is −0.3182 square units.

Calculations without formulas

Knowing how to calculate the integral or derivative of a signal using formulas is mainly useful in theoretical maths. In everyday signal processing, we would be unlikely to be working on a signal for which we were given the formula. In this section, we will go through the steps for finding the derivative and an anti-derivative of a wave without using calculus.

For the examples in this section, we will look at the wave of "y = sin (2πt)", where Sine is working in radians. We will take y-axis values from the Sine wave at intervals of 0.01 seconds, and then use those points to calculate the derivative and one of the anti-derivatives. The following table shows a list of the y-axis values from our Sine wave, rounded to four significant decimal places. The first column shows the time in seconds. The second column shows the y-axis value at that time.

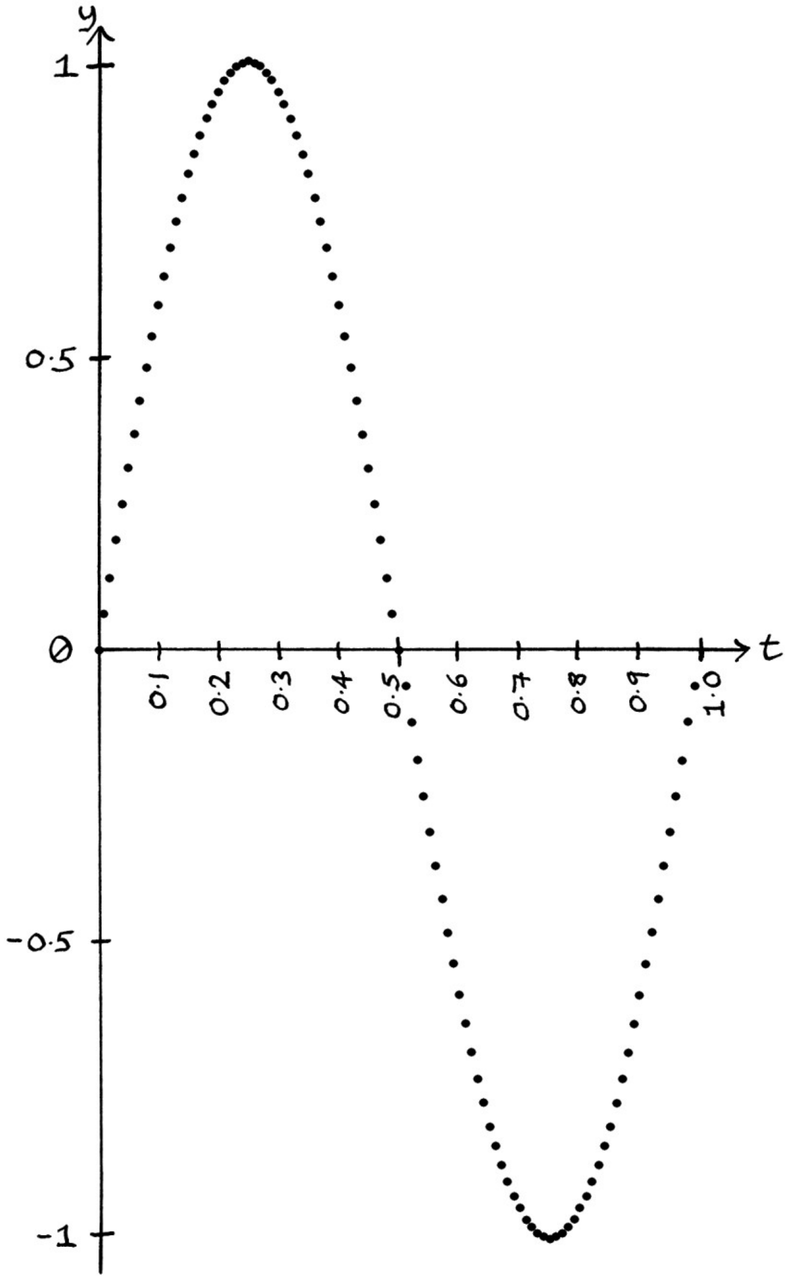
Time (seconds)	**y-axis value of the curve at this time**
0.00	0
0.01	0.06279
0.02	0.1253
0.03	0.1874
0.04	0.2487
0.05	0.3090
0.06	0.3681
0.07	0.4258
0.08	0.4818
0.09	0.5358
0.10	0.5878
0.11	0.6374
0.12	0.6845
0.13	0.7290
0.14	0.7705
0.15	0.8090
0.16	0.8443
0.17	0.8763

0.18	0.9048
0.19	0.9298
0.20	0.9511
0.21	0.9685
0.22	0.9823
0.23	0.9921
0.24	0.9980
0.25	1
0.26	0.9980
0.27	0.9921
0.28	0.9823
0.29	0.9686
0.30	0.9511
0.31	0.9298
0.32	0.9048
0.33	0.8763
0.34	0.8443
0.35	0.8090
0.36	0.7705
0.37	0.7290
0.38	0.6845
0.39	0.6374
0.40	0.5878
0.41	0.5358
0.42	0.4818
0.43	0.4258
0.44	0.3681
0.45	0.3090
0.46	0.2487
0.47	0.1874
0.48	0.1253
0.49	0.06279
0.50	0
0.51	−0.06279
0.52	−0.1253
0.53	−0.1874
0.54	−0.2487
0.55	−0.3090
0.56	−0.3681
0.57	−0.4258
0.58	−0.4818

0.59	−0.5358
0.60	−0.5878
0.61	−0.6374
0.62	−0.6845
0.63	−0.7290
0.64	−0.7705
0.65	−0.8090
0.66	−0.8443
0.67	−0.8763
0.68	−0.9048
0.69	−0.9298
0.70	−0.9511
0.71	−0.9686
0.72	−0.9823
0.73	−0.9921
0.74	−0.9980
0.75	−1
0.76	−0.9980
0.77	−0.9921
0.78	−0.9823
0.79	−0.9686
0.80	−0.9511
0.81	−0.9298
0.82	−0.9048
0.83	−0.8763
0.84	−0.8443
0.85	−0.8090
0.86	−0.7705
0.87	−0.7290
0.88	−0.6845
0.89	−0.6374
0.90	−0.5878
0.91	−0.5358
0.92	−0.4818
0.93	−0.4258
0.94	−0.3681
0.95	−0.3090
0.96	−0.2487
0.97	−0.1874
0.98	−0.1253
0.99	−0.06279

The next entry, which is the start of the next cycle, would be:
1.00 0

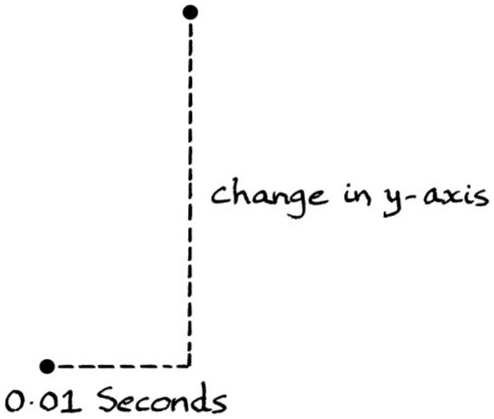
When drawn on a graph with the axes to the same scale as each other, the points look like this:



If we joined up the points, we would have the full wave.

