Waves I: The Wave Equation

The basics of waves are covered in chapter 18 of Halliday and Resnick, and the rest of the material needed for Olympiad physics can be found in chapter 13 of Wang and Ricardo, volume 1. For more on Fourier series, see chapter 6 of French, and for waves in general, see chapters 7 and 8. For a more advanced perspective, see chapter 16 of Taylor. For many physical examples, see chapters 4 and 6 of Crawford. For more about the physics of music, see these lecture notes. For more fun, see chapters I-47 through I-50 of the Feynman lectures. There is a total of **74** points.

1 Traveling Waves

Waves is a vast subject, ranging from the humble wave on a string to electromagnetic waves, gravitational waves, and quantum matter waves. The math used to analyze waves will appear in just about every physics class you'll ever take. But more importantly, the subject is rich in examples, because waves are the physics of the everyday world.

Example 1

Consider a string with mass density μ under tension T. The transverse displacement of the string is given by the wave function $y(x, t)$, and for simplicity we assume the wave is shallow, i.e. $\partial y/\partial x \ll 1$. What's the equation of motion for y?

Solution

Consider a segment of length Δx. At each end of the segment, the tension provides horizontal and vertical forces

$$T_x = \frac{T}{\sqrt{1+y'^2}} \approx T, \quad T_y = \frac{Ty'}{\sqrt{1+y'^2}} \approx Ty'$$

where we're expanding to first order in y'. Therefore the total force is

$$F_y = \Delta T_y = Ty'' \Delta x.$$

This mass of this segment is $\mu \Delta x$, again to first order, so by Newton's Second Law,

$$T\Delta x \, \frac{\partial^2 y}{\partial x^2} = \mu \Delta x \, \frac{\partial^2 y}{\partial t^2}$$

Cleaning this up a bit, we have the wave equation

$$\frac{\partial^2 y}{\partial t^2} = v^2 \, \frac{\partial^2 y}{\partial x^2}, \quad v^2 = \frac{T}{\mu}.$$

Physically, this simply says the string tries to straighten out curvature (represented by $\partial^2 y/\partial x^2$). The wave equation is the simplest possible equation of motion for waves. Even in more complicated situations, we often start with this equation and treat the extra terms as perturbations. The wave equation thus occupies a position like that of the simple harmonic oscillator.

Idea 1

We may factor the wave equation as a difference of squares,

$$(\partial_t^2 - v^2 \partial_x^2) y = (\partial_t - v\partial_x)(\partial_t + v\partial_x) y = 0.$$

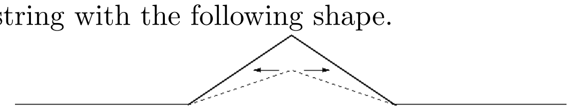
Therefore, functions that satisfy $(\partial_t \pm v\partial_x) y = 0$ solve the wave equation. It is simple to verify that these are functions of the form

$$y(x,t) = f(x \pm vt).$$

Since the wave equation is linear, superpositions of solutions to the wave equation are also solutions to the wave equation. The general solution is of the form $f(x - vt) + g(x + vt)$ for arbitrary functions f and g.

[1] **Problem 1.** Waves of the form $y(x,t) = f(x \pm vt)$ simply translate with uniform velocity v. Does a wave of the form $y(x,t) = f(x + vt)$ move to the left or the right?

[2] **Problem 2.** Consider a string with the following shape.



(a) If this is a traveling wave moving to the right with velocity v, carefully draw the velocity and acceleration of every point on the string.

(b) Now suppose the string is held in place, with zero velocity. If it is suddenly relased, sketch the subsequent behavior of the string.

[2] **Problem 3** (HRK)**.** A uniform circular hoop of string is rotating clockwise in the absence of gravity. The tangential speed is v. Find the speed of waves on this string.

[2] **Problem 4.** A uniform rope of mass m and length L hangs from a ceiling.

(a) Show that the time it takes for a transverse wave pulse to travel from the bottom of the rope to the top is approximately $2\sqrt{L/g}$. Under what circumstances is this approximation good?

(b) Does the pulse get longer or shorter as it travels?

[3] **Problem 5.** [A] At time $t = 0$, the position and transverse velocity of a string obeying the wave equation are given by $y(x)$ and $v_y(x)$. Find an explicit expression for $y(x,t)$ in terms of these functions; this is called d'Alembert's solution. (Hint: construct solutions with initial position $y(x)$ and zero initial velocity, and vice versa, and add them together.)

Idea 2

A sinusoidal wave has the form

$$y(x,t) = A\cos(kx - \omega t + \phi), \quad v = \frac{\omega}{k}$$

where k is the wavenumber and ω is the angular frequency. They are related to the wavelength and period by

$$k = \frac{2\pi}{\lambda}, \quad \omega = \frac{2\pi}{T}.$$

Sinusoidal waves will be especially useful because the wave equation is linear. Fourier analysis tells us that *any* initial condition can be written in terms of a sum of sinusoids, so if we know what happens to the sinusoids, we know what happens in general by superposition. This is just a generalization of ideas we've seen in **M4** and **E6**. Just as we saw there, it can also be useful to promote y to a complex number, where the physical value of y is the real part; for a sinusoidal wave we would have $y(x,t) = y_0 e^{i(kx-\omega t)}$.

Remark

Physicists almost universally use k and ω rather than λ, f, and T. A nice way of thinking of these variables is that they represent how quickly the phase ϕ changes, in space or time,

$$k = \frac{d\phi}{dx}, \quad \omega = \frac{d\phi}{dt}.$$

If we use a little special relativity, we can even combine these into a single equation,

$$k^\mu = \partial^\mu \phi.$$

The fundamental relation between particle and wave properties in quantum mechanics is

$$p^\mu = \hbar k^\mu.$$

These are the de Broglie relations, which we'll cover in **X1**.

[4] **Problem 6.** For a wave on a string, there are two contributions to the energy: potential energy from stretching, and kinetic energy from transverse motion.

(a) Find the kinetic and potential energy density of the string in terms of T, μ, y, and its derivatives.

(b) Evaluate the above quantities for $y = A\cos(kx - \omega t)$. Is the total energy density uniform?

(c) Show that for a general traveling wave of the form $y = f(x - vt)$, the total kinetic and potential energy are equal.

(d) Show that for any wave function y, total energy is conserved. This will require some integration by parts, as well as the wave equation itself; you should assume y goes to zero at infinity.

(e) Compute the energy of the static configuration in problem 2(b), assuming the triangle has height h and base L, where $h \ll L$.

One warning: as we saw in **E6**, energy is quadratic, so it does *not* obey the superposition principle. Locally, the amount of energy can be more or less than the sum of the energies of the superposed waves, due to interference.

Remark

How can we account for damping in the wave equation? The simplest thing would be to add a force proportional to v_y, which e.g. could be due to air drag. Then

$$\partial_t^2 y = v^2 \partial_x^2 y + A\,\partial_t y.$$

But what if the string is in a vacuum? Then the simplest kind of damping would be due to the energy lost in bending and unbending of the string, which takes the form

$$\partial_t^2 y = v^2 \partial_x^2 y + A\,\partial_t \partial_x^2 y$$

because $\partial_x^2 y$ describes the bending. This is called Kelvin–Voigt damping.

In both cases, it's straightforward to handle the damping since the wave equation remains linear; we just plug in a solution of the form $e^{i(kx-\omega t)}$ and find the new relation between ω and k. If we pick k to be a real number, we will generally find ω to be complex, with its imaginary part corresponding to exponential decay of the wave over time.

[3] **Problem 7.** [A] With a little vector calculus, the results above can be generalized to an arbitrary number of dimensions. For example, ideal waves in three dimensions obey

$$\frac{\partial^2 \psi}{\partial t^2} = v^2 \left(\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} + \frac{\partial^2 \psi}{\partial z^2} \right) = v^2 \nabla^2 \psi$$

where the function $\psi(\mathbf{r}, t)$ could stand for a variety of things, such as the pressure, density, or temperature (for a sound wave) or the electric or magnetic field (for an electromagnetic wave).

(a) For simplicity, let's restrict to waves which have spherical symmetry, so that ψ only depends on r and t. Plug such a spherical wave into the wave equation, and simplify until you get an equation only in terms of the partial derivatives of $\psi(r, t)$.

(b) Because the area of a sphere goes as r^2, we expect the energy density of a spherical wave to fall as $1/r^2$, and therefore expect the amplitude to fall as $1/r$. Therefore, it is useful to consider the quantity $r\psi$, which has this falloff factored out. By considering the differential equation that $r\psi$ obeys, find the general solution for $\psi(r, t)$.

2 Standing Waves

Idea 3

A standing wave is a solution to the wave equation of the form

$$y(x, t) = f(x) \cos(\omega t).$$

Typically, only discrete values of ω are possible, with the allowed values depending on the boundary conditions. If the setup is translationally symmetric, then $f(x)$ will be sinusoidal. If you want to get some intuition, try playing with this PhET simulation.

[2] **Problem 8.** Consider a string of length L and wave speed v.

(a) Suppose the ends of the string are fixed, i.e. $y(x, t) = 0$ at $x = 0$ and $x = L$. Find the standing wave angular frequencies and sketch the configurations.

(b) Do the same if the ends of the string are free, i.e. $\partial y/\partial x = 0$ at $x = 0$ and $x = L$.

(c) Do the same if one end is fixed and one end is free.

[2] **Problem 9.** USAPhO 1997, problem A1.

Idea 4

When a musical instrument plays a note, typically multiple standing waves are excited, so the resulting sound is composed of multiple frequencies. As you saw in problem 8, often the standing wave frequencies are all multiples of a single, lowest frequency. This frequency f_0 is called the fundamental, or first harmonic, while the multiple nf_0 is called the n^{th} harmonic. The fundamental frequency determines the pitch we perceive, while the distribution of energy among the harmonics determines the timbre, or tonal quality, of the instrument.

[2] **Problem 10** (Feynman)**.** Pinch a single length of rubber band about 5 cm long between the fingernails of your two hands, and twang it in the middle to observe the pitch. Then stretch it to several times its original length, observing the pitch as you proceed. Make a simple physical model to explain the observed results.

[5] **Problem 11.** Some instruments, such as xylophones and marimbas, are made with rigid rods instead of strings. The equation that describes transverse vibrations is instead

$$\frac{\partial^2 y}{\partial t^2} = -A \frac{\partial^4 y}{\partial x^4}$$

for a constant A that depends on the material and cross-sectional area.

(a) For a xylophone bar of length L, find the standing wave solutions and their angular frequencies. For simplicity, pretend that the solutions are sinusoidal in space, and that the bar has free ends just like a string, even though this is not true in reality.

(b) When the bar in part (a) is hit, a certain note is sounded. What is the length of the bar that makes a note one octave higher?

(c) [A] $\star$ The actual boundary conditions for a free bar are

$$\frac{\partial^2 y}{\partial x^2} = \frac{\partial^3 y}{\partial x^3} = 0$$

at the endpoints, and the solutions aren't purely sinusoidal in space. Compute the lowest few standing wave angular frequencies and compare them to those you found in part (a). You'll have to use a calculator or computer to do this.

(d) A guitar or piano string satisfies the wave equation with a small additional fourth-order term,

$$\frac{\partial^2 y}{\partial t^2} = v^2 \frac{\partial^2 y}{\partial x^2} - A \frac{\partial^4 y}{\partial x^4}.$$

Show that the standing wave frequencies are not linearly spaced, as they would be for an ideal string, but instead are slightly more spaced out. This effect is called inharmonicity. (Hint: the spatial profiles of the standing waves are still sinusoidal.)

We perceived two notes to be "in tune" when the component frequencies in the notes line up with each other. But since the frequencies are more spread out than ideal harmonics, a piano feels more in tune when the fundamental frequencies are spread out a little bit more. This "stretch tuning" is significant and adds up to about an entire semitone across the piano!

Remark: Beam Theory

Where does the strange fourth-order equation for transverse vibrations above come from? Since force is the derivative of energy, it's easier to think about how the energy stored in a rigid rod differs from that of a string. When a string with tension T, mass per length λ, and length ℓ is plucked, giving it a transverse displacement y, then

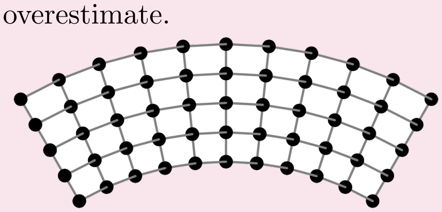
$$\frac{\text{kinetic energy}}{\text{length}} \sim \lambda \dot{y}^2, \quad \frac{\text{potential energy}}{\text{length}} \sim \frac{T \Delta \ell}{\ell} \sim \frac{T y^2}{\ell^2}$$

where our expression for $\Delta \ell$ comes from the Pythagorean theorem. As we know from **M4**, the ratio of the coefficients gives ω^2, so $\omega \ell \sim \sqrt{T/\lambda}$. For a general wave with wavenumber k, we would replace ℓ with $1/k$ above, giving $\omega \propto k$ as expected.

Now, a rod is characterized by a Young's modulus Y, mass density ρ, length ℓ, width w, and vertical thickness h. If the transverse displacement is y, then

$$\frac{\text{kinetic energy}}{\text{volume}} \sim \rho \dot{y}^2, \quad \frac{\text{potential energy}}{\text{volume}} \sim Y (\text{strain})^2.$$

The tricky part is understanding the strain. If you naively used the same logic as for the string, then all parts of the rod would be stretched, with typical strain $(y/\ell)^2$. This is correct in the limit of large displacements, $y \gg h$, where the rod's thickness is negligible. But for small displacements, it's an overestimate.



As the rod is displaced vertically, it slightly shrinks horizontally. As a result, there is a "neutral line" in the middle of the rod that is neither stretched or compressed. Bonds above the line are stretched, and bonds below the line are compressed.

The neutral line has radius of curvature $R \sim \ell^2/y$. Thus, the strain at the top and bottom of the rod is of order $h/R \sim hy/\ell^2$. Plugging this in gives

$$\omega \sim \sqrt{\frac{Y}{\rho}} \frac{h}{\ell^2}.$$

Again, for a general wavenumber we would replace ℓ with $1/k$, giving the $\omega \propto k^2$ scaling. For a derivation of this result by dimensional analysis, see section 9.2.3 of The Art of Insight.

There's another neat bit of physics we can get here. Consider a horizontal rod with one end fixed at a wall. What is the vertical deflection of the other end of the rod, due to its own weight? The gravitational and internal potential energy densities both have "reasonable", power-law dependence on the deflection y. Thus, when their derivatives match, so that forces balance, their absolute values should match within an order of magnitude,

$$\frac{\text{elastic potential energy}}{\text{volume}} \sim \frac{\text{gravitational potential energy}}{\text{volume}} \sim \rho g y.$$

Solving for the deflection gives

$$y \sim \frac{\rho g \ell^4}{Y h^2}$$

which is the fundamental result of Euler–Bernoulli beam theory. (For a proper derivation in terms of force and torque balance, see chapters 9 and 10 of Lautrup.)

Example 2

How are the sounds of a violin, a trumpet, and a person different in a room full of helium?

Solution

As we saw in **T3**, the speed of sound in air is $\sqrt{\gamma p/\rho}$. When the air is replaced with helium, ρ decreases, increasing the speed of sound.

The standing wave frequencies of a violin are determined by properties of the strings, which aren't affected by the helium. The only difference will be that the way the violin's sound reverberates will be subtly changed.

For the trumpet, the standing wave frequencies are proportional to v/L where L is the length of the air column inside the trumpet. Thus, the standing wave frequencies go up, and the trumpet makes higher-pitched notes.

The human voice is more subtle. A wind instrument works by exciting standing waves inside it. But the source of the human voice is the vibrations of the vocal folds, whose fundamental vibration frequency is directly controlled by your muscles. The entire rest of your vocal tract does not affect what frequencies are present, but rather affects how energy is distributed between those frequencies. (For instance, vowels are characterized by having extra energy near two particular frequencies, called formants.) Helium changes the resonant frequencies of the vocal tract and thus changes which frequencies emitted by the vocal folds are emphasized. It thus changes the timbre, but not the pitch.

[3] **Problem 12.** Some questions about musical instruments.

(a) A piano makes sound by quickly striking a string with a hammer. The seventh harmonic doesn't fit in with the rest that well. If you want to eliminate the seventh harmonic, at what point(s) can you put the hammer?

(b) A violinist can make the note from an open string sound an octave higher by lightly touching it at a point while bowing it somewhere else. Which point(s) should be touched?

(c) Suppose a string has its ends attached to walls. A person can set up a standing wave by holding the string at some point and moving it side to side, sinusoidally with fixed *amplitude*. At which point(s) should the string be driven to maximize the amplitude of a given standing wave? Assume the string experiences very little damping.

[5] **Problem 13.** EuPhO 2017, problem 1. (Hint: don't try to use fancy math here. EuPhO problems are designed to be solved with only elementary math and graph reading.)

Idea 5

Standing wave solutions also exist for waves in more than one spatial dimension. In the special case where the wave medium is uniform, and shaped like a rectangle (in two dimensions) or a rectangular prism (in three dimensions), all the standing wave solutions can be found by separation of variables. That is, they can all be written as

$$\psi(x, y, z, t) = f(x)g(y)h(z)\cos(\omega t)$$

where f, g, and h are sinusoids.

[3] **Problem 14.** The top of a drum is like a string, in that it has a uniform surface mass density σ and surface tension γ.

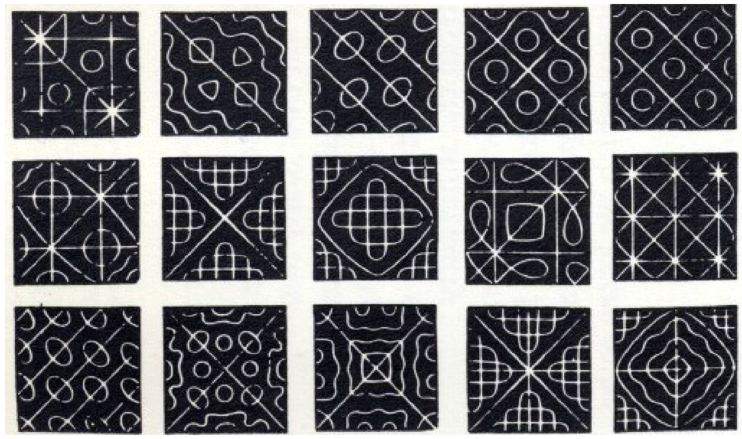
(a) Waves on the drum can be described by its height $z(x, y, t)$. Find the wave equation for a drum. What is the speed of traveling waves?

(b) Consider a square drum of side length L, where the boundaries are fixed to $z = 0$. Find the standing wave solutions and the corresponding ω. What's the lowest standing wave angular frequency?

The frequencies will not be multiples of a fundamental frequency, so they are called overtones, rather than harmonics; that's why drums don't sound like they're playing notes. (Special examples, such as the timpani, are designed to mostly excite the harmonic frequencies.)

(c) Why does a drum sound different if you hit it near the edge, versus at the center?

[4] **Problem 15.** When sand is sprinkled on a vibrating metal plate, it forms Chladni patterns. Suppose we (unrealistically) model the plate as a square elastic membrane, as in problem 14, of side length L obeying the wave equation with wave speed v. Unlike in problem 14, we now assume the boundaries of the plate are free.



(a) Do Chladni patterns form at the nodes or antinodes of a standing wave?

(b) Find the general standing wave solutions $z(x, y, t)$ and their angular frequencies.

(c) The plate is also fixed in the middle by the support, so $z = \partial z/\partial x = \partial z/\partial y = 0$ there, which removes many of the standing wave solutions. Find the lowest and second-lowest angular frequencies of allowed standing waves.

(d) Sketch the Chladni pattern for the lowest standing wave frequency.

(e) For the second-lowest standing wave frequency, there will be two independent standing waves with that frequency. What superpositions of them will yield Chladni patterns with 90° rotational symmetry? (If you want to see these patterns, you'll need a computer.)

Remark: Plate Theory

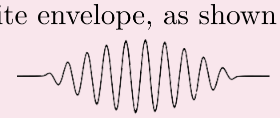
The treatment of problem 15 is inaccurate because the restoring force in a metal plate is rigidity, not tension. The waves actually satisfy the two-dimensional analogue of the fourth-order equation considered in problem 11, which is called the biharmonic equation,

$$-\frac{\partial^2 z}{\partial t^2} \propto \nabla^4 z = \nabla^2 \nabla^2 z = \left(\partial_x^2 + \partial_y^2\right)\left(\partial_x^2 + \partial_y^2\right) z = \frac{\partial^4 z}{\partial x^4} + 2\frac{\partial^4 z}{\partial x^2 \partial y^2} + \frac{\partial^4 z}{\partial y^4}.$$

Of course, I didn't ask you to consider this, because it would have been quite a slog! But if you want to learn more about this thrilling subject, see *Plates*, by Bhaskar and Varadan.

Remark: Wavepackets

Purely sinusoidal traveling waves of the form $e^{i(kx-\omega t)}$ are unrealistic, because they have infinite spatial extent. A realistic alternative is a wavepacket, which looks like a sinusoid with wavenumber k but with a finite envelope, as shown below.



To understand how sinusoids are constructed, consider the superposition of two traveling waves with wavenumbers $k \pm \Delta k$. The wavefunction is

$$e^{i((k-\Delta k)x-(\omega-\Delta\omega)t)} + e^{i(k+\Delta k)x-(\omega+\Delta\omega)t)} = 2e^{i(kx-\omega t)} \cos(\Delta k\, z - \Delta\omega\, t).$$

This is simply a sinusoid of wavenumber k with a slowly varying envelope, whose characteristic size is $1/\Delta k$, reflecting how the two component waves slowly move in and out of phase. The wave is still infinite in size, but this can be remedied by superposing infinitely many wavenumbers; in this case the component sinusoids never get back in phase again.

If the wavenumbers occupy a region Δk, then the size of the envelope is of order $1/\Delta k$, because this is the distance required for the component waves to get out of phase with each other. This yields an "uncertainty principle" for waves,

$$\Delta x\, \Delta k \gtrsim 1.$$

In quantum mechanics, particles are described by waves with $p = \hbar k$. Substituting this in immediately gives the Heisenberg uncertainty principle; it fundamentally holds because one cannot get a finite wave without superposing different wavenumbers.

Alternatively, if we had worked with angular frequencies instead, we would have had

$$\Delta t\, \Delta\omega \gtrsim 1.$$

This is an "acoustic uncertainty principle", also important in digital signal processing, where it is called the Gabor limit. Upon using the de Broglie relations, one finds the energy-time uncertainty principle.

Idea 6

The dispersion relation of a system is the function $\omega(k)$ relating the angular frequency and wavenumber of sinusoidal waves. The phase and group velocity

$$v_p = \frac{\omega}{k}, \quad v_g = \frac{d\omega}{dk}$$

describe the velocities of sinusoidal waves of wavenumber k and the envelopes of wavepackets built from sinusoids near wavenumber k, respectively. We can see the latter result from the remark above: the peak of the envelope is the point where the components are in phase, and this point travels at speed $\Delta\omega/\Delta k \approx d\omega/dk$.

For ideal waves, the dispersion relation is linear, the group and phase velocities are constant and equal, and waves travel while maintaining their shape. When the dispersion relation isn't linear, the group and phase velocities depend on k, so wavepackets gradually fall apart (i.e. they disperse). For more discussion of these topics, see chapter 6 of Morin.

Remark

In **R1**, you learned that nothing can go faster than the speed of light. But the phase velocity can exceed it; for instance, in problem 16 you will find a phase velocity that can be infinite! This is compatible with relativity, because the phase velocity isn't the speed of an actual object. It's just a formal quantity, namely the rate of change of the position of points of constant phase in an infinite plane wave. To reinforce the point, suppose we arranged to stand at different places and clap at the same time. Then we could say "the clap moved from me to you at infinite speed", but clearly nothing about this contradicts relativity.

