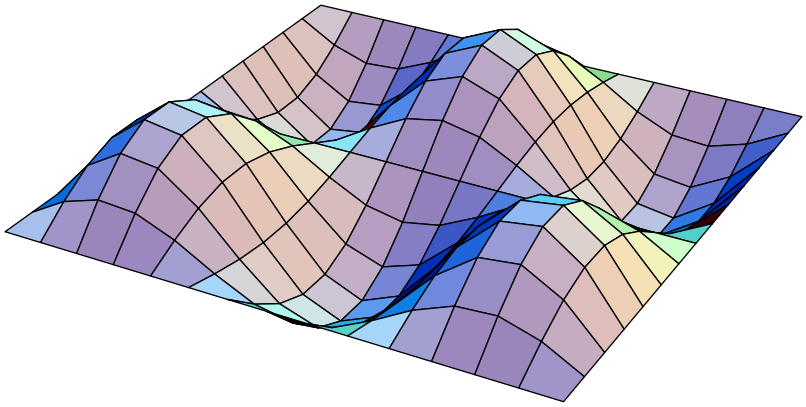
Linear Systems

John A. Scales



Samizdat
Press

Linear Systems

John A. Scales

Colorado School of Mines
jscales@mines.edu

Samizdat Press Golden · White River Junction

Published by the Samizdat Press

Center for Wave Phenomena
Department of Geophysics
Colorado School of Mines
Golden, Colorado 80401

and
New England Research
76 Olcott Drive
White River Junction, Vermont 05001

Release 4, April 1, 2001

Samizdat Press publications are available via FTP
from http://samizdat.mines.edu

Contents

List of Figures

List of Tables

Chapter 1

Simple Harmonic Motion: From Springs to Waves

You will remember from your elementary physics courses[1] that if you want to know the electric field produced by a collection of point charges, you can figure this out by adding the field produced by each charge individually. That is, if we have n charges $\{q_i\}_{i=1,n}$, then the total electric field is (neglecting constant factors)

$$\mathbf{E}(q_1 + q_2 + \cdots q_n) = \sum_{i=1}^{n} q_i \frac{(\mathbf{r} - \mathbf{r_i})}{|\mathbf{r} - \mathbf{r_i}|^3} \tag{1.0.1}$$

where $\mathbf{r}$ is the observation point and $\mathbf{r_i}$ is the position vector of the ith charge. The ith term in the sum on the right hand side,

$$\mathbf{E}(q_i) = q_i \frac{(\mathbf{r} - \mathbf{r_i})}{|\mathbf{r} - \mathbf{r_i}|^3}, \tag{1.0.2}$$

is the electric field of the ith point charge (Coulomb's law). This property, whereby we can analyze a complicated system (in this case the total electric field $\mathbf{E}(q_1 + q_2 + \cdots q_n)$) by breaking it into its constituent pieces (in this case $\mathbf{E}(q_i)$) and then adding up the results is known as **linearity**. There are three reasons why linear systems are so important.

1. We can solve them. There are systematic mathematical techniques that let us tackle linear problems.

2. Lots of physics is linear. Maxwell's equations of electricity and magnetism for instance, or the elastic and acoustic wave equations.

3. Lots of physics that is nonlinear is approximately linear at low energies or for small displacements from equilibrium.

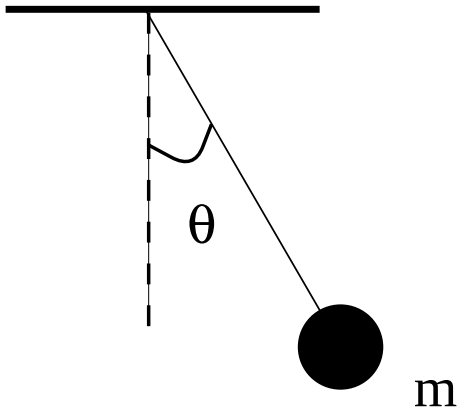


Figure 1.1: A pendulum. The restoring force is the component of the gravitational force acting perpendicular to the wire supporting the mass. This is $-mg\sin(\theta)$. Assuming the wire support is rigid, the acceleration of the mass is in the θ direction, so $ma = m\ell\ddot{\theta}$ and we have from Newton's second law: $\ddot{\theta} + \frac{g}{\ell}\sin(\theta) = 0$. This is a nonlinear equation except for small θ, in which case $\sin(\theta) \approx \theta$.