Calculating the derivative signal

If we want to calculate the derivative signal, we have to calculate the gradient between every single value in the list. Outside of calculus, when we calculate the gradient of the hypotenuse of a right-angled triangle, we divide the opposite side by the adjacent side. If we are dealing with a line that is not the hypotenuse of a right-angled triangle, we treat it in the same way, and divide the change in the y-axis by the change in the x-axis. We divide the "rise by the tread". For our list of y-axis values, we do the same – for any one y-axis value, we divide the rise since the previous one, and then divide it by the length of time that has passed between the two. The time between each value is 0.01 seconds. Therefore, we divide the y-axis rise by 0.01.



The rule for calculating the gradient between any two points in the list is to calculate the difference and divide it by 0.01. To put this mathematically:

$g = (y_1 - y_2) \div 0.01$
... where:

- "g" is the gradient that we are calculating. It will become the y-axis value of the derivative wave at this time.
- "y_1" is a y-axis value from the original wave.
- "y_2" is the next y-axis value from the original wave.
- 0.01 is the time in seconds between each y-axis value from the original wave, and also the time between each y-axis value of the future derivative wave.

As we saw earlier, the first few y-axis values of our original Sine wave are as so:
0
0.06279
0.1253
0.1874
0.2487
0.3090
0.1253
0.1874
...

The first thing to notice is that we can only start calculating the gradient on the second y-axis value because there is nothing with which to compare the first one. This means that our resulting derivative wave will consist of 99 points and not 100.

To calculate the first gradient, we calculate:
(0.06279 – 0) ÷ 0.01 = 6.279. This will be the first point on our derivative wave. The point will be at t = 0 and y = 6.279.

The next gradient is:
(0.1253 – 0.06279) ÷ 0.01 = 6.251. This is the second point on our derivative wave. The point is at t = 0.01 and y = 6.251.

The next gradient is:
(0.1874 – 0.1253) ÷ 0.01 = 6.21. This is the third point on our derivative wave. It is at t = 0.02 and y = 6.21.

We continue like this until we reach the end of the list. One moment of interest is at 0.25 and 0.26 seconds, when the original y-axis values are:
1
0.9980

We still divide the increase in the y-axis value by the length of time between them. However, as the increase is negative, this means that the gradient becomes negative at this point:
(0.9980 – 1) ÷ 0.01 = −0.2

The last two y-axis values of the cycle (at t = 0.98 and t = 0.99 seconds) are:
−0.1253
−0.06279

The gradient between them is:
(−0.06279 – −0.1253) ÷ 0.01 = 6.251.

The y-axis value of −0.06279 was the 100th y-axis value in our list, but the gradient between it and the 99th y-axis value is the 99th gradient. There are 100 points in our original wave, but only 99 points in our derivative wave. As each gradient point required two values from the list, the derivative wave has one point less than the original wave. The derivative points are all between the original points.

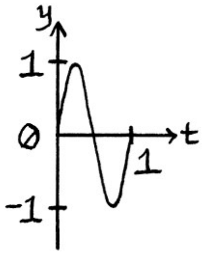
The start of a table showing the original y-axis values and the gradients between them is as so:

Time (seconds)	**y-axis value**	**Gradient between these values**
0.00	0	
		(0.06279 – 0) ÷ 0.01 = 6.279
0.01	0.06279	
		(0.1253 – 0.06279) ÷ 0.01 = 6.251
0.02	0.1253	
		(0.1874 – 0.1253) ÷ 0.01 = 6.21
0.03	0.1874	
		(0.2487 – 0.1874) ÷ 0.01 = 6.13
0.04	0.2487	
		(0.3090 – 0.2487) ÷ 0.01 = 6.03
0.05	0.3090	
		(0.3681 – 0.3090) ÷ 0.01 = 5.91
0.06	0.3681	

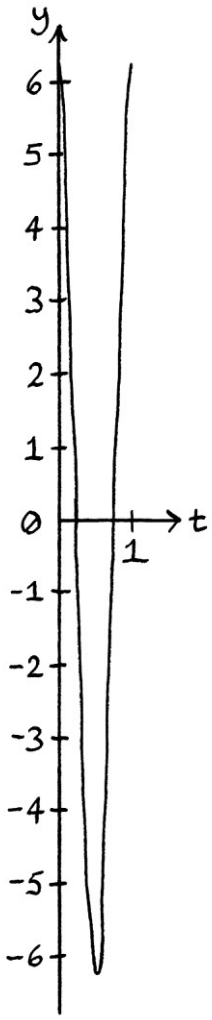
... and so on.

We could plot these points on a graph, join them up, and we would end up with the derivative wave. We would find out that the derivative wave has an amplitude of 6.279. This means that it reaches much further up and down the y-axis than our original wave did. This makes it much harder to draw on a graph if the y-axis and t-axis have the same sized units.

To demonstrate the difference in sizes, here is one cycle of the *original* wave with the dots connected, drawn to a smaller scale:

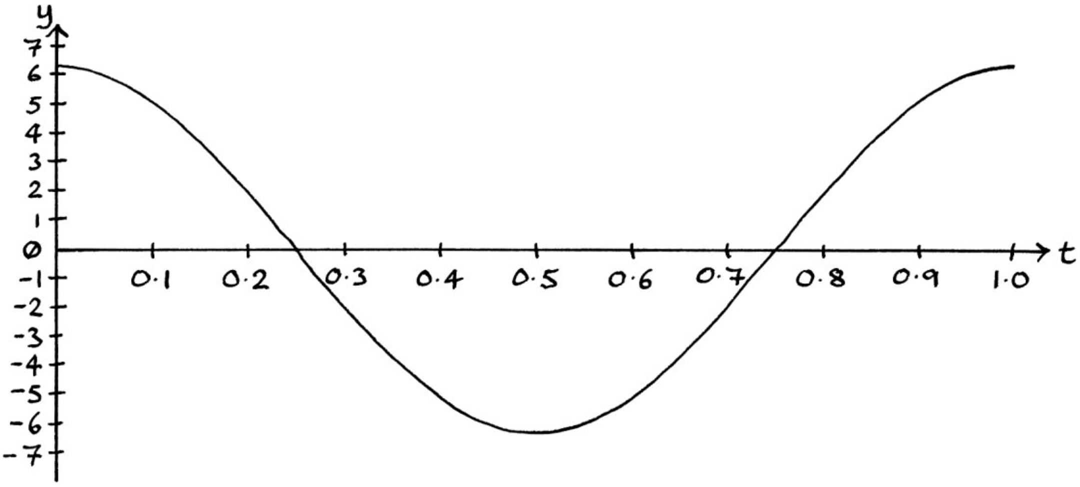


Here is the derivative wave created from joining up the points we just calculated, drawn to the same scale:



It is much harder to see the curve in the above picture.