In some textbooks, you'll read that while the phase velocity can be faster than light, the group velocity can't be, because it's the speed of an actual pulse. But that's not quite true in general either, because that result follows from an approximation. For instance, in materials with really weird dispersion relations, a single pulse can split up into two, in which case the speed of "the" peak or "the" envelope isn't even well-defined. Accordingly, in these cases the group velocity can be formally faster than light, but it doesn't contradict relativity because the group velocity ceases to have its intuitive meaning.

If you're mathematically minded, you might be bothered by the argument that a superluminal phase velocity is okay because no "actual object" moves faster than light, since it seems hard to rigorously define the term "actual object". Luckily, there's a simple and perfectly rigorous definition of the light speed limit: the observable effects of an action must lie in the future light cone of the action. Suppose you change the value of a field at the origin, at time $t = 0$. Then at time t, the field at all points $r > ct$ must be the same as if you didn't make the change at all. The maximum speed at which *changes* of the field propagate is called the signal velocity, and it can never exceed c.

[3] **Problem 16.** Consider transverse waves on a horizontal string with tension T and mass density μ. The string is attached to the ceiling by a large number of vertical springs, so that if the entire string is pulled down, it will oscillate with angular frequency ω_0.

(a) Find the wave equation for waves on this string.

(b) By guessing sinusoidal solutions, find $\omega(k)$ and the minimum possible angular frequency.

(c) Compute the phase and group velocity for wavepackets of angular frequency ω.

(d) What actually happens if you grab one end of the string and try to wiggle it at a frequency below the minimum possible frequency?

If we treat the string as a quantum system, excitations of the string are particles with $E(p)$ determined by the function $\omega(k)$ you found, along with the de Broglie relations $E = \hbar\omega$ and $p = \hbar k$. Therefore, there is a minimum energy for excitations. In a relativistic and quantum context, this means that all the particles must be massive; the minimum energy is mc^2. This is a toy model for how the Higgs field gives particles mass.

[2] **Problem 17.** The motion of ripples of short wavelength (less than 1 cm) on water is controlled by the surface tension γ and density ρ.

(a) Use dimensional analysis to constrain the phase velocity v_p of ripples with wavenumber k.

(b) Show that $v_g = (3/2)v_p$.

3 Reflection and Transmission

When we considered standing waves in the previous section, we were only considering "steady state" behavior. Now we consider the dynamics of a wave hitting an obstacle more explicitly.

Example 3

Suppose a string defined for $x < 0$ ends at a hard wall at $x = 0$. Show that any wave directed towards the wall will be reflected back upside-down.

Solution

We suppose that we send in a wave of the form

$$y_{\text{in}}(x,t) = f(kx - \omega t).$$

Let the reflected wave be a general wave traveling backward,

$$y_r(x,t) = g(-kx - \omega t).$$

Both of these expressions only have physical meaning for $x < 0$, since the string only exists there. Now, the boundary condition is $y(0,t) = 0$, so we have

$$f(-\omega t) + g(-\omega t) = 0.$$

This tells us precisely that $g = -f$, so the wave is reflected upside-down but otherwise unchanged.

There's an easy way to visualize what's going on here. We can imagine that there really is string for $x > 0$, but that the point $x = 0$ stays fixed for some reason. Then this situation corresponds to an incoming wave coming from the left, and a flipped wave coming from the right. The two meet and cancel at $x = 0$, and the flipped wave continues on going to the left, where the physical string is. Fundamentally, this story works for the same reason as the method of images in electromagnetism: as long as you satisfy the boundary conditions, you can do whatever you want beyond the boundary.

[1] **Problem 18.** Another type of boundary condition is the "soft" boundary condition, which requires $dy/dx = 0$ at $x = 0$. Show that waves are reflected from this boundary but not flipped.

[3] **Problem 19.** Consider the triangular "plucked" shape of problem 2 again, but suppose that the string starts at rest, and its two outer corners are always held fixed.

(a) Sketch what happens after the string is released. What is the period of the motion?

(b) Confirm explicitly that the initial potential energy of the string is equal to the kinetic energy of the string when it is purely horizontal.

(c) What would prevent a real string from achieving this ideal motion? What will the string look like after a few oscillations?

(d) Try to sketch what happens if the string begins with the "pluck" off-center.

[4] **Problem 20.** [A] The general, turn-the-crank method to find the time evolution of an arbitrary wave on a string of length L is Fourier series. In this method, we write the initial shape $y_0(x)$ of the wave as a combination of standing waves,

$$y_0(x) = \sum_n c_n \sin \frac{\pi n x}{L}, \qquad 0 \leq x \leq L.$$

We know how each standing wave oscillates in time, so by linearity, the entire wave evolves as

$$y(x,t) = \sum_n c_n \sin \frac{\pi n x}{L} \cos(\omega_n t)$$

where ω_n is the angular frequency of the n^{th} harmonic.

(a) The coefficients c_n can be extracted by integrating $y_0(x)$ against another sine,

$$c_n \propto \int_0^L dx\, y_0(x) \sin \frac{\pi n x}{L}.$$

Explain why this works, and find the constant of proportionality.

(b) Now let's consider the plucked string considered in part (a) of problem 19. If the pluck is centered at the middle of the string and has height h, find the coefficients c_n. (If you're so inclined, you can use a computer to see how the resulting $y(x,t)$ approaches the answer to problem 19 as more terms are included.)

(c) Argue that in general, we have

$$\int_0^L y_0^2(x)\, dx = \frac{L}{2} \sum_n |c_n|^2.$$

By applying this result to the plucked string, show that the Riemann zeta function has value

$$\zeta(4) = \sum_{n=1}^{\infty} \frac{1}{n^4} = \frac{\pi^4}{90}.$$

In fact, this is one of the simplest ways to compute $\zeta(4)$.

We'll use the idea of Fourier series to illustrate some conceptual points in **W2**.

Idea 7

More generally, the relation between the incoming and reflected waves may depend on the exact form of the incoming wave. In this case, it's useful to consider sinusoidal solution. Let

$$y_{\text{in}}(x,t) = e^{i(kx - \omega t)}.$$

Almost all boundary conditions will state that something at the boundary is constant in time, which is only possible if the reflected wave has the same frequency. So in general we have

$$y_r(x,t) = r e^{i(-kx - \omega t)}$$

where r is the reflection coefficient. If the medium exists for $x > 0$, there is also a transmitted wave there, of the form

$$y_t(x,t) = te^{i(k'x-\omega t)}$$

where k' might differ from k, and t is the transmission coefficient. In general, both r and t may depend on k as well as the boundary conditions. Note that the phases of r and t depend on the conventions we used to define $y_r(x,t)$ and $y_t(x,t)$, though the magnitudes don't.

[4] **Problem 21.** Suppose the string at $x < 0$ has a tension T_1 and mass density μ_1, while the string at $x > 0$ has a tension T_2 and mass density μ_2. (If you were doing this at home, it would be difficult to have $T_1 \neq T_2$ since the whole setup would accelerate longitudinally. But for the sake of the problem, suppose the two strings are attached at $x = 0$ by a massless ring which slides on a vertical frictionless pole, so that the normal force from the pole balances the longitudinal force $T_2 - T_1$.) As above, let $y_{\text{in}}(x,t) = e^{i(kx-\omega t)}$.

(a) Write down k' and the boundary conditions at $x = 0$.

(b) Show that the reflection and transmission coefficients are

$$r = \frac{Z_1 - Z_2}{Z_1 + Z_2}, \quad t = \frac{2Z_1}{Z_1 + Z_2}, \quad Z_i = \sqrt{\mu_i T_i}.$$

The quantity Z_i is called the impedance.

(c) What limiting cases correspond to hard and soft boundary conditions? Verify that the reflection coefficients match the results above.

(d) Suppose the incoming wave has the exponential form above, but only lasts for a long but finite time τ. After a long time, the incoming wave is gone, and we have a reflected and transmitted wave. Verify that energy has been conserved. (Be careful: it's not simply $|r|^2 + |t|^2 = 1$.)

The great thing about the coefficients r and t is that they contain all the information about the reflection and transmission. For complicated problems with multiple interfaces, it's best to work purely in terms of r and t, as solving the wave equation as a whole can get messy.

Remark

In the previous problem, you found that a discontinuity in the wave medium can cause reflection, if the impedances of the two sides are different. This reflection is often wasteful, and can be reduced by insertion of "impedance matching" devices which soften the discontinuity. For example, a conical megaphone helps match the air column of the mouth and throat to the air outside the mouth.

The language of "impedance matching" comes from circuitry. In **E7** you found that transmission lines have a characteristic impedance Z. When two transmission lines are attached, wave reflection occurs if the impedances mismatch. The point here is that the same ideas apply to many kinds of waves, as long as one generalizes the notion of impedance.

Remark

You can generalize the methodology of the previous problem to a large variety of similar problems. For example, suppose the ring at $x = 0$ wasn't massless. Then the boundary conditions would have been changed; instead of the transverse force on the ring vanishing, the transverse force would have had to be equal to its mass times its transverse acceleration. (You may recall that setup from the preliminary problem set.) You could even put the ring on a spring, or give it a damping force (in which case the wave energy is no longer conserved). In all cases, the technique is just to take exponential solutions on both sides and apply the relevant boundary conditions. I won't assign such problems, since they usually involve lots of messy algebra, but the idea is very important in physics.

4 Interference

Idea 8

The intensity of a wave is proportional to its amplitude squared, so if two waves with amplitudes A_1 and A_2 are superposed, the resultant intensity is

$$I \propto (A_1 + A_2)^2.$$

This differs from the sum of the intensities by an interference term,

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\theta$$

where θ is the phase difference between the waves.

[4] **Problem 22.** Consider two thin imperfect mirrors, each with reflection and transmission coefficients r and t from both sides, placed a distance L apart, with air in between them and outside them. This system is called a Fabry–Perot interferometer. A wave with wavenumber k hits the apparatus; we want to find the reflection and transmission coefficients r_{net} and t_{net} of the entire system.

(a) Draw all paths that the light could take to be reflected, and to be transmitted.

(b) By applying the principle of superposition and summing a geometric series, show that

$$r_{\text{net}} = r + \frac{rt^2 e^{2ikL}}{1 - r^2 e^{2ikL}}, \quad t_{\text{net}} = \frac{t^2 e^{ikL}}{1 - r^2 e^{2ikL}}.$$

Note that your answers may differ by phases, depending on your conventions for r_{net} and t_{net}.

(c) Show that all the light is transmitted for some special values of k, even if $r \approx 1$. That is, nearly ideal mirrors can become perfectly transparent! This is called resonant transmission, and it occurs because the reflected waves perfectly destructively interfere. (Hint: be careful, and don't forget that r and t are complex numbers.)

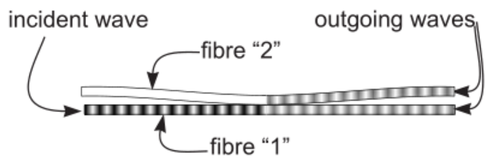
(d) For this system, the statement of energy conservation is

$$|r_{\text{net}}|^2 + |t_{\text{net}}|^2 = 1.$$

Use this fact to get a relation between r and t. (Hint: evaluating this in general is very messy, so pick an appropriate special value of L to simplify things.)

[3] **Problem 23.** USAPhO 2004, problem A3.

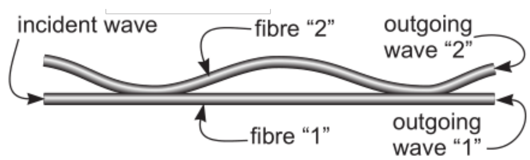
[3] **Problem 24** (Kalda)**.** In fiber optics, devices called equal ratio splitters are often used; these are devices where two optical fibers are brought into such a contact so that if an electromagnetic wave is propagating in one fiber, it splits into two equal amplitude waves traveling in each of the fibers. Assume that all waves propagate with the same polarization, i.e. that all electric fields are parallel.



(a) Show that whenever a wave enters the splitter, from either fiber, one of the outgoing waves is advanced in phase by $\pi/4$, while the other is retarded by $\pi/4$.

(b) From part (a) alone, it's ambiguous which wave is advanced and which wave is retarded. Let's suppose that the fibers are set up so that, when a wave enters along fiber 1, the wave that exits along fiber 1 is advanced. If a wave enters along fiber 2, is the wave that exits along fiber 1 advanced or retarded?

(c) Now consider two sequentially positioned, identical equal ratio splitters, as shown.



This is called a Mach–Zehnder interferometer. The optical path difference between the inter-splitter segments of the two fibers is $30\,\mu$m. Assuming the wavelength of the incoming monochromatic light varies from 610 nm to 660 nm, for what wavelengths is all the light energy directed into fiber 2?

[4] **Problem 25.** APhO 2003, problem 2. A nice, though somewhat clunkily worded question on an interferometer for measuring rotation.

Remark: Interference and Energy Conservation

People sometimes get the impression that interference violates energy conservation, but it doesn't. For instance, in the double slit experiment, you get destructive interference in some places, and constructive interference in other places, so that the total energy stays the same.

A natural followup question is: what if you could engineer waves to have destructive interference *everywhere*? Wouldn't that unambiguously violate energy conservation? Actually, it still won't, but the reason is a bit subtle and depends on the details.

For simplicity, suppose we start with a long string at rest. You hold one end, and your friend holds the string some distance away. You wiggle your hand, using energy E, to produce a wave pulse traveling towards your friend. Then you ask your friend to wiggle their hand in the exact "opposite" way when the wave passes by them, which should also require energy

E, but which should create a wave which perfectly destructively interferes with yours. So doesn't an energy $2E$ just vanish into nowhere?

The subtlety is that your friend will be trying to move the string at the precise moment that your wave pulse is passing by them. There are two simple limiting cases we can consider.

- If you created the wave by exerting a vertical force profile $F(t)$, then your friend exerts a force $-F(t)$. But in this case, your friend will be doing *negative* work on the string, because it'll be moving opposite the force they exert. They're just absorbing the pulse you put in, so conservation of energy is satisfied because $E - E = 0$.
- If you created the wave by displacing the rope vertically by $y(t)$, then your friend displaces it by $-y(t)$. But in this case, the net displacement of the rope at your friend's hand will just be *zero*, because their displacement cancels with the displacement of your wave pulse passing by. In this case, your friend is acting like a hard wall boundary condition. They don't do any work, since their hand doesn't move. The forward-moving pulse is indeed completely destroyed, but it is replaced with a reflected pulse of equal energy, so conservation of energy is still satisfied because $E + 0 = E$.

We can also try to route around this issue. For example, suppose you and your friend tied together some strings into a Y shape, and you each held one of the prongs of the Y, and made opposite pulses at the same time. Now there's no issue like the one above, and once the pulses meet at the vertex, they'll perfectly destructively interfere, leaving no energy in the "neck" of the Y. But the waves will also reflect off the vertex, and transmit from one prong to the other. If you carry out the analysis, you'll find that all the energy will get redirected into waves going back up the prongs. Similar arguments hold for electromagnetic waves encountering optical elements, like beam splitters.

All of this is not surprising, because interference comes from wave equations, which in turn are derived from Newton's laws or Maxwell's equations, which obey energy conservation.

Waves I: The Wave Equation

The basics of waves are covered in chapter 18 of Halliday and Resnick, and the rest of the material needed for Olympiad physics can be found in chapter 13 of Wang and Ricardo, volume 1. For more on Fourier series, see chapter 6 of French, and for waves in general, see chapters 7 and 8. For a more advanced perspective, see chapter 16 of Taylor. For many physical examples, see chapters 4 and 6 of Crawford. For more about the physics of music, see these lecture notes. For more fun, see chapters I-47 through I-50 of the Feynman lectures. There is a total of **74** points.

1 Traveling Waves

Waves is a vast subject, ranging from the humble wave on a string to electromagnetic waves, gravitational waves, and quantum matter waves. The math used to analyze waves will appear in just about every physics class you'll ever take. But more importantly, the subject is rich in examples, because waves are the physics of the everyday world.

Example 1

Consider a string with mass density μ under tension T. The transverse displacement of the string is given by the wave function $y(x, t)$, and for simplicity we assume the wave is shallow, i.e. $\partial y/\partial x \ll 1$. What's the equation of motion for y?

Solution

Consider a segment of length Δx. At each end of the segment, the tension provides horizontal and vertical forces

$$T_x = \frac{T}{\sqrt{1 + y'^2}} \approx T, \quad T_y = \frac{Ty'}{\sqrt{1 + y'^2}} \approx Ty'$$

where we're expanding to first order in y'. Therefore the total force is

$$F_y = \Delta T_y = Ty'' \Delta x.$$

This mass of this segment is $\mu \Delta x$, again to first order, so by Newton's Second Law,

$$T \Delta x \frac{\partial^2 y}{\partial x^2} = \mu \Delta x \frac{\partial^2 y}{\partial t^2}$$

Cleaning this up a bit, we have the wave equation

$$\frac{\partial^2 y}{\partial t^2} = v^2 \frac{\partial^2 y}{\partial x^2}, \quad v^2 = \frac{T}{\mu}.$$

Physically, this simply says the string tries to straighten out curvature (represented by $\partial^2 y/\partial x^2$). The wave equation is the simplest possible equation of motion for waves. Even in more complicated situations, we often start with this equation and treat the extra terms as perturbations. The wave equation thus occupies a position like that of the simple harmonic oscillator.

Idea 1

We may factor the wave equation as a difference of squares,

$$(\partial_t^2 - v^2\partial_x^2)y = (\partial_t - v\partial_x)(\partial_t + v\partial_x)y = 0.$$

Therefore, functions that satisfy $(\partial_t \pm v\partial_x)y = 0$ solve the wave equation. It is simple to verify that these are functions of the form

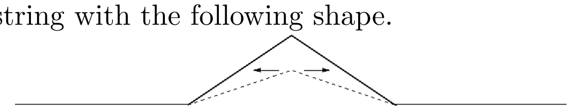
$$y(x,t) = f(x \pm vt).$$

Since the wave equation is linear, superpositions of solutions to the wave equation are also solutions to the wave equation. The general solution is of the form $f(x-vt) + g(x+vt)$ for arbitrary functions f and g.

[1] **Problem 1.** Waves of the form $y(x,t) = f(x \pm vt)$ simply translate with uniform velocity v. Does a wave of the form $y(x,t) = f(x+vt)$ move to the left or the right?

Solution. This wave moves towards the left. To see this, note that at time $t = 0$ the wave profile is $f(x)$, while a small time later it looks like $f(x + v\Delta t)$. This is the graph of $f(x)$ shifted to the left.

[2] **Problem 2.** Consider a string with the following shape.



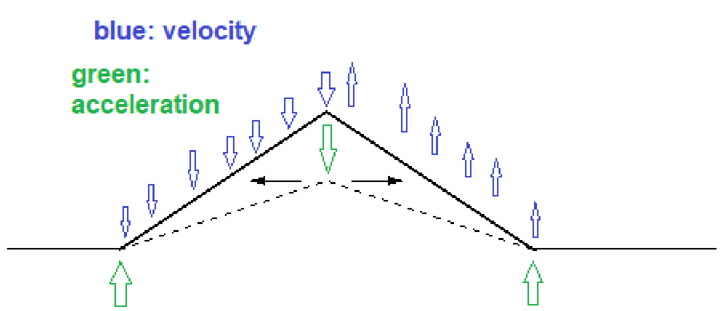
(a) If this is a traveling wave moving to the right with velocity v, carefully draw the velocity and acceleration of every point on the string.

(b) Now suppose the string is held in place, with zero velocity. If it is suddenly relased, sketch the subsequent behavior of the string.

Solution. (a) You can figure out the velocity in two different ways. First, since the wave is proportional to $f(x-vt)$, the vertical velocity is

$$\frac{\partial y}{\partial t} = -vf' = -v\frac{\partial y}{\partial x}$$

so it is proportional to the slope of the string. Alternatively, you can think about how the string has to move so that a moment later, its shape is the same but translated to the right. The result is shown below:



To derive the acceleration, you can think about how the velocity profile has to change as the string moves, or you can think about how it comes about from the tension in the string. In general, the net force depends on the concavity $\partial^2 y/\partial x^2$ of the string. In this case, it's only nonzero at the three kinks.

(b) To keep the string in that position, we must hold it at three points. Since information can't travel faster than the speed of waves, only the bits of string near those three points can move right after release, because they're the only ones that know about the release. The direction of motion can be found with the wave equation (the middle goes down, the ends go up).

Now, for the general solution, note that a solution to the wave equation with zero initial velocity may be written in the form $f(x+vt)+f(x-vt)$. Here, the function f has the same shape as the wave, but half the height. Evidently, two traveling waves split off in opposite directions.

[2] **Problem 3** (HRK)**.** A uniform circular hoop of string is rotating clockwise in the absence of gravity. The tangential speed is v. Find the speed of waves on this string.

Solution. Let the tension in the string be T, and let the mass density be μ and radius be r. By considering force balance on a small angle $d\theta$ of the string,

$$\frac{(\mu r\,d\theta)v^2}{r} = T\,d\theta$$

which gives us

$$T = \mu v^2.$$

The speed of waves on the string is thus

$$v' = \sqrt{\frac{T}{\mu}} = \sqrt{v^2} = v.$$

In other words, waves travel with the exact same speed as the tangential speed! This means that, in an inertial frame, waves on this string look like they're moving with speed 0 or $2v$, depending on whether they're moving with or against the string's rotation. In the former case, the waves look "frozen in place" in the inertial frame.

[2] **Problem 4.** A uniform rope of mass m and length L hangs from a ceiling.

(a) Show that the time it takes for a transverse wave pulse to travel from the bottom of the rope to the top is approximately $2\sqrt{L/g}$. Under what circumstances is this approximation good?

(b) Does the pulse get longer or shorter as it travels?

Solution. (a) The velocity is

$$v = \sqrt{T/\mu} = \sqrt{TL/m}$$

and $T = xmg/L$ where x is the distance from the bottom, so

$$v = \sqrt{xg}.$$

At the most naive level, a wave pulse just travels along the string with a varying velocity, and takes a total time

$$t = \int dt = \int \frac{dt}{dx}\, dx = \int_0^L \frac{dx}{v} = \frac{1}{\sqrt{g}} \int_0^L \frac{dx}{\sqrt{x}} = 2\sqrt{\frac{L}{g}}.$$

This approximation makes sense as long as the wave pulse can't "see" that the velocity is actually changing, which means it works if the pulse has a length much smaller than the length of the rope itself. This idea is discussed further for the case of quantum matter waves in **X1**.

(b) Because the tension is higher at higher points on the rope, the top part of the pulse is always traveling a bit faster than the bottom. Thus, the pulse gets longer as it travels up.

[3] **Problem 5.** [A] At time $t = 0$, the position and transverse velocity of a string obeying the wave equation are given by $y(x)$ and $v_y(x)$. Find an explicit expression for $y(x,t)$ in terms of these functions; this is called d'Alembert's solution. (Hint: construct solutions with initial position $y(x)$ and zero initial velocity, and vice versa, and add them together.)

Solution. When $v_y(x) = 0$, then the waveform $y(x)$ will travel to both directions. Then the waveform contributed by $v_y(x)$, call that $g(x,t)$, can be added on,

$$y(x,t) = \frac{y(x-vt) + y(x+vt)}{2} + g(x,t).$$

The initial conditions state that $\partial_t y(x,t)$ at $t = 0$ is $v_y(x)$, and differentiating the above equation yields

$$\partial_t y(x,t)|_{t=0} = \frac{1}{2}(-vy(x) + vy(x)) + \partial_t g(x,t)|_{t=0}, \quad v_y(x) = \dot{g}(x,t)|_{t=0}.$$

The waveform should go left and right, so

$$\dot{g}(x,t) = \frac{v_y(x-vt) + v_y(x+vt)}{2}.$$

Integrating to find $g(x,t)$, knowing that $g(x,0) = 0$ due to the first equation, gives

$$g(x,t) = \int_0^t \dot{g}(x,t) dt = \frac{1}{2} \int_0^t (v_y(x-vt) + v_y(x+vt))\, dt.$$

Thus, the full solution is,

$$y(x,t) = \frac{1}{2}\left(y(x-vt) + y(x+vt) + \int_0^t (v_y(x-vt) + v_y(x+vt))\, dt \right)$$

Idea 2

A sinusoidal wave has the form

$$y(x,t) = A\cos(kx - \omega t + \phi), \quad v = \frac{\omega}{k}$$

where k is the wavenumber and ω is the angular frequency. They are related to the wavelength and period by

$$k = \frac{2\pi}{\lambda}, \quad \omega = \frac{2\pi}{T}.$$

Sinusoidal waves will be especially useful because the wave equation is linear. Fourier analysis tells us that *any* initial condition can be written in terms of a sum of sinusoids, so if we know what happens to the sinusoids, we know what happens in general by superposition. This is just a generalization of ideas we've seen in **M4** and **E6**. Just as we saw there, it can also be useful to promote y to a complex number, where the physical value of y is the real part; for a sinusoidal wave we would have $y(x,t) = y_0 e^{i(kx-\omega t)}$.