Here we introduce a nice, simplifying notation for derivatives of functions of time

$$\frac{df(t)}{dt} = \dot{f}(t) = \dot{f}$$
$$\frac{d^2f(t)}{dt^2} = \ddot{f}(t) = \ddot{f}. \tag{1.0.3}$$

For functions of some other variable we write

$$\frac{dh(x)}{dx} = h'(x) = h'$$
$$\frac{d^2h(x)}{dx^2} = h''(x) = h''. \tag{1.0.4}$$

For instance, the motion of a plane pendulum of length ℓ (Figure 1.1) is governed by

$$\ddot{\theta} + \frac{g}{\ell}\sin(\theta) = 0. \tag{1.0.5}$$

Equation 1.0.5 is not linear in θ. To solve it exactly we have to resort to elliptic functions or numerical methods. However, for small displacements ($\theta << 1$), $\sin(\theta)$ is approxi-

[1] My treatment of elementary simple harmonic motion is standard in most introductory physics textbooks. My favorite is Feynman's *Lectures on Physics* [3].

mately equal to θ. So for small displacements, the equation for the pendulum is

$$\ddot{\theta} + \frac{g}{\ell}\theta = 0. \tag{1.0.6}$$

Think of this equation as being the result of an **operator** acting on θ. Let's call the operator L, in which case Equation 1.0.6 becomes $L(\theta) = 0$. Why bother? Now we can see that

$$\begin{aligned} L(\theta_1 + \theta_2) &= \frac{d^2(\theta_1 + \theta_2)}{dt^2} + \frac{g}{\ell}(\theta_1 + \theta_2) \\ &= \left(\frac{d^2\theta_1}{dt^2} + \frac{g}{\ell}\theta_1\right) + \left(\frac{d^2\theta_2}{dt^2} + \frac{g}{\ell}\theta_2\right) \\ &= L(\theta_1) + L(\theta_2). \end{aligned} \tag{1.0.7}$$

A formal definition of a linear operator L, acting on numbers or vectors x and y is that

$$L(ax + by) = aL(x) + aL(y) \tag{1.0.8}$$

for any a and b. Thus the equation of motion for the pendulum is linear in θ when θ is small.

This is not an unusual situation. Suppose we have a point mass m constrained to move along the x-axis under the influence of some force F. Then Newton's second law is $m\ddot{x} = F(x)$. Suppose x_0 is an equilibrium point of the system, so $F(x_0) = 0$. Then expanding F in a Taylor series about x_0 we have

$$m\ddot{x} = F'(x_0)(x - x_0) + \frac{1}{2}F''(x_0)(x - x_0)^2 + \cdots \tag{1.0.9}$$

If x is close to x_0, then $x - x_0$ is small and we can drop all the terms past the first.[2] If we do drop the higher order terms, Equation 1.0.9 becomes linear in $x - x_0$. Since we said nothing at all about F itself, only that x is close to an equilibrium position, this argument applies to any F and is therefore one reason that nature appears linear so much of the time.

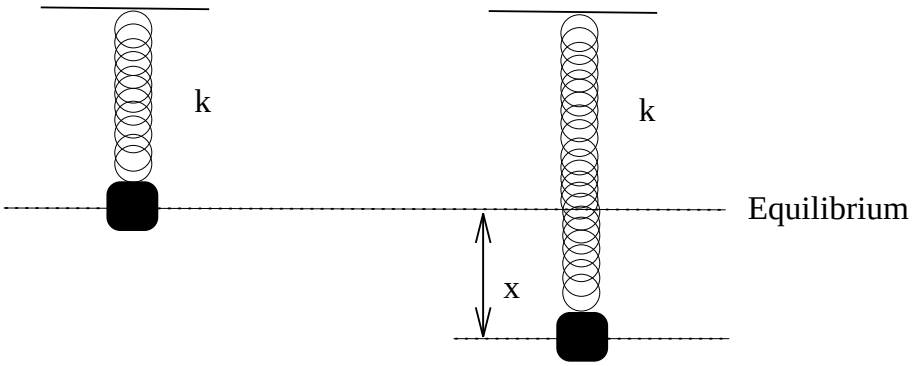


Figure 1.2: A linear spring satisfies Hooke's law: the force applied by the spring to a mass is proportional to the displacement of the mass from its equilibrium, with the proportionality being the spring constant. Since the spring wants to return to its equilibrium, the force must have the opposite sign as the displacement. Thus mass times acceleration $m\ddot{x} = -kx$.