The calculated curve is clearer if we draw the t-axis to a different scale to the y-axis, in which case, if the points are joined up, it looks like this:



Using our knowledge of derivatives, we know that the derivative wave *should* be:
"y = 2π sin ((2π * t) + 0.5π)"
... which we could also phrase as a Cosine wave as so:
"y = 2π cos (2π * t)"

We can test that our method worked by checking some values. For example, the first gradient we calculated was 6.279. The first y-axis value of the wave:
"y = 2π sin ((2π * t) + 0.5π)"
... is 2π (when t = 0), which is 6.2832. Given that we were only using 100 y-axis values per second and we rounded each of them to four significant decimal places, the first gradient was a good approximation. If we had used more y-axis values, we would have had more accurate results.

This method of calculating the derivative signal works for any given signal, whether it is a wave, a straight line, or any curve of any nature. If you can program arrays in any programming language, it is not difficult to automate this process. [You could also do this in a spreadsheet program such as Microsoft Excel.]

Calculating an anti-derivative

We will now calculate one of the anti-derivatives of our original Sine wave using the list of y-axis values. To calculate an anti-derivative signal for our Sine wave, we have to calculate which two y-axis values would have a gradient equal to each y-axis value in our list.

Our original list of Sine wave y-axis values began with these points:
0
0.06279
0.1253
0.1874
0.2487
0.3090
0.3681
0.4258
... and so on.

We will start with the first value, which is zero. We need to find the two points that between them have a gradient of zero. Of course, there is an infinite number of points for which this is true. For example, the gradient between the points y = 7 and y = 7 is zero, as is the gradient between the points y = 10,000,000 and y = 10,000,000. Supposing the first y-axis value were a value other than zero, we would still have an infinite number of possibilities for the two points that between them would have that gradient. As we are finding an anti-derivative, and all the anti-derivatives have the same curve, but start at different y-axis values, it does not actually matter which two values we start with. The resulting curve's shape will always be the same, and therefore, its derivative will be the same. From the point of view of having an anti-derivative wave graph that fits on a piece of paper, it is easiest to start with the first y-axis value of the anti-derivative as being the arbitrary value of zero. We then use that as a starting point. [If we want to, we can change the mean level of the anti-derivative wave after we have calculated it, and it will still remain an anti-derivative wave for the original formula.]

If the first point on the anti-derivative wave is y = 0, and we want the gradient from the first to the second point to be 0, then the second point on the anti derivative wave must be zero too. The first two points on the anti-derivative wave are, therefore:
y = 0
... and:
y = 0.

The next y-axis value in our original list of Sine wave values is y = 0.06279. Therefore, we want the gradient from the latest anti-derivative point to the next to be equal to 0.06279. As the latest anti-derivative point was 0, to achieve a gradient of 0.06279, the next anti-derivative point must be higher by an amount equal to 0.06279 over 1 unit (1 second). As the interval of time between each point is 0.01 seconds, the next anti-derivative point will be: 0.06279 ÷ 100 = 0.0006279. The step from 0 to 0.0006279 over 0.01 seconds has a gradient of 0.06279.

Our first three points on the anti-derivative wave are now:
y = 0
y = 0
y = 0.0006279

The next y-axis value on our original Sine wave is 0.1253. Therefore, the line from the third point to the fourth point on our anti-derivative wave will have a gradient of 0.1253. If it has this gradient, then it must rise at a rate of 0.1253 y-axis units over 1 t-axis unit, which is 0.1253 y-axis units per second. As the fourth point is 0.01 seconds from the third point, this means it must be 0.1253 ÷ 100 = 0.001253 units higher than 0.0006279, which is 0.001881. Our list of points on the anti-derivative wave is now:
y = 0
y = 0
y = 0.0006279
y = 0.001881

The next y-axis value from the original Sine wave is 0.1874. Therefore, on the anti-derivative wave, we will have a line at a gradient of 0.1874 from 0.0018799. This means the next point on the anti-derivative wave will be 0.1874 ÷ 100 = 0.001874 units higher. Therefore, it will be at 0.001881 + 0.001874 = 0.003755.

Our list of points on the anti-derivative wave is now:
y = 0
y = 0
y = 0.0006279
y = 0.001881
y = 0.003755

We continue in this way until we reach the end of our list of y-axis values.

The formula to work out each consecutive anti-derivative value is:

$a_2 = a_1 + (y * 0.01)$
... where:

- "a_2" is the y-axis value of the anti-derivative point that we are calculating.
- "a_1" is the y-axis value of the anti-derivative point that we have just calculated. If we are starting, we set this, arbitrarily, to zero.
- "y" is the current y-axis value from the original signal, which in this case, is our original Sine wave.
- 0.01 is the time in seconds between each y-axis value from the original signal, and also the time between each y-axis value for the anti-derivative signal.

The formula expressed as words is:

"The next anti-derivative value will be the previous anti-derivative value added to 'the next value from the original Sine wave multiplied by 0.01'."

As an example of this in use, the next value from the Sine wave is 0.2487. Therefore, we calculate:

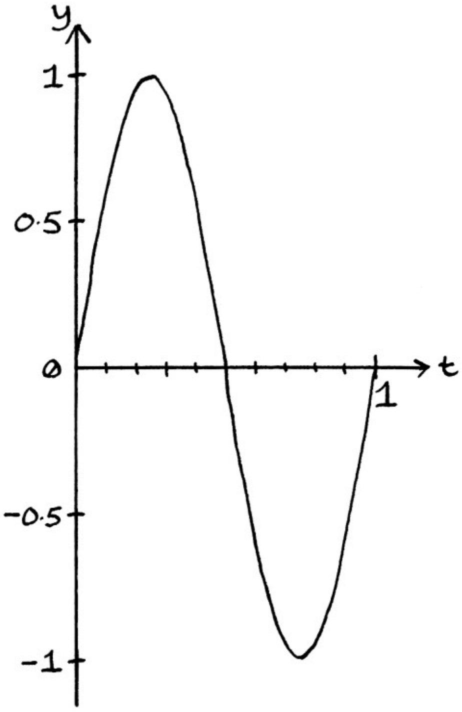
0.003755 + (0.2487 * 0.01) = 0.006242.

Therefore, 0.006242 is the next anti-derivative value.

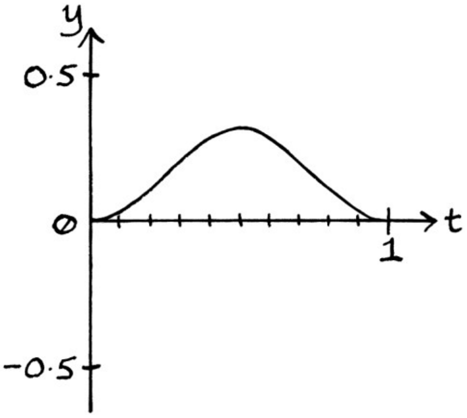
It is worth noting that because the original Sine wave values were only listed to four significant decimal places, and because every anti-derivative value is dependent on its predecessor, our results will become less and less accurate as we go through the list. Ideally, we would have many more decimal places in all our calculations. Having said that, in this particular example, our calculated results so far are exactly what they should be, but rounded to 4 significant decimal places. The method to calculate the values works perfectly, but it is only as good as the data with which we have been supplied.

We will end up with 101 values in the anti-derivative wave, even though the original wave only had 100 values. This is because the anti-derivative wave has values between each value in the original wave, as well as a starting value of 0. If we found the derivative of the anti-derivative, we would find the gradient between each point, and so end up with exactly 100 values.