Remark

Physicists almost universally use k and ω rather than λ, f, and T. A nice way of thinking of these variables is that they represent how quickly the phase ϕ changes, in space or time,

$$k = \frac{d\phi}{dx}, \quad \omega = \frac{d\phi}{dt}.$$

If we use a little special relativity, we can even combine these into a single equation,

$$k^\mu = \partial^\mu \phi.$$

The fundamental relation between particle and wave properties in quantum mechanics is

$$p^\mu = \hbar k^\mu.$$

These are the de Broglie relations, which we'll cover in **X1**.

[4] **Problem 6.** For a wave on a string, there are two contributions to the energy: potential energy from stretching, and kinetic energy from transverse motion.

(a) Find the kinetic and potential energy density of the string in terms of T, μ, y, and its derivatives.

(b) Evaluate the above quantities for $y = A\cos(kx - \omega t)$. Is the total energy density uniform?

(c) Show that for a general traveling wave of the form $y = f(x - vt)$, the total kinetic and potential energy are equal.

(d) Show that for any wave function y, total energy is conserved. This will require some integration by parts, as well as the wave equation itself; you should assume y goes to zero at infinity.

(e) Compute the energy of the static configuration in problem 2(b), assuming the triangle has height h and base L, where $h \ll L$.

One warning: as we saw in **E6**, energy is quadratic, so it does *not* obey the superposition principle. Locally, the amount of energy can be more or less than the sum of the energies of the superposed waves, due to interference.

Solution. (a) We will assume that the displacements of the string is small, and take the lowest order terms. Using $\frac{1}{2}mv^2$ for kinetic energy of a piece moving in the transverse direction gets

$$\Delta K = \frac{1}{2}\Delta m \dot{y}^2 = \frac{1}{2}(\mu \Delta x \sqrt{1 + y'^2})\dot{y}^2,$$

$$\frac{dK}{dx} = \frac{1}{2}\mu\dot{y}^2\sqrt{1+y'^2} \approx \frac{1}{2}\mu\dot{y}^2.$$

For the potential energy, the work done on stretching the string is $\Delta U = T\Delta\ell$ where $\Delta\ell = \sqrt{1+y'^2}\,\Delta x - \Delta x \approx \frac{1}{2}y'^2\,\Delta x$ since the displacements are small. Thus

$$\frac{dU}{dx} = \frac{1}{2}Ty'^2.$$

(b) We have

$$\frac{dK}{dx} = \frac{1}{2}\mu A^2\omega^2\sin^2(kx-\omega t), \quad \frac{dU}{dx} = \frac{1}{2}TA^2k^2\sin^2(kx-\omega t).$$

Here, we can see that the total energy density is not uniform, but rather comes in "lumps". This is also true for electromagnetic waves.

(c) The densities are

$$\frac{dK}{dx} = \frac{1}{2}\mu v^2 f'^2, \quad \frac{dU}{dx} = \frac{1}{2}Tf'^2$$

and for a wave traveling in one direction, these densities are exactly equal because $v^2 = T/\mu$, so the total kinetic and potential energy are equal. This is also true for ideal simple harmonic oscillations.

(d) The total energy is

$$E = \int_{-\infty}^{\infty}\left(\frac{1}{2}\mu\dot{y}^2 + \frac{1}{2}Ty'^2\right)dx.$$

Taking the time derivative and applying the wave equation,

$$\frac{dE}{dt} = \int_{-\infty}^{\infty}\mu\dot{y}\ddot{y} + T\dot{y}'y'\,dx \propto \int_{-\infty}^{\infty}\dot{y}y'' + \dot{y}'y'\,dx$$

where we used the wave equation in the second equality. Integrating the first term by parts,

$$\int_{-\infty}^{\infty}\dot{y}y''\,dx = \dot{y}y'\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty}\dot{y}'y'\,dx$$

and the boundary term vanishes by our assumptions. The remaining term is just the opposite of the other term in dE/dt, so $dE/dt = 0$ as desired.

(e) Since the string was initially held steady, there is only potential energy. The amount of potential energy is just T times the total length the string is stretched, so

$$U = T(\sqrt{4h^2+L^2} - L) \approx \frac{2Th^2}{L}$$

where we used $h \ll L$ in the last step.

Remark

How can we account for damping in the wave equation? The simplest thing would be to add a force proportional to v_y, which e.g. could be due to air drag. Then

$$\partial_t^2 y = v^2\partial_x^2 y + A\,\partial_t y.$$

But what if the string is in a vacuum? Then the simplest kind of damping would be due to the energy lost in bending and unbending of the string, which takes the form

$$\partial_t^2 y = v^2 \partial_x^2 y + A\, \partial_t \partial_x^2 y$$

because $\partial_x^2 y$ describes the bending. This is called Kelvin–Voigt damping.

In both cases, it's straightforward to handle the damping since the wave equation remains linear; we just plug in a solution of the form $e^{i(kx-\omega t)}$ and find the new relation between ω and k. If we pick k to be a real number, we will generally find ω to be complex, with its imaginary part corresponding to exponential decay of the wave over time.

[3] **Problem 7.** [A] With a little vector calculus, the results above can be generalized to an arbitrary number of dimensions. For example, ideal waves in three dimensions obey

$$\frac{\partial^2 \psi}{\partial t^2} = v^2 \left(\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} + \frac{\partial^2 \psi}{\partial z^2} \right) = v^2 \nabla^2 \psi$$

where the function $\psi(\mathbf{r}, t)$ could stand for a variety of things, such as the pressure, density, or temperature (for a sound wave) or the electric or magnetic field (for an electromagnetic wave).

(a) For simplicity, let's restrict to waves which have spherical symmetry, so that ψ only depends on r and t. Plug such a spherical wave into the wave equation, and simplify until you get an equation only in terms of the partial derivatives of $\psi(r, t)$.

(b) Because the area of a sphere goes as r^2, we expect the energy density of a spherical wave to fall as $1/r^2$, and therefore expect the amplitude to fall as $1/r$. Therefore, it is useful to consider the quantity $r\psi$, which has this falloff factored out. By considering the differential equation that $r\psi$ obeys, find the general solution for $\psi(r, t)$.

Solution. (a) To do this, we need to simplify the partial derivatives with respect to x, y, and z. We have

$$\frac{\partial \psi}{\partial x} = \frac{\partial \psi}{\partial r} \frac{\partial r}{\partial x}$$

where these partial derivatives are all keeping the other spatial variables y and z constant. To evaluate $\partial r / \partial x$, we note that

$$r^2 = x^2 + y^2 + z^2$$

and take $\partial / \partial x$ of both sides, giving

$$2r \frac{\partial r}{\partial x} = 2x$$

from which we conclude

$$\frac{\partial r}{\partial x} = \frac{x}{r}.$$

Plugging this back in, we conclude

$$\frac{\partial \psi}{\partial x} = \frac{x}{r} \frac{\partial \psi}{\partial r}.$$

Of course, we actually want the second spatial derivative, which is

$$\frac{\partial^2 \psi}{\partial x^2} = \frac{1}{r} \frac{\partial \psi}{\partial r} - \frac{x^2}{r^3} \frac{\partial \psi}{\partial r} + \frac{x}{r} \frac{\partial}{\partial x} \frac{\partial \psi}{\partial r} = \frac{1 - x^2/r^2}{r} \frac{\partial \psi}{\partial r} + \frac{x^2}{r^2} \frac{\partial^2 \psi}{\partial r^2}$$

where we used the product rule. By similar logic for the other components, we conclude

$$\frac{\partial^2\psi}{\partial t^2} = v^2\left(\frac{3-(x^2+y^2+z^2)/r^2}{r}\frac{\partial\psi}{\partial r} + \frac{x^2+y^2+z^2}{r^2}\frac{\partial^2\psi}{\partial r^2}\right) = v^2\left(\frac{2}{r}\frac{\partial\psi}{\partial r} + \frac{\partial^2\psi}{\partial r^2}\right).$$

Of course, a shortcut to this result would be to just look up the formula for the Laplacian in spherical coordinates, but it's good to go through this once in your life.

(b) We notice that

$$\frac{\partial^2}{\partial t^2}(r\psi) = v^2\frac{\partial^2}{\partial r^2}(r\psi)$$

by the product rule. That is, the quantity $r\psi$ obeys the ordinary, one-dimensional wave equation, for which we already know the general solution! We thus conclude

$$\psi(r,t) = \frac{f(r-vt)+g(r+vt)}{r}.$$

If we also want the wave not to blow up at $r=0$, we additionally need $f(-vt)+g(vt)=0$.

2 Standing Waves

Idea 3

A standing wave is a solution to the wave equation of the form

$$y(x,t) = f(x)\cos(\omega t).$$

Typically, only discrete values of ω are possible, with the allowed values depending on the boundary conditions. If the setup is translationally symmetric, then $f(x)$ will be sinusoidal. If you want to get some intuition, try playing with this PhET simulation.

[2] **Problem 8.** Consider a string of length L and wave speed v.

(a) Suppose the ends of the string are fixed, i.e. $y(x,t)=0$ at $x=0$ and $x=L$. Find the standing wave angular frequencies and sketch the configurations.

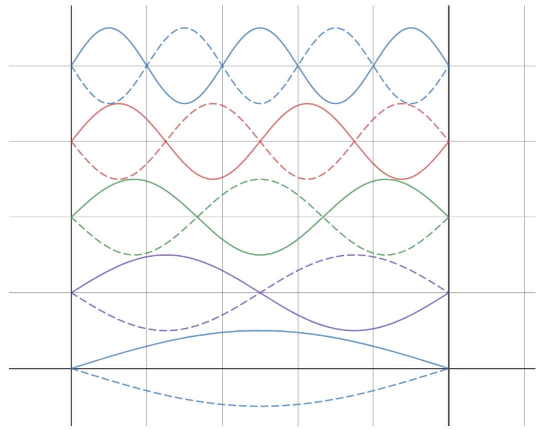
(b) Do the same if the ends of the string are free, i.e. $\partial y/\partial x = 0$ at $x=0$ and $x=L$.

(c) Do the same if one end is fixed and one end is free.

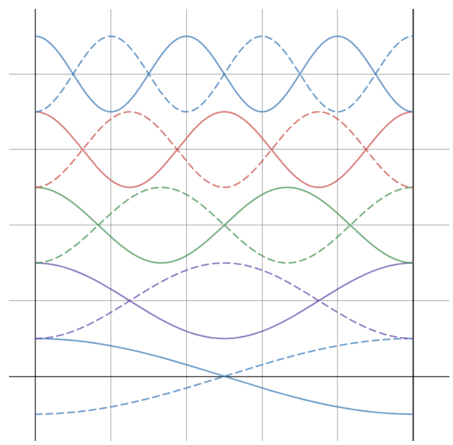
Solution. (a) The standing wave equations are $y(x,t)=A\sin(kx)\cos(\omega t)$ and $y(0,t)=y(L,t)=0$. Thus $kL=\pi n$, giving the angular frequencies

$$\omega_n = \frac{\pi v n}{L}$$

for $n\geq 1$. The waves will look like this:



(b) We can replace the sine with a cosine in the above solution (so integrating or differentiating the above solution with respect to x gets the solutions to this problem). Thus the angular frequencies are the same, $\omega_n = \pi v n/L$, and the waves look like this:

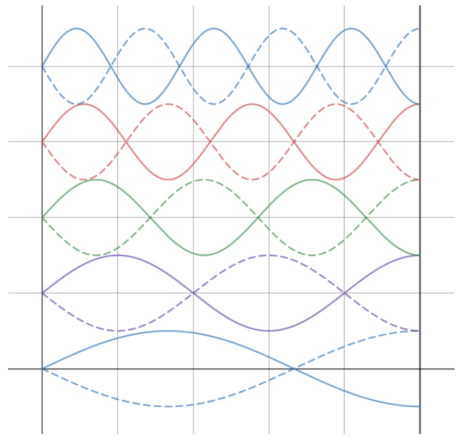


Technically, while the boundary conditions in part (a) required $n \geq 1$, here we can actually take $n \geq 0$. The $n = 0$ solution just corresponds to the whole string being moved up or down and staying there, with zero frequency. But this trivial solution is not typically called a "standing wave", so it's conventional to say the lowest frequency is at $n = 1$.

(c) Let $x = 0$ be fixed and $x = L$ be free. Then for $y(x,t) = A\sin(kx)\cos(\omega t)$, we have $kL = \pi(n + 1/2)$, so

$$\omega_n = \frac{\pi v}{L}(n + 1/2)$$

for $n \geq 0$. Some standing wave solutions are shown below, though note that the diagram unfortunately doesn't include $n = 0$, which is a legitimate standing wave.



[2] **Problem 9.** USAPhO 1997, problem A1.

Idea 4

When a musical instrument plays a note, typically multiple standing waves are excited, so the resulting sound is composed of multiple frequencies. As you saw in problem 8, often the standing wave frequencies are all multiples of a single, lowest frequency. This frequency f_0 is called the fundamental, or first harmonic, while the multiple nf_0 is called the n^{th} harmonic. The fundamental frequency determines the pitch we perceive, while the distribution of energy among the harmonics determines the timbre, or tonal quality, of the instrument.

[2] **Problem 10** (Feynman)**.** Pinch a single length of rubber band about 5 cm long between the fingernails of your two hands, and twang it in the middle to observe the pitch. Then stretch it to several times its original length, observing the pitch as you proceed. Make a simple physical model to explain the observed results.

Solution. You should find that the pitch varies relatively little, even as the length varies by a factor of 5 or more. This pitch depends on the frequency of the lowest frequency standing wave. Now suppose we model the rubber band as a spring with nearly zero relaxed length. Then the tension $T \propto L$ and the linear mass density $\mu \propto 1/L$, so $f \propto v/L \propto \sqrt{T/\mu}/L$ is independent of L, explaining the result.

[5] **Problem 11.** Some instruments, such as xylophones and marimbas, are made with rigid rods instead of strings. The equation that describes transverse vibrations is instead

$$\frac{\partial^2 y}{\partial t^2} = -A\frac{\partial^4 y}{\partial x^4}$$

for a constant A that depends on the material and cross-sectional area.

(a) For a xylophone bar of length L, find the standing wave solutions and their angular frequencies. For simplicity, pretend that the solutions are sinusoidal in space, and that the bar has free ends just like a string, even though this is not true in reality.

(b) When the bar in part (a) is hit, a certain note is sounded. What is the length of the bar that makes a note one octave higher?

(c) [A] ⋆ The actual boundary conditions for a free bar are

$$\frac{\partial^2 y}{\partial x^2} = \frac{\partial^3 y}{\partial x^3} = 0$$

at the endpoints, and the solutions aren't purely sinusoidal in space. Compute the lowest few standing wave angular frequencies and compare them to those you found in part (a). You'll have to use a calculator or computer to do this.

(d) A guitar or piano string satisfies the wave equation with a small additional fourth-order term,

$$\frac{\partial^2 y}{\partial t^2} = v^2 \frac{\partial^2 y}{\partial x^2} - A\frac{\partial^4 y}{\partial x^4}.$$

Show that the standing wave frequencies are not linearly spaced, as they would be for an ideal string, but instead are slightly more spaced out. This effect is called inharmonicity. (Hint: the spatial profiles of the standing waves are still sinusoidal.)

We perceived two notes to be "in tune" when the component frequencies in the notes line up with each other. But since the frequencies are more spread out than ideal harmonics, a piano feels more in tune when the fundamental frequencies are spread out a little bit more. This "stretch tuning" is significant and adds up to about an entire semitone across the piano!

Solution. (a) If we assume that the solutions are $y(x, t) = \cos(kx)\cos(\omega t)$ where $kL = \pi n$, using the differential equation will get

$$-\omega^2 = -Ak^4, \quad \omega = \sqrt{A}\,\frac{\pi^2 n^2}{L^2}.$$

(b) The fundamental frequency is proportional to $1/L^2$, and going up by an octave doubles this, so we need a length of $L/\sqrt{2}$.

(c) This is rather involved; see here for a solution.

(d) Plugging $y(x, t) = \cos(kx)\cos(\omega t)$ into the differential equation and setting $kL = \pi n$ gives

$$-\omega^2 = -v^2 k^2 - Ak^4$$

which gives solutions

$$\omega_n = k\sqrt{v^2 + Ak_n^2} = \frac{\pi n}{L}\sqrt{v^2 + \frac{A\pi^2 n^2}{L^2}}$$

where the extra spacing comes from the $A\pi^2 n^2/L^2$ term. With household equipment, you can check that this holds for real strings, such as piano strings and guitar strings.

Remark: Beam Theory

Where does the strange fourth-order equation for transverse vibrations above come from? Since force is the derivative of energy, it's easier to think about how the energy stored in a rigid rod differs from that of a string. When a string with tension T, mass per length λ, and length ℓ is plucked, giving it a transverse displacement y, then

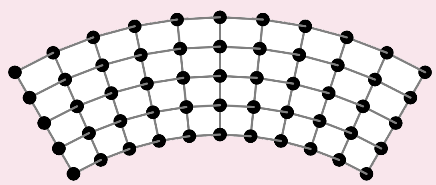
$$\frac{\text{kinetic energy}}{\text{length}} \sim \lambda \dot{y}^2, \quad \frac{\text{potential energy}}{\text{length}} \sim \frac{T\Delta\ell}{\ell} \sim \frac{Ty^2}{\ell^2}$$

where our expression for $\Delta\ell$ comes from the Pythagorean theorem. As we know from **M4**, the ratio of the coefficients gives ω^2, so $\omega\ell \sim \sqrt{T/\lambda}$. For a general wave with wavenumber k, we would replace ℓ with $1/k$ above, giving $\omega \propto k$ as expected.

Now, a rod is characterized by a Young's modulus Y, mass density ρ, length ℓ, width w, and vertical thickness h. If the transverse displacement is y, then

$$\frac{\text{kinetic energy}}{\text{volume}} \sim \rho \dot{y}^2, \quad \frac{\text{potential energy}}{\text{volume}} \sim Y(\text{strain})^2.$$

The tricky part is understanding the strain. If you naively used the same logic as for the string, then all parts of the rod would be stretched, with typical strain $(y/\ell)^2$. This is correct in the limit of large displacements, $y \gg h$, where the rod's thickness is negligible. But for small displacements, it's an overestimate.



As the rod is displaced vertically, it slightly shrinks horizontally. As a result, there is a "neutral line" in the middle of the rod that is neither stretched or compressed. Bonds above the line are stretched, and bonds below the line are compressed.

The neutral line has radius of curvature $R \sim \ell^2/y$. Thus, the strain at the top and bottom of the rod is of order $h/R \sim hy/\ell^2$. Plugging this in gives

$$\omega \sim \sqrt{\frac{Y}{\rho}} \frac{h}{\ell^2}.$$

Again, for a general wavenumber we would replace ℓ with $1/k$, giving the $\omega \propto k^2$ scaling. For a derivation of this result by dimensional analysis, see section 9.2.3 of The Art of Insight.

There's another neat bit of physics we can get here. Consider a horizontal rod with one end fixed at a wall. What is the vertical deflection of the other end of the rod, due to its own weight? The gravitational and internal potential energy densities both have "reasonable", power-law dependence on the deflection y. Thus, when their derivatives match, so that forces balance, their absolute values should match within an order of magnitude,

$$\frac{\text{elastic potential energy}}{\text{volume}} \sim \frac{\text{gravitational potential energy}}{\text{volume}} \sim \rho g y.$$

Solving for the deflection gives

$$y \sim \frac{\rho g \ell^4}{Y h^2}$$

which is the fundamental result of Euler–Bernoulli beam theory. (For a proper derivation in terms of force and torque balance, see chapters 9 and 10 of Lautrup.)

Example 2

How are the sounds of a violin, a trumpet, and a person different in a room full of helium?

Solution

As we saw in **T3**, the speed of sound in air is $\sqrt{\gamma p/\rho}$. When the air is replaced with helium, ρ decreases, increasing the speed of sound.

The standing wave frequencies of a violin are determined by properties of the strings, which aren't affected by the helium. The only difference will be that the way the violin's sound reverberates will be subtly changed.

For the trumpet, the standing wave frequencies are proportional to v/L where L is the length of the air column inside the trumpet. Thus, the standing wave frequencies go up, and the trumpet makes higher-pitched notes.

The human voice is more subtle. A wind instrument works by exciting standing waves inside it. But the source of the human voice is the vibrations of the vocal folds, whose fundamental vibration frequency is directly controlled by your muscles. The entire rest of your vocal tract does not affect what frequencies are present, but rather affects how energy is distributed between those frequencies. (For instance, vowels are characterized by having extra energy near two particular frequencies, called formants.) Helium changes the resonant frequencies of the vocal tract and thus changes which frequencies emitted by the vocal folds are emphasized. It thus changes the timbre, but not the pitch.

[3] **Problem 12.** Some questions about musical instruments.

(a) A piano makes sound by quickly striking a string with a hammer. The seventh harmonic doesn't fit in with the rest that well. If you want to eliminate the seventh harmonic, at what point(s) can you put the hammer?

(b) A violinist can make the note from an open string sound an octave higher by lightly touching it at a point while bowing it somewhere else. Which point(s) should be touched?

(c) Suppose a string has its ends attached to walls. A person can set up a standing wave by holding the string at some point and moving it side to side, sinusoidally with fixed *amplitude*. At which point(s) should the string be driven to maximize the amplitude of a given standing wave? Assume the string experiences very little damping.

Solution. (a) In order to avoid exciting a given standing wave, we should strike the piano at a node of that wave, so that there is zero "overlap" between the wave and the strike. So we can place it $1/7, 2/7, \ldots, 6/7$ of the way down the string.

(b) The midpoint of the string should be touched. This removes all harmonics that don't have a node at the midpoint, i.e. all odd harmonics. The remaining frequencies are even multiples of the fundamental f_0, and since these are all multiples of $2f_0$, we perceive $2f_0$ as the pitch.

(c) Suppose a standing wave has a spatial profile $f(x)$ and frequency f, and to excite it we drive at frequency f and amplitude A at a point x_0. Assuming the damping is small, then in the steady state the spatial profile will be $f(x)(A/f(x_0))$, which is largest if we drive near a *node* of the harmonic, not an antinode!

This may be somewhat unintuitive. The point is that driving at an antinode maximizes the rate at which you initially put energy into the harmonic; this is what we cared about in part (a). But driving at the node maximizes the eventual steady state amplitude, which is what matters here. A real-world example of this is playing jump rope: your hands driving the rope are near at the nodes, because they need to move much less than the middle of the rope does.

[5] **Problem 13.** EuPhO 2017, problem 1. (Hint: don't try to use fancy math here. EuPhO problems are designed to be solved with only elementary math and graph reading.)

Solution. See the official solutions here.

Idea 5

Standing wave solutions also exist for waves in more than one spatial dimension. In the special case where the wave medium is uniform, and shaped like a rectangle (in two dimensions) or a rectangular prism (in three dimensions), all the standing wave solutions can be found by separation of variables. That is, they can all be written as

$$\psi(x, y, z, t) = f(x)g(y)h(z)\cos(\omega t)$$

where f, g, and h are sinusoids.

[3] **Problem 14.** The top of a drum is like a string, in that it has a uniform surface mass density σ and surface tension γ.

(a) Waves on the drum can be described by its height $z(x, y, t)$. Find the wave equation for a drum. What is the speed of traveling waves?

