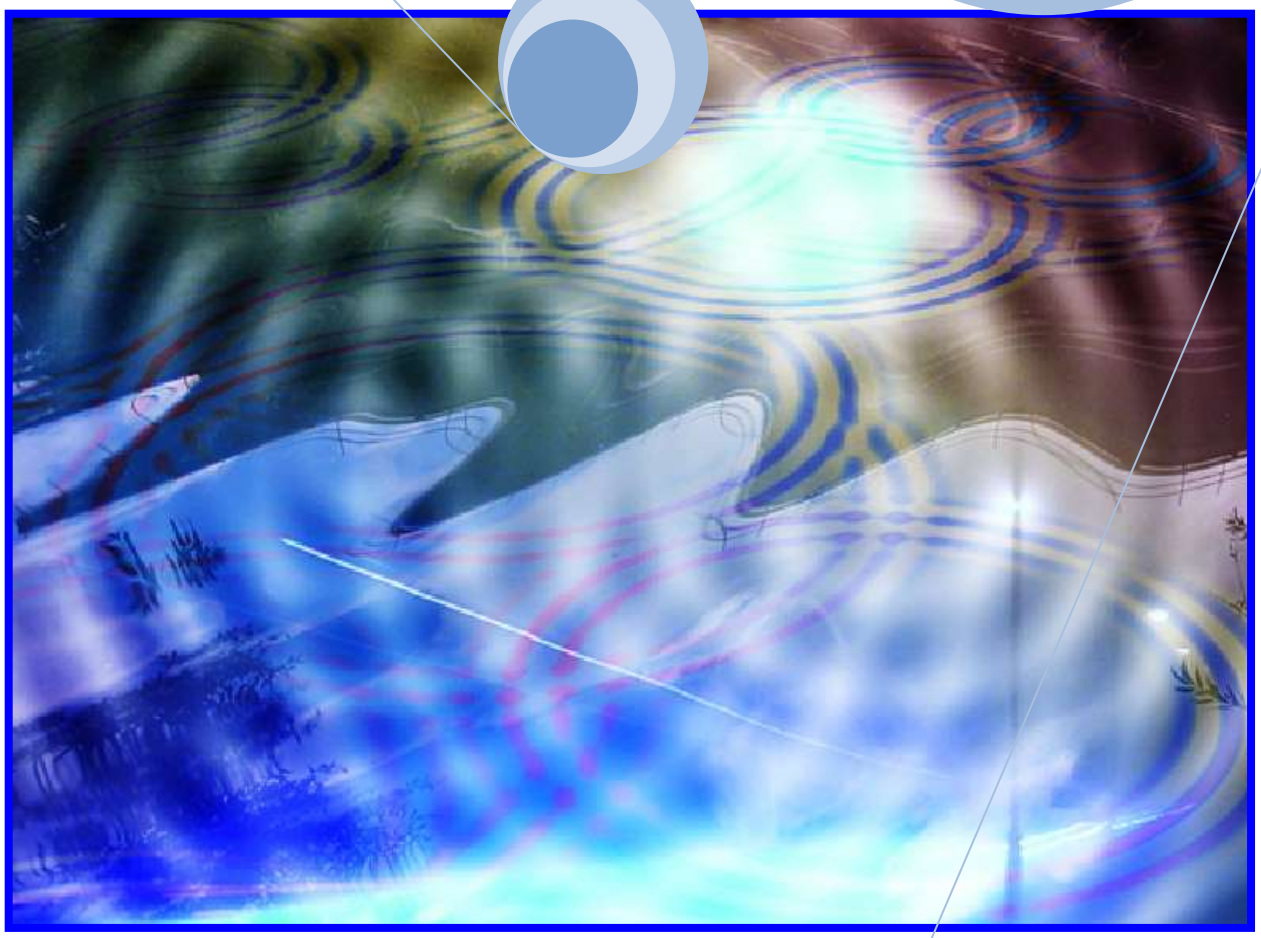


Revision Guide

Chapter 6 – Wave Behaviour

Contents

Revision Checklist

I can show my understanding of effects, ideas and relationships by describing and explaining:

how standing waves are formed by sets of wave travelling in opposite directions *e.g. by drawing diagrams to show what happens*	
how waves passing through two slits combine and interfere (superpose) to produce a wave / no-wave pattern	
what happens when waves pass through a single narrow gap (diffraction)	
how a diffraction grating works in producing a spectrum	

I can use the following words and phrases accurately when describing effects and observations:

wave, standing wave, frequency, wavelength, amplitude, phase, phasor	
path difference, interference, diffraction, superposition, coherence	

I can sketch and interpret diagrams:

illustrating the propagation of waves	
showing how waves propagate in two dimensions using Huygens' wavelets	
showing how waves that have travelled to a point by different paths combine to produce the wave amplitude at that point *i.e. by adding together the different phases of the waves, using phasors*	

I can calculate:

wavelengths, wave speeds and frequencies by using (and remembering) the formula $v = f\lambda$	
wavelengths of standing waves *e.g. in a string or a column of air*	
path differences for waves passing through double slits and diffraction gratings	
the unknown quantity when given other relevant data in using the formula $n\lambda = d \sin\theta$	

I can show my ability to make better measurements by:

measuring the wavelength of light	

I can show an appreciation of the growth and use of scientific knowledge by:

commenting on how and why ideas about the nature of light have changed	

Revision Notes

Wave characteristics

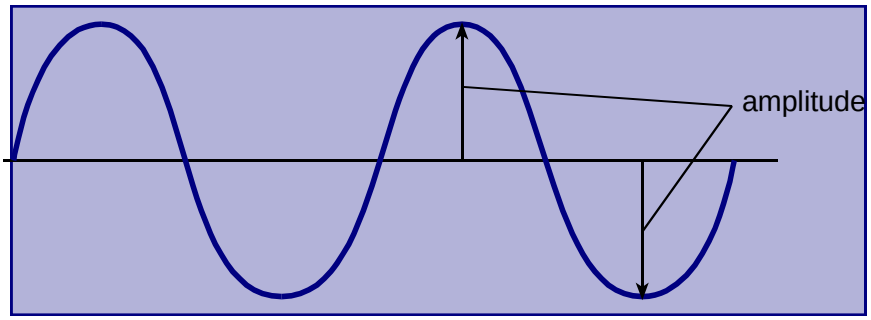
Waves are characterised by several related parameters: amplitude (how big they are); frequency (how rapidly they oscillate); wavelength (the distance over which they repeat) and their speed of travel.

Displacement

Displacement, x is the distance in metres of a point on a wave from its undisturbed (equilibrium) position. It can be positive or negative.

Amplitude

The amplitude, A of a wave at a point is the maximum displacement from its equilibrium position.



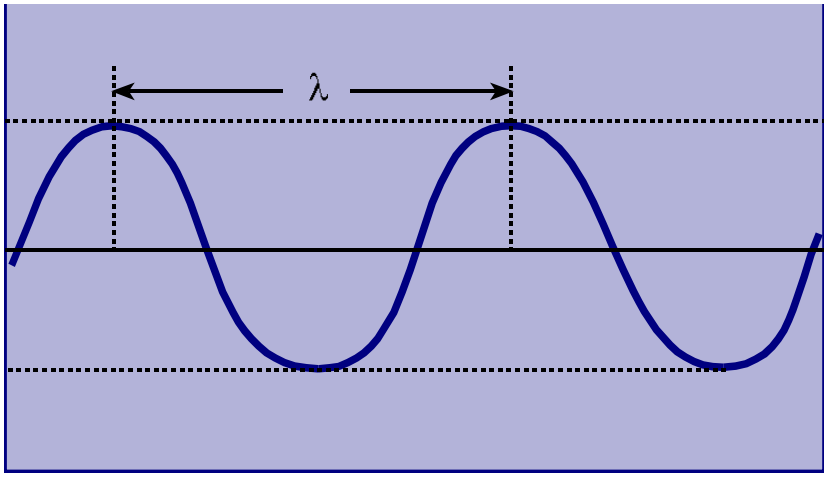
Frequency and period

The period, T of an oscillation is the time taken in seconds for one complete oscillation. The frequency, f of an oscillation is the number of complete cycles of oscillation each second. The SI unit of frequency is the hertz (Hz), equal to one complete cycle per second.

Wavelength

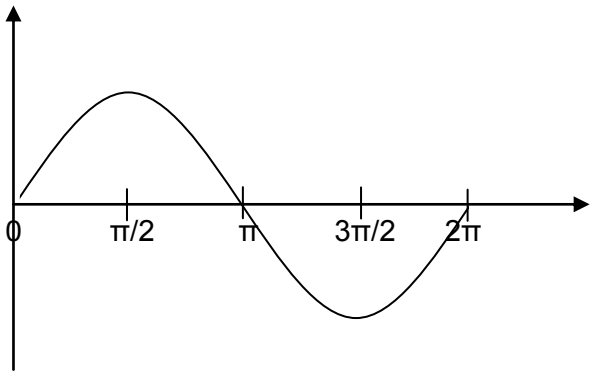
The wavelength, λ of a wave is the distance along the direction of propagation between adjacent points where the motion at a given moment is identical, for example from one wave crest to the next.