1.1 A Spring and a Mass

1.1.1 Simple harmonic oscillation

Figure 1.2 shows a mass suspended from a spring. On the left, the mass and spring are in equilibrium. On the right the system has been displaced a distance x from equilibrium.[3] The spring is said to be linear if it satisfies Hooke's law: the force applied to the mass is proportional to its displacement from equilibrium.

> Robert Hooke (Born: July 1635, Isle of Wight, Died: March 1703 in London) was an English natural scientist. He first published his law in 1676 as a Latin anagram: *ceiiinosssttuv*. Three years later he published the solution: *ut tensio sic vis*, more or less "as the force, so the displacement". Hooke attempted to prove that the Earth moves in an ellipse round the Sun and conjectured an inverse square law for gravity. Unfortunately, no portrait of Hooke is known to exist.

The constant of proportionality is called the spring constant and is usually denoted by k. Since the force wants to restore the mass to its equilibrium, it must have the opposite

[2]This is an example of a common sort of approximation. Unless $x = x_0$ all the powers appearing in the Taylor series are potentially nonzero, but their importance decreases with increasing power; the most important being the **lowest** or **leading** order. In this case the leading order term is $x - x_0$. If $x - x_0$ is .1 for instance, then by dropping terms beyond the first we are introducing an error of order .01. Conversely, if we want to achieve an accuracy of a given order, then we can work out how many terms in the Taylor series are needed to achieve that accuracy.

[3]In the next section we will take gravity into account.

sign of the displacement. Thus mass times acceleration $m\ddot{x} = -kx$:

$$\text{Hooke's Law } m\ddot{x} = F = -kx. \tag{1.1.1}$$

Equation 1.1.1 is a simple example of a linear, constant coefficient, ordinary differential equation. You learned how to solve this equation by mathematics in your differential equations class. Here we think about the physics. Since Equation 1.1.1 is supposed to model the motion of the spring/mass system, let's see experimentally what that motion really is. I went to the hardware store and bought a spring of unknown properties and attached a 375 g mass to the end. Pull the mass down a little bit and let it go. Now count the number of complete oscillations in, say, 15 seconds. For this spring we get 19 complete oscillations. That's just under 1.3 oscillations per second, or 1.3 Hz.

Now pull the mass down twice as far and let it go. Count the oscillations. We still get 19. This is an interesting fact: **the frequency of oscillation is independent of the amplitude of the motion.** That means that the solution of Equation 1.1.1 must be a function that is periodic, with a fixed period of 1/1.3, since every 1/1.3 seconds the motion repeats. So $x(t) = \cos(\omega t)$ or $x(t) = \sin(\omega t)$ where $\omega = 2\pi$ 1.3 radians per second.[4]

Hz is Short for Hertz, after the German physicist Heinrich Hertz. Hertz was born in Hamburg in 1857 and studied at the University of Berlin. Hertz is best known as the first to broadcast and receive radio waves: a profoundly influential discovery. Less well known is the fact that Hertz also discovered the photoelectric effect, while pursuing his research into radio waves. A modest man, Hertz died at the age of 37 in 1894 from blood poisoning. His brief career as a professor was spent at the Universities of Karlsruhe and Bonn.

Since there is no absolute scale of time, we are free to choose the origin of the time axis however we wish. In this experiment it seems reasonable to choose the point $t = 0$ to correspond to the time we let the mass go. That being the case, the solution must be $x(t) = A\cos(\omega t)$ since $\sin(0) = 0$. Now $x(0) = A$ so the constant A corresponds to the amplitude of displacement. If we plug $x(t) = A\cos(\omega t)$ into Equation 1.1.1, then this is indeed a solution for any A provided that $\omega^2 = k/m$. This makes sense qualitatively, since increasing the mass should decrease the frequency: the greater mass stretches the spring more and so takes longer to complete each oscillation. On the other hand, if we increase k we are making the spring stiffer, so it oscillates faster.

But let's see if this analysis holds up experimentally. Let's increase the mass a little bit by adding a 60 g magnet to it. Count the oscillations again. This time we get only

[4]Frequencies in Hz are usually denoted by an f. ω is almost universally used for circular frequencies. Just remember that since the motion repeats itself once every $1/f$ seconds, the argument of the sine function must increase by 2π during this time.