(b) Consider a square drum of side length L, where the boundaries are fixed to $z = 0$. Find the standing wave solutions and the corresponding ω. What's the lowest standing wave angular frequency?

The frequencies will not be multiples of a fundamental frequency, so they are called overtones, rather than harmonics; that's why drums don't sound like they're playing notes. (Special examples, such as the timpani, are designed to mostly excite the harmonic frequencies.)

(c) Why does a drum sound different if you hit it near the edge, versus at the center?

Solution. (a) A piece of mass $dm = \sigma\, dx\, dy$ will experience a net force from differing forces from the sides. Consider the force on the dy side, where the force from surface tension is $\gamma\, dy$, and the vertical component for small displacements is $\gamma\, dy\, \frac{\partial z}{\partial x}$ (since we are going perpendicularly away from the dy side). To find the change in this vertical force across dx, we will take the differential again to get $dF_x = \gamma\, dy\, dx\, \frac{\partial^2 z}{\partial x^2}$. Adding the force from the y direction gets

$$dF_x + dF_y = \sigma\, dx\, dy\, \frac{\partial^2 z}{\partial t^2}$$

which gives the wave equation

$$\frac{\partial^2 z}{\partial t^2} = \frac{\gamma}{\sigma}\left(\frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2}\right)$$

and hence a wave speed of

$$v = \sqrt{\frac{\gamma}{\sigma}}.$$

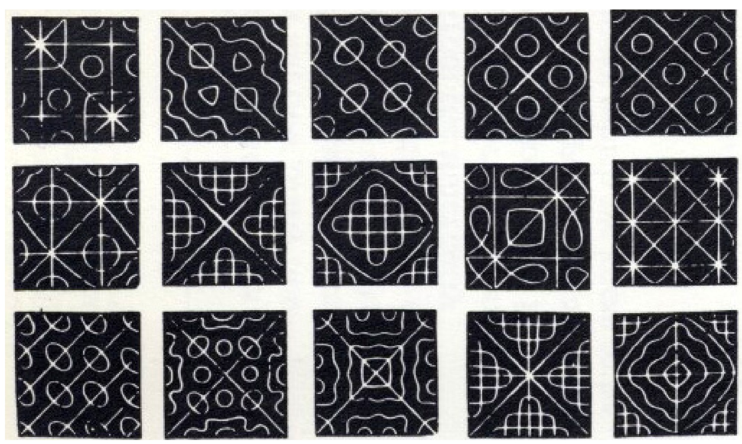
(b) Let $z(x, y, t) = \sin(k_x x)\sin(k_y y)\cos(\omega t)$. The boundary conditions give $k_x = \pi n/L$ and $k_y = \pi m/L$ by the same logic as in problem 8. Plugging this into the wave equation the simplifying gives

$$\omega^2 = \frac{\gamma}{\sigma}(k_x^2 + k_y^2), \quad \omega_{n,m} = \sqrt{n^2 + m^2}\,\frac{\pi v}{L}.$$

Note that neither n or m can be zero, because then the standing wave solution just becomes zero. So the lowest frequency corresponds to $n = m = 1$, where $\omega = \sqrt{2}\,\pi v/L$.

(c) When a drum is hit at the center, it primarily excites the fundamental and some of the lower modes. When it's hit near the edge, it doesn't excite these modes, because it's close to a node for them. Higher modes are excited instead, leading to a "higher", "thinner" sound.

[4] **Problem 15.** When sand is sprinkled on a vibrating metal plate, it forms Chladni patterns. Suppose we (unrealistically) model the plate as a square elastic membrane, as in problem 14, of side length L obeying the wave equation with wave speed v. Unlike in problem 14, we now assume the boundaries of the plate are free.



(a) Do Chladni patterns form at the nodes or antinodes of a standing wave?

(b) Find the general standing wave solutions $z(x, y, t)$ and their angular frequencies.

(c) The plate is also fixed in the middle by the support, so $z = \partial z/\partial x = \partial z/\partial y = 0$ there, which removes many of the standing wave solutions. Find the lowest and second-lowest angular frequencies of allowed standing waves.

(d) Sketch the Chladni pattern for the lowest standing wave frequency.

(e) For the second-lowest standing wave frequency, there will be two independent standing waves with that frequency. What superpositions of them will yield Chladni patterns with 90° rotational symmetry? (If you want to see these patterns, you'll need a computer.)

Solution. (a) The sand can sit still at the nodes, while it gets bounced away from everywhere else. So the Chladni pattern shows the nodes. (The true story is a bit more complicated.

Very fine dust is substantially affected by the air currents created by the vibrating plate. It turns out that this causes dust to accumulate near antinodes instead. To avoid confusion, real demonstrations are often performed with sifted sand, which does not contain dust.)

(b) This is very similar to the result of problem 14. For concreteness, let's put the origin at the bottom-left of the plate. For the boundary condition to be satisfied at the bottom and left edges of the plate, the standing waves should be proportional to cosines,

$$z(x, y, t) = \cos(\omega t) \cos(k_x x) \cos(k_y y).$$

For the boundary conditions to be satisfied at the opposite edges of the plate, we require

$$k_x = \frac{\pi n}{L}, \quad k_y = \frac{\pi m}{L}$$

from which we conclude

$$\omega_{nm} = \sqrt{n^2 + m^2}\, \frac{\pi v}{L}.$$

(c) Because of these additional restrictions, both n and m have to be odd. This means the lowest frequency standing wave corresponds to $(n, m) = (1, 1)$ and $\omega = \sqrt{2}\,\pi v/L$. The next lowest corresponds to $(n, m) = (1, 3)$ and $(3, 1)$ and thus $\omega = \sqrt{10}\,\pi v/L$.

(d) In this case, the Chladni pattern is a centered plus sign.

(e) Setting $\pi/L = 1$ for convenience, the standing wave profiles are

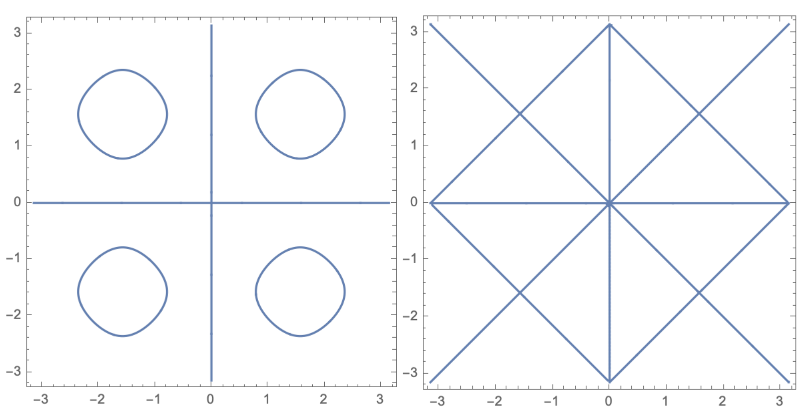
$$f(x, y) = \cos(3x) \cos(y), \quad g(x, y) = \cos(x) \cos(3y).$$

Neither of these has 90° rotational symmetry, but the combinations

$$h_\pm(x, y) = f(x, y) \pm g(x, y)$$

either stay the same, or flip sign upon a 90° rotation. Thus, their Chladni patterns both have 90° rotational symmetry.

The two Chladni patterns, computed with Mathematica with the origin moved to the center of the plate, are shown below.



Remark: Plate Theory

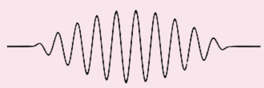
The treatment of problem 15 is inaccurate because the restoring force in a metal plate is rigidity, not tension. The waves actually satisfy the two-dimensional analogue of the fourth-order equation considered in problem 11, which is called the biharmonic equation,

$$-\frac{\partial^2 z}{\partial t^2} \propto \nabla^4 z = \nabla^2 \nabla^2 z = \left(\partial_x^2 + \partial_y^2\right)\left(\partial_x^2 + \partial_y^2\right) z = \frac{\partial^4 z}{\partial x^4} + 2\frac{\partial^4 z}{\partial x^2 \partial y^2} + \frac{\partial^4 z}{\partial y^4}.$$

Of course, I didn't ask you to consider this, because it would have been quite a slog! But if you want to learn more about this thrilling subject, see *Plates*, by Bhaskar and Varadan.

Remark: Wavepackets

Purely sinusoidal traveling waves of the form $e^{i(kx-\omega t)}$ are unrealistic, because they have infinite spatial extent. A realistic alternative is a wavepacket, which looks like a sinusoid with wavenumber k but with a finite envelope, as shown below.



To understand how sinusoids are constructed, consider the superposition of two traveling waves with wavenumbers $k \pm \Delta k$. The wavefunction is

$$e^{i((k-\Delta k)x-(\omega-\Delta\omega)t)} + e^{i(k+\Delta k)x-(\omega+\Delta\omega)t)} = 2e^{i(kx-\omega t)}\cos(\Delta k\, z - \Delta\omega\, t).$$

This is simply a sinusoid of wavenumber k with a slowly varying envelope, whose characteristic size is $1/\Delta k$, reflecting how the two component waves slowly move in and out of phase. The wave is still infinite in size, but this can be remedied by superposing infinitely many wavenumbers; in this case the component sinusoids never get back in phase again.

If the wavenumbers occupy a region Δk, then the size of the envelope is of order $1/\Delta k$, because this is the distance required for the component waves to get out of phase with each other. This yields an "uncertainty principle" for waves,

$$\Delta x\, \Delta k \gtrsim 1.$$

In quantum mechanics, particles are described by waves with $p = \hbar k$. Substituting this in immediately gives the Heisenberg uncertainty principle; it fundamentally holds because one cannot get a finite wave without superposing different wavenumbers.

Alternatively, if we had worked with angular frequencies instead, we would have had

$$\Delta t\, \Delta\omega \gtrsim 1.$$

This is an "acoustic uncertainty principle", also important in digital signal processing, where it is called the Gabor limit. Upon using the de Broglie relations, one finds the energy-time uncertainty principle.

Idea 6

The dispersion relation of a system is the function $\omega(k)$ relating the angular frequency and wavenumber of sinusoidal waves. The phase and group velocity

$$v_p = \frac{\omega}{k}, \quad v_g = \frac{d\omega}{dk}$$

describe the velocities of sinusoidal waves of wavenumber k and the envelopes of wavepackets built from sinusoids near wavenumber k, respectively. We can see the latter result from the remark above: the peak of the envelope is the point where the components are in phase, and this point travels at speed $\Delta\omega/\Delta k \approx d\omega/dk$.

For ideal waves, the dispersion relation is linear, the group and phase velocities are constant and equal, and waves travel while maintaining their shape. When the dispersion relation isn't linear, the group and phase velocities depend on k, so wavepackets gradually fall apart (i.e. they disperse). For more discussion of these topics, see chapter 6 of Morin.

Remark

In **R1**, you learned that nothing can go faster than the speed of light. But the phase velocity can exceed it; for instance, in problem 16 you will find a phase velocity that can be infinite! This is compatible with relativity, because the phase velocity isn't the speed of an actual object. It's just a formal quantity, namely the rate of change of the position of points of constant phase in an infinite plane wave. To reinforce the point, suppose we arranged to stand at different places and clap at the same time. Then we could say "the clap moved from me to you at infinite speed", but clearly nothing about this contradicts relativity.

In some textbooks, you'll read that while the phase velocity can be faster than light, the group velocity can't be, because it's the speed of an actual pulse. But that's not quite true in general either, because that result follows from an approximation. For instance, in materials with really weird dispersion relations, a single pulse can split up into two, in which case the speed of "the" peak or "the" envelope isn't even well-defined. Accordingly, in these cases the group velocity can be formally faster than light, but it doesn't contradict relativity because the group velocity ceases to have its intuitive meaning.

If you're mathematically minded, you might be bothered by the argument that a superluminal phase velocity is okay because no "actual object" moves faster than light, since it seems hard to rigorously define the term "actual object". Luckily, there's a simple and perfectly rigorous definition of the light speed limit: the observable effects of an action must lie in the future light cone of the action. Suppose you change the value of a field at the origin, at time $t = 0$. Then at time t, the field at all points $r > ct$ must be the same as if you didn't make the change at all. The maximum speed at which *changes* of the field propagate is called the signal velocity, and it can never exceed c.

[3] **Problem 16.** Consider transverse waves on a horizontal string with tension T and mass density μ. The string is attached to the ceiling by a large number of vertical springs, so that if the entire string is pulled down, it will oscillate with angular frequency ω_0.

(a) Find the wave equation for waves on this string.

(b) By guessing sinusoidal solutions, find $\omega(k)$ and the minimum possible angular frequency.

(c) Compute the phase and group velocity for wavepackets of angular frequency ω.

(d) What actually happens if you grab one end of the string and try to wiggle it at a frequency below the minimum possible frequency?

If we treat the string as a quantum system, excitations of the string are particles with $E(p)$ determined by the function $\omega(k)$ you found, along with the de Broglie relations $E = \hbar\omega$ and $p = \hbar k$. Therefore, there is a minimum energy for excitations. In a relativistic and quantum context, this means that all the particles must be massive; the minimum energy is mc^2. This is a toy model for how the Higgs field gives particles mass.

Solution. (a) There is now an additional acceleration of $-\omega^2 z$ due to the springs, so the wave equation is

$$\frac{\partial^2 z}{\partial t^2} = \frac{T}{\mu}\frac{\partial^2 z}{\partial x^2} - \omega_0^2 z.$$

(b) Guessing a sinusoidal solution gets

$$\omega^2 = \frac{T}{\mu}k^2 + \omega_0^2$$

and the minimum possible angular frequency is $\omega = \omega_0$.

(c) The phase velocity $v_p = \omega/k$ is

$$v_p = \frac{\omega}{\sqrt{\omega^2 - \omega_0^2}}\sqrt{\frac{T}{\mu}}$$

The group velocity is $d\omega/dk$, and we have

$$2\omega\frac{d\omega}{dk} = \frac{T}{\mu}(2k)$$

so therefore

$$v_g = \frac{T}{\mu}\frac{1}{\sqrt{T/\mu + \omega_0^2/k^2}} = \sqrt{\frac{T}{\mu}}\sqrt{\frac{\omega^2 - \omega_0^2}{\omega^2}}.$$

(d) In this case, you won't manage to create any propagating waves. The part of the string near you will just move up and down, following your hand, analogous to how the position of a mass on a spring simply follows the force if the driving frequency is much lower than the resonant frequency. Or, to say this more formally, the wave solutions of the frequency corresponding to your hand's driving are exponentially decaying, rather than oscillating and propagating, formally because the solution for k is imaginary.

[2] **Problem 17.** The motion of ripples of short wavelength (less than 1 cm) on water is controlled by the surface tension γ and density ρ.

(a) Use dimensional analysis to constrain the phase velocity v_p of ripples with wavenumber k.

(b) Show that $v_g = (3/2)v_p$.

Solution. (a) By dimensional analysis, the only dispersion relation we can write down is

$$\omega(k) \propto \sqrt{\frac{\gamma k^3}{\rho}}.$$

This tells us that

$$v_p = \frac{\omega}{k} \propto \sqrt{\frac{\gamma k}{\rho}}.$$

(b) We have

$$v_g = \frac{d\omega}{dk} \propto \frac{3}{2}\sqrt{\frac{\gamma k}{\rho}}$$

with the same constant of proportionality as in part (a), giving the desired result. So interestingly, this is a case where dimensional analysis can give us a numeric prefactor!

3 Reflection and Transmission

When we considered standing waves in the previous section, we were only considering "steady state" behavior. Now we consider the dynamics of a wave hitting an obstacle more explicitly.

Example 3

Suppose a string defined for $x < 0$ ends at a hard wall at $x = 0$. Show that any wave directed towards the wall will be reflected back upside-down.

Solution

We suppose that we send in a wave of the form

$$y_{\text{in}}(x, t) = f(kx - \omega t).$$

Let the reflected wave be a general wave traveling backward,

$$y_r(x, t) = g(-kx - \omega t).$$

Both of these expressions only have physical meaning for $x < 0$, since the string only exists there. Now, the boundary condition is $y(0, t) = 0$, so we have

$$f(-\omega t) + g(-\omega t) = 0.$$

This tells us precisely that $g = -f$, so the wave is reflected upside-down but otherwise unchanged.

There's an easy way to visualize what's going on here. We can imagine that there really is string for $x > 0$, but that the point $x = 0$ stays fixed for some reason. Then this situation corresponds to an incoming wave coming from the left, and a flipped wave coming from the right. The two meet and cancel at $x = 0$, and the flipped wave continues on going to the left, where the physical string is. Fundamentally, this story works for the same reason as the method of images in electromagnetism: as long as you satisfy the boundary conditions, you can do whatever you want beyond the boundary.

[1] **Problem 18.** Another type of boundary condition is the "soft" boundary condition, which requires $dy/dx = 0$ at $x = 0$. Show that waves are reflected from this boundary but not flipped.

Solution. Let the reflected wave again be a general wave traveling backward,

$$y_r(x, t) = g(-kx - \omega t).$$

The boundary condition is

$$\frac{\partial}{\partial x}\left(f(kx - \omega t) + g(-kx - \omega t)\right)|_{x=0} = kf'(-\omega t) - kg'(-\omega t) = 0$$

which tells us that $f = g$ up to a constant. Of course, that constant is just the initial height of the string at $x = 0$, which we set to zero. Thus, $f = g$, so the wave is reflected without a sign flip.

[3] **Problem 19.** Consider the triangular "plucked" shape of problem 2 again, but suppose that the string starts at rest, and its two outer corners are always held fixed.

(a) Sketch what happens after the string is released. What is the period of the motion?

(b) Confirm explicitly that the initial potential energy of the string is equal to the kinetic energy of the string when it is purely horizontal.

(c) What would prevent a real string from achieving this ideal motion? What will the string look like after a few oscillations?

(d) Try to sketch what happens if the string begins with the "pluck" off-center.

Solution. (a) The answer is shown in the early stages of this video. This generally shows how standing waves are produced. From an initial pluck, the reflections from the ends naturally create the oppositely-moving waves needed to form a standing wave. Since the wave becomes inverted after bouncing off the ends, it needs to travel a distance of $2L$ before it reaches its original state, giving a period of $2L/v = 2L\sqrt{\mu/T}$.

(b) Let the string have a height h initially, so that $y(x, t) = f(x - vt) + f(x + vt)$ where $f(x)$ a triangle of height $h/2$ and base L. Since $f(x)$ was in the form of $f(x) = x(h/2)/(L/2) = hx/L$, then $\dot{y} = hv/L$. When the string is straight, each waveform has its peak at the end with the other half reflected, so the velocity adds up to give a velocity of $2hv/L$ at every point uniformly across the string, so $K = \frac{1}{2}(\mu L)(2hv/L)^2$. The potential energy can be found with $U/L = \frac{1}{2}Ty'^2 = \frac{1}{2}T(2h/L)^2$, which gives $U = 2Th^2/L = 2\mu v^2h^2/L = K$.

(c) Of course, the amplitude decreases over time. But also, sharp features in the string are damped faster, due to both energy losses in the bending string, and drag with the air. As a result, after a while, the initially sharp features get smoothed out. You can start to see this happen in the latter part of the video linked in part (a).

(d) You can check your answer here.

[4] **Problem 20.** [A] The general, turn-the-crank method to find the time evolution of an arbitrary wave on a string of length L is Fourier series. In this method, we write the initial shape $y_0(x)$ of the wave as a combination of standing waves,

$$y_0(x) = \sum_n c_n \sin\frac{\pi n x}{L}, \qquad 0 \le x \le L.$$

We know how each standing wave oscillates in time, so by linearity, the entire wave evolves as

$$y(x,t) = \sum_n c_n \sin\frac{\pi n x}{L} \cos(\omega_n t)$$

where ω_n is the angular frequency of the n^{th} harmonic.

(a) The coefficients c_n can be extracted by integrating $y_0(x)$ against another sine,

$$c_n \propto \int_0^L dx\, y_0(x) \sin\frac{\pi n x}{L}.$$

Explain why this works, and find the constant of proportionality.

(b) Now let's consider the plucked string considered in part (a) of problem 19. If the pluck is centered at the middle of the string and has height h, find the coefficients c_n. (If you're so inclined, you can use a computer to see how the resulting $y(x,t)$ approaches the answer to problem 19 as more terms are included.)

(c) Argue that in general, we have

$$\int_0^L y_0^2(x)\, dx = \frac{L}{2}\sum_n |c_n|^2.$$

By applying this result to the plucked string, show that the Riemann zeta function has value

$$\zeta(4) = \sum_{n=1}^{\infty} \frac{1}{n^4} = \frac{\pi^4}{90}.$$

In fact, this is one of the simplest ways to compute $\zeta(4)$.

We'll use the idea of Fourier series to illustrate some conceptual points in **W2**.

Solution. (a) This works because all the other terms in $y_0(x)$ will cancel out, as seen here:

$$\int_0^L dx \sin\frac{\pi n x}{L}\sin\frac{\pi m x}{L} = \frac{1}{2}\int_0^L dx\left(\cos\frac{\pi x}{L}(m-n) - \cos\frac{\pi x}{L}(m+n)\right)$$

which will always equal to 0 when $m \neq n$, since the arguments of sine (the anti-derivative of cosine) will always be an integer multiple of π. To find the constant of proportionality, we only need to look at the n^{th} term of the expansion of $y_0(x)$:

$$\int_0^L dx\, y_0(x)\sin\frac{\pi n x}{L} = c_n \int_0^L dx\, \sin^2\frac{\pi n x}{L} = c_n L/2$$

Thus the constant of proportionality is $2/L$.

(b) The equation of the plucked string is

$$y_0(x) = h - 2h|x - L/2|/L, \quad 0 < x < L.$$

By symmetry, if n is even, $c_n = 0$ since $y_0(x)$ is even about $x = L/2$ and $\sin(\pi nx/L)$ is odd about $x = L/2$, so the integral will be 0. Thus we will consider the case where n is odd,

$$c_n = \frac{2}{L}\int_0^{L/2} \frac{2hx}{L}\sin\left(\frac{\pi nx}{L}\right)\,dx + \frac{2}{L}\int_{L/2}^{L}(2h - 2hx/L)\sin\left(\frac{\pi nx}{L}\right)\,dx$$

Using symmetry again, the two integrals above are equal, so we only need to evaluate the first. We have

$$c_n = \frac{4h}{L^2}\left(-\frac{L}{\pi n}x\cos\left(\frac{\pi nx}{L}\right)\Big|_0^{L/2} + \frac{L}{\pi n}\int_0^{L/2}\cos\left(\frac{\pi nx}{L}\right)dx\right) + \frac{4h}{L}\int_{L/2}^{L}\left(1-\frac{x}{L}\right)\sin\left(\frac{\pi nx}{L}\right)dx$$

Considering only odd n, the first term will vanish since $\cos(\pi n/2) = 0$ for odd n. Also, by symmetry the integral that goes from $L/2$ to L should be equal to the one that goes from 0 to $L/2$. Thus, for odd n,

$$c_n = \frac{8h}{\pi^2 n^2}\sin\left(\frac{\pi n}{2}\right)$$

while $c_n = 0$ for even n.