The SI unit of wavelength is the metre.



Phase

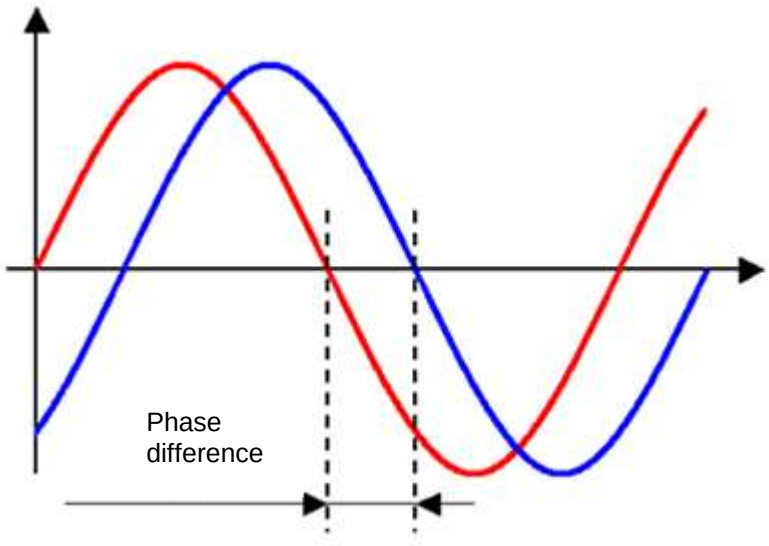
The phase of a wave is a measure of how far through an oscillation the wave is at a particular time or position.



Phase is measured in radians or degrees. For the wave above the phase at the peak is π/2 radians(90□), the phase at the trough is 3π/2 radians (270□), etc.

Phase difference

This measures how far out of step two waves (or the oscillations of two points on an individual wave) are. It is usually measured in degrees or radians. In the diagram the phase difference is about 50□



Relationships

Frequency *f* and period *T*

$$f = \frac{1}{T}$$

$$T = \frac{1}{f}$$

Frequency *f* , wavelength λ and wave speed *v*

$$v = f\lambda$$

Displacement *s* at any one point in a wave, where ϕ is the phase.

$$s = A\sin(2\pi ft + \phi)$$

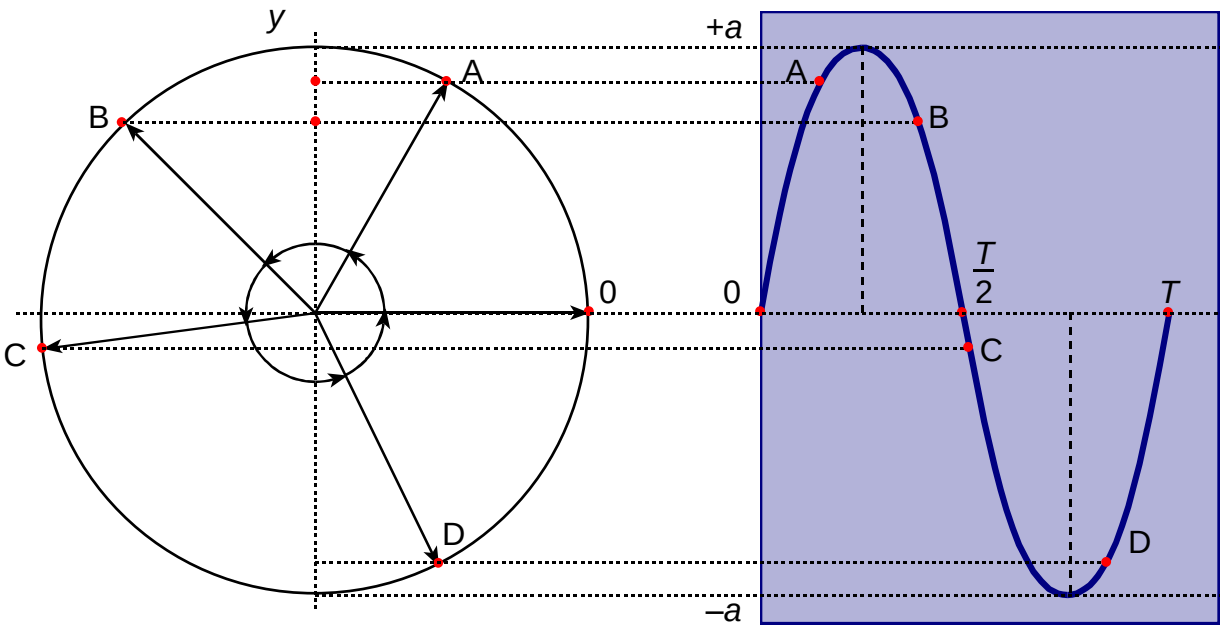
Phasors

Phasors are used to represent amplitude and phase in a wave. A phasor is a rotating arrow used to represent a sinusoidally changing quantity.

Suppose the amplitude s of a wave at a certain position is $s = a \sin(2\pi ft)$, where a is the amplitude of the wave and f is the frequency of the wave. The amplitude can be represented as the projection onto a straight line of a vector of length a rotating at constant frequency f, as shown in the diagram. The vector passes through the $+x$-axis in an anticlockwise direction at time $t = 0$ so its projection onto the y-axis at time t later is $a \sin(2\pi ft)$ since it turns through an angle $2\pi ft$ in this time.

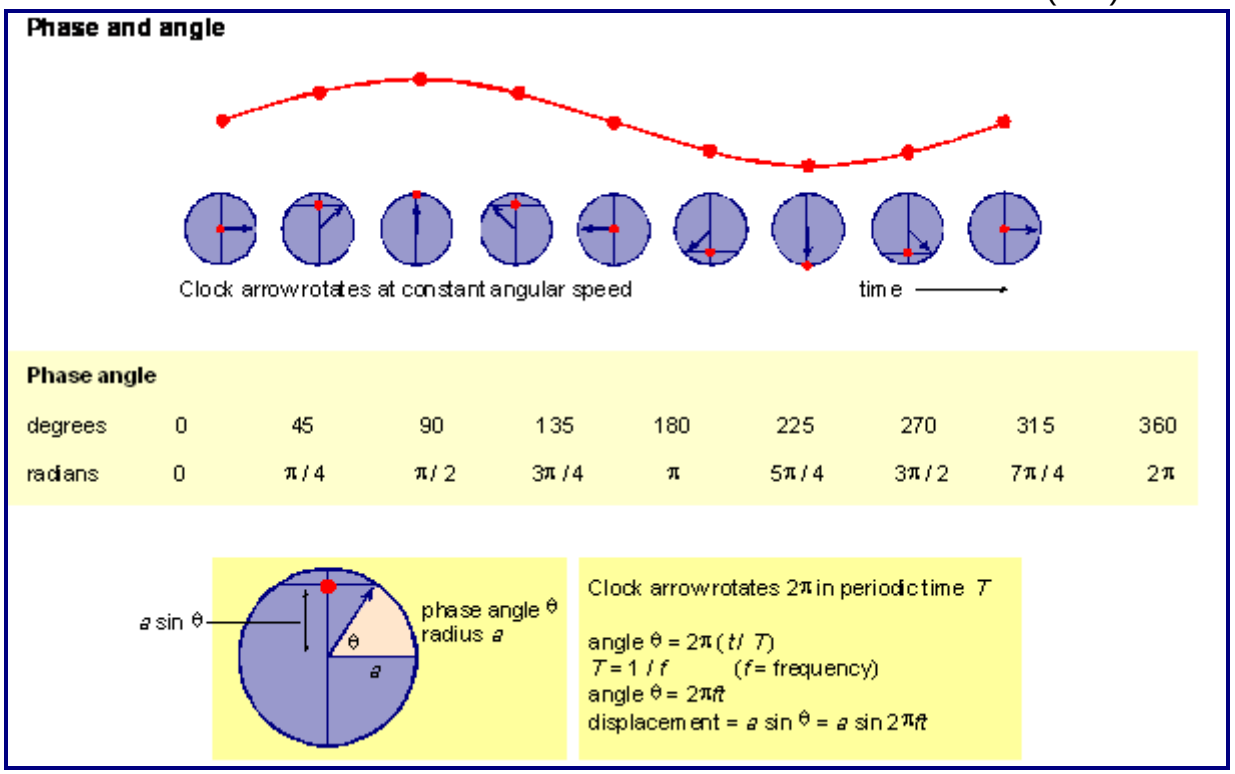
Phasors can be used to find the resultant amplitude when two or more waves superpose. The phasors for the waves at the same instant are added together 'tip to tail' to give a resultant phasor which has a length that represents the resultant amplitude. If all the phasors add together to give zero resultant, the resultant amplitude is zero at that point.