18 oscillations in 15 seconds, or 1.2 Hz. So by increasing the mass by about 16% we've decreased the frequency of oscillation by about 8%. That means the frequency must go as one over the square root of the mass! Not convinced? We can use a Taylor series to do a perturbation analysis (using m_0 to denote the unperturbed mass):

$$\begin{aligned}
\omega(m) &\approx \omega(m_0) + \omega'(m_0)\delta m \\
&= \omega_0 - \frac{1}{2}\sqrt{\frac{k}{m_0^3}}\,\delta m \text{ defining } \omega_0 = \omega(m_0) \\
&= \omega_0 - \frac{1}{2}\omega_0\frac{\delta m}{m_0} \text{ since } \omega_0 = \sqrt{\frac{k}{m_0}} \\
&= \omega_0(1 - \frac{1}{2}\frac{\delta m}{m_0}) \\
&= 2\pi\ 1.3(1 - \frac{1}{2}\frac{60}{375}) = 2\pi\ 1.3 \times .92 \approx 2\pi\ 1.2\text{r/s}. \qquad (1.1.2)
\end{aligned}$$

So the theory holds water. NB. The argument of the sinusoid is an angle. Therefore during one complete oscillation, the angle goes through 2π radians.

No matter how we start the spring/mass system going, it always oscillates with the frequency $\sqrt{\frac{k}{m}}$. So this is its **natural** or **characteristic** frequency. Let's continue to refer to this characteristic frequency as ω_0 to emphasize the fact that it is a constant for a given spring/mass system.

Exercise on spring constants

First compute the spring constant of the spring using the data above. You'll see that with or without the added 60 g mass, the spring constant is about 25 N/m. Now consider the "spring constant" of a diatomic molecule. Look up the mass of a nitrogen or oxygen molecule, for example. You can assume that the resonant frequency is in the infrared (why?), which makes it about 10^{13} Hz. What you will find is that the spring constant is within a factor of 2 or 3 the same as the spring we used in class!

Spring/mass or spring + mass?

Scenario 1: We suspend the spring without a mass. It has some relaxed length l. Now attach the mass m. It stretches the spring a little bit Δl. This stretching is caused by the weight (gravitational force) of the mass: mg. So $mg = k\Delta l$. (Since in equilibrium the total force is zero: $mg + (-k\Delta l) = 0$.) Now pull the mass down a little bit and let it go. The total displacement from the equilibrium position of the spring alone is $\Delta l + x$. So the restoring force of the spring is $-k(\Delta l + x)$. Thus the total force on the mass (spring + gravity, but no damping for now) is $mg - k(\Delta l + x)$. But since $mg = k\Delta l$, this reduces to $-kx$. Thus, even if we explicitly include gravity and the original relaxed length of

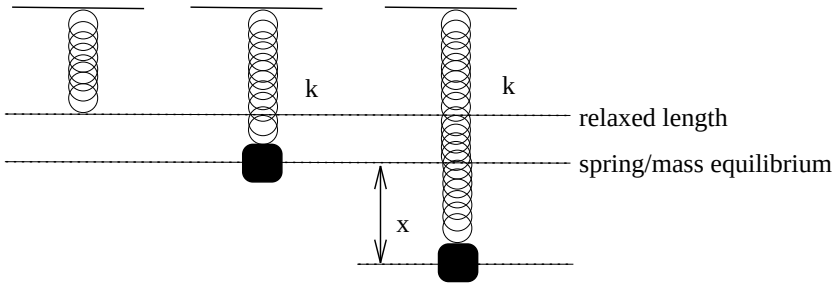


Figure 1.3: The relaxed length of the spring with no mass attached (left). Adding the mass increases the relaxed length but does not change the motion, provided the spring is linear.

the spring, we end up with $m\ddot{x} + kx = 0$ as the equations of motion. This is shown in Figure 1.3.

As an aside, the fact that $mg = k\Delta l$ gives us a connection between g and ω_0: $g/\Delta l = \omega_0^2$. For the spring we used above Δl was about 9 cm. The frequency is $\sqrt{\frac{980}{9}}/2\pi \approx 1.7$. This is a bit higher than measured, but our theory is approximate since we've neglected the finite mass of the spring itself.

Scenario 2: We suspend the spring with the mass and take the relaxed position of the combined spring/mass system as the equilibrium state and measure displacements from this position. Ignoring gravity, we still get $m\ddot{x} + kx = 0$. In effect what we've done is to use the mass m to increase the relaxed length of the spring. Since the spring is linear, a constant change in the equilibrium position has no effect on the motion.