(c) Again, integrating sinusoids with different values of n will get 0, and the same value will get $L/2$. Thus when representing $y_0(x)$ as a sum of sinusoids and having the integral of all the cross terms go to 0, we get that

$$\int_0^L y_0(x)\,dx = \int_0^L dx \sum_n c_n^2 \sin^2\frac{\pi nx}{L} = \sum_n c_n^2 \int_0^L dx\,\sin^2\frac{\pi nx}{L} = \frac{L}{2}\sum_n c_n^2.$$

Using $y_0(x) = h - 2h|x - L/2|/L$, and symmetry about $x = L/2$, we get

$$\int_0^L y_0^2(x)\,dx = 2\int_0^{L/2} y_0^2(x)\,dx = 2\int_0^{L/2}\frac{4h^2x^2}{L^2}\,dx = \frac{1}{3}h^2 L.$$

To find the sum of c_n, we use our answer above and consider the odd terms:

$$\frac{L}{2}\sum_n |c_n|^2 = \frac{L}{2}\sum_n \frac{64h^2}{\pi^4(2n+1)^4} = \frac{32h^2L}{\pi^4}\sum_n \frac{1}{(2n+1)^4}$$

To relate that sum to $\zeta(4)$, define the sums for the even and odd numbers as E and O so that

$$\zeta(4) = E + O, \quad E = \sum_n \frac{1}{(2n)^4} = \frac{1}{16}\zeta(4), \quad O = \zeta(4) - E = \frac{15}{16}\zeta(4)$$

Now equating our expressions will get

$$\frac{1}{3}h^2 L = \frac{32h^2L}{\pi^4}\frac{15}{16}\zeta(4)$$

from which we conclude

$$\zeta(4) = \frac{\pi^4}{90}.$$

Idea 7

More generally, the relation between the incoming and reflected waves may depend on the exact form of the incoming wave. In this case, it's useful to consider sinusoidal solution. Let

$$y_{\text{in}}(x,t) = e^{i(kx-\omega t)}.$$

Almost all boundary conditions will state that something at the boundary is constant in time, which is only possible if the reflected wave has the same frequency. So in general we have

$$y_r(x,t) = re^{i(-kx-\omega t)}$$

where r is the reflection coefficient. If the medium exists for $x > 0$, there is also a transmitted wave there, of the form

$$y_t(x,t) = te^{i(k'x-\omega t)}$$

where k' might differ from k, and t is the transmission coefficient. In general, both r and t may depend on k as well as the boundary conditions. Note that the phases of r and t depend on the conventions we used to define $y_r(x,t)$ and $y_t(x,t)$, though the magnitudes don't.

[4] **Problem 21.** Suppose the string at $x < 0$ has a tension T_1 and mass density μ_1, while the string at $x > 0$ has a tension T_2 and mass density μ_2. (If you were doing this at home, it would be difficult to have $T_1 \neq T_2$ since the whole setup would accelerate longitudinally. But for the sake of the problem, suppose the two strings are attached at $x = 0$ by a massless ring which slides on a vertical frictionless pole, so that the normal force from the pole balances the longitudinal force $T_2 - T_1$.) As above, let $y_{\text{in}}(x,t) = e^{i(kx-\omega t)}$.

(a) Write down k' and the boundary conditions at $x = 0$.

(b) Show that the reflection and transmission coefficients are

$$r = \frac{Z_1 - Z_2}{Z_1 + Z_2}, \quad t = \frac{2Z_1}{Z_1 + Z_2}, \quad Z_i = \sqrt{\mu_i T_i}.$$

The quantity Z_i is called the impedance.

(c) What limiting cases correspond to hard and soft boundary conditions? Verify that the reflection coefficients match the results above.

(d) Suppose the incoming wave has the exponential form above, but only lasts for a long but finite time τ. After a long time, the incoming wave is gone, and we have a reflected and transmitted wave. Verify that energy has been conserved. (Be careful: it's not simply $|r|^2 + |t|^2 = 1$.)

Solution. (a) By continuity of the frequency, we have

$$k' = \frac{v_1}{v_2}k = \frac{\sqrt{T_1/\mu_1}}{\sqrt{T_2/\mu_2}}k.$$

The boundary conditions are continuity of the string, and continuity of the transverse component of the tension at $x = 0$, so that the forces on the ring balance.

(b) Continuity of the string requires

$$1 + r = t.$$

Continuity of the transverse component of tension requires

$$T_1(k - rk) = T_2 t k'.$$

Combining this with the result of part (a) gives

$$1 - r = \frac{\sqrt{T_2\mu_2}}{\sqrt{T_1\mu_1}}\, t = \frac{Z_2}{Z_1}\, t.$$

Combining this with the continuity condition and solving gives the desired results.

(c) A hard boundary can be modeled by setting $T_1 = T_2$ and $\mu_2 \to \infty$, which is equivalent to $Z_2/Z_1 \to \infty$. In this limit, $r = -1$ and $t = 0$ as expected.

A soft boundary can be modeled by setting $T_1 = T_2$ and $\mu_2 \to 0$, which sets $Z_2/Z_1 \to 0$. In this case we have $r = 1$ as expected. Oddly we also have $t = 2$, but this isn't really physical because in a soft boundary, the string at $x > 0$ doesn't exist. (It makes no difference from the standpoint of the reflection coefficient whether or not the string at $x > 0$ exists, because the wave carries no energy in the limit $\mu_2 \to 0$.)

(d) First, we need to find the energy for a wave of given amplitude. The kinetic and potential energies in a wave are equal, so we can look at either. The potential energy per unit length is proportional to $Ty'^2 \propto T(Ak)^2$ where A is the amplitude and k is the wavenumber. Since the waves in this problem have fixed frequency, and $v = \omega/k$, we have $k \propto 1/v$. Finally, the total duration T of the wave is fixed, meaning the total length is $L = v\tau \propto v$. Combining these results,

$$U \propto LTA^2k^2 \propto vTA^2/v^2 = \frac{T}{v}A^2 = ZA^2.$$

In other words, the impedance determines the energy per unit time in a wave pulse of given amplitude. That gives some intuition for why the transmission of energy is perfect when $Z_1 = Z_2$. If you want there to be no reflection, then the amplitudes of the transmitted and incoming waves have to match by continuity. This is only consistent with energy conservation if the impedances are matched too.

Therefore, the statement of energy conservation is

$$Z_1 = Z_1|r|^2 + Z_2|t|^2.$$

Plugging in the expressions above and doing the algebra confirms this.

The great thing about the coefficients r and t is that they contain all the information about the reflection and transmission. For complicated problems with multiple interfaces, it's best to work purely in terms of r and t, as solving the wave equation as a whole can get messy.

Remark

In the previous problem, you found that a discontinuity in the wave medium can cause reflection, if the impedances of the two sides are different. This reflection is often wasteful, and can be reduced by insertion of "impedance matching" devices which soften the discontinuity. For example, a conical megaphone helps match the air column of the mouth and throat to the air outside the mouth.

The language of "impedance matching" comes from circuitry. In **E7** you found that transmission lines have a characteristic impedance Z. When two transmission lines are attached, wave reflection occurs if the impedances mismatch. The point here is that the same ideas apply to many kinds of waves, as long as one generalizes the notion of impedance.

Remark

You can generalize the methodology of the previous problem to a large variety of similar problems. For example, suppose the ring at $x = 0$ wasn't massless. Then the boundary conditions would have been changed; instead of the transverse force on the ring vanishing, the transverse force would have had to be equal to its mass times its transverse acceleration. (You may recall that setup from the preliminary problem set.) You could even put the ring on a spring, or give it a damping force (in which case the wave energy is no longer conserved). In all cases, the technique is just to take exponential solutions on both sides and apply the relevant boundary conditions. I won't assign such problems, since they usually involve lots of messy algebra, but the idea is very important in physics.

4 Interference

Idea 8

The intensity of a wave is proportional to its amplitude squared, so if two waves with amplitudes A_1 and A_2 are superposed, the resultant intensity is

$$I \propto (A_1 + A_2)^2.$$

This differs from the sum of the intensities by an interference term,

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\theta$$

where θ is the phase difference between the waves.

[4] **Problem 22.** Consider two thin imperfect mirrors, each with reflection and transmission coefficients r and t from both sides, placed a distance L apart, with air in between them and outside them. This system is called a Fabry–Perot interferometer. A wave with wavenumber k hits the apparatus; we want to find the reflection and transmission coefficients r_{net} and t_{net} of the entire system.

(a) Draw all paths that the light could take to be reflected, and to be transmitted.

(b) By applying the principle of superposition and summing a geometric series, show that

$$r_{\text{net}} = r + \frac{rt^2 e^{2ikL}}{1 - r^2 e^{2ikL}}, \quad t_{\text{net}} = \frac{t^2 e^{ikL}}{1 - r^2 e^{2ikL}}.$$

Note that your answers may differ by phases, depending on your conventions for r_{net} and t_{net}.

(c) Show that all the light is transmitted for some special values of k, even if $r \approx 1$. That is, nearly ideal mirrors can become perfectly transparent! This is called resonant transmission, and it occurs because the reflected waves perfectly destructively interfere. (Hint: be careful, and don't forget that r and t are complex numbers.)

(d) For this system, the statement of energy conservation is

$$|r_{\text{net}}|^2 + |t_{\text{net}}|^2 = 1.$$

Use this fact to get a relation between r and t. (Hint: evaluating this in general is very messy, so pick an appropriate special value of L to simplify things.)

Solution. (a) Light can go in between the mirrors and be reflected any number of times before leaving the system through either mirror, or be immediately reflected from the first mirror.

(b) For a wave getting reflected n times within the system, it will be transmitted once and then travel a distance of nL immediately after the last reflection. The y value of that wave will be

$$y = tr^n e^{iknL} = t\left(re^{ikL}\right)^n = t\alpha^n, \quad \alpha = re^{ikL}.$$

In order for the wave to be reflected by the system, it must be reflected immediately or $2n+1$ times for $n \geq 0$, then travel a distance of L and then transmitted again.

$$\begin{aligned} r_{\text{net}} &= r + t^2 e^{ikL}\sum_{n=0}^{\infty}\alpha^{2n+1} = r + \alpha t^2 e^{ikL}\sum_{n=0}^{\infty}(\alpha^2)^n \\ &= r + \frac{\alpha t^2 e^{ikL}}{1-\alpha^2} \end{aligned}$$

Similarly for t, it must be transmitted, reflected $2n$ times, travel to the other side, and transmitted again.

$$t_{\text{net}} = \sum_{n=0}^{\infty} t^2(\alpha^2)^n e^{ikL} = \frac{t^2 e^{ikL}}{1-\alpha^2}$$

Then

$$r_{\text{net}} = r + \frac{rt^2 e^{2ikL}}{1-r^2e^{2ikL}}, \quad t_{\text{net}} = \frac{t^2 e^{ikL}}{1-r^2e^{2ikL}}.$$

(c) The transmission is

$$T = |t_{\text{net}}|^2 = \frac{|t|^2}{|1-r^2e^{2ikL}|}$$

so the maximum value is achieved when $r^2 e^{2ikL}$ is real and positive, giving

$$T = \frac{|t|^2}{1-|r|^2} = 1$$

which is perfect transmission.

This is a striking result: you can put two nearly perfect mirrors next to each other, and light of the right color will still go right through. This is because the light that goes go through the first can bounce around inside many times, eventually completely canceling the zeroth order reflected wave. This phenomenon is called resonant transmission.

(d) This is quite tricky if one tries to evaluate the desired expression directly. However, note that we only want to get a relationship between r and t. Such a relationship should be independent of L, so we can adjust it to a convenient value. We choose a value so that $r^2 e^{2ikL}$ is real and positive, i.e. so that

$$r^2 e^{2ikL} = |r|^2.$$

In this case, we have $|t_{\text{net}}|^2 = 1$, so we need $r_{\text{net}} = 0$, which means

$$0 = r + \frac{(t^2/r) r^2 e^{2ikL}}{1 - r^2 e^{2ikL}} = r + \frac{t^2 |r|^2/r}{1 - |r|^2} = r + \frac{t^2 |r|^2/r}{|t|^2}.$$

A little rearrangement gives

$$\frac{r^2}{|r|^2} = -\frac{t^2}{|t|^2}$$

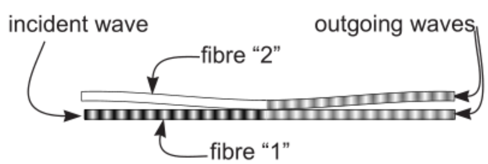
or alternatively,

$$\frac{r}{|r|} = \pm i \frac{t}{|t|}$$

which shows that the phase of r and t differ by $\pi/2$. Note that this derivation only applies to the r and t of an object which is infinitely thin, and has the same material on both sides of it, which is why it doesn't apply to the interface in problem 21.

[3] **Problem 23.** USAPhO 2004, problem A3.

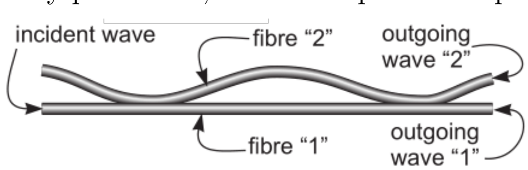
[3] **Problem 24** (Kalda)**.** In fiber optics, devices called equal ratio splitters are often used; these are devices where two optical fibers are brought into such a contact so that if an electromagnetic wave is propagating in one fiber, it splits into two equal amplitude waves traveling in each of the fibers. Assume that all waves propagate with the same polarization, i.e. that all electric fields are parallel.



(a) Show that whenever a wave enters the splitter, from either fiber, one of the outgoing waves is advanced in phase by $\pi/4$, while the other is retarded by $\pi/4$.

(b) From part (a) alone, it's ambiguous which wave is advanced and which wave is retarded. Let's suppose that the fibers are set up so that, when a wave enters along fiber 1, the wave that exits along fiber 1 is advanced. If a wave enters along fiber 2, is the wave that exits along fiber 1 advanced or retarded?

(c) Now consider two sequentially positioned, identical equal ratio splitters, as shown.



This is called a Mach–Zehnder interferometer. The optical path difference between the inter-splitter segments of the two fibers is $30\,\mu$m. Assuming the wavelength of the incoming monochromatic light varies from $610\,$nm to $660\,$nm, for what wavelengths is all the light energy directed into fiber 2?

Solution. (a) Let the ingoing electric field be E_{in}, and let the outgoing fields be E_1 and E_2. If we regard all three of these as phasors, in the sense of **E6**, then

$$E_{\text{in}} = E_1 + E_2, \quad |E_{\text{in}}|^2 = |E_1|^2 + |E_2|^2$$

from continuity of the electric field, and energy conservation. Thus, by the Pythagorean theorem, E_1 and E_2 must differ in phase by $\pi/2$. For the equal ratio splitter relevant to this problem, one of them is advanced in phase by $\pi/4$, while the other is delayed in phase by $\pi/4$.

(b) Consider sending in waves with the same phase and equal amplitude E_{in} along *both* fibers 1 and 2 simultaneously. If the wave that exits along fiber 1 is always advanced, then the final amplitudes are

$$E_1 = (e^{i\pi/4} + e^{i\pi/4})\frac{E_0}{\sqrt{2}} = \sqrt{2}e^{i\pi/4}E_0, \quad E_2 = (e^{-i\pi/4} + e^{-i\pi/4})\frac{E_0}{\sqrt{2}} = \sqrt{2}e^{-i\pi/4}E_0.$$

On the other hand, if a wave that enters along fiber 2 exits along fiber 1 retarded instead, the final amplitudes are

$$E_1 = (e^{i\pi/4} + e^{-i\pi/4})\frac{E_0}{\sqrt{2}} = E_0, \quad E_2 = (e^{-i\pi/4} + e^{i\pi/4})\frac{E_0}{\sqrt{2}} = E_0.$$

Only the second option respects energy conservation, $|E_1|^2 + |E_2|^2 = 2|E_0|^2$, so that one must occur. (This is another example, like the one in problem 21, where carefully applying energy conservation gives more information.) In other words, we have:

$$1 \to 1, 2 \to 2 : \text{advanced}, \quad 1 \to 2, 2 \to 1 : \text{retarded}$$

(c) Let's consider the two components of the waves that eventually exit along fiber 1.

- Part of the incident wave stays in fiber 1 at the first splitter, getting advanced by $\pi/4$. It picks up some phase between the two splitters, then gets advanced by $\pi/4$ again at the second splitter.
- Part of the incident wave goes into fiber 2 at the first splitter, getting retarded by $\pi/4$. It picks up some phase between the two splitters, then (by the result of part (b)) gets retarded by $\pi/4$ again at the second splitter.

For all the light to come out along fiber 2, these two components that come out along fiber 1 have to cancel out. That means they need opposite phases, which implies

$$\pi/4 + k\ell + \pi/4 - (-\pi/4 + k(\ell + \Delta\ell) - \pi/4) = (2n+1)\pi.$$

This simplifies to $k\Delta\ell = 2\pi n$, or $n\lambda = \Delta\ell = 30\,\mu\text{m}$, from which we conclude

$$n \in \{46, 47, 48, 49\}, \quad \lambda \in \{612, 625, 638, 652\}\,\text{nm}.$$

[4] **Problem 25.** APhO 2003, problem 2. A nice, though somewhat clunkily worded question on an interferometer for measuring rotation.

Remark: Interference and Energy Conservation

People sometimes get the impression that interference violates energy conservation, but it doesn't. For instance, in the double slit experiment, you get destructive interference in some places, and constructive interference in other places, so that the total energy stays the same.

A natural followup question is: what if you could engineer waves to have destructive interference *everywhere*? Wouldn't that unambiguously violate energy conservation? Actually, it still won't, but the reason is a bit subtle and depends on the details.

For simplicity, suppose we start with a long string at rest. You hold one end, and your friend holds the string some distance away. You wiggle your hand, using energy E, to produce a wave pulse traveling towards your friend. Then you ask your friend to wiggle their hand in the exact "opposite" way when the wave passes by them, which should also require energy E, but which should create a wave which perfectly destructively interferes with yours. So doesn't an energy $2E$ just vanish into nowhere?

The subtlety is that your friend will be trying to move the string at the precise moment that your wave pulse is passing by them. There are two simple limiting cases we can consider.

- If you created the wave by exerting a vertical force profile $F(t)$, then your friend exerts a force $-F(t)$. But in this case, your friend will be doing *negative* work on the string, because it'll be moving opposite the force they exert. They're just absorbing the pulse you put in, so conservation of energy is satisfied because $E - E = 0$.
- If you created the wave by displacing the rope vertically by $y(t)$, then your friend displaces it by $-y(t)$. But in this case, the net displacement of the rope at your friend's hand will just be *zero*, because their displacement cancels with the displacement of your wave pulse passing by. In this case, your friend is acting like a hard wall boundary condition. They don't do any work, since their hand doesn't move. The forward-moving pulse is indeed completely destroyed, but it is replaced with a reflected pulse of equal energy, so conservation of energy is still satisfied because $E + 0 = E$.

We can also try to route around this issue. For example, suppose you and your friend tied together some strings into a Y shape, and you each held one of the prongs of the Y, and made opposite pulses at the same time. Now there's no issue like the one above, and once the pulses meet at the vertex, they'll perfectly destructively interfere, leaving no energy in the "neck" of the Y. But the waves will also reflect off the vertex, and transmit from one prong to the other. If you carry out the analysis, you'll find that all the energy will get redirected into waves going back up the prongs. Similar arguments hold for electromagnetic waves encountering optical elements, like beam splitters.

All of this is not surprising, because interference comes from wave equations, which in turn are derived from Newton's laws or Maxwell's equations, which obey energy conservation.

Waves II: Interference and Diffraction

Interference and diffraction are covered in chapters 41, 42, and 43 of Halliday and Resnick, and chapter 14 of Wang and Ricardo, volume 1. For more mathematical detail, see chapter 9 of Crawford. For lighter reading with neat examples, see chapters I-29 and I-30 of the Feynman lectures. For much more, see chapters 9 and 10 of *Optics* by Hecht, a well-written but somewhat long-winded university level text. There is a total of **80** points.

1 Double Slit Interference

Remark: Coherence

We introduced interference in **W1**. However, in many cases interference can't be observed at all, because the relative phase between the two waves will oscillate rapidly, making the interference term cancel out. This will even occur with two light sources which are both seemingly the same color; the phase always has an extra "jitter" on top. In practice, one can only observe interference between rays of light from the same source; the phase fluctuations of the two will cancel out.

Even this is not enough, because light emitted at different angles will typically have independent "jitter". In the 19th century, the usual technique for avoiding this was to put a small pinhole between the light source and the actual experiment; if the pinhole is smaller than the light source's coherence length, then the light going through the pinhole is coherent with itself and suitable for use. For sunlight, such pinholes need to be a fraction of a millimeter in size, explaining why the double slit experiment took so long to be carried out properly.

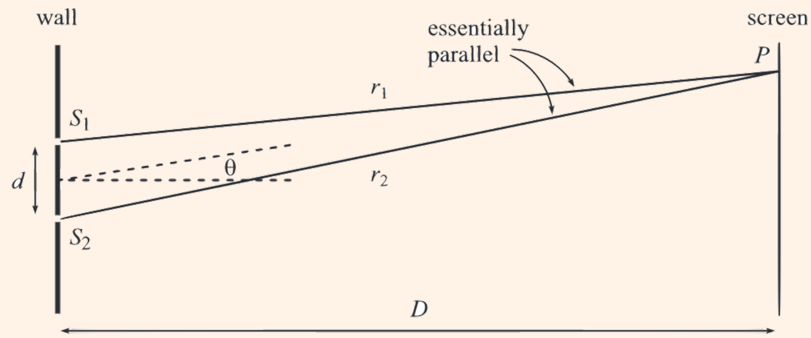
Today, one often uses lasers, which are much more coherent than other light sources; there is no need for the pinhole. Using lasers, it's possible, though challenging, to observe interference between two *different* lasers of the same color. For the problems below we'll generally assume coherence is perfect, but one should remember it's an important practical consideration in real setups.

Example 1

Derive the far-field intensity pattern for the double slit experiment.

Solution

We suppose a plane wave with wavenumber k is incident on two small slits. Using Huygens' principle, the screen absorbs all wavelets except the ones at the slits, so it's as if we have two spherical waves coming from each slit. They produce an interference pattern on a screen a distance D away because they have a phase difference depending on their path length difference.



The amplitude of the light on the screen at some height y is

$$A(t) \propto e^{i(kr_1-\omega t)} + e^{i(kr_2-\omega t)}$$

where we've switch to complex notation, and the quantities r_1 and r_2 depend on y. In principle the two terms don't need to have the same magnitude, e.g. if the two slits are not equal in size. Additionally, as you saw in **W1**, for spherical waves the amplitude actually falls as $1/r$. However, in the far-field limit we have $r_1 \approx r_2$, so we neglect this effect.

At each point on the screen, the amplitude will be some constant times the time-varying phase,

$$A(t) \propto (e^{ikr_1} + e^{ikr_2})e^{-i\omega t} = A_0 e^{-i\omega t}.$$

The brightness of the light we see at that point is proportional to $|A_0|^2$, using the results about wave energy from **W1**. Since only the coefficient squared matters, we can factor out a common phase to get

$$A_0 \sim e^{-ik\Delta r/2} + e^{ik\Delta r/2}, \quad \Delta r = r_2 - r_1$$

and the intensity is

$$I \propto |A_0|^2 \propto |e^{-ik\Delta r/2} + e^{ik\Delta r/2}|^2 = |2\cos(k\Delta r/2)|^2 \propto \cos^2(k\Delta r/2).$$

Next, we find the path length difference Δr. For far-field diffraction, using $D \gg d$,

$$\Delta r = r_2 - r_1 \approx d \sin\theta$$

where θ is the angle from the slits to the point on the screen we're looking at. Therefore,

$$I(\theta) \propto \cos^2(kd(\sin\theta)/2).$$

This yields a series of bright and dark bands on the screen. Most of the time, we'll only be concerned with the few most central minima and maxima, which for typical parameter values lie at $\theta \ll 1$. Then we may use the small angle approximation $\sin\theta \approx \tan\theta = y/D$, giving

$$I(y) \propto \cos^2 \frac{kdy}{2D}.$$

We have a periodic pattern of dark and light fringes. The separation between the minima and between the maxima on the screen is $D\lambda/d$.

The set of approximations made above gives us the theory of Fraunhofer, or far field diffraction. For example, when we calculated the path length difference, which was of order d, we neglected a higher order term of order $d(d/D)$. This is acceptable as long as this is small compared to the wavelength λ. Thus, defining the Fresnel number $F = d^2/D\lambda$, we have far field (Fraunhofer) diffraction when $F \ll 1$. When $F \gg 1$ we can ignore diffraction entirely (why?) and use geometrical optics, as covered in **W3**, while for $F \sim 1$ we must use the more complicated near field (Fresnel) diffraction.

[2] **Problem 1.** Sketch the intensity on the screen as a function of θ if $kd = 8$. What are the bounds on θ? How many completely dark points exist?

[5] **Problem 2.** Some basic but important questions about the double slit result.