Generating a sine wave



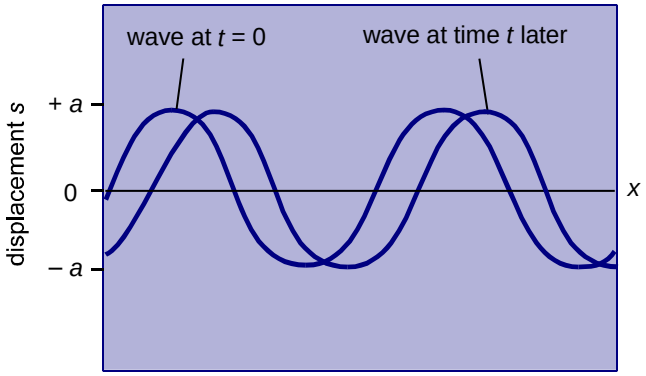
phasor rotating at constant frequency f

$s = a \sin(2\pi ft)$



Travelling waves

Travelling waves propagate through space or through a substance.

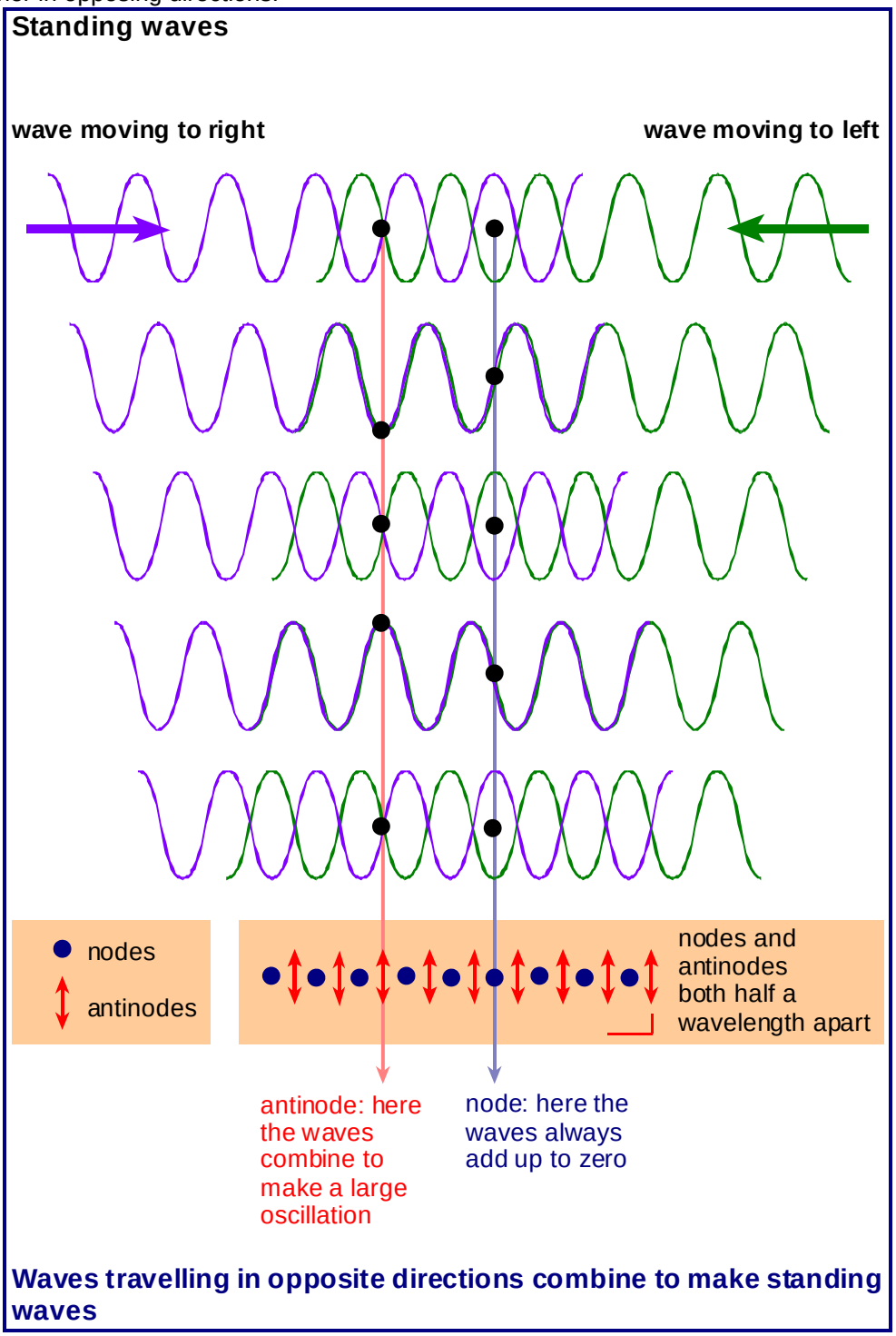


The speed of a travelling wave is related to frequency and wavelength by the formula:

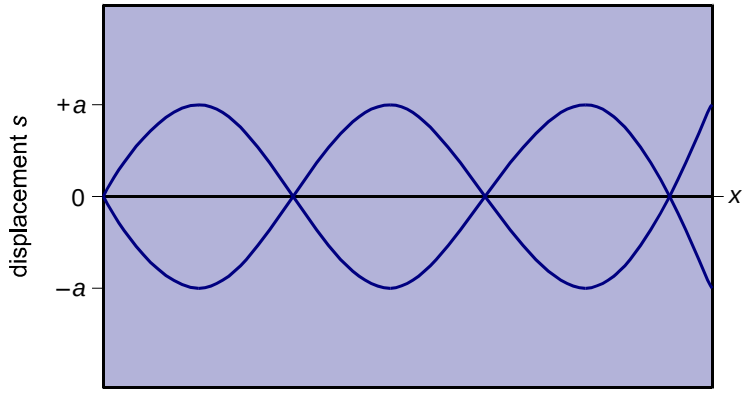
speed = frequency × wavelength or $(v = f\lambda)$

Standing Waves

Stationary or **standing** waves are produced when travelling waves of the same frequency and amplitude pass through one another in opposing directions.



A standing wave



The resultant wave has the same frequency of oscillation at all points. The wave does not travel. Its amplitude varies with position. Positions of minimum amplitude are called displacement nodes and positions of maximum amplitude are called displacement antinodes.

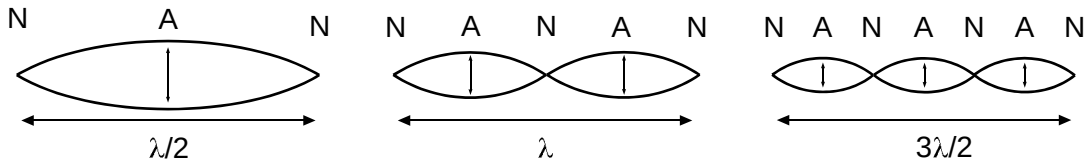
Nodes and antinodes alternate in space. The nodes (and the antinodes) are half a wavelength apart.

Standing waves are an example of **wave superposition**. The waves on a guitar string or in an organ pipe are standing waves.

Standing or stationary waves are a superposition effect, which occurs when two identical waves travelling in opposite directions add together.

With a string or spring of fixed length there will be certain frequencies where a standing wave pattern is produced as shown below. These frequencies will be those which give a whole number of half wavelengths within the fixed length. Nodes which are points of zero displacement, and antinodes are points of maximum displacement.

The first few possible patterns for transverse waves on a string or spring are shown below.



The separation of adjacent nodes or antinodes is half a wavelength. The lowest frequency (on the left) is called the fundamental. The others are whole number multiples of the fundamental frequency and are called harmonics. Because of the phase change on reflection at the fixed ends the ends are both nodes.

The wavelength can be found by measuring the node separation and doubling it. This gives a method for finding the speed of sound or electromagnetic waves using $v = f\lambda$. For example, if a standing wave is set up using 1 GHz radio waves and the node separation is found to be 0.15 m the wavelength will be 0.30 m and the wave speed is given by:

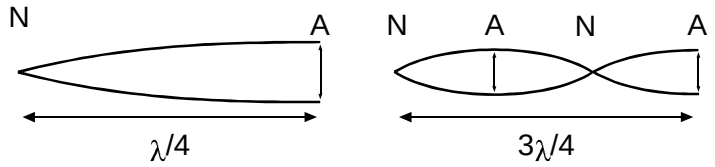
$$c = f\lambda = (1 \times 10^9 \text{ Hz}) \times 0.30 \text{ m} = \mathbf{3 \times 10^8 \text{ ms}^{-1}}$$

Standing Waves in Musical Instruments

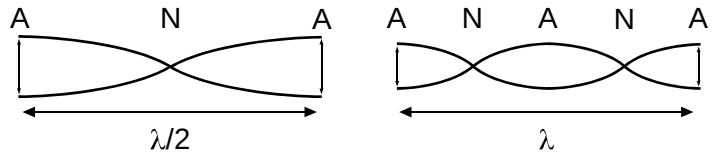
Stringed instruments such as violins and pianos produce standing waves as described above. The pitch of the fundamental and therefore of the harmonics may be increased by reducing the length of the string, by increasing the tension in it or by using a lighter string.