Sine or Cosine?

At first glance it seems odd that something as arbitrary as the choice of the origin of time could influence our solution. Once the mass is oscillating we could just as easily suppose that the time at which it passes through the origin is $t = 0$. But then the "initial" displacement would be zero (since the mass is at the origin), while the initial velocity would be non-zero. That would make the displacement proportional to a sine function, say $x(t) = B\sin(\omega_0 t)$. Then $x(0) = 0$ and $\dot{x}(0) = \omega_0 B$. So B must equal whatever velocity the mass has when it zips through the origin, divided by the characteristic frequency ω_0. So it looks like we can have either

$$x(t) = x(0)\cos(\omega_0 t), \tag{1.1.3}$$

taking $t = 0$ as the time at which we let the mass go, or

$$x(t) = \frac{\dot{x}(0)}{\omega_0}\sin(\omega_0 t), \tag{1.1.4}$$

taking $t = 0$ as the time at which the spring passes through the origin with velocity $\dot{x}(0)$.

Mathematically this amounts to saying that we must specify both $x(0)$ and $\dot{x}(0)$ since the equations of motion are second order. Physically this amounts to noticing that the motion is the same no matter how we choose the origin of time. We can cover all the bases by writing the displacement as

$$x(t) = A\cos(\omega_0(t + t_0)) = A\cos(\omega_0 t + \Delta) \tag{1.1.5}$$

where t_0 is an arbitrary time shift. It's a little cleaner if we absorb the product of ω_0 and t_0 as a single, dimensionless **phase** constant Δ. Using the law of addition of cosines, this last expression can be written

$$\begin{aligned} x(t) &= A\cos(\omega_0 t + \Delta) = A\cos(\omega_0 t)\cos(\Delta) - A\sin(\omega_0 t)\sin(\Delta) \\ &= a\cos(\omega_0 t) + b\sin(\omega_0 t), \end{aligned} \tag{1.1.6}$$

where $a = A\cos(\Delta)$ and $b = -A\sin(\Delta)$. No matter how we write it, we must specify two constants: $x(0), \dot{x}(0)$; A, t_0; A, Δ; a, b.

Energy is conserved

So far we have not accounted for the damping of the spring. Theoretically, once started in motion it should oscillate at ω_0 Hz forever. Later we will take into account the actual dissipation of energy in the spring, but for now as a check we should verify that the solution we have obtained really does conserve energy. The total energy of the spring/mass system is a combination of the kinetic energy of the mass $1/2m\dot{x}^2$ and the potential energy of the spring. The force $-kx$ is minus the derivative of $1/2kx^2$, so this must be the potential energy. It is generally true that if energy is conserved, the force is minus the gradient (d/dx in the one-dimensional case) of a potential energy function.

OK, so the total energy of the spring/mass system is the sum of the kinetic and potential energies:

$$E = T + U = \frac{1}{2}m\dot{x}^2 + \frac{1}{2}kx^2. \tag{1.1.7}$$

Using the general solution $x(t) = A\cos(\omega_0 t + \Delta)$ we have

$$\begin{aligned} E &= \frac{1}{2}m\omega_0^2 A^2 \sin^2(\omega_0 t + \Delta) + \frac{1}{2}kA^2\cos^2(\omega_0 t + \Delta) \\ &= \frac{1}{2}kA^2\sin^2(\omega_0 t + \Delta) + \frac{1}{2}kA^2\cos^2(\omega_0 t + \Delta) \\ &= kA^2. \end{aligned} \tag{1.1.8}$$

Since the t disappears, we see that the energy is constant with time, and thus energy is conserved. The reasoning is somewhat circular however, since we can't really take the force to be the gradient of a potential unless energy is conserved.