If our original Sine wave (with the points joined up) is drawn to this scale:



... then our calculated anti-derivative wave (with the points also joined up) will look like this:



[It is not centred around y = 0 because of the way we calculated the first y-axis value as y = 0, and based all the later calculations on that. If we wanted, we could re-centre it around y = 0, and it would still be an anti-derivative wave for our original wave.]

Supposing we had had more accuracy in our original readings, our list of y-axis values for the calculated anti-derivative wave would have been:
0
0
0.00062791
0.00188124
0.00375505
0.00624195
0.00933212
... and so on.

For interest's sake, we will calculate the gradient between each of these points, and so retrieve the original Sine wave.

The gradient between 0 and 0 is 0.

The gradient between 0 and 0.00062791 is the rise divided by the amount of time it takes for that rise, so is:
0.00062791 ÷ 0.01 = 0.062791.

The gradient between 0.00062791 and 0.00188124 is the rise divided by the amount of time it takes for that rise, so is:
(0.00188124 – 0.0006279) ÷ 0.01 = 0.125334

The gradient between 0.00188124 and 0.00375505 is:
(0.00375505 – 0.00188124) ÷ 0.01 = 0.187381

All of these results match the y-axis values of the Sine wave that we started with (but with slightly more accuracy), which shows that everything works correctly.

Calculating the area

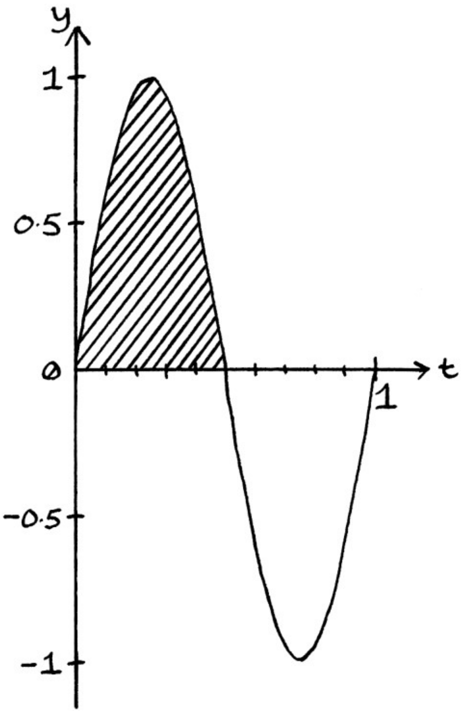
We can use our list of Sine wave y-axis values to calculate the area between the curve and the t-axis.

As we saw before, the y-axis values of our original Sine wave are as so:

Time (seconds)	**y-axis value of the curve at this time**
0.00	0
0.01	0.06279
0.02	0.1253
0.03	0.1874
0.04	0.2487
0.05	0.3090
0.06	0.3681
0.07	0.4258

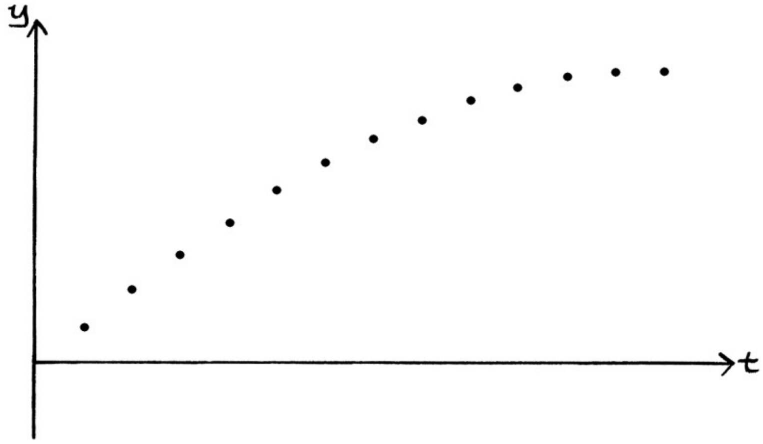
... and so on.

We will calculate the area from t = 0 seconds to t = 0.49 seconds.

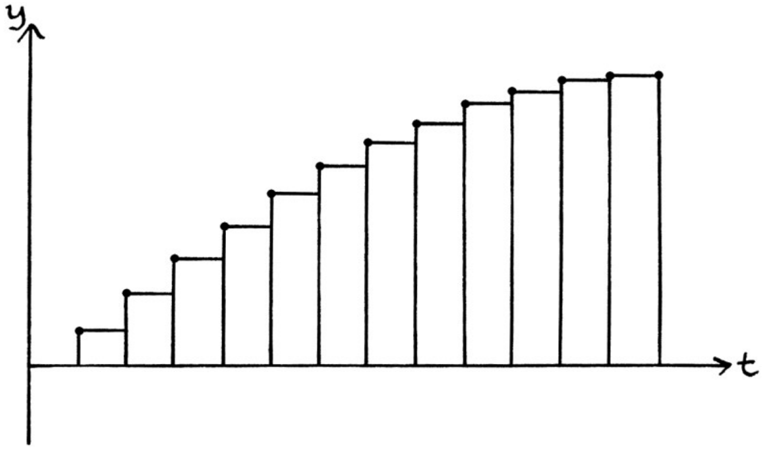


Calculating the area is actually very easy as we just calculate the area of narrow rectangles and add them all up. In fact, with the data that we have been given, this is the only way that we can calculate the area.

If the start of our series of points looked like this:

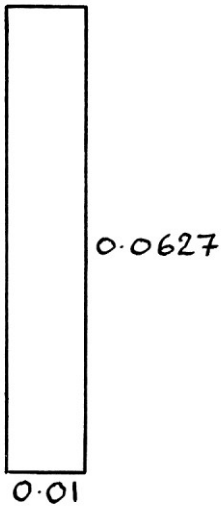


... then we would be calculating the areas of these rectangles:



For our actual data, the first rectangle has a height of zero units and a base that is 0.01 seconds long. Its area is 0 square units.

The second rectangle has a height of 0.06279 units and a base that is 0.01 seconds long.



The area of this rectangle is 0.06279 * 0.01 = 0.0006279 square units.

Our third rectangle has a height of 0.1253 units and a base that is 0.01 seconds long. The area of this rectangle is 0.1253 * 0.01 = 0.001253 square units.

Our fourth rectangle has a height of 0.1874 units and a base of 0.01 seconds. Its area is 0.1874 * 0.01 = 0.001874.

We continue going through the list of values, turning each one into a rectangle until we reach the y-axis value at 0.49 seconds, which is the fiftieth value (because we started at t = 0). That rectangle has an area of 0.06279 * 0.01 = 0.0006279 square units. [This is the same as the first rectangle because the first half of the Sine wave is symmetrical.]

The list around this time looks like this:

time	**y-axis value**
0.47	0.1874
0.48	0.1253
0.49	0.06279

Now, we add up all the areas that we calculated, and we will have the total area underneath the curve from t = 0 to t = 0.49 seconds. This is 0.3182 square units.

Supposing we had used more y-axis values per second, we would have had a more accurate result. We would have had more rectangles, but they would have been thinner.