Wind instruments work by setting up longitudinal standing waves in tubes. Examples are organ pipes, trumpets and clarinets.

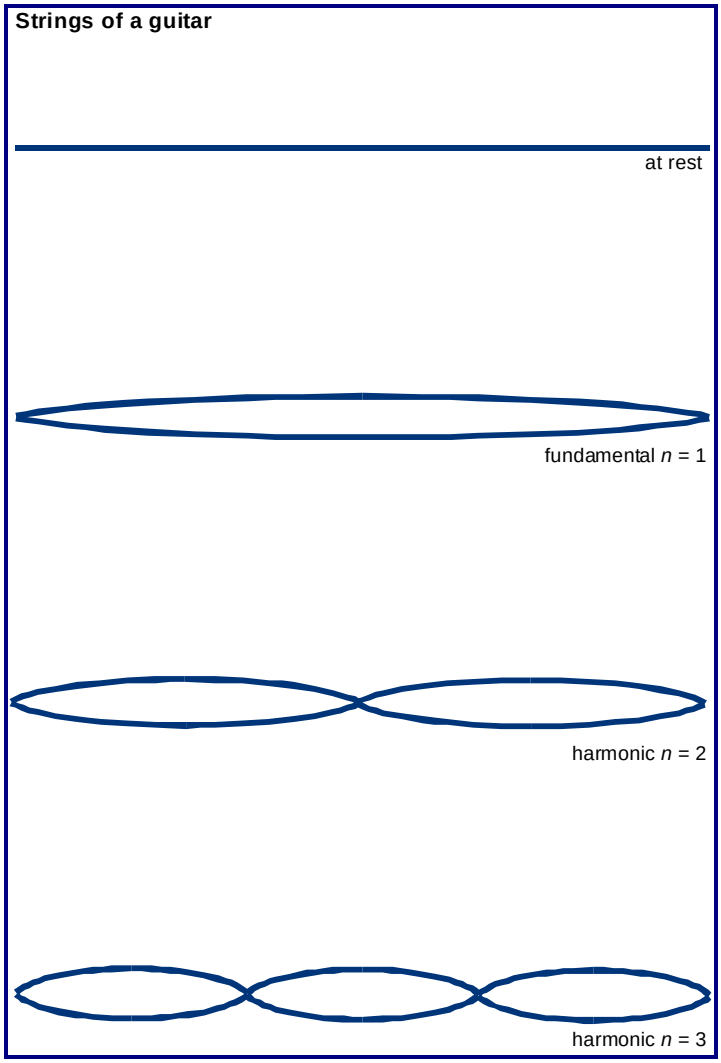
If the tube is open at one end and closed at the other there will be a node at the closed end and an antinode at the open end. At the fundamental frequency the length of the tube is therefore ¼ of a wavelength. Harmonics occur at frequencies where the length of the tube is ¾, $^5/_4$, $^7/_4$ etc wavelengths.



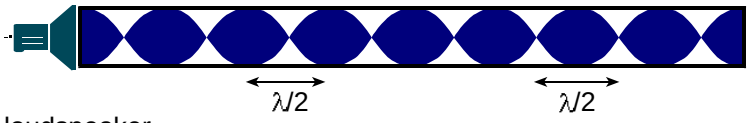
If the tube is open at both ends both ends have an antinode and the fundamental occurs when the tube length is half a wavelength. Harmonics are at whole number multiples of the fundamental frequency.



Standing waves on a guitar



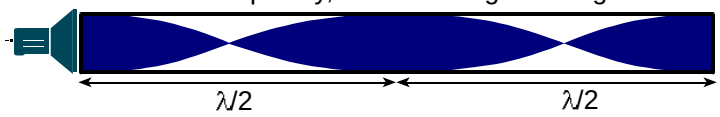
Closed pipes

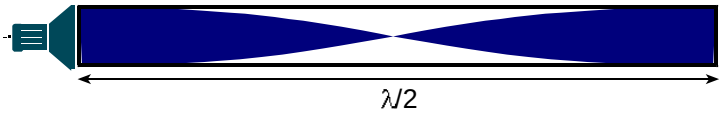


loudspeaker

A loudspeaker sends a sound into a long tube. Dust in the tube can show nodes and antinodes. Nodes are half a wavelength apart. So are antinodes. Dark colour shows maximum pressure variation and minimum motion of air (pressure antinode). Light colour shows minimum pressure variation and maximum motion of air (pressure node).

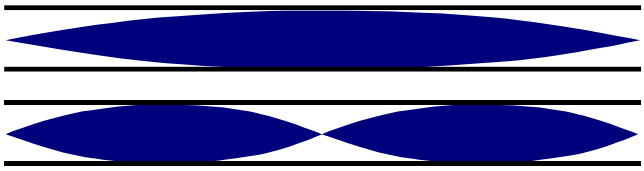
At a lower frequency, the wavelength is longer





The fundamental: The lowest frequency which can form a standing wave has wavelength equal to twice the length of the tube

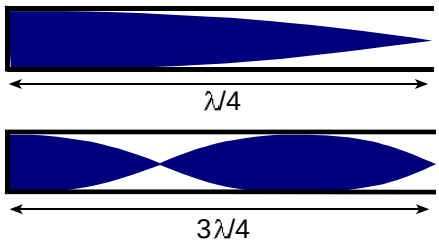
Pipes open at both ends



Sound can be reflected from an open end as well as from a closed end.

This is how open organ pipes and flutes work.

Pipes closed at one end



Pipes closed at one end are shorter, for the same note.

A clarinet is like this. An oboe is too, but with a tapered tube.

Some organ pipes are stopped at one end.

Frequencies of standing waves

	pipes open or closed at both ends strings fixed at both ends	pipes open at one end
length L	$L = n\lambda / 2$	$L = (2n–1)\ \lambda / 4$
fundamental	$f = v / 2L$	$f = v / 4L$
harmonics	$2f$ $3f$ ... nf	$3f$ $5f$ ... $(2n–1)f$

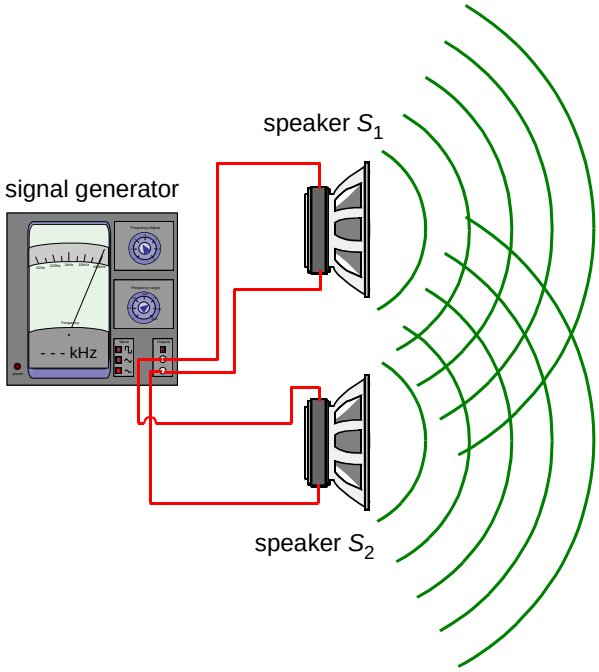
The frequencies sounded by a pipe depend on the pipe's length

Interference

When waves overlap, the resultant displacement will be equal to the sum of the individual displacements at that point and at that instant (if the waves superpose linearly).

Interference is produced if waves from two coherent sources overlap or if waves from a single source are divided and then reunited.

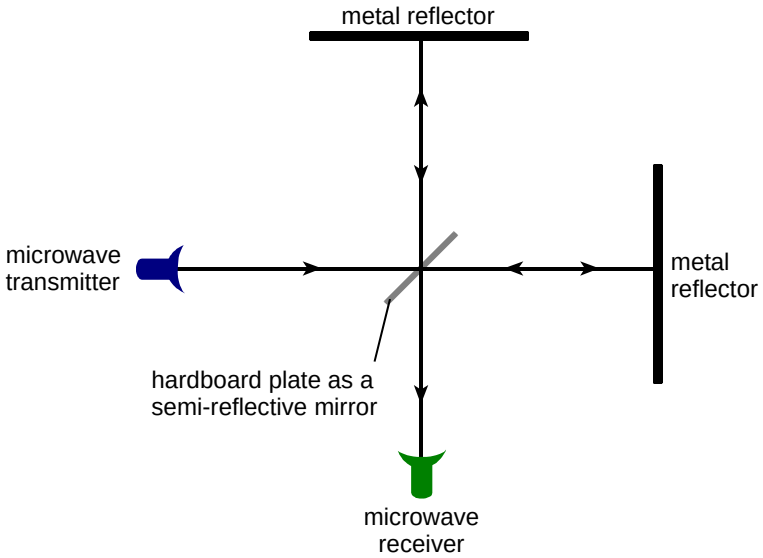
Interference of sound



Interference using sound waves can be produced by two loudspeakers connected together to an oscillator. If you move about where the waves overlap you will detect points of reinforcement (louder) and of cancellation (quieter).