(a) Explain intuitively how the pattern on the screen changes if the slits are brought closer together. Do the same if the wavelength of the light is increased.

(b) Suppose the light going through the top slit has a small additional phase shift of θ, e.g. because it passes through a piece of glass in the slit. Describe how the interference pattern changes.

(c) Now suppose the top slit is twice as wide. Describe the interference pattern.

(d) Suppose that coherent white light is passed through the slits. Describe the interference pattern.

(e) It's important to be able to rederive these results quickly. Without looking at the above results, find the spacing between the minima on the screen.

(f) Usually, the double slit experiment is done by shining a laser on both slits. Now suppose a different laser was aimed on each slit. Assume, somewhat unrealistically, that the lasers have very precisely calibrated but slightly different frequencies, say $\Delta f = 1\,\text{Hz}$. Describe what the interference pattern would look like.

[3] **Problem 3.** USAPhO 1999, problem A3. A triple slit experiment.

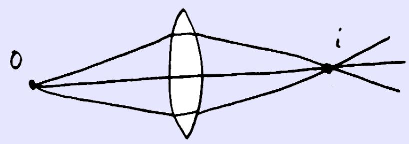
[2] **Problem 4.** A pair of slits is separated by a distance d_1, and two of these pairs are separated by a larger distance d_2, so that $d_2 \gg d_1 \gg \lambda$. Sketch the intensity pattern on the screen for this four-slit apparatus. (Hint: to avoid a complicated computation, factor the expression for the amplitude.)

Idea 1: Image Sources

Some interference problems have complex arrangements of mirrors and lenses. In these cases, actually computing the path length differences can be a nightmare. For instance, you'd have to account for the detailed shape of every lens. Also, you won't just have to compute the path length, but rather than optical path length, which is the ordinary path length weighted by the index of refraction. This is because the index of refraction affects the wavelength and hence the phase difference.

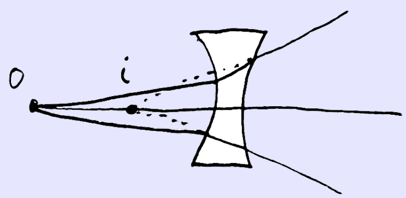
However, there's a trick which makes everything much simpler: any point image can be treated like its own light source. That means you can compute path length differences by starting from the images, rather than having to go all the way back to the original objects.

For real images, there's a very simple way to see why this works. For instance, consider the setup below, where an object o is focused with a lens to an image i.



Fermat's principle of least time tells us that all of the paths shown take the same time, and since phases are directly related to time by $\Delta\phi = \omega \Delta t$, it means that all of the rays arrive at the image with the same phase. That means they leave the image with the same phase, so the image can be treated just like a coherent source. (That is, the phase of the light coming from the image doesn't depend on the direction it comes out.) To find the phase of the waves at i, you can pick any of the paths, most conveniently the one on the symmetry axis.

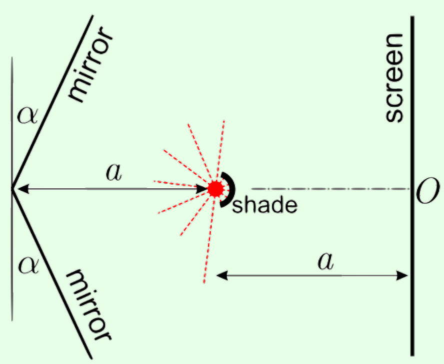
We can also consider virtual images, as shown below.



Here, the Fermat's principle argument doesn't work because the rays never actually meet at i, but we can use Huygens' principle. The key ideas are that (1) light locally propagates perpendicularly to wavefronts, and (2) the phase on a wavefront is always uniform, by definition. The first point implies the outgoing wavefronts are spheres centered on i. The second point implies that the phase only depends on the distance from i, so it can again be treated just like a source. In this way, seemingly impossible questions can be solved instantly.

Example 2: Kalda 17

Consider the optical setup shown below.



Many light and dark bands appear on the screen, with dark bands separated by distance d. Assuming that $\alpha \ll 1$, find the wavelength λ of the light.

Solution

This is actually just a double slit interference problem! Each mirror produces a (virtual) image source reflected behind it, and the pattern on the screen results from the interference between the two image sources, just as if there were two slits at those points.

Specifically, let the light source have coordinates $(0, 0)$, with the screen at $x = a$. Then the image sources are located at $(-a, \pm 2\alpha a)$, so we have a double slit setup with sources $4\alpha a$ apart from each other, a distance $3a$ from the screen. Using our existing results,

$$\lambda = (4\alpha a)\,\frac{d}{3a} = \frac{4\alpha d}{3}.$$

Note that reflection from a mirror changes the phase by π, but that didn't matter in this problem, because both image sources pick up the same phase.

[3] **Problem 5.** USAPhO 2020, problem B2. A problem on interference with images.

2 Thin Film Interference

Idea 2

In general, the phase of a wave is unaffected by reflection from a rarer medium, and flipped by 180° when reflected from a denser medium; here a "denser" medium is defined as one where the wave speed is lower, e.g. one with a higher index of refraction for light. This is analogous to the result for wave reflection in a string derived in **W1**, and is derived starting from Maxwell's equations in **E8**.

Example 3

A very thin soap film on a wire loop looks dark when viewed from above, while a very thin oil film floating on water appears bright when viewed from above. Explain the difference.

Solution

For the soap film, we consider interference between two paths for the light: bouncing off the top surface, or transmitting through and bouncing off the bottom surface. These have almost the same phase, since the soap film is thin, but the former has an extra 180° phase shift. So the two destructively interfere, making the soap film look dark.

For the oil film, the analysis is similar, but both paths have a 180° phase shift, so they interfere constructively, making the oil film look bright.

Remark

This is the usual high school textbook analysis, but the real situation is a bit more subtle. First off, there are actually infinitely many possible paths for the light to take, and sometimes many of these paths are important, as you saw in **W1** for the Fabry-Perot interferometer for $r \approx 1$. However, in both cases above, the two paths we chose were indeed the most important by far.

Second off, the amplitudes upon reflection and transmission must be computed using the results you found in **W1**, and generally won't have the same magnitude. This means that generically we don't get complete destructive interference for the soap film case, just a lowered intensity. Finally, these reflection and transmission coefficients will vary significantly with angle according to Fresnel's equations, as shown in **E8**. We typically ignore this by focusing on normal incidence.

By the way, you might be wondering why we're specializing to thin films; why isn't there thick film interference? Technically there could be interference fringes, but they would be too close to see even if everything was perfect. And in reality, they would then get blurred together due to imperfections in the surfaces, and the spread of frequencies and incidence directions in the incoming light.

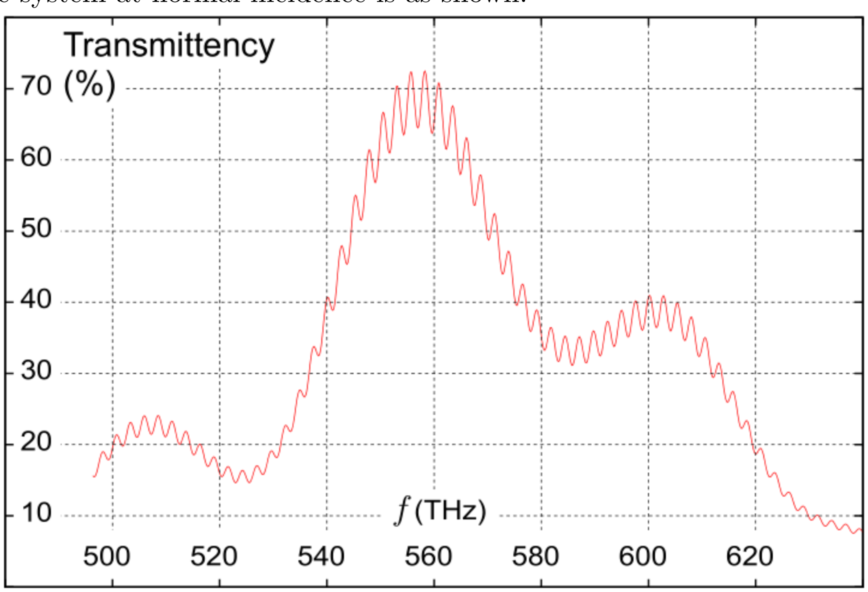
[2] **Problem 6.** AuPhO 2008, problem 15. A simple data analysis problem.

[2] **Problem 7.** The anti-reflection coating on your glasses consists of a thin layer of material whose index of refraction is between that of air ($n = 1$) and glass ($n = 1.5$). The coating is designed to eliminate the reflection of green light, $\lambda = 550\,\text{nm}$. **(improve question)**

(a) Accounting for only the two most significant paths for the light, find the minimum possible thickness of the coating, and its index of refraction. (You'll have to use results for reflection and transmission coefficients derived in **W1**.)

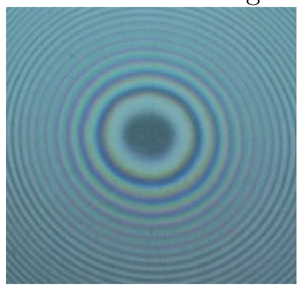
(b) Roughly how much error do we make in neglecting the next most significant path?

[2] **Problem 8** (EFPhO 2004)**.** A thick glass plate is coated by a thin transparent film. The emission spectrum of the system at normal incidence is as shown.



The refractive index of the film is $n = 1.3$. Find the thickness of the film.

[2] **Problem 9.** Newton's rings are an interference pattern formed when a lens is placed on a flat glass surface and illuminated from above by light of wavelength λ. For concreteness, suppose the side of the lens touching the surface is spherical, with radius of curvature R. When viewed from above, one sees an interference pattern with circular fringes.



The most important reflection paths are (1) reflection off the top surface of the lens, (2) reflection off the bottom surface of the lens, and (3) reflection off the flat surface. However, the first reflection path has a very different path length from the other two, which means it won't give rise to visible interference fringes, as explained in the remark above. Instead, the first reflection path just adds some background intensity everywhere, preventing the dark fringes from being perfectly dark. Thus, in this problem we'll only consider the second and third paths.

(a) Explain why the center of the pattern is dark.

(b) Find the radii of the bright and dark fringes, i.e. the values of r where there is a local minimum or maximum of the intensity. For simplicity, assume $r \ll R$.

[3] **Problem 10** (Kalda)**.** A hall of a contemporary art installation has white walls and a white ceiling, lit with a monochromatic green light of wavelength $\lambda = 550\,\text{nm}$. The floor of the hall is made of flat transparent glass plates. The lower surfaces of the glass plates are matte and painted black; the upper surfaces are polished and covered with thin transparent film. A visitor standing in the room will see circular concentric bright and dark strips on the floor, centered around himself. A curious visitor observes that the stripe pattern depends on their height, and upon lowering themselves, sees a maximum of 20 stripes. The film's index of refraction is 1.4 and the glass's is 1.6. Determine the thickness of the film.

3 Diffraction

Next, we turn to diffraction.

Example 4

Find the interference pattern of a diffraction grating, a set of N identical slits in a row, each separated by a distance d.

Solution

Defining $\Delta r = d \sin\theta$ as before, the amplitude is

$$A \sim 1 + e^{ik\Delta r} + e^{2ik\Delta r} + \ldots + e^{(N-1)ik\Delta r} = \frac{e^{ikN\Delta r} - 1}{e^{ik\Delta r} - 1}.$$

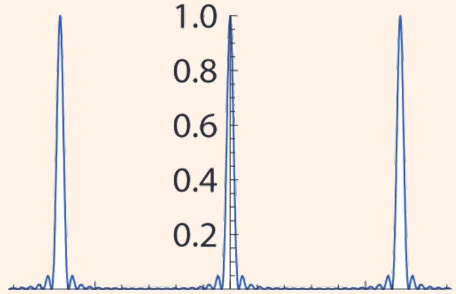
Factoring out a common phase, we have

$$A \sim \frac{e^{ikN\Delta r/2} - e^{-ikN\Delta r/2}}{e^{ik\Delta r/2} - e^{-ik\Delta r/2}} = \frac{\sin(Nk\Delta r/2)}{\sin(k\Delta r/2)}$$

so the intensity is

$$I \propto \frac{\sin^2(Nk\Delta r/2)}{\sin^2(k\Delta r/2)}.$$

The normalized intensity is plotted below as a function of θ.



The numerator yields rapid oscillations which aren't very visible; their envelope is given by the slow oscillations in the denominator. These slow oscillations are the ones we care about; they are the diffraction peaks and occur when

$$\frac{k\Delta r}{2} = n\pi, \quad d\sin\theta = \frac{2\pi n}{k} = n\lambda, \quad n \in \mathbb{Z}.$$

This is intuitive, because at the maxima, the contributions from each slit are in phase, as the path length difference between adjacent slits is a multiple of λ.

[3] **Problem 11.** For practical applications of diffraction gratings, we usually focus on the intensity maxima. However, around each maximum there are also secondary maxima.

(a) Argue that the first minimum occurs when there is a phase difference of $2\pi/N$ between adjacent slits, then compute its angle. This is easiest to see using phasors, i.e. by drawing the individual terms in the amplitude A as vectors in the complex plane.

(b) Show that each secondary maximum is *half* as wide as the central maximum.

(c) Let the intensity at the central maximum be I_0. Assuming $N \gg 1$, use phasors to show that the intensity at the k^{th} adjacent secondary maximum is roughly $I_0/((k+1/2)^2\pi^2)$.

This final result shows that in general, the secondary maxima are much dimmer, and can be neglected; we will ignore them for almost all problems below.

Example 5

Use the previous example to get the interference pattern for a single wide slit of width a.

Solution

We can think of a single slit as the limit of a diffraction grating with $Nd = a$, where $N \to \infty$ and $d \to 0$. Taking these two limits simultaneously is a bit delicate. Starting with our previous result

$$A \sim \frac{\sin(Nkd\sin\theta/2)}{\sin(kd\sin\theta/2)}$$

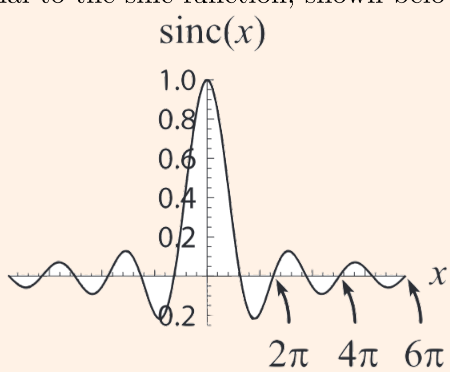
we may substitute $Nd = a$ in the numerator. Only d remains in the denominator, so the $d \to 0$ limit allows us to use the small angle approximation. We thus have

$$A \sim \frac{\sin(ka\sin\theta/2)}{kd\sin\theta/2}.$$

But as $d \to 0$ the expression blows up, because we're taking the number of slits to infinity while keeping the amplitude from each slit constant. To get a consistent limit, we normalize by dividing the amplitude by N, giving $Nd = a$ in the denominator for

$$A \sim \frac{\sin\beta}{\beta} \quad \beta = \frac{ka\sin\theta}{2}$$

The amplitude is proportional to the sinc function, shown below.



What we're really doing here is zooming in on the central maximum of the diffraction grating; the other maxima have been removed by sending the slit spacing to zero.

Remark: Uncertainty Principle

There's a neat way to rephrase our results. In the far field limit and small angle approximation, an opening at height z gives a wave with amplitude $e^{i(k/D)yz}$ at height y on the screen. If we think of a "slit function" $f(z)$ which is equal to one at holes and zero elsewhere, then the amplitude at the screen is simply the Fourier transform of y! Phasors are just a visual way to compute the Fourier transform.

We won't use this language explicitly below, but it can add some intuition if you know it. For example, we know from **W1** that the products of the widths of any function and its Fourier transform are bounded. For example, a wavepacket of width Δx with Fourier components of width Δk obeys $\Delta x\, \Delta k \gtrsim 1$.

In this case, the Fourier pair is screen height y and the scaled slit height $(k/D)z$. (Don't get confused with the notation here; now k is fixed while z varies.) Hence the uncertainty principle says

$$\Delta y\,\Delta z \gtrsim \frac{D}{k} \sim D\lambda$$

which you can check holds for all the examples we've seen so far. The uncertainty principle provides a simple explanation for why making the slits narrower makes the pattern wider.

In fact, this is equivalent to the Heisenberg uncertainty relation $\Delta y \Delta p_y \gtrsim \hbar$ for photons passing through the slit, as you can verify. This makes sense, as we should be able to calculate the diffraction pattern in terms of either the whole light wave, or in terms of what happens to each of the photons in the light wave.

Remark

If we take the limit $a \to \infty$ for the single slit, the central maximum becomes an infinitely sharp, bright point. But in reality, a light will just uniformly illuminate the screen.

The problem is that when a gets too high, the approximations of Fraunhofer diffraction break down, and we must switch to Fresnel diffraction. Fresnel diffraction augments Fraunhofer diffraction with two additional effects.

1. The amplitude of each wavelet falls off as $1/r$.
2. The amplitude of each wavelet is proportional to the "obliquity factor" $(1 + \cos\theta)/2$, where θ is the angle from its original forward direction of propagation. (Strictly speaking, this factor appears in Fraunhofer diffraction too, but in that case it's not too important, because all the wavelets that reach a given point of the screen have about the same θ.)

Both effects matter, but it suffices to consider the first to fix the problem. This amplitude falloff implies that in the case $a \gg D$, the illumination at each point on the screen mostly comes from points on the slit within a distance D, not from the entire slit. Since every point on the screen can see such a range of points, the screen is uniformly illuminated.

For points on the screen near the edge of the slit, there is a gradual shadow, along with some interference bands from "edge diffraction". In the limit $D \gg \lambda$ these residual diffraction peaks get very close and blur together, leaving only a smooth shadow. This is just as expected, as in this case we have $F \gg 1$ and geometrical optics should apply.

For a derivation of Fresnel diffraction starting from the wave equation, see section 10.4 of Hecht. Incidentally, Fresnel diffraction came first historically, since reaching the simpler Fraunhofer regime $F \ll 1$ requires manufacturing tiny optical instruments. (That's why the "Arago spot" is always called a prediction of Fresnel diffraction, even though it also appears in Fraunhofer diffraction; it was one of the first successful predictions of *any* diffraction theory.) This is yet another example of how the textbook treatment we enjoy today is easier. We can start with the simple case, but the pioneers had to get it all right at once.

Remark: Interference vs. Diffraction

Interference is the name for the fact that when waves superpose, their energy doesn't just add; it can become larger than the sum (constructive interference) or smaller (destructive interference). Diffraction is the name for the fact that waves do not need to keep going in a straight line when they hit an obstacle.

What's confusing is that double slit "interference", single slit "diffraction", and a many-slit "diffraction" grating all involve *both* interference and diffraction. Why is it called diffraction when there are one or many holes, but not when there's two? I was very confused about this in high school, but I'm pretty sure there is no difference; it's just historical convention. (However, one pattern is that things that have maxima at larger angles tend to be called "diffraction", because it's more apparent that the direction of the light has been changed.)

[2] **Problem 12.** Comparing the single and double slit.

(a) Show that the minima for the single slit occur when

$$a \sin\theta = n\lambda, \quad n \neq 0.$$

(b) Note that this looks almost identical to the result for the *maxima* of a double slit with separation a. Explain the difference using phasors.

(c) In our analysis of the double slit, we didn't account for the small but nonzero width of each slit. Using the idea of problem 4, sketch the diffraction pattern accounting for this.

[2] **Problem 13.** Practical measurements using diffraction tools.

(a) Suppose you are given a diffraction grating with known parameters and a light source with unknown wavelength. Describe how you would determine this wavelength as accurately as possible. (Be specific: exactly what features of the diffraction pattern would you measure?)

(b) Now suppose you are given a light source with known wavelength and a single slit with unknown width. Describe how you would determine this width as accurately as possible.

[3] **Problem 14** (MPPP 123)**.** Some imperfect diffraction gratings. For this problem, you can ignore secondary maxima. Use the small angle approximation throughout, and neglect any diffraction effects from the finite widths of the slits.

(a) In an imperfect diffraction grating, the slits have equal widths, but the distances between the slits are alternately d and $3d$. Sketch the resulting diffraction pattern, indicating the relative heights of the maxima.

(b) In another imperfect diffraction grating, the slits are evenly spaced, but their widths are alternatively a and b, where $a \approx b$. Sketch the resulting diffraction pattern, indicating the relative heights of the maxima.

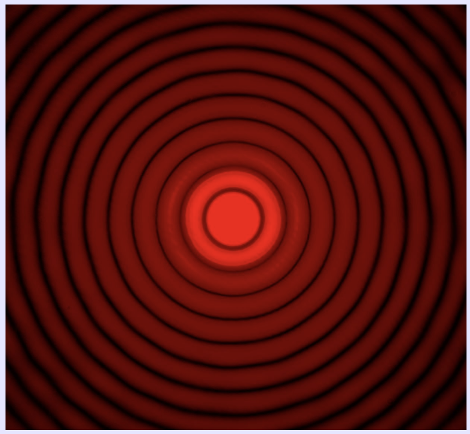
It may be useful to refer to the technique introduced in problem 4.

4 Higher Dimensions

In these problems we tackle interference and diffraction effects in more than one dimension.

Idea 3

We've shown that in a single narrow, tall slit, the central maximum is a tall band, bounded by minima at $\theta = \pm\lambda/d$. If we instead had a circular slit, the central maximum is a circle, bounded by minima at $\theta \approx 1.22\lambda/d$. The radii of the higher-order minima then get closer and closer spaced as one moves outward. The resulting pattern is called an Airy disc.



You can straightforwardly write down an integral that gives the intensity $I(r)$, but the integral can only be performed in terms of special functions, called Bessel functions.

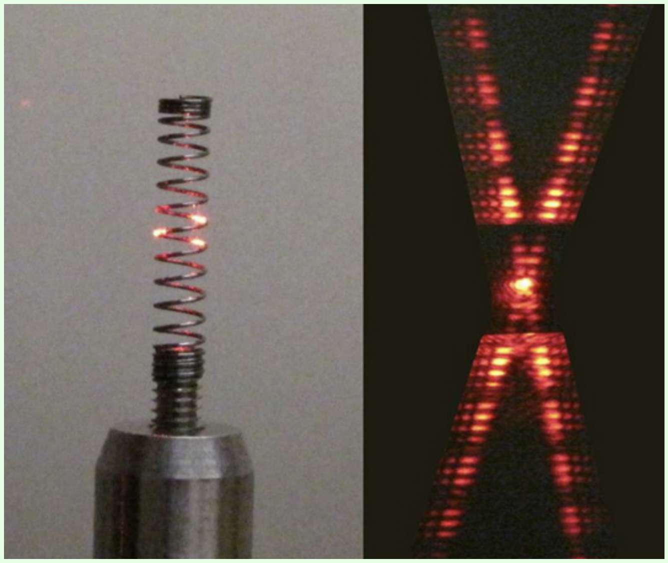
Idea 4: Babinet's Principle

Consider all of the rays R that strike a point P on the screen. If the intensity at P is zero, then the rays must completely destructively interfere. That means that if we split R into two sets of rays R_1 and R_2 in any way, then the amplitudes due to the rays R_1 and the rays R_2 must be equal and opposite, which means either set of rays alone would produce the same intensity at P. This is Babinet's principle.

As a concrete example, consider shining a laser pointer at a wall. There will be a bright spot on the wall at the exact location the laser hits, and darkness everywhere else. Consider some dark point P. If we had instead passed the laser through two slits, we would only get the rays R_1 going through the slits, and we would generally get some nonzero intensity at P, due to the double slit interference pattern. Babinet's principle tells us we would get the exact same intensity at P if we put two slit-shaped obstacles in the way, because then we would get precisely the rays R_2 which don't hit the slits.

Example 6: BPhO 2016.5

When a laser pointer hits a spring, the following pattern is produced on a screen behind it.