Supposing we had used calculus to work out the area, we would have solved this:

$$\int_{0}^{0.5} \sin{(2\pi t)}\, dt$$

This involves calculating the following:

$$C\ +\ \frac{1}{2\pi}\, sin\ ((2\pi * 0.5) - 0.5\pi)$$

... minus:

$$C\ +\ \frac{1}{2\pi}\, sin\ ((2\pi * 0) - 0.5\pi)$$

... which is:
0.15915494 + C – (−0.15915494 + C)
... which is:
0.15915494 + C + 0.15915494 − C)
... which is:
0.3183 square units (rounded to 4 decimal places)

This shows that our manual method was reasonably accurate.

Thoughts

The methods that we used in this section work no matter what the type of signal that we have been given. The methods are not as accurate as using calculus, but if we have enough readings, they are good enough for most purposes. If we do not have the formula for a signal, then calculus cannot be used, while these methods will always work. Calculating the processes by hand helps clarify what calculus is supposed to be doing.

When we calculate the area between a curve and the x-axis with a list of y-axis values, we just add up a series of thin rectangles. The more y-axis values we use, the more rectangles there will be, and the thinner they will be. Integration to find an area is conceptually equivalent to having an infinite number of rectangles that are infinitely thin. Given that, we can now understand why the stretched letter "S" is used as the integral formula. It is suggesting a *sum* of tiny pieces. The "dx" [or "dt", "dθ", or similar] essentially means "a tiny piece of "x". Therefore, we can think of a formula such as:

$$\int_0^{100} x^3 \, dx$$

... as meaning the sum of rectangles made up of infinitesimally tiny parts of "x" from x = 0 to x = 100 for the curve of "x^3".

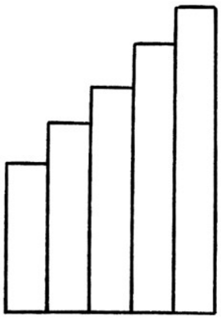
Various other calculus matters

In this section, we will look at various ideas related to calculus.

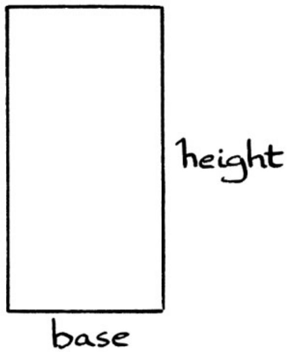
Mean levels

As we know, integration can be used to calculate the area between a wave's curve and the x-axis (or θ-axis or t-axis). The concept of this area is very similar to the concept of mean level.

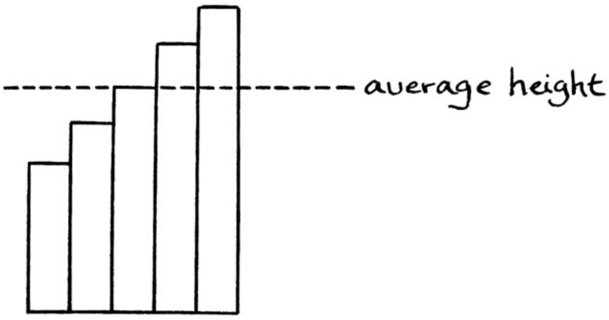
We can think of definite integrals as being the sum of very narrow rectangles. We can think of mean levels as being the average height of these rectangles.



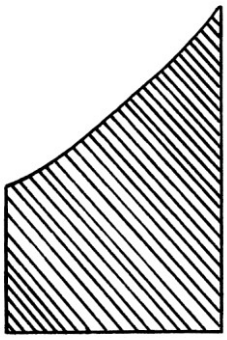
If we have the area of a rectangle, we can divide it by the length of the base, and we will have the height of the rectangle. This is because the area is defined as the length of the base multiplied by the height.

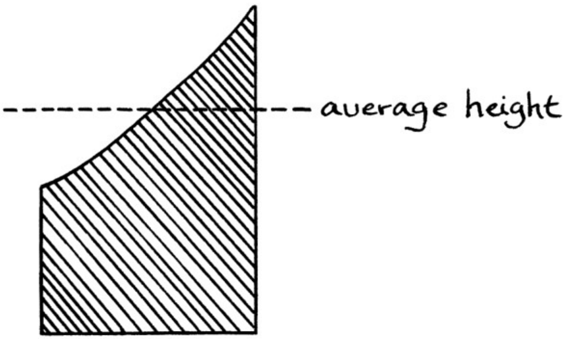


If we have the sum of the areas of many rectangles, we can divide it by the sum of the bases, and we will end up with the average height.



Similarly, if we have an area between a *curve* and the x-axis (or θ-axis or t-axis), we can divide it by the length of the x-axis under the area, and we will have the average height:





The reason that this works, even if we have a curve, is easier to understand if you imagine the area being made up of countless narrow rectangles. An area with a curved top is essentially the same as an area made up of an infinite number of rectangles.

Similarly again, if we have the area between a wave's curve and the x-axis (or θ-axis or t-axis) for one cycle, we can divide it by the x-axis length of that cycle and we will have the average height, which is another name for the mean level. The mean level is equal to the area for one cycle divided by the length of that cycle. The same is true for any periodic signal, whether it is a pure wave or not.

We can say that the mean level of one cycle of a wave or signal is equal to the definite integral for that cycle divided by the length of that cycle. Putting this slightly more mathematically, for a time-based signal, we can say that:

$$mean\ level = \frac{1}{length\ of\ cycle} * \int_{start\ of\ cycle}^{end\ of\ cycle} Our\ signal\ dt$$

We can phrase this in terms of period, with the understanding that we are starting at the beginning of the signal, as so:

$$mean\ level = \frac{1}{period} * \int_{0}^{one\ period\ later} Our\ signal\ dt$$

We can rephrase this with the symbol "T" to represent the word "period":

$$mean\ level = \frac{1}{T}\int_{0}^{T} Our\ signal\ dt$$

The above formula is saying that the mean level of our signal is equal to the area between the curve and the t-axis over the length of one period (which is the same thing as saying one cycle), all divided by the length of that period.

Derivatives and gradients

As we know, the gradient of a line shows its steepness. The gradient also shows the rate of change of the line. For example, if we have the formula "y = 2x", the gradient of the line is 2 for all values of "x". The derivative formula, which shows the gradients of every point, is "y = 2". We know that no matter what the value of "x", the gradient of "y = 2x" will be 2. Another way of thinking about the formulas is that the derivative formula shows the *rate of change* of "y = 2x". For every increase in "x", there is double that increase for "y". The value of "y" increases at a rate of twice "x".

As a less abstract example, we will think about a vehicle that travels from a particular place at a constant speed. We will say that its distance from the start over time can be given by "y = 10t", where "y" is the distance from the starting place in metres, and "t" is the time in seconds. The derivative formula is "y = 10". This means that whatever the time, the gradient of "y = 10t" will be 10. It also means that no matter what the time is, the vehicle's *speed* will be 10 metres per second. The gradient is the speed. The derivative formula shows the gradient at any moment in time, which is also the speed at any one time.

It is the case that:

- The speed of a vehicle is the derivative of its distance.
- The acceleration of a vehicle is the derivative of its speed.
- The frequency of a wave can be thought of as the derivative of a changing phase. This idea will be explained more fully in Chapter 35 of this book.