Another way to produce interference is to divide the waves from one source and then recombine them. The diagram below shows this being done for microwaves, sending part of the wave along one path and part along another. The receiver gives a minimum response when the paths differ by half a wavelength.

Division of amplitude



Other examples of interference include the ‘blooming’ of camera lenses, and the colours of oil films and soap bubbles.

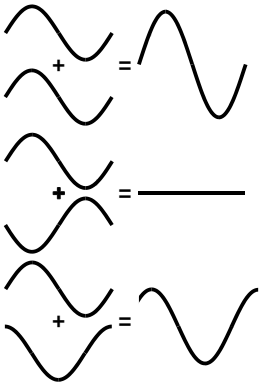
Superposition

When two or more waves arrive at the same place at the same time they combine with each other according to the principle of superposition. This states that the resultant displacement at any point is equal to the sum of the displacements of the individual waves. Note that a displacement may be negative so addition of two displacements may be smaller than either of the original displacements.

When two waves of equal amplitude and wavelength meet in phase they interfere constructively to produce a new wave of twice the amplitude but the same wavelength.

When two waves of equal amplitude and wavelength meet in antiphase (180º or π radians out of phase) they interfere destructively to cancel each other out.

When two waves of equal amplitude and wavelength meet with any other phase difference the resultant will have an intermediate amplitude.



Path difference

The path difference between two waves will determine what happens when they superpose.

If the path difference between two wavefronts is a whole number of wavelengths, the waves reinforce.

If the path difference is a whole number of wavelengths plus or minus one half of a wavelength, the waves cancel.

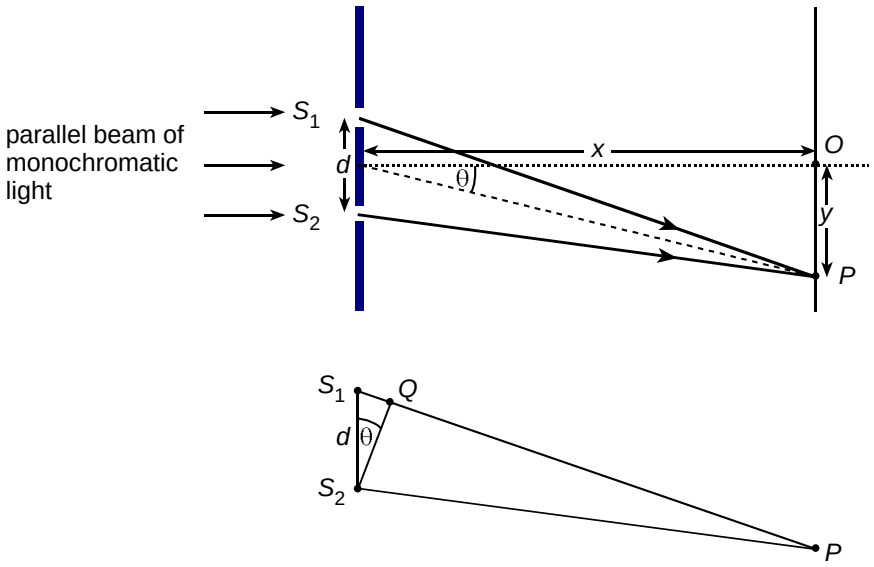
The importance of a path difference is that it introduces a time delay, so that the phases of the waves differ. It is the difference in phase that generates the superposition effects.

Double-slit interference

The double-slit experiment with light requires light from a narrow source to be observed after passing through two closely spaced slits. A pattern of alternate bright and dark fringes is observed.

In the diagram below the path difference from the two slits to a point P on the screen is equal to $d \sin\theta$, where d is the spacing between the slit centres and θ is the angle between the initial direction of the beam and the line from the centre of the slits to the point P.

Double slit interference



$QP = S_2P$

$S_1Q = S_1P - S_2P$

since $S_1Q = d\sin\theta$

then $S_1P - S_2P = d\sin\theta$ = path difference

Bright fringes: the waves arriving at P **reinforce** if the path difference is a whole number of wavelengths, i.e. $d \sin\theta = n\lambda$, where λ is the wavelength of the light used and n is an integer.

Dark fringes: the waves arriving at P **cancel** if the path difference is a whole number of wavelengths plus a half wavelength, i.e. $d \sin\theta = (n + ½)\lambda$, where λ is the wavelength of the light used and n is an integer.

The angle $\sin\theta = y / L$ where y is the distance OP to the fringe and L is the distance from the fringe to the centre of the slits. However, y is very small so L does not differ appreciably from X, the distance from the centre of the fringe pattern to the slits. Hence, for a bright fringe:

$$\sin\theta = \frac{y}{X} = \frac{n\lambda}{d}$$

which gives

$$\frac{y}{X} = \frac{n\lambda}{d}$$

Adjacent fringes have values of n equal to n and n+1. Thus the spacing between pairs of adjacent bright (or dark) fringes = $\frac{\lambda X}{d}$. Or:

$$\frac{\text{fringe width}}{\text{slit to screen distance}} = \frac{\text{wavelength}}{\text{slit separation}}$$

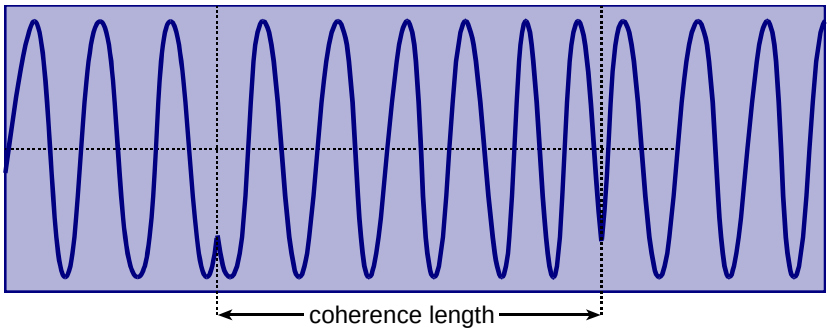
Coherence

Coherence is an essential condition for observing the interference of waves.

Two sources of waves are coherent if they emit waves with a constant phase difference. Two waves arriving at a point are said to be coherent if there is a constant phase difference between them as they pass that point.

The coherence length of light from a given source is the average length of a wave train between successive sudden phase changes.

Coherence along a wave

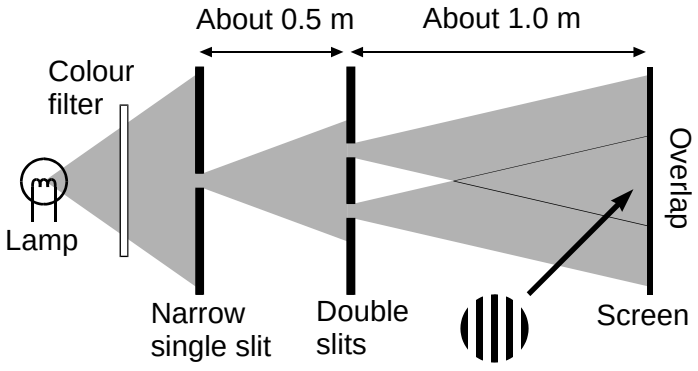


To see interference with light, the two sets of waves need to be produced from a single source, so that they can be coherent. For this, the path difference must not be larger than the coherence length of the source.

Is an essential condition for a stable interference pattern of waves. Two sources of waves are described as coherent if they emit waves with a constant phase difference (note that the waves do not necessarily have to be *in* phase). Interference cannot be observed with light - or any form of electromagnetic radiation - from two separate light sources. This is because the phase difference between light waves from the two sources changes randomly so the points of cancellation and reinforcement move about at random. The two waves are generally produced by dividing the wave from a single source into two.