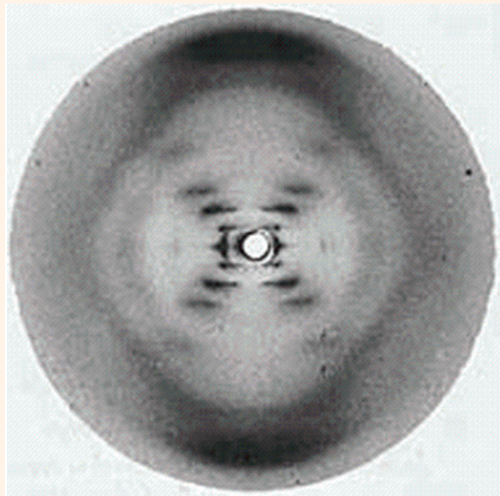
Explain why this happens, and what we can learn about the spring.

Solution

If we look at the spring along the direction the laser pointer is going, it's essentially two sets of obstructions, one going up and to the right (the front of the spring, in the picture), and one going up and to the left (the back of the spring). By Babinet's principle, the resulting diffraction pattern should be the same as if we had two sets of *openings* instead. Thus, we expect to see two independent diffraction patterns, one due to each of these obstructions. The angle between these two patterns is twice the angle that the spiral path of the spring makes with the horizontal.

Now focus on one interference pattern. By Babinet's principle, it's basically a single slit pattern, which is indeed what we see. However, from the reflection of the laser in the picture, we see that the laser beam is wide enough to hit two separate turns of the spring. The result is a double slit pattern multiplied by a single slit pattern, where the latter yields a minimum at approximately every 5 double slit minima. The spacing between the single slit minima tells us the thickness of the wire, and the spacing between the double slit minima tells us the spacing between the turns in the spring. Combining this with what we know about the angle of the spring tells us about the radius of the spring. Thus, we can figure out essentially everything about its 3D shape.

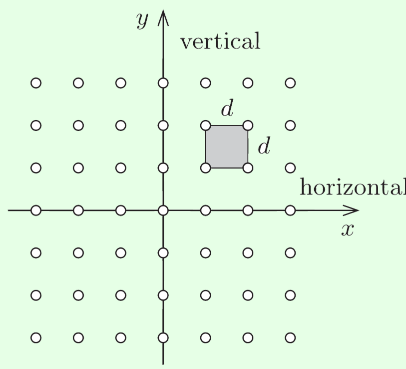
The coolest thing about this problem is that all the logic here applies to diffraction off any helical structure. For example, below is the X-ray diffraction pattern of DNA, taken by Rosalind Franklin in 1952.



As discussed in detail here, such data was used to discover the structure of DNA.

Example 7: MPPP 125

An opaque sheet is perforated by many small holes arranged in a square grid of side length d. It is illuminated by light of wavelength λ, and a screen lies a distance D behind it.



Assuming $D \gg d \gg \lambda$, find the locations of the primary diffraction maxima on the screen.

Solution

Let (x, y) denote coordinates on the sheet, and (x', y') denote coordinates on the screen, with the same center. When we considered one-dimensional diffraction gratings, we found that light which originates from y and hits point y' on the screen has a path length difference yy'/D relative to light that came from point $y = 0$. A similar argument shows light which comes from (x', y') gets a path length difference

$$\Delta \ell = \frac{xx' + yy'}{D}$$

relative to light coming from $x = y = 0$.

For a square grid, $(x, y) = (nd, md)$ for integers n and m, giving a path length difference

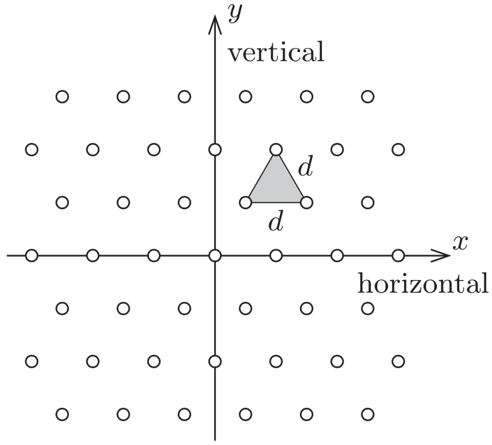
$$\Delta \ell_{n,m} = \frac{d}{D}(nx' + my').$$

We get a diffraction maximum at (x', y') when the light from each hole arrives in phase, which means this quantity must be a multiple of λ for *all* n and m. This occurs precisely when

$$x' = \frac{\lambda D}{d} n', \quad y' = \frac{\lambda D}{d} m'$$

for integers n' and m'. That is, the diffraction maxima also form a square grid of side length $\lambda D/d$. Notice again that the diffraction pattern is "inverse" to the pattern on the sheet. It gets bigger when the sheet gets smaller; for instance, if the sheet is compressed horizontally, the maxima on the screen are stretched horizontally.

[3] **Problem 15** (MPPP 126)**.** Continuing on the previous example, suppose the holes are instead arranged in a triangular grid with side length d.



Find the primary diffraction maxima on the screen. What kind of grid do they form?

[3] **Problem 16.** AuPhO 2015, problem 14. Print out the relevant part of the answer booklet as well. This problem contains several instructive examples of higher-dimensional diffraction patterns.

[5] **Problem 17.** APhO 2021, experimental problem 2.

Remark

The previous few problems illustrate how X-ray crystallography can be used to infer the structure of crystals by diffraction. For an absurdly difficult multi-dimensional diffraction problem, far beyond the scope of the Olympiad, see Physics Cup 2019, problem 5.

5 Technological Applications

Example 8

Gratings split light into its components. If a grating can just resolve the two wavelengths λ and $\lambda + \Delta\lambda$, its resolving power is $R = \lambda/\Delta\lambda$. Compute the resolving power of a diffraction grating with N slits at the n^{th} order maximum.

Solution

Conventionally, we say that two diffraction peaks are distinguishable if the maximum of one falls outside the first minimum of the other. We know the n^{th} order maximum for wavelength λ occurs when $d \sin\theta = n\lambda$, as here the path length difference between adjacent slits is $n\lambda$. Furthermore, the first minimum around this maximum occurs when there is a extra net path length difference of λ across the entire diffraction grating, i.e. when

$$Nd \sin\theta = Nn\lambda + \lambda.$$

Setting the value of $\sin\theta$ equal to that for wavelength $\lambda + \Delta\lambda$, we see that we can just resolve these two wavelengths if

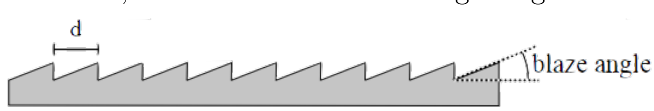
$$\frac{n(\lambda + \Delta\lambda)}{d} = \frac{(Nn+1)\lambda}{Nd}, \quad R = \frac{\lambda}{\Delta\lambda} = Nn.$$

Note that the resolving power is also the number of wavelengths by which the longest and shortest possible paths to the diffraction maximum differ (i.e. the paths through the very top and very bottom slits). The fact that a larger distance may be used to resolve smaller wavelength differences is another manifestation of the uncertainty principle we met in **W1**.

[3] **Problem 18.** Now we consider some more realistic, reflective diffraction gratings.

(a) We may crudely model a reflective diffraction grating as a mirror with N notches, spaced a distance d apart. The notches do not reflect light, but the rest of the mirror serves as a source of Huygens' wavelets when light is incident on the grating. Show that, unlike the transmission gratings we considered above, the zeroth order maximum of a reflective grating is much brighter than the others.

(b) This feature is undesirable because the zeroth order maximum is useless for distinguishing different wavelengths. Instead, most modern diffraction gratings as blazed, as shown.



Suppose that light is incident vertically. How should the blaze angle γ be chosen so that the m^{th} order maximum is the brightest?

In general, reflective diffraction gratings are much easier to make, more flexible, and more common than the transmission diffraction gratings we introduced above. (CDs/DVDs, "holographic" trading cards and stickers, and the pieces of chocolate in the linked video are all examples.) The reason textbooks talk about transmission gratings is just that they make the diagrams a little cleaner.

[2] **Problem 19** (PPP 127)**.** When a particular line spectrum is examined using a diffraction grating with 300 lines/mm with the light at normal incidence, it is found that a line at $24.46°$ contains both red ($640-750$ nm) and blue/violet ($360-490$ nm) components. Are there any other angles at which the same would be observed?

[3] **Problem 20.** Diffraction limits the resolution of optical instruments.

(a) Suppose that light of wavelength λ enters through an aperture of width D. As a result, the light diffracts, which causes the angle of the light's propagation to pick up an additional spread of order θ. Estimate θ.

(b) The diameter of a human pupil is about 3 mm. Estimate the size of the smallest text that a human being could read from 20 ft away.

(c) A typical amateur telescope has an aperture of order 10 cm. The Sun has a radius of 7×10^8 m. Estimate the furthest possible distance, in light years, that such a telescope could resolve a Sun-sized star. (Stars further away than this will just show up as blurry points.)

Remark

The effect described in problem 20 motivates astronomers to build ever larger telescopes. The largest examples are radio telescopes, such as the Arecibo observatory that collapsed in 2020, because diffraction effects are more significant the larger the wavelength is. However, the telescope doesn't have to be one big piece. An array of telescopes can effectively be combined into a telescope whose radius, for the purposes of the diffraction limit, is the total width of the array, as in the Very Large Array. The Event Horizon Telescope was able to resolve a black hole 5×10^7 ly away because it combined telescopes spaced around the entire Earth.

Example 9

How close does a Sun-like star have to be in order to see it with the naked eye in daylight?

Solution

Let the distance to the Sun be d, and the distance to the star be D. Then the ratio of intensities of the two is naively

$$\frac{I_{\text{star}}}{I_{\text{Sun}}} = \left(\frac{d}{D}\right)^2.$$

This suggests the star is hard to see if $D > d$, which is always true. But this is too pessimistic, because the light from the Sun comes from all directions in the sky, while the light from the star comes from only a single direction. The actual ratio we want to calculate is

$$\frac{I_{\text{star}}/\Omega_{\text{star}}}{I_{\text{Sun}}/2\pi} = \left(\frac{d}{D}\right)^2 \frac{2\pi}{\Omega_{\text{star}}}$$

where Ω_{star} is the apparent solid angle of the star in the sky.

This in turn is given by the diffraction limit: if your pupils have radius r, then

$$\Omega_{\text{star}} \sim (\Delta\theta)^2 \sim (\lambda/r)^2.$$

The star should be visible above daylight if the ratio above is at least one or so, which means the maximum distance is

$$D \sim \frac{r}{\lambda} d \sim \frac{3\,\text{mm}}{600\,\text{nm}}(1\,\text{AU}) \sim 5 \times 10^3\,\text{AU} \sim 0.1\,\text{ly}.$$

This is still closer than the closest other star, so you would need a telescope to see any.

Notice how this differs from a microscope! Microscopes are used to resolve finer details on a small object. But most telescopes can't resolve *any* of the details of any but the nearest stars. Increasing the size of the telescope has two benefits: increasing the amount of light that goes through, and improving the contrast due to decreasing the blurring due to diffraction.

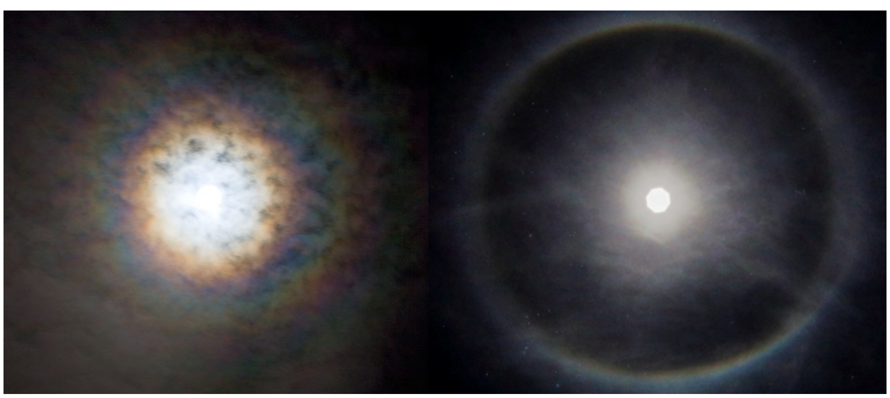
[3] **Problem 21** (PPP 126). A compact disc contains approximately 650 MB of information. Estimate the size of one bit on a CD using an ordinary ruler. Confirm your estimate using a laser pointer. (If you can't find a CD, a DVD will also work.)

[4] **Problem 22.** EFPhO 2005, problem 5. A subtle interference problem.

6 Real World Examples

These questions are not neat and self contained – they illustrate real physical phenomena, for which you'll have to guess an appropriate physical model. Of course, you have the massive advantage of knowing that all of the problems involve interference and diffraction (or do they?).

[4] **Problem 23.** This problem is about some neat atmospheric phenomena. For some parts, it will be useful to use Babinet's principle: for all directions except for $\theta = 0$, the diffraction pattern from an obstacle is precisely the same as the diffraction pattern from an identically shaped slit.



(a) On a foggy night, there are many tiny water droplets in the air. On such nights one can see a ring around the moon, called a lunar corona, shown at left above. The ring is usually reddish in color. If one looks very carefully on a good night, one can see a blue ring outside the red ring and a blueish-white region inside the red ring. On other nights, one can only see a white haze around the moon. Explain these observations.

(b) The size of the corona depends on the atmospheric conditions. Estimate the size of the water droplets in the air if the red ring around the moon appears to have a diameter 1.5 times that of the moon. The angular diameter of the moon in the sky is 0.5°.

(c) On a cold night, there are many thin hexagonal ice crystals in the air. On such nights one can see a much larger, sharper ring around the moon, called a 22° halo, shown at right above. The size of the halo does not depend on the size of the crystals. Explain these observations.

(d) In the photo used in part (c), the moon is shaped like an octagon. Why?

(e) On a cold and exceptionally calm night, the results will be different.

Instead of a circle, one will see two "moon dogs", bright spots displaced about 22° from the moon horizontally. In addition, lights on the ground will produce vertical "light pillars". Explain these observations.

[2] **Problem 24** (Povey)**.** Consider a reflective metal tube, such as a length of copper pipe, with length L and radius r. If you place a diffuse light source at one end of the tube, on the axis of symmetry, and look at it from the other end, with your eye also on the axis of symmetry, then you will see both the light source and bright circular rings around it. Why does this happen? Assuming the light has wavelength λ, calculate the angles of the bright rings.

[2] **Problem 25.** Take a pair of glasses, exhale on them to fog them up, and put them on and look at a light. You should see something strange; why does it happen?

[2] **Problem 26.** Some rare clouds display iridescence, as shown.

What is the explanation for this phenomenon? Why do only a few rare clouds display it? While you're at it, what's going on with the sun in this photo?

[3] **Problem 27.** While under lockdown, a student noticed an odd pattern when the light from a streetlamp shined through their open window. The window is fitted with a metal mesh screen and a curtain. Photos were taken with the curtain up (left) and down (right).

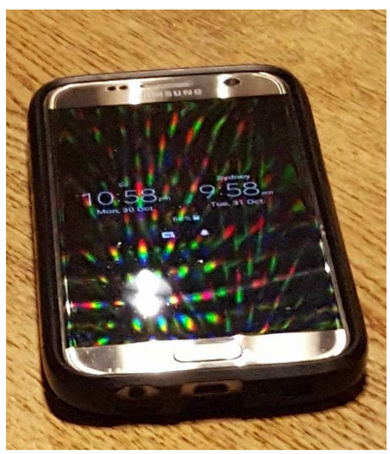
For scale, the streetlamp was about 30 m away, the distance between the metal wires was 1.4 mm, the diameter of each wire was 0.4 mm, and the curtain was woven from fibers whose width was comparable to that of a human hair.

(a) Explain everything you can about the pattern on the left.

(b) Explain everything you can about the pattern on the right.

(c) What, if anything, can we learn about the streetlamp from either picture?

[5] **Problem 28.** One day, somebody sent me a photo of a weird pattern on their phone.



The phone was about a third of a meter away from the camera.

(a) Is the desk light fluorescent (sharp peaks in spectrum), or LED (roughly continuous spectrum)?

(b) Qualitatively explain everything about the pattern seen. In particular, explain the geometrical pattern of the colored dots, the way the colors are distributed, why the colored dots cover the entire phone even though the white central dot only covers a small part of it, and why the central dot is white.

(c) How are the pixels on the phone laid out?

(d) Estimate the pixel spacing on the phone, looking up or estimating any numbers necessary.

(e) Using a suitable source of light, such as a laser pointer, determine the resolution of your own phone screen as accurately as possible. (You can look up relevant wavelengths of light; also note that it won't work with all phones; older ones may fare better.) Prepare a lab report with a data table and an uncertainty estimate, as explained in **P2**, and compare your result against the advertised value.

Waves III: Specific Waves

Sound waves are covered in chapter 19 of Halliday and Resnick, while light waves are covered in chapters 39 and 44. For more about light waves, see chapter 9 of Purcell. For water waves, and many other neat wave phenomena, see chapter I-51 of the Feynman lectures and chapter 7 of Crawford. For more about polarization, see chapter I-33 of the Feynman lectures, or for more detail, chapter 8 of Hecht. Basic geometrical optics is covered in chapter 40 of Halliday and Resnick. There is a total of **87** points.

1 Sound and Longitudinal Waves

[4] **Problem 1.** In this problem, you'll work through Newton's slick derivation of the speed of sound. This derivation avoids thinking about how individual parcels of gas move, which can be confusing, by instead considering the motion of a piston at the end of the gas.

(a) Consider a cylinder of gas of length L and area A, closed on one end with a movable piston on the other end. Suppose the gas exerts a force F on the piston when in equilibrium. We may define an equivalent spring constant by $K = -dF/dL$. Show that for the gas,

$$K = -A^2 \frac{dp}{dV}.$$

(b) Argue that the speed v of longitudinal waves on a spring of mass M, spring constant K, and length L obeys

$$v^2 = \frac{KL^2}{M}.$$

(c) If we assume the sound waves are adiabatic, show that for the gas,

$$v^2 = \frac{\gamma p}{\rho}.$$

Check this answer is reasonable by evaluating the result for air. If each gas molecule has mass m, write the result in terms of γ, T, and m.

Next, we consider some limitations of this result.

(d) In an ideal gas, we assume the particles are noninteracting: they pass right through each other. But for sound waves to propagate, adjacent packets of ideal gas must exert pressure on each other. How is this possible? Use this observation to estimate the maximum possible frequency of sound in a gas in terms of the number density $n = N/V = P/k_BT$, the radius r of a gas molecule, and the speed of sound v.

(e) Our analysis also breaks down if the pressure variations are no longer adiabatic. The rate of heat conduction in a gas with thermal conductivity k_t across a surface of area A is

$$\frac{dQ}{dt} = -Ak_t \frac{dT}{dx}$$

Taking a sinusoidal temperature variation, show that the adiabatic approximation holds when $\omega \ll pk_B/mk_t$. Is this approximation good for audible sound in air, where $k_t \approx 25\,\mathrm{mW/m\,K}$?

(f) In a more traditional derivation of the speed of sound, such as the one in Halliday and Resnick, one would show that

$$v^2 = \frac{B}{\rho}$$

where B is the bulk modulus, the pressure per fractional change in volume,

$$B = -V \frac{dP}{dV}.$$

Check that this result is compatible to your result in (c) for adiabatic compression.

Liquids and solids typically have a much higher B and ρ than gases, and B is high enough so that the speed of sound in liquids and solids is typically greater as well.

Remark

Phase shifts upon reflection for sound waves can be a bit tricky. Recall from **W1** that a hard boundary for a transverse string wave $y(x, t)$ sets y to zero. As a result, upon reflection, y flips sign, but $v_y = \partial y/\partial t$ stays the same.

When a sound wave hits a hard wall, the wall sets the displacement $\xi(x, t)$ to zero. Then upon reflection, the displacement flips sign, while the pressure variation $\delta P(x, t) \propto \partial \xi/\partial x$ stays the same. In standing waves, a hard wall is thus a node for ξ and an antinode for δP. Similarly, when sound waves in a tube reflect off an open end, the end sets δP to zero, so it flips sign. An open end is thus a node for δP and an antinode for ξ.

The rule is always the same: whatever quantity gets fixed to zero by the boundary gets flipped in sign upon reflection, and for a standing wave, that quantity has a node at the boundary. But it's confusing enough that several common high school textbooks get it wrong. Some even state, in their confusion, that "hard boundaries flip transverse waves but not longitudinal ones", which is definitely not true in general.

[3] **Problem 2** (HRK)**.** Some conceptual questions about sound waves.

(a) What is larger for a sound wave, the relative density variations $\Delta\rho/\rho$ or the relative pressure variations $\Delta P/P$? Or does it depend on the situation?

(b) What is larger, the velocity of a sound wave v or the amplitude of the velocity variations Δu of the underlying particles? Or does it depend on the situation?

(c) Bats and porpoises each emit sound waves of frequency about 100 kHz. However, bats can detect objects as small as insects but porpoises only small fish. Why the difference?

[3] **Problem 3.** Consider a rubber rope with unstretched length L_0, which is stretched to length L.

(a) Find the ratio of the speeds of transverse and longitudinal waves.

(b) Experimentally, it is found that the longitudinal waves are much more strongly damped. (You can check this at home, by making such a rope by tying together cut rubber bands.) Can you explain why, by considering the molecular structure of rubber?

Idea 1: Doppler Effect

Working in one dimension with speed of sound c, if a source of sound at frequency f_0 travels at velocity v_s while an observer to their right travels at velocity v_o, the observed frequency is

$$f = \frac{c - v_o}{c - v_s} f_0.$$

Example 1

A speaker is between two perfectly reflective walls and emits a sound of frequency f_0. If you carry the speaker and walk with small speed v towards one of the walls, what do you hear?

Solution

In this solution we'll work to lowest order in v/c everywhere. The wall you're walking toward experiences a sound of approximate frequency $f_0(1+v/c)$ by the Doppler effect, and this is the frequency it reflects. Since you're walking towards the wall, a second Doppler effect occurs, causing you to hear frequency $f_0(1+2v/c)$. We also saw this "double Doppler shift" back in **R1**. By similar reasoning, you hear sound of frequency $f_0(1-2v/c)$ from the wall behind you.

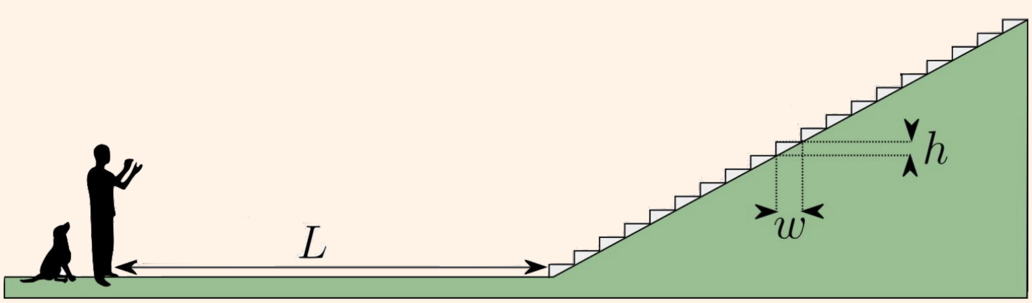
In practice, if v/c is small enough, this means you hear sound of frequency f_0, but with "beats" of frequency $8v/c$. Incidentally, there are many ways of producing beats in everyday life; try to find one!

Example 2

In my former college at Oxford, there is a long staircase that is said to "quack" when one claps at it. What is the explanation of this phenomenon?

Solution

A diagram of the staircase is given below, courtesy of Felix Flicker, fellow of New College.



The key is that each clap reflects off a stair individually. When the echoes arrive back at the listener, they arrive quickly enough to be heard as a pitch.

The width and height of the steps are $w = 30\,\text{cm}$ and $h = 16\,\text{cm}$. Suppose one claps at a distance $L \gg w, h$. The path length differences for reflections off the bottom few steps are approximately $2w$, giving the frequency

$$f = \frac{v}{2w} = 570\,\text{Hz}$$

where we used $v = 343\,\text{m/s}$. The quack then continues, due to reflections off higher and higher stairs. Once the stairs are much further away than L, path length differences for subsequent reflections are approximately $2\sqrt{w^2+h^2}$, giving frequency

$$f = \frac{v}{2\sqrt{w^2+h^2}} = 500\,\text{Hz}.$$

Hence the quack consists of a pitch that starts high and then falls slightly lower as it fades away. For further discussion, see the article *How the Mound got its Quack*.