Sums of waves

As we saw earlier in this chapter, if we have a sum of items, we can find the derivative or integral by splitting the sum into parts, and working on each part in turn. We can do the same thing with waves. For example, if wanted to find the derivative of:
"y = sin (2π * 3t) + 3 sin (2π * 4t)"
... we would find the derivative of "sin (2π * 3t)" and add that to the derivative of "3 sin (2π * 4t)".

As we saw in Chapter 18, any periodic signal is equal to, or approximately equal to, a sum of pure waves of various amplitudes, phases and frequencies. This means that we could perform calculus on any periodic signal by finding its constituent waves, performing calculus on each of them in turn, and then adding up the results. How accurately the constituent waves added up to make the original signal would determine the accuracy of the calculus.

Other ideas

There are other simple calculus ideas that are useful to know, but we will explore them later in this book when they become relevant.

Calculus’s relevance to the subject of waves

In Chapter 18, I explained how to work out which waves were added to make up a periodic signal. The summary of the process as I gave it in Chapter 18 is as follows:

- Remove the mean level from the signal, if any, and make a note of it for later.

- Find the frequency of the signal, and make a list of test frequencies that are sequential integer multiples of this number.

- For every frequency in the list, multiply the signal by each of these test waves in turn, swapping “f” for the frequency being tested:
  “y = 2 sin (2π * f * t)”
  ... and:
  “y = 2 sin ((2π * f * t) + 0.5π)”
  [Note how I am now giving these formulas in terms of *radians*, when in Chapter 18, I gave the formulas in terms of *degrees*. For the method of Fourier series analysis that I described in Chapter 18, it makes no difference as long as we are consistent. For methods that involve calculus, it *does* make a difference, and we need to use radians.]

- Calculate the mean levels of the signals resulting from these multiplications. If both mean levels are zero, that frequency does not exist in the signal. Otherwise, the amplitude of the wave for that frequency will be the square root of the sum of the square of these mean levels. The phase of the wave for that frequency will be the arctan of the second mean level divided by the first mean level (and we need to check that the arctan result is the one that we want).

- Do not forget the mean level, if any, that was removed from the signal at the start.

In most maths books, the summary of the process is given much more cryptically than in my explanation. Normally, maths books describe the process in the form of integrals.

The very first step, which calculates the mean level of the original signal, would be described *in essence* as so:

$$mean\ level\ of\ the\ original\ signal = \frac{1}{T}\int_0^T "the\ original\ signal"\ dt$$

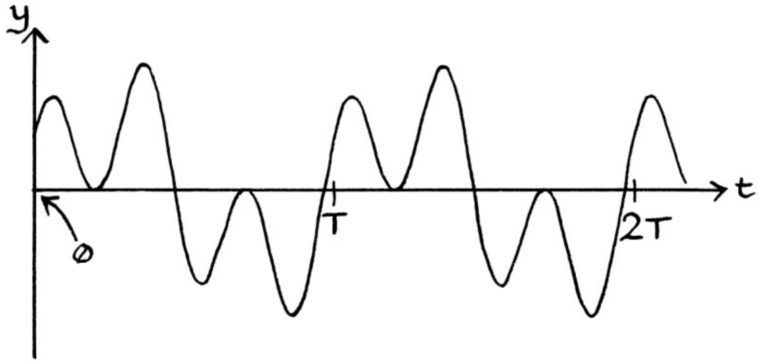
It would be more common to see it given more mathematically as so:

$$mean\ level\ of\ the\ original\ signal = \frac{1}{T}\int_0^T f(t)\ dt$$

... where:

- “T” is the period of the original signal. In other words, it is 1 divided by the frequency. It is the duration of one repetition of the signal that we are analysing. To start with, it might be confusing having an upper-case “T” representing the length of the period, and a lower-case “t” representing time, all within the same formula, but you will quickly become used to such things.
- “f(t)” is shorthand for the signal as a whole. It can be thought of as meaning “the function that we are discussing that works with time”.
- We are calculating the area between the curve and the t-axis between t = 0 and t = T. This is the same as saying we are finding the area over one cycle (one period) of the original signal.

Using the period “T” instead of the frequency can allow us to mark the start of each repetition on the time axis as being a multiple of “T”. “T” is the duration of one repetition of our signal. Using “T” makes what we are doing slightly more abstract, but means that our formulas can be used more generally:



The integral formula as a whole means that we find the area between the curve and the t-axis over one cycle (from 0 to T), *and then divide that by the length of one cycle* ("T"). This is just a roundabout way of finding the mean level, and a way that appears much more complicated and convoluted.

We can then subtract that mean level from the original signal and start analysing the result for its constituent waves. We make a list of test waves, and then set about multiplying the signal by each of these. For every frequency in the list, we multiply the signal by each of these test waves in turn.

Supposing we were testing the frequency of 3 cycles per second, we would multiply our signal by:
"y = 2 sin (2π * 3t)"
... and:
"y = 2 sin ((2π * 3t) + 0.5π)"
... and then pay attention to the resulting mean levels.

When it comes to the standard "calculus" way of portraying this task, it could be described as so:

$$first\ mean\ level = \frac{1}{T}\int_0^T Our\ Signal * (2\ sin\ (2\pi * 3t))\ dt$$

... where:

- "T" is the period of the signal that we are analysing, which is also 1 divided by the frequency.
- "Our Signal" is the signal that we are analysing (after we have removed any mean level in the first step).
- 2 sin (2π * 3t) is the first of the pair of test waves with which we multiply our signal.
- The integration means that we find the area between the curve and t-axis of the result of the multiplication of the signal and the test wave over one cycle.

The calculation as a whole finds the area between the curve and the t-axis of the result of the multiplication for the length of one cycle, and then divides that area by the length of one cycle. This has exactly the same effect as finding the mean level over one cycle.

For the second mean level for 3 cycles-per-second, we would have the following:

$$second\ mean\ level = \frac{1}{T}\int_0^T Our\ Signal * (2\ sin\ ((2\pi * 3t) + 0.5\pi))\ dt$$

This works in the same way.

After we have calculated both mean levels, we check if they are both zero, and if they are not, we use Pythagoras's theorem to calculate the amplitude, and we use arctan to calculate the phase (after checking that the result is the correct arctan result of the two possible ones).

As it is more common to write a Sine wave with a 0.5π radian phase (90-degree phase) as a Cosine wave, the second equation would more commonly be written as:

$$second\ mean\ level = \frac{1}{T}\int_0^T Our\ Signal * (2\ cos\ (2\pi * 3t))\ dt$$

As both our test waves have amplitudes of 2 units, those twos can be transferred on to the division by the period, and the calculations will have the same meaning:

$$first\ mean\ level = \frac{2}{T}\int_0^T Our\ Signal * (sin\ (2\pi * 3t))\ dt$$

$$second\ mean\ level = \frac{2}{T}\int_0^T Our\ Signal * (cos\ (2\pi * 3t))\ dt$$

We can make the formula slightly more mathematical by changing the term "Our Signal" into "f(t)", where "f(t)" is just shorthand for the signal with which we are dealing *after we have removed the original mean level*.

$$first\ mean\ level = \frac{2}{T}\int_0^T f(t) * (sin\ (2\pi * 3t))\ dt$$

$$second\ mean\ level = \frac{2}{T}\int_0^T f(t) * (cos\ (2\pi * 3t))\ dt$$

We can make the formulas relate to any pair of test frequencies, and not just ones for 3 cycles per second:

$$first\ mean\ level = \frac{2}{T}\int_0^T f(t) * (sin\ (2\pi ft))\ dt$$

$$second\ mean\ level = \frac{2}{T}\int_0^T f(t) * (cos\ (2\pi ft))\ dt$$

[This can be mildly confusing as we have “f(t)” to indicate the signal, and then “f” to indicate the frequency. Again, you will become used to such things.]