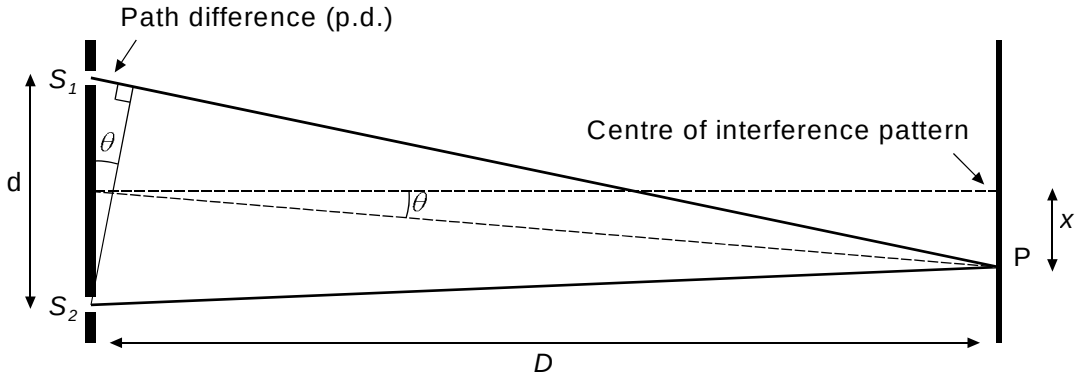
Young's Slits

This effect is named after Thomas Young who first observed it in the early nineteenth century. He observed the interference of light waves using a double slit arrangement.



A single narrow slit transmits light from a lamp and the light is allowed to fall on the double slit arrangement. ***Diffraction*** (the spreading out of waves when they pass though a narrow gap) causes the light from the two slits to overlap and bright and dark 'fringes' are seen on the screen. Dark fringes are formed where waves from one slit arrive exactly out of phase with those from the other slit. This destructive interference occurs when the path difference is an odd number of half wavelengths. Bright fringes occur where waves from the two slits arrive in phase, giving constructive interference. This occurs when the path difference for the waves from the two slits is a whole number of wavelengths.

The diagram below is not to scale as the distance *D* is very much bigger than the slit separation d.



The extra distance travelled by the light from the top slit is the path difference. By trigonometry its value is d $\sin\theta$. For a bright fringe this must be a whole number of wavelengths, $n\lambda$.

Hence $n\lambda = d \sin\theta$ at a bright fringe.

x is the distance from the centre of the pattern to the n^{th} bright fringe.

Again by trigonometry $\tan\theta = x/D$.

If θ is small $\tan\theta \cong \sin\theta$, so $\sin\theta = x/D$ and $\sin\theta = \lambda/d$ *if n=1,* so $x/D = \lambda/d$.

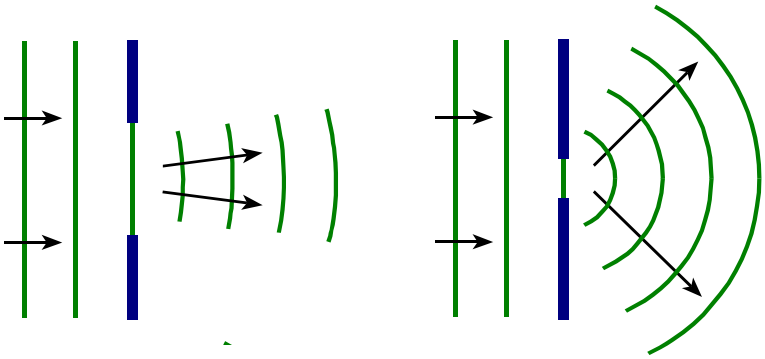
Re-arranging we obtain $x = \lambda D/d$.

If white light is used the different colours will have different fringe spacings and the fringes will become jumbled up once you move a small number of fringes away from the centre. Laser light is highly monochromatic and will give a single series of evenly spaced fringes.

Diffraction

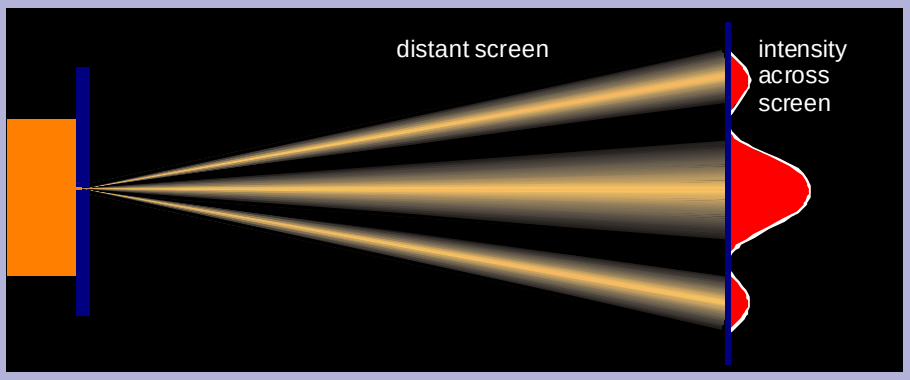
Diffraction is the spreading of waves after passing through a gap or past the edge of an obstacle.

The spreading increases if the gap is made narrower or if the wavelength of the waves is increased.



The amount of diffraction that occurs depends on the aperture size. Optimum diffraction occurs when the aperture width is equal to the wavelength. When the aperture is reduced to less than the wavelength very little energy will pass through the aperture.

Single slit



Monochromatic light passing through a single narrow slit produces a pattern of bright and dark fringes. Intensity **minima** are observed at angles θ given by the equation $d \sin\theta = n\lambda$, where d is the gap width, n is a positive integer and θ is the angle between the incident direction and the direction of diffraction.

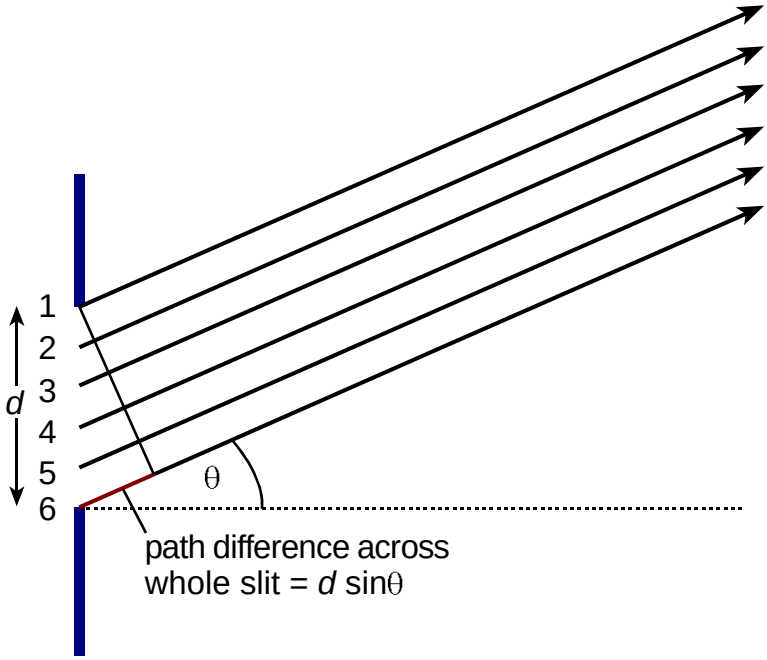
If the distance across the gap is taken to be a large number of equally spaced point sources, **1**, **2**, **3**, etc, the phasor due to **1** will be a certain fraction of a cycle behind the phasor due to **2,** which will be the same fraction behind the phasor due to **3** etc. The resultant phasor is therefore zero at those positions where the tip of the last phasor meets the tail of the first phasor.

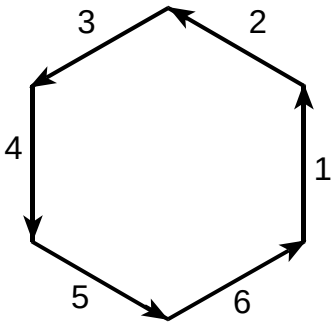
The path difference between the top and bottom of the slit is $d \sin\theta$. If this path difference is equal to a whole number of wavelengths, $n\lambda$, and if the last and first phasors join tip to tail minima occur when