[3] **Problem 4.** USAPhO 1998, problem B1.

[3] **Problem 5.** USAPhO 2016, problem A1.

[2] **Problem 6.** Some problems about sound waves in everyday life.

(a) Get a coffee cup with a handle and tap on the rim with a spoon. You will hear two distinct pitches, e.g. if you tap directly above the handle, or 45° away from this point. Investigate what happens for different angles. Can you explain why this happens?

(b) The octave key on an oboe forces the resonance mode from the fundamental to the first overtone. It does this by opening a small hole on the back of the clarinet. Should this hole be placed at a pressure node or antinode for the fundamental, or somewhere else entirely?

(c) According to introductory textbooks, the fundamental mode for a pipe of length L and radius $r \ll L$, closed at one end and open at the other, has wavelength $4L$. In reality, it's a little bit different because the radius is nonzero. Is the wavelength actually higher or lower than $4L$?

[3] **Problem 7.** BPhO 2008, problem 2.

2 Polarization

Now we'll introduce polarization for light waves, putting the results of **E7** to work.

Idea 2

The polarization of a light wave refers to the direction of its electric field; the light waves we saw in **E7** were linearly polarized. For example, a light wave traveling along $\hat{\mathbf{z}}$ with its polarization an angle θ from the x-axis has electric field

$$E_x(z,t) = (E_0 \cos\theta)\cos(kz - \omega t), \quad E_y(z,t) = (E_0 \sin\theta)\cos(kz - \omega t).$$

A polarizing filter lets only light of a certain linear polarization through; if light with a linear polarization an angle θ from this axis passes through it, then a fraction $\cos^2\theta$ of the energy is transmitted. Just as light can be incoherent, it can be unpolarized; unpolarized light hitting a polarization filter loses half its energy.

[3] **Problem 8** (HRK)**.** Some basic questions about polarization.

(a) A simple polarizing filter consists of a large number of very thin, closely spaced vertical wires. Does this filter produce vertically or horizontally polarized light?

(b) Suppose we shine vertically polarized light through one slit of a double slit apparatus and horizontally polarized light through the other. What does the intensity pattern look like?

(c) A stack of N polarizing sheets is arranged so that the angle between any two adjacent sheets is θ/N. What is the fraction of light that passes through the stack in the limit $N \to \infty$?

Idea 3

It's useful to specify more complicated polarizations with the Jones unit vector,

$$\mathbf{e} = e_x \hat{\mathbf{x}} + e_y \hat{\mathbf{y}}$$

so that the electric field is

$$\mathbf{E}(z, t) = \mathrm{Re}(\mathbf{e}\, e^{i(kz-\omega t)}).$$

For example, horizontal, vertical, and 45° diagonal polarizations are described by

$$\mathbf{e} = \hat{\mathbf{x}}, \quad \mathbf{e} = \hat{\mathbf{y}}, \quad \mathbf{e} = \frac{\hat{\mathbf{x}} + \hat{\mathbf{y}}}{\sqrt{2}}.$$

We could also multiply any of these by a phase, which would just shift the phase of the wave.

When linear polarizations are combined with a relative phase, the result is circular (or more generally, elliptical) polarization. For example, when $\mathbf{e} = (\hat{\mathbf{x}} + i\hat{\mathbf{y}})/\sqrt{2}$, we have

$$E_x(z, t) = \frac{E_0}{\sqrt{2}} \cos(kz - \omega t), \quad E_y(z, t) = \frac{E_0}{\sqrt{2}} \sin(kz - \omega t)$$

which is a circularly polarized light wave; the electric field at a fixed point rotates in a circle over time, and if one draws the electric field vectors in a line along $\hat{\mathbf{k}}$, they trace out a spiral. Birefringent materials, which have different indices of refraction in different directions, cause such phase shifts, and thus can convert linear polarizations into other polarizations.

Example 3

A plane wave with horizontal polarization $\hat{\mathbf{x}}$ enters an optical device, which does not absorb or reflect any energy. When the plane wave exits the device, it has circular polarization $(\hat{\mathbf{x}} + i\hat{\mathbf{y}})/\sqrt{2}$. What does the device do to light with vertical polarization?

Solution

It seems at first that there isn't enough information to solve the problem. The key is to use energy conservation. If the device doesn't absorb or reflect any energy, all of it must come out. So if we put in a plane wave with horizontal polarization and energy E, a plane wave with circular polarization and energy E comes out. Similarly, if we put in a plane wave with vertical polarization and energy E', a plane wave of unknown polarization and energy E'

comes out.

Now consider superposing waves of energy E and E'. The total energy is

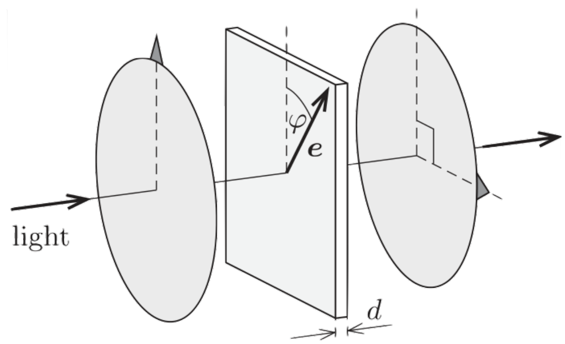
$$\sqrt{E^2 + E'^2 + 2EE' \cos^2 \theta}$$

where θ is the angle between the Jones vectors. (This is just interference, as in **W2**.) Now consider passing a superposition of horizontal and vertical polarization through the device. These polarizations are orthogonal, so the energy put in is $E + E'$. And the energy that comes out must also be $E + E'$, so the outgoing polarizations must also be orthogonal. Thus, a vertical polarization must be converted into circular polarization of the opposite helicity,

$$\hat{\mathbf{x}} \to \frac{\hat{\mathbf{x}} + i\hat{\mathbf{y}}}{\sqrt{2}}, \quad \hat{\mathbf{y}} \to \frac{\hat{\mathbf{x}} - i\hat{\mathbf{y}}}{\sqrt{2}}.$$

By superposition, this completely specifies what the device does.

[3] **Problem 9** (MPPP 127). A birefringent material is placed between two orthogonal polarizing filters. The material has thickness d, and has an index of refraction of n_1 for light linearly polarized along the axis $\mathbf{e}$, and n_2 for light polarized about an orthogonal axis.



If the system is illuminated with light of wavelength λ, give a value for d and orientation of $\mathbf{e}$ that maximizes the transmitted light.

[2] **Problem 10** (HRK). A quarter-wave plate is a birefringent plate that causes a $\pi/2$ phase shift between light polarized along $\mathbf{e}$ and perpendicular to $\mathbf{e}$. Similarly, a half-wave plate causes a π phase shift. Suppose you are given a supply of quarter-wave plates, half-wave plates, and linear polarizers. Given another object, which may be any one of these things, or an semiopaque disk of ordinary glass, describe how you could identify what the object is.

[3] **Problem 11** (MPPP 128). In the first 3D movies, spectators would wear glasses with one eye tinted blue and the other tinted red. This was quickly abandoned in favor of a system that used the polarization of light.

(a) If you wear an old pair of 3D movie glasses, close one eye, and look in the mirror, then you can only see the open eye. Explain how these glasses employ light polarization. What disadvantages might this system have?

(b) If you wear a new pair of 3D movie glasses and do the same, then you can only see the closed eye. Explain why.

[2] **Problem 12** (HRK)**.** A polarizing sheet and a quarter-wave plate are glued together so that, if the combination is placed with face A against a shiny coin, the face of the coin can be seen when illuminated by light of appropriate wavelength. When the combination is placed with face A away from the coin, the coin cannot be seen. Which component is on face A and what is the relative orientation of the components?

[4] **Problem 13.** IZhO 2021, problem 3. A problem on the propagation of light through a waveguide, unifying material from **E7** and **W1**.

3 Water Waves

Water waves are the most familiar examples of waves in everyday life, but you won't find them mentioned often in introductory textbooks, because they're far more complicated than any other kind of wave we'll consider. In all the problems below, we will completely neglect viscosity, surface tension, and compressibility of the water. Despite this, our results will still only be approximate.

[4] **Problem 14.** In this problem we consider shallow water waves, the case where the water depth is much less than the wavelength. Let the water has density ρ and depth d. Assume the hydrostatic pressure formula applies almost everywhere. Furthermore, assume that if the wave is traveling along the x direction, the velocity of the water molecules is solely directed along the x direction and only depends on x, almost everywhere. The height $h(x, t)$ of the wave is a sinusoid, as is the horizontal velocity $v(x, t)$ of the water.

(a) Find a relation between $h(x, t)$ and $v(x, t)$ using Newton's second law.

(b) Find a relation between $h(x, t)$ and $v(x, t)$ using conservation of mass.

(c) Combining these two relations, find the speed of shallow water waves with wavenumber k.

Now let's consider what happens when a shallow water wave created at sea approaches the shore, and the depth d gradually decreases.

(d) Explain why waves always arrive at the shore moving perpendicular to the shoreline.

(e) If the depth is slowly halved, by what factor is the height of the wave multiplied? This phenomenon is known as shoaling.

For general depths, the motion of the water molecules is much more complicated. The problem is that in general the waves are neither fully transverse nor fully longitudinal. We know there is a transverse component because the water surface moves up and down, while we can observe the longitudinal component by seeing how people in a wave pool move forward and backward as the wave passes by. In fact, in general the water molecules move in ellipses! In the limit of a deep water wave, $d \gg \lambda$, the ellipses reduce to circles.

Remark

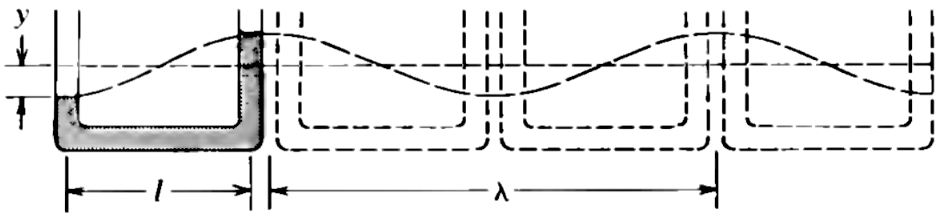
Textbooks commonly say that liquids can't support transverse waves, because they don't support shear stresses. But the waves considered in problem 14 are clearly transverse. This is possible because the textbook statement only applies to the internal forces of water *alone*.

At the surface of the water, gravity provides the transverse restoring force; that's why these waves are also commonly called "gravity waves".

[3] **Problem 15** (French 7.20). In this problem we'll make a crude model for a deep water wave by using an artificial setup that forces the water to move purely vertically.

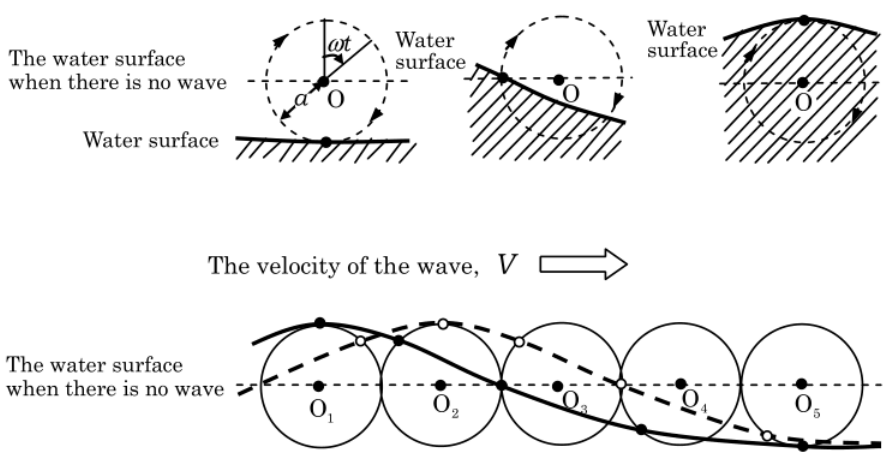
(a) Consider a U-tube of uniform cross section with two vertical arms, so that the horizontal section of the tube, with length l, is much longer than the water depth. Show that the period of oscillation is approximately $\pi\sqrt{2l/g}$.

(b) Now suppose a succession of such tubes is placed next to each other and set oscillating to define a succession of crests and troughs, as shown.



This may be interpreted as a water wave of wavelength $\lambda = 2l$. Conclude that $v \approx \sqrt{2/\pi}\sqrt{g/k}$. Is this a phase or group velocity?

[3] **Problem 16** (Japan). In this problem, we'll give another heuristic treatment of deep water waves, accounting for the circular motion. For simplicity, we assume that the water molecules at the surface of the wave move in uniform circular motion with radius a and angular velocity ω, as shown.



(a) Express the velocity v of the wave in terms of the amplitude a and the angular frequency ω.

(b) Now work in a frame of reference moving to the right with velocity v. In this frame, the surface of the water is completely stationary, while molecules travel along the surface. Consider a small parcel of water which travels from a valley to a peak. By applying conservation of energy, derive a relationship between v, ω, and g.

(c) Combine these results to show that $v = \sqrt{g/k}$, the *exact* result for deep water waves. Again, is this a phase or group velocity? Use this result to calculate the other velocity.

Of course, this derivation is quite incomplete since we haven't shown why circular motion occurs; you can find a complete derivation here.

As you can see, water waves are quite complex. A diagram of the speeds of nine different limiting cases of water waves can be seen in section 8.4 of The Art of Insight.

4 Reflection and Refraction

Now we'll introduce reflection and refraction with some real-world applications.

Idea 4

If a wave hits an interface, while traveling at an angle θ_1 to the normal to the interface, then it will generically both reflect and refract. The angle of the reflected ray is $\theta_2 = \theta_1$, and the angle of the refracted ray obeys $n_1 \sin\theta_1 = n_2 \sin\theta_2$. If there is no solution for θ_2 in the latter equation, then only reflection occurs.

These results follow directly from Huygens' principle, so they are very general, applying to light waves, sound waves, water waves, and so on, as long as the index of refraction n_i is always defined to be inversely proportional to the wave speed in each medium.

[2] **Problem 17.** Some conceptual questions about reflection and refraction.

(a) Does the index of refraction determine the phase velocity or the group velocity?

(b) Does a light beam of finite width get wider or narrower upon passing from air to water? Assume the light enters at an angle to the normal.

(c) Three mutually perpendicular mirrors intersect so as to form an internal right-angled corner. If a light ray strikes all three mirrors, show that it ends up traveling exactly opposite to its original direction. Can you think of a practical application of such a "corner reflector"?

[3] **Problem 18** (Feynman)**.** Practical measurements of the index of refraction.

(a) How would you measure the index of refraction of a polished rectangular prism of glass?

(b) How would you measure the index of refraction of a polished plate of black obsidian?

Example 4

Let the index of refraction at height h above the Earth's surface be $n(h)$. In terms of $n(0)$ and the Earth's radius R, what should dn/dh be at the surface so that light rays orbit in circles around the Earth, with constant height?

Solution

First, let's ignore the curvature of the Earth. Consider a light ray moving slightly upward, at a small angle θ to the horizontal, experiencing index of refraction n. Over a horizontal distance L, it goes up by a height $L\theta$. At this point, it will have a different angle θ' to the horizontal, and experience index of refraction $n + L\theta\, dn/dh$. Snell's law says

$$n \cos\theta = \left(n + L\theta \frac{dn}{dh}\right) \cos\theta'$$

and expanding to lowest order in the small angles θ and θ' gives

$$\frac{n}{2}(\theta'^2 - \theta^2) = L\theta \frac{dn}{dh}.$$

Approximating again to lowest order gives

$$\theta - \theta' \approx -\frac{L}{n}\frac{dn}{dh}.$$

Thus, the light ray turns through an angle of $(1/n)\, dn/dh$ per unit horizontal distance. For the light ray to stay at a constant height over the curved Earth, this must equal $1/R$, giving

$$\frac{dn}{dh} = -\frac{n(0)}{R}.$$

More generally, this calculation shows that light bends towards the direction with higher n. In the case of air, where $n - 1 \ll 1$, we can rewrite this as

$$\frac{d(n-1)}{dh} \approx -\frac{1}{R}$$

which can plausibly occur on Earth, due to the nice coincidence that $n-1$ and H/R (where H is the typical scale height of the atmosphere) are both of order 10^{-3}.

Remark: Mirages

There are two classes of mirages.

- When $dn/dh < 0$, light rays bend down. If there is a distant object at the horizon, its image will appear *above* the horizon. This is called a superior mirage, or "fata morgana".
- When $dn/dh > 0$, light rays bend up. Then a distant object at the horizon will appear *below* the horizon, forming an inferior mirage. This also applies to the sky near the horizon, producing the illusion of water on the ground sometimes seen in deserts.

In air, the refractive index is close to 1, and $n - 1 \propto \rho \propto P/T$, where ρ is the air density and the second step used the ideal gas law. Usually we have $d\rho/dh < 0$, since $dP/dh < 0$ in hydrostatic equilibrium, but it depends on the value of dT/dh.

- In normal conditions, the Sun warms the ground and the hot air rises and adiabatically mixes the atmosphere (as discussed in **T1**), so that $dT/dh < 0$. This partially cancels the effect of the pressure variation, so that dn/dh is still negative but has small magnitude, so that mirage effects aren't apparent.
- In rare "thermal inversion" conditions, we have $dT/dh > 0$, so that dn/dh is negative with large magnitude, leading to strong superior mirage effects. If dn/dh is negative

enough, it can match the value computed in example 4, allowing an observer to see arbitrarily far along the horizon despite the curvature of the Earth. This was the reason the famous Bedford Level experiment concluded the Earth was flat.

- In hot deserts, the air near the ground is very hot, so that $dT/dh < 0$ with a large magnitude. (A strongly negative dT/dh also occurs in cold days above water, since the water stays warmer than the air above it.) Here the temperature gradient overpowers the pressure gradient, so that $dn/dh > 0$ and inferior mirages can occur.

Proponents of the flat Earth hypothesis claim that the Earth only seems curved due to atmospheric refraction. But they have it backwards: in almost all conditions $dn/dh < 0$, which makes the Earth look *less* curved than it actually is.

[4] **Problem 19.** IPhO 1995, problem 2. Refraction in the presence of a linearly varying wave speed. (This is a classic setup with a neat solution, also featured in IPhO 1974, problem 2.)

[3] **Problem 20.** INPhO 2019, problem 1. Another exercise on refraction, with an uglier solution.

[3] **Problem 21.** IPhO 2003, problem 3B. An exercise on refraction and radiation pressure.

[4] **Problem 22.** IPhO 1993, problem 2. Another exercise on the same theme.

5 Ray Tracing

Idea 5

A pointlike object emits light rays in all directions. When those light rays subsequently converge at some other point, that point is the object's real image. If they don't actually converge, but all propagate outward with a common center, that point is the object's virtual image. In general, if we're given that an image exists, we can find its location by following the paths of selected rays from the object and looking for intersections.

[2] **Problem 23.** AuPhO 2020, section C. A cute series of real-world examples.

[2] **Problem 24.** A pinhole camera is a simplified camera with no lens. It simply consists of a small hole with a screen behind it.

(a) Explain how the pinhole camera works by ray tracing.

(b) What are the disadvantages of having an especially small hole, or an especially large hole?

(c) Assuming the object being photographed is very bright, estimate the optimal aperture size for taking a clear picture with a pinhole camera.

[2] **Problem 25.** AuPhO 2013, problem 11. A simple problem which tests intuition for everyday optics. Write your answers on the official answer booklet.

[3] **Problem 26.** AuPhO 2019, problem 12. Another nice question involving drawing good sketches. Write your answers on the official answer booklet.

[3] **Problem 27.** IZhO 2020, problem 1.3. A tricky test of your intuition for 3D ray tracing.

Idea 6

Conic sections have some simple properties under reflection.

- Light rays emitted from one focus of an ellipse will all be reflected to its other focus.
- Light rays emitted from one focus on a hyperbola will all be reflected so that the resulting rays all travel radially outward from the other focus.
- Parallel light rays entering a parabola along its symmetry axis (i.e. the axis perpendicular to the directrix) will all be reflected to its focus.

In the language of idea 5, if the foci of an ellipse/hyperbola are called F_1 and F_2, then an object at F_1 produces a real/virtual image at F_2. Note that a parabola is simply an ellipse in the limit where F_1 becomes very far away, so that the rays from the object which make it near F_2 are approximately parallel.

[2] **Problem 28** (Povey). The mirascope is a toy consisting of two parabolic mirrors, pointing toward each other, so that the focus of each one is at the vertex of the other.

(a) When an object is placed at the bottom vertex, a real image appears at the top vertex. Why?

(b) How is the image oriented relative to the object?

Idea 7: Paraxial Approximation

If a light ray hits a thin lens of focal length f at a shallow angle, and at a distance $y \ll f$ above the lens's center, then it will exit the lens bent vertically by an angle $\pm y/f$, where the sign depends on whether the lens is converging or diverging. (For example, any light ray going straight through the lens's center isn't bent at all.) This is the paraxial approximation, which only holds for light rays incident at shallow angles near the center of the lens.

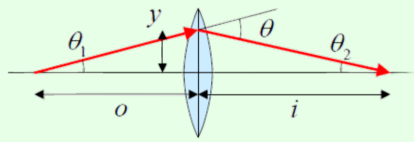
Conversely, if you don't know the focal length of a system, you can use this idea to find it. For example, the lensmaker's equation, giving the focal length of a lens of radii of curvature R_1 and R_2 and thickness d, is

$$\frac{1}{f} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2} + \frac{(n-1)\,d}{nR_1R_2}\right)$$

and can be seen by computing the bending of the light ray at each interface.

Example 5

An object is placed a distance o behind a thin converging lens with focal length f.



An image is formed a distance i in front of the lens. How are o, i, and f related?

Solution

The horizontal light ray goes straight through, so let's consider another light ray which emerges at a small angle θ_1 to the horizontal. Then we read off

$$\theta_1 \approx \frac{y}{o}, \quad \theta_2 \approx \frac{y}{i}$$

but their sum is the deflection y/f, from which we conclude

$$\frac{1}{o} + \frac{1}{i} = \frac{1}{f}.$$

This is the familiar thin lens equation.

Example 6

A candle is placed behind a converging lens. An image is formed on a screen on the other side of the lens. Now suppose that the *top* half of the lens is covered with a black cloth. Describe how the image changes.

Solution

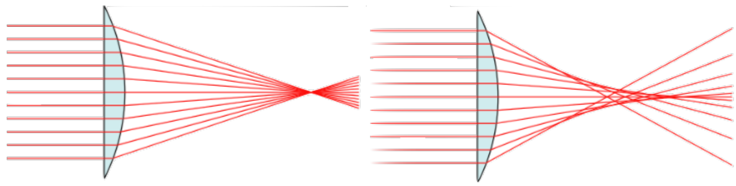
It is tempting to say that half of the candle's image disappears, but that's not right. Ray tracing shows that you can get a complete image of the candle, since there are always rays that pass through the bottom half of the lens. Instead, by blocking half the lens, the image gets half as bright.

[3] **Problem 29.** USAPhO 2024, problem A3. A series of optics exercises relevant for real cameras.

Idea 8: Fermat's Principle

For fixed starting and ending points, light always takes the path of least time. This implies that if light from point P is all focused at point P', then all the relevant paths from P to P' take the same time. This principle is completely equivalent to the laws of reflection and refraction above, but may be more useful in certain situations.

[3] **Problem 30.** Find the shape of a lens that will focus parallel incoming rays to a point, as shown at left. Concretely, suppose that the rays coming in along the $+\hat{\mathbf{x}}$ direction, the left edge of the lens is the y axis, and the right edge is described by the function $x(y)$. What kind of curve is $x(y)$?



The fact that $x(y)$ isn't exactly a circle is responsible for spherical aberration. That is, a spherical lens will fail to focus all incoming horizontal light to a point, as shown at right. However, since spheres are much easier to deal with, most of the lenses we consider below will be spherical.