We can make the formulas even more mathematical by changing the “first mean level” and “second mean level” text into variables instead:

$$a = \frac{2}{T}\int_0^T f(t) * (sin\ (2\pi ft))\ dt$$

$$b = \frac{2}{T}\int_0^T f(t) * (cos\ (2\pi ft))\ dt$$

We now have two formulas that calculate the mean levels resulting from multiplying our signal by each of a pair of test waves. The values “a” and “b” will be the first and second mean levels, respectively, and we can use them to calculate the amplitude and phase of the constituent waves. The amplitude of the constituent wave will be the square root of “$a^2 + b^2$”; the phase will be “arctan (b / a)”.

We go through each calculation for every test frequency that we are going to use.

Before reading this chapter, these formulas would have appeared extremely complicated. By going through the steps to create them, they should be much less intimidating. They express in a mathematical form, something that is much clearer and simpler with words. However, by expressing them in this form, we can extend the ideas to much more complicated concepts, should we need to do so in the future.

Fourier series analysis using calculus

A summary of the process for finding the constituent waves in a signal when using calculus is as follows:

- Use integration to calculate the area between the signal's curve and the t-axis for one cycle of the signal. Divide this by the length of one cycle (the period) to obtain the mean level. Remove this amount from the signal, and make a note of it for later.

- Find the frequency of the signal, and make a list of test frequencies that are sequential integer multiples of this number.

- For every frequency in the list, multiply the signal by each of these test waves in turn, swapping "f" for the frequency being tested:
  "y = 2 sin (2π * f * t)"
  ... and:
  "y = 2 cos (2π * f * t)"

- Use integration to calculate the area between the *first* resulting signal's curve and the t-axis for one cycle. Divide this by the length of one cycle to obtain the first mean level.

- Use integration to calculate the area between the *second* resulting signal's curve and the t-axis for one cycle. Divide this by the length of one cycle to obtain the second mean level.

- If both mean levels are zero, that frequency does not exist in the signal. Otherwise, the amplitude of the wave for that frequency will be the square root of the sum of the square of these mean levels. The phase of the wave for that frequency will be the arctan of the second mean level divided by the first mean level (and we need to check that the arctan result is the one that we want).

- Do not forget the mean level, if any, that was removed from the signal at the start.

Using calculus to discover the constituent waves requires that we know the formula of the signal being analysed. If the formula is known, it is possible to calculate the areas using integration, and any results will be perfect. Otherwise, if we still wanted to proceed in the spirit of calculus, some other method would be needed to calculate the areas. However, if any effort is going to be spent calculating areas, it might be quicker just to calculate the mean levels directly instead.

Descriptive (synthesis) formulas

In Chapter 18, we saw a variety of formulas that expressed that any periodic signal can be said to be made up of the sum of waves with various amplitudes and phases, and different frequencies. All the formulas were of two basic types:

$$f(t) = \sum_{n=0}^{\infty} a_n \, sin \, ((2\pi * f_n * t) + \phi_n)$$

... or:

$$f(t) = \sum_{n=0}^{\infty} a_n \, sin \, ((2\pi * f * n * t) + \phi_n)$$

From this chapter onwards, we will count with the letter “k” instead of the letter “n” to avoid confusion when we deal with discrete signals later in this book. Therefore, we will rewrite the formulas as:

$$f(t) = \sum_{k=0}^{\infty} a_k \, sin \, ((2\pi * f_k * t) + \phi_k)$$

... and:

$$f(t) = \sum_{k=0}^{\infty} a_k \, sin \, ((2\pi * f * k * t) + \phi_k)$$

In the first type of formula, the amplitude, frequency and phase are unspecified, but there is the unmentioned rule that all the frequencies will be different from each other, and that one of the frequencies will be zero to act as a mean level. In the second type of formula, the amplitude and phase are unspecified, but the “f” specifically refers to the frequency of the signal being analysed. Therefore, by running through every possible value of "k", we will have integer multiples of the frequency of the signal as well as one wave with a frequency of zero cycles per second to act as the mean level. This formula implies that many of the amplitudes will be zero. The first formula can be interpreted as only counting waves with frequencies that are relevant, in which case, none of the amplitudes need be zero [although that is not actually stated in the formula.]

Both formulas and their variations only *describe* how a periodic signal is made up of the sum of pure waves. They do not give any information as to how to calculate those waves. For this reason, it would make sense to call such a formula a "descriptive formula". In practice, it is more common for a "descriptive formula" to be called a "synthesis formula". The formulas show how a signal can be *synthesised*, in the sense that they show how parts can be combined to create a signal. The term "descriptive" is itself more descriptive of what such formulas are doing, whereas the term "synthesis" requires an explanation for it to make sense. However, to fit in with typical maths teaching, I will refer to such formulas as "synthesis" formulas.

On their own, these "synthesis" formulas are not particularly useful – first, they are not strictly true, and second, they rely on knowledge that is not expressed in the formulas.

The full process

When a synthesis formula is paired with the integral formulas, we have a description and a process together. In such situations, the synthesis formula becomes much more useful [although, it still relies on unstated information for it to make sense.]

In the following synthesis formula, we are using "c" instead of "a" for the amplitude – this means that we can continue using "a" and "b" for the results of the integrals.

Note also how we use "f(t)" to refer to the original signal with its mean level, and "g(t)" to refer to that signal after the mean level has been removed. The symbol "f(t)" means "the function we are discussing that is working with time", so it essentially means "the signal we are working with". The symbol "g(t)" is just a variation of "f(t)". It still means "the signal we are working with", but it refers to a different signal to that referred to by "f(t)".

The synthesis formula, together with how the values can be calculated, is as follows:

$$f(t) = \sum_{k=0}^{\infty} c_k \, sin\, ((2\pi * f * k * t) + \phi_k)$$

... where:

“f(t)” is the periodic signal that we are analysing.

“k” is always an integer, and increases in steps of 1.

“c” is the amplitude of each wave.

“f” is the frequency of the signal we are analysing. Some people would call it the “fundamental frequency”.