$$d \sin\theta = n\lambda$$

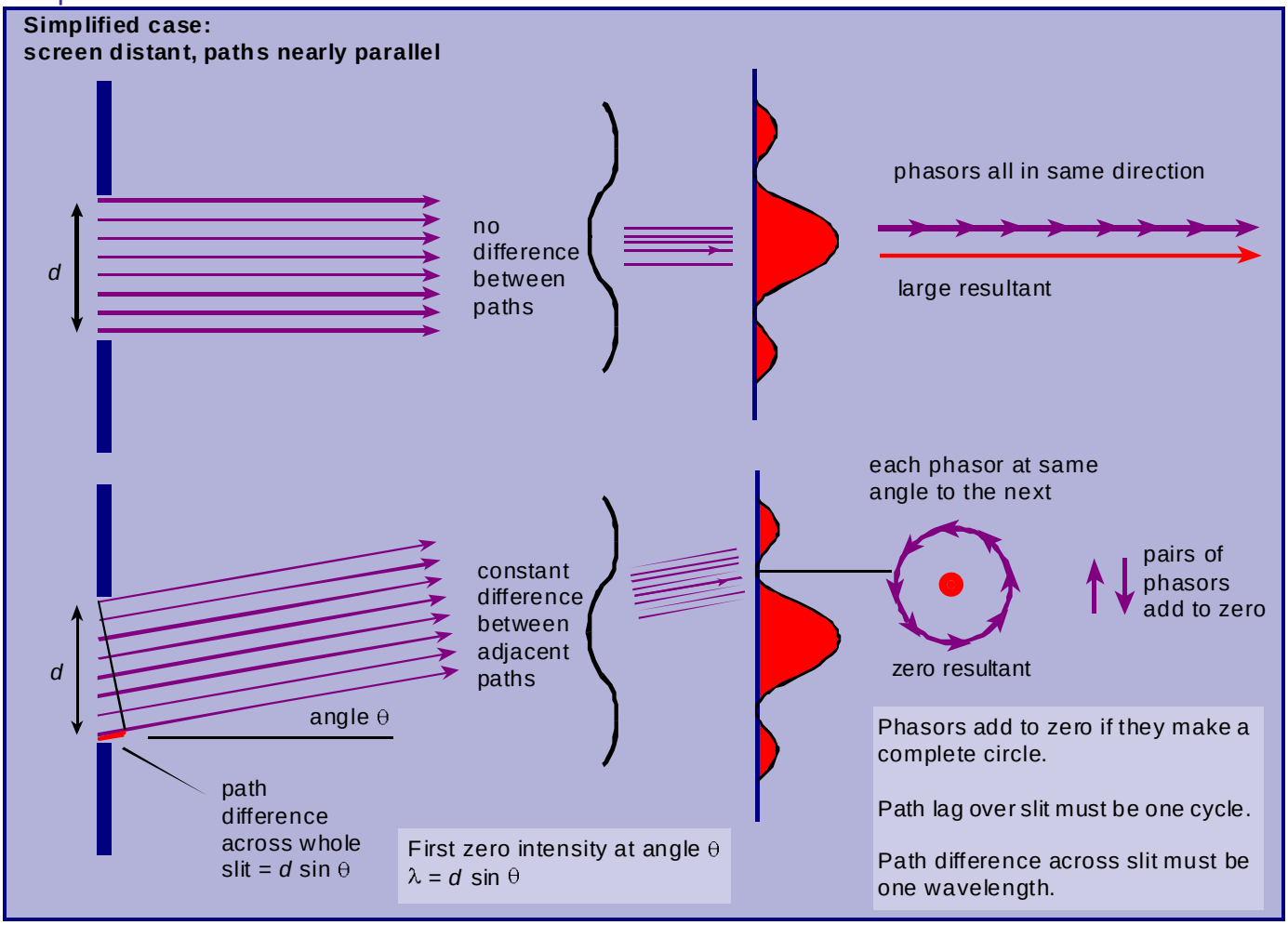
For small angles, $\sin\theta = \theta$ giving an angular width $2\lambda / d$ for the central maximum.

Single slit diffraction



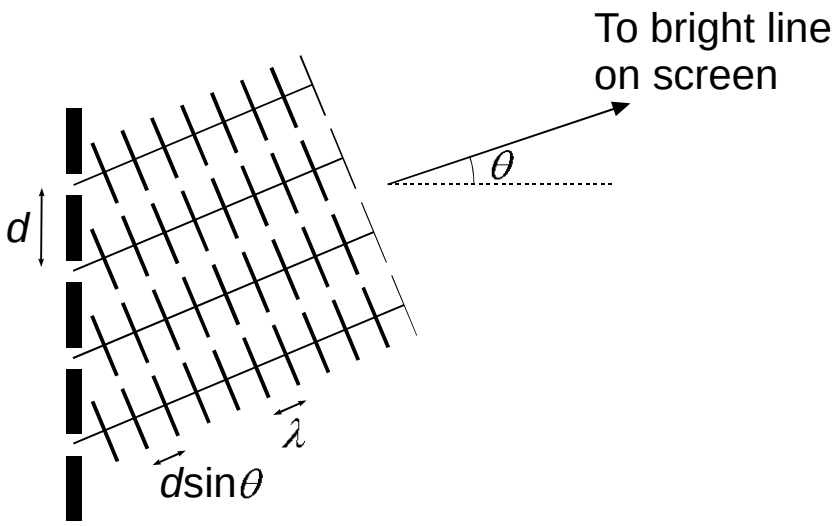


phasors add to zero


Simplified case:
screen distant, paths nearly parallel
d
no difference between paths
phasors all in same direction
large resultant
each phasor at same angle to the next
pairs of phasors add to zero
zero resultant
d
angle θ
constant difference between adjacent paths
path difference across whole slit = d sin θ
First zero intensity at angle θ
λ = d sin θ
Phasors add to zero if they make a complete circle.
Path lag over slit must be one cycle.
Path difference across slit must be one wavelength.

Diffraction gratings

A diffraction grating is a very large number of closely and evenly spaced slits. In practice this is achieved by having a very lines ruled on a glass or clear plastic slide. An incident beam of light is transmitted at and diffracted by each of the slits.



The diagram shows that the path difference between light from adjacent slits travelling in a direction which makes an angle θ with the straight through direction is $d \sin\theta$.

Constructive interference will therefore occur in directions where

$n\lambda = d \sin\theta$

and a bright spot will be seen in these directions.

Because of interference between light from slits in all parts of the grating the diffraction maxima become increasingly sharp as the number of slits illuminated increases. Light of a given wavelength is therefore directed in a very specific direction and no light of that wavelength goes in most other directions. Different wavelengths are diffracted in different directions so the diffraction grating can be used to divide up the component wavelengths in any source of light. This gives the spectrum of the light source.

Diffraction maxima for which $n = 1$ are referred to as first order maxima. Further maxima with $n = 2$, 3, etc may be observed. These are second, third, etc order diffraction maxima. No maxima will be observed with $\theta > 90°$. As the separation of the slits on a grating decreases the angles θ at which the light is diffracted will increase.

Gratings and spectra

A grating is a plate with a large number of parallel grooves ruled on it. A transmission grating transmits and diffracts light into spectra.

When a narrow beam of monochromatic light is directed normally at a transmission grating, the beam passes through and is diffracted into well-defined directions (ie there are **maxima** in theses directions) given by $d \sin\theta = n\lambda$, where d, the grating spacing, is the distance between adjacent slits and n is an integer called the spectral order. The path difference between waves from adjacent slits is $d \sin\theta$ and this must be equal to a whole number n of wavelengths for reinforcement.

Using a white light source, a continuous spectrum is observed at each order, with blue nearer the centre and red away from the centre. This is because blue light has a smaller wavelength than red light so is diffracted less. Spectra at higher orders begin to overlap because of the spread.

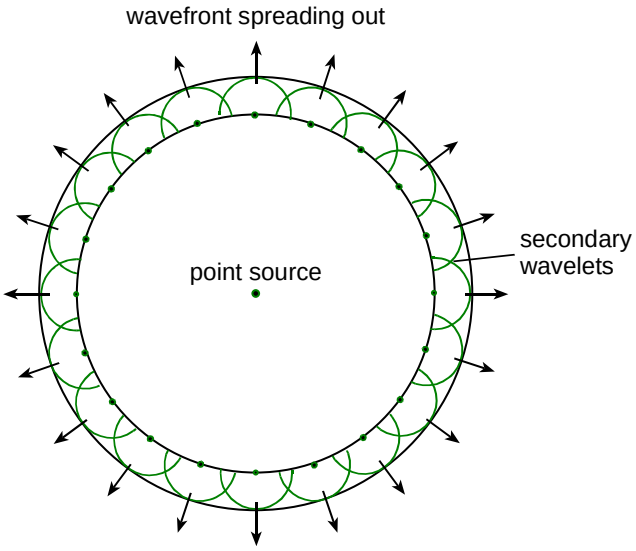
Using light sources that emit certain wavelengths only, a line emission spectrum is observed which is characteristic of the atoms in the light source.

Huygens' wavelets

Huygens' wavelet theory can be used to explain reflection, refraction, diffraction and interference (or superposition) of light.

Huygens' theory of wavelets considers each point on a wavefront as a secondary emitter of wavelets. The wavelets from the points along a wavefront create a new wavefront, so that the wave propagates.

Huygens' wavelets



Accuracy and precision

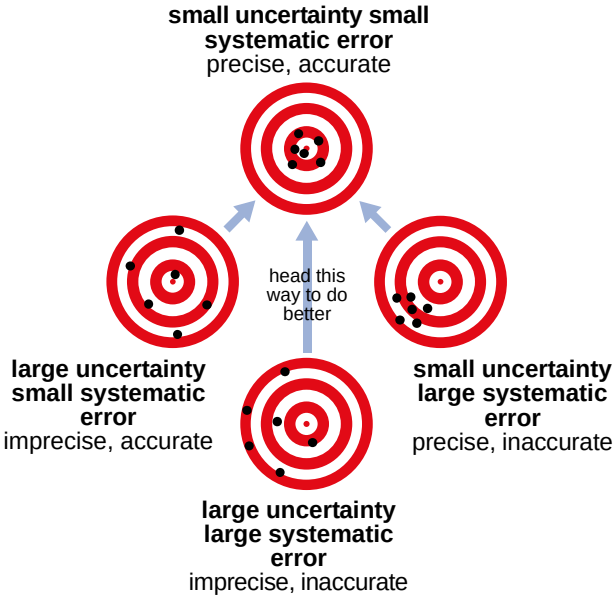
A measurement is accurate if it is close to the true value. A measurement is precise if values cluster closely, with small uncertainty.