When “k” is zero:

$$c_0 = \frac{1}{T}\int_0^T f(t)\ dt$$

$$\phi_0 = 0.5\pi$$

For all the other occasions, when “k” is non-zero:

$$g(t) = f(t) - \ c_0 \, sin\, ((2\pi * f * 0 * t) + \phi_0)$$

[In other words, “g(t)” is the signal we are analysing *after we have removed the original mean level*. The original mean level is the wave created when “k” is zero.]

$$a_k = \frac{2}{T}\int_0^T g(t) * (sin\ (2\pi f k t))\, dt$$

$$b_k = \frac{2}{T}\int_0^T g(t) * (cos\ (2\pi f k t))\, dt$$

$$c_k = \sqrt{(a_k)^2 + (b_k)^2}$$

$$\phi_k = arctan\ (b_k\ /\ a_k)$$

Notes:

- The first wave, as portrayed with symbols, will have the formula:
  "$c_0 \sin ((2\pi * f * 0 * t) + \phi_0)$"
  It will act as the mean level for the whole signal. Therefore, it needs to have a frequency of zero cycles per second, a phase of 0.5π radians, and an amplitude equal to the mean level of the entire signal. The mean level of the signal, and therefore, also the amplitude of the wave, are calculated as the area under the curve for one period divided by the length of time for one period. A more succinct formula would be:
  "$c_0 \sin (0.5\pi)$"
  [If it were a Cosine wave, the phase would be zero radians.]
  The wave's characteristics need to be calculated separately to the rest of the analysis procedure.

- The waves after the first wave (when "k" is 1 or more) will be based on the signal after it has had its mean level removed. Whereas we used "f(t)" to refer to the signal before the mean level had been removed, we are using "g(t)" to refer to the signal *after* the mean level has been removed. Without this distinction, the process would calculate incorrect results. We could also have indicated that it is a different signal by putting a suffix on the "f", such as "$f_2(t)$", but this could be confusing when we have so many suffixes already.

- The amplitude of each wave is represented by the symbol "c_k" (such as c_0, c_1, c_2, c_3 and so on). For every wave except the first wave (the mean level), we need to calculate "c_k" by calculating "a_k" and "b_k" first.

- When we use arctan to calculate the phase, we need to check the result to make sure it is the one we want out of the two possible ones. [If we were automating the process by writing a computer program, we could use the programming command "atan2" to remove any ambiguity.]

- If "a_k" or "b_k" are zero, we do not need to use arctan because we know what the phase would be by thinking about the position of the phase point on a circle ("a_k" would be the x-axis value and "b_k" would be the y-axis value). We could also think about a triangle with either no height or no width ("a_k" would be the adjacent side and "b_k" would be the opposite side).

- If we were automating the process by writing a computer program, we would need to make sure that we did not accidentally perform a division by zero when calculating the phase.

- [If we were automating the process with a computer program, we would be using discrete signals and discrete waves (where the signal or wave consists of a list of y-axis values at evenly spaced moments in time). In which case, it would be easier to calculate the mean levels directly instead of bothering with calculating the areas first.]

No notes

The same formulas without notes are as so:

$$f(t) = \sum_{k=0}^{\infty} c_k \, sin \, ((2\pi * f * k * t) + \phi_k)$$

... where:

$$c_0 = \frac{1}{T}\int_0^T f(t) \; dt$$

$$\phi_0 = 0.5\pi$$

$$g(t) = f(t) - \; c_0 \, sin \, ((2\pi * f * 0 * t) + \phi_0)$$

$$a_k = \frac{2}{T}\int_0^T g(t) * (sin \, (2\pi fkt)) \, dt$$

$$b_k = \frac{2}{T}\int_0^T g(t) * (cos \, (2\pi fkt)) \, dt$$

$$c_k = \sqrt{(a_k)^2 + (b_k)^2}$$

$$\phi_k = arctan \; (b_k \, / \, a_k)$$

A standard formula

There has become a sort of "traditional" group of formulas that often appears in the backs of textbooks. These usually have each wave given in terms of the sum of Sine and Cosine waves *with zero phases*. This means that we do not end up with the actual constituent waves in the signal, and the process is left unfinished.

Usually, no mention is made that the waves after k = 0 will be working with the signal after the mean level has been removed. Therefore, instead of seeing "g(t)", there is only mention of "f(t)" again.

One such standard group of formulas is as so:

$$f(t) = \sum_{k=0}^{\infty} a_k \, sin\,(2\pi * f * k * t) + b_k \, cos\,(2\pi * f * k * t)$$

... where:

$$a_0 = 0$$

$$b_0 = \frac{1}{T}\int_0^T f(t)\ dt$$

$$a_k = \frac{2}{T}\int_0^T f(t) * (sin\,(2\pi f k t))\,dt$$

$$b_k = \frac{2}{T}\int_0^T f(t) * (cos\,(2\pi f k t))\,dt$$

Note that "a_0" is zero because the mean level can be expressed using just a Cosine wave with zero phase.

It is very common to see this group of formulas or its variations. However, giving the formulas in this way without any explanation means that what the symbols represent, and what we are actually doing, cannot be deduced from just the formulas. Without already knowing what the formulas and symbols mean, there is no way to use them to analyse a signal.

Thoughts

By combining the “synthesis” (“descriptive”) formula with the process of calculating the constituent waves, we have everything we need to perform Fourier series analysis on a signal. Although the process can be more succinctly summarised with words than with formulas, the formulas are a stepping stone to more complicated methods of analysing signals.

Generally, in maths as a subject, the mean level is the “forgotten” attribute of a wave. This is most obvious in how it is often referred to by the cumbersome and inaccurate name, “the DC component”. [The term “DC component” implies we are working with electrical current.] It is much more common to see integration used to calculate areas than it is to see other maths used to calculate mean levels. If we know the formula for a signal, integration can be useful. However, sometimes even when we know the formula, or in situations when we do not know the formula, or if we are working with a discrete signal, then integration adds an unnecessary level of complexity to the task and the concept of analysing a periodic signal. Having said all of that, the standard “maths book” method of Fourier series analysis uses integration, so it pays to become used to it.

There are countless variations of the above formulas. Some are reasonably easy to understand; some are unnecessarily complicated and obscure.

Thoughts on calculus

Normally, an explanation of calculus would take up more than one chapter of a book. For the purposes of this book, I think it is better to understand the basis of calculus than to spend unnecessary time on all its variations. It is much easier to learn calculus if you have a desire to do so, as opposed to learning it for learning’s sake. If you have understood this chapter, you will be able to understand any calculus that appears in the rest of this book.

In the study of waves and signals, calculus is most often used in the form of integration to find the area under a curve as a step towards finding the mean level. Therefore, even if you do not know how to *solve* every formula you see, you will still be able to understand what a formula is intending to achieve. If you understand that a definite integral finds the area under a curve, and that, when studying signals, the usual reason for finding an area is to calculate the mean level, then you will know enough for most purposes.

Calculus is also used to find an anti-derivative signal for the purposes of frequency modulation. We will explore this idea in Chapters 37 and 38.

If you want to learn more about calculus, the two best books of which I am aware are:

- "Calculus Made Easy" by Silvanus Thompson, published in 1914 by The Macmillan Company.

- "Calculus For Dummies" by Mark Ryan, published in 2003 by Wiley Publishing, Inc.

Any mistakes in this chapter are my fault and not theirs.

Coming next

In the second part of this book, we will be looking at:

- Real-world waves such as sound and light waves
- Modulation
- Discrete signals

To download the latest version of the second part of this book, visit: www.timwarriner.com