A watch with an accuracy of 0.1% could be up to five minutes astray within a few days of being set. A space probe with a trajectory accurate to 0.01 % could be more than 30 km off target at the Moon.

Think of the true value as like the bullseye on a target, and measurements as like arrows or darts aimed at the bullseye.

Uncertainty and systematic error

Think of measurements as shots on a target. Imagine the 'true value' is at the centre of the target



An accurate set of measurements is like a set of hits that centre on the bullseye. In the diagram above at the top, the hits also cluster close together. The uncertainty is small. This is a measurement that gives the true result rather precisely.

On the left, the accuracy is still good (the hits centre on the bullseye) but they are more scattered. The uncertainty is higher. This is a measurement where the average still gives the true result, but that result is not known very precisely.

On the right, the hits are all away from the bullseye, so the accuracy is poor. But they cluster close together, so the uncertainty is low. This is a measurement that has a systematic error, giving a result different from the true result, but where other variations are small.

Finally, at the bottom, the accuracy is poor (systematic error) and the uncertainty is large.

A statement of the result of a measurement needs to contain two distinct estimates:

1. The best available estimate of the value being measured.
2. The best available estimate of the range within which the true value lies.

Note that both are statements of belief based on evidence, not of fact.

For example, a few years ago discussion of the 'age-scale' of the Universe put it at 14 plus or minus 2 thousand million years. Earlier estimates gave considerably smaller values but with larger ranges of uncertainty. The current (2008) estimate is 13.7 ± 0.2 Gy. This new value lies within the range of uncertainty for the previous value, so physicists think the estimate has been improved in precision but has not fundamentally changed.

Fundamental physical constants such as the charge of the electron have been measured to an astonishing small uncertainty. For example, the charge of the electron is $1.602\ 173\ 335 \times 10^{-19}$ C to an uncertainty of $0.000\ 000\ 005 \times 10^{-19}$ C, better than nine significant figures.

There are several different reasons why a recorded result may differ from the true value:

1. **Constant systematic bias**, such as a zero error in an instrument, or an effect which has not been allowed for.

   Constant systematic errors are very difficult to deal with, because their effects are only observable if they can be removed. To remove systematic error is simply to do a better experiment. A clock running slow or fast is an example of systematic instrument error. The effect of temperature on the resistance of a strain gauge is an example of systematic experimental error.

2. **Varying systematic bias**, or drift, in which the behaviour of an instrument changes with time, or an outside influence changes.

   Drift in the sensitivity of an instrument, such as an oscilloscope, is quite common in electronic instrumentation. It can be detected if measured values show a systematic variation with time. Another example: the measured values of the speed of light in a pipe buried in the ground varied regularly twice a day. The cause was traced to the tide coming in on the nearby sea-shore, and compressing the ground, shortening the pipe a little.

3. **Limited resolution of an instrument**. For example the reading of a digital voltmeter may change from say 1.25 V to 1.26 V with no intermediate values. The true potential difference lies in the 0.01 V range 1.25 V to 1.26 V.

   All instruments have limited resolution: the smallest change in input which can be detected. Even if all of a set of repeated readings are the same, the true value is not exactly equal to the recorded value. It lies somewhere between the two nearest values which can be distinguished.

4. **Accidental momentary effects**, such as a 'spike' in an electrical supply, or something hitting the apparatus, which produce isolated wrong values, or 'outliers'.

   Accidental momentary errors, caused by some untoward event, are very common. They can often be traced by identifying results that are very different from others, or which depart from a general trend. The only remedy is to repeat them, discarding them if further measurements strongly suggest that they are wrong. Such values should never be included in any average of measurements, or be used when fitting a line or curve.

5. **Human errors**, such as misreading an instrument, which produce isolated false recorded values.

   Human errors in reading or recording data do occur, such as placing a decimal point wrongly, or using the wrong scale of an instrument. They can often be identified by noticing the kinds of mistake it is easy to make. They should be removed from the data, replacing them by repeated check observations.

6. **Random fluctuations**, for example noise in a signal, or the combined effect of many unconnected minor sources of variation, which alter the measured value unpredictably from moment to moment.

   Truly random variations in measurements are rather rare, though a number of unconnected small influences on the experiment may have a net effect similar to random variation. But because there are well worked out mathematical methods for dealing with random variations, much emphasis is often given to them in discussion of the estimation of the uncertainty of a measurement. These methods can usually safely be used when inspection of the data suggests that variations around an average or a fitted line or curve are small and unsystematic. It is important to look at visual plots of the variations in data before deciding how to estimate uncertainties.

Systematic error

Systematic error is any error that biases a measurement away from the true value.

All measurements are prone to systematic error. A systematic error is any biasing effect, in the environment, methods of observation or instruments used, which introduces error into an experiment. For example, the length of a pendulum will be in error if slight movement of the support, which effectively lengthens the string, is not prevented, or allowed for.

Incorrect zeroing of an instrument leading to a **zero error** is an example of systematic error in instrumentation. It is important to check the zero reading during an experiment as well as at the start.

Systematic errors can change during an experiment. In this case, measurements show trends with time rather than varying randomly about a mean. The instrument is said to show **drift** (e.g. if it warms up while being used).

Systematic errors can be reduced by checking instruments against known standards. They can also be detected by measuring already known quantities.

The problem with a systematic error is that you may not know how big it is, or even that it exists. The history of physics is littered with examples of undetected systematic errors. The only way to deal with a systematic error is to identify its cause and either calculate it and remove it, or do a better measurement which eliminates or reduces it.

Uncertainty

The uncertainty of an experimental result is the range of values within which the true value may reasonably be believed to lie. To estimate the uncertainty, the following steps are needed.

1. Removing from the data **outlying** values which are reasonably suspected of being in serious error, for example because of human error in recording them correctly, or because of an unusual external influence, such as a sudden change of supply voltage. Such values should not be included in any later averaging of results or attempts to fit a line or curve to relationships between measurements.
2. Estimating the possible magnitude of any **systematic error**. An example of a constant systematic error is the increase in the effective length of a pendulum because the string's support is able to move a little as the pendulum swings. The sign of the error is known (in effect increasing the length) and it may be possible to set an upper limit on its magnitude by observation. Analysis of such systematic errors points the way to improving the experiment.
3. Assessing the **resolution** of each instrument involved, that is, the smallest change it can detect. Measurements from it cannot be known to less than the range of values it does not distinguish.
4. Assessing the magnitude of other small, possibly random, unknown effects on each measured quantity, which may include human factors such as varying speed of reaction. Evidence of this may come from the spread of values of the measurement conducted under what are as far as possible identical conditions. The purpose of repeating measurements is to decide how far it appears to be possible to hold conditions identical.
5. Determining the combined effect of possible **uncertainty** in the result due to the limited resolution of instruments (3 above) and uncontrollable variation (4 above).

To improve a measurement, it is essential to identify the largest source of uncertainty. This tells you where to invest effort to reduce the uncertainty of the result.

Having eliminated accidental errors, and allowed for systematic errors, the range of values within which the true result may be believed to lie can be estimated from (a) consideration of the resolution of the instruments involved and (b) evidence from repeated measurements of the variability of measured values.

Most experiments involve measurements of more than one physical quantity, which are combined to obtain the final result. For example, the length L and time of swing T of a simple pendulum may be used to determine the local acceleration of free fall, g , using

$$T=2\pi\sqrt{\frac{L}{g}}$$

so that

$$g=\frac{4\pi^2 L}{T^2}.$$

The range in which the value of each quantity may lie needs to be estimated. To do so, first consider the resolution of the instrument involved – say ruler and stopwatch. The uncertainty of a single measurement cannot be better than the resolution of the instrument. But it may be worse. Repeated measurements under supposedly the same conditions may show small and perhaps random variations.

If you have repeated measurements, ‘plot and look’, to see how the values vary. A simple estimate of the variation is the spread $=\pm\frac{1}{2}$ range .

A simple way to see the effect of uncertainties in each measured quantity on the final result is to recalculate the final result, but adding or subtracting from the values of variables the maximum possible variation of each about its central value. This is pessimistic because it is unlikely that ‘worst case’ values all occur together. However, pessimism may well be the best policy: physicists have historically tended to underestimate uncertainties rather than overestimate them. The range within which the value of a quantity may reasonably be believed to lie may be reduced somewhat by making many equivalent measurements, and averaging them. If there are N independent but equivalent measurements, with range R, then the range of their average is likely to be approximately R divided by the factor $\sqrt{N}$. These benefits are not automatic, because in collecting many measurements conditions may vary.