1.1.2 Forced motion

When the mass shown in Figure 1.2 is bouncing up and down, we can tap it gently from below. Notice that if you tap it at just the natural frequency ω_0, your taps are synchronized with the motion, so the energy you apply goes directly into increasing the amplitude of the oscillation. The same thing happens when you're swinging in a swing. If you swing your legs back and forth with the natural frequency of the swing, you'll get a big amplification of your motion. This is called a resonance. To model it we need to add another term to the equation of motion of the spring/mass.

$$m\ddot{x}(t) + kx(t) = F(t). \tag{1.1.9}$$

Because the motion of the mass is oscillatory, the easiest sorts of forces to deal with will be oscillatory too. Later we will see that it is no loss to treat sinusoidal forces; the linearity of the equations will let us build up the result for arbitrary forces by adding a bunch of sinusoids together. But for now, let's just suppose that the applied force has the same form as the unforced motion of the mass:

$$m\ddot{x}(t) + kx(t) = F_0 \cos(\omega t). \tag{1.1.10}$$

The forcing function doesn't know anything about the natural frequency of the system and there is no reason why the forced oscillation of the mass will occur at ω_0. Of course, we will be especially interested in the solution when $\omega = \omega_0$. To keep the algebra simple, let's take the phase Δ equal to zero and look for solutions of the form

$$x(t) = A\cos(\omega t). \tag{1.1.11}$$

Plugging this into Equation 1.1.10 we have

$$(-\omega^2 + \omega_0^2)A\cos(\omega t) = \frac{F_0}{m}\cos(\omega t). \tag{1.1.12}$$

The cosines cancel and we are left with an equation for the amplitude of motion:

$$A = \frac{F_0}{m}\frac{1}{\omega_0^2 - \omega^2}. \tag{1.1.13}$$

Notice especially what happens if we force the system at the natural frequency: $\omega = \omega_0$ and the amplitude blows up. In practice the amplitude never becomes infinite. In the first place the spring would stretch to the point of breaking; but also, dissipation, which we have neglected, would come into play. Nevertheless, the idea is sound. If we apply a force to a system at its characteristic frequency we should expect a big effect.

Forced and free oscillations

The motion of the mass with no applied force is an example of a **free** oscillation. Otherwise the oscillations are **forced**. An important example of a free oscillation is the motion

of the entire earth after a great earthquake. Free oscillations are also called **transients** since for any real system in the absence of a forcing term, the damping will cause the motion to die out

1.1.3 Complex numbers and constant coefficient differential equations

We solved the equations of the simple spring/mass system $\ddot{x}(t) + \omega_0^2 x = 0$ by thinking about the physics. This is an example of a constant coefficient differential equation. The coefficients are the terms multiplying the derivatives of the independent variable. (x is the zeroth derivative of x.) The most general linear nth order constant coefficient differential equation is

$$a_n \frac{d^n x}{dt^n} + a_{n-1} \frac{d^{n-1} x}{dt^{n-1}} + \cdots a_1 \frac{dx}{dt} + a_0 x = 0. \tag{1.1.14}$$

These constant coefficient differential equations have a very special property: they reduce to polynomials for exponential x. To see this, plug an exponential e^{pt} into Equation 1.1.14. The ith derivative with respect to time is $p^i e^{pt}$, so Equation 1.1.14 becomes

$$\left(a_n p^n + a_{n-1} p^{n-1} + \cdots + a_1 p + a_0\right) e^{pt} = 0. \tag{1.1.15}$$

Canceling the overall factor of e^{pt} we see that solving Equation 1.1.14 reduces to finding the roots of an nth order polynomial. For $n = 2$, we know the formula for the roots

$$p = \frac{-a_1 \pm \sqrt{a_1^2 - 4a_0 a_2}}{2a_2}. \tag{1.1.16}$$

When n is greater than 2 life becomes more complicated; fortunately most equations in physics are first or second order.

For example, suppose

$$\ddot{x} + x = 0, \tag{1.1.17}$$

so $a_2 = 1$, $a_1 = 0$, and $a_0 = 1$. Then $p = \pm\sqrt{-1}$. $\sqrt{-1} = i$ is called the pure imaginary number. Here we see the main reason for complex numbers. Without i, not even a simple equation such as $\ddot{x}+x = 0$ has a solution. With i **every** algebraic equation can be solved.

Complex numbers are things of the form $a+ib$ where a and b are real numbers. We can think of 1 and i as being **basis** vectors in a two-dimensional Cartesian space. Addition of complex numbers is component-wise: $(a+ib)+(p+iq) = (a+p)+i(b+q)$. Multiplication is as you would expect, but with $ii = -1$. So $(a + ib)(p + iq) = (ab - bq) + i(bp + aq)$. In the complex plane, multiplication by i acts as a rotation by $\pi/4$. $i1 = i$, $ii = -1$, $i(-1) = -i$ and $i(-i) = 1$. (Figure 1.4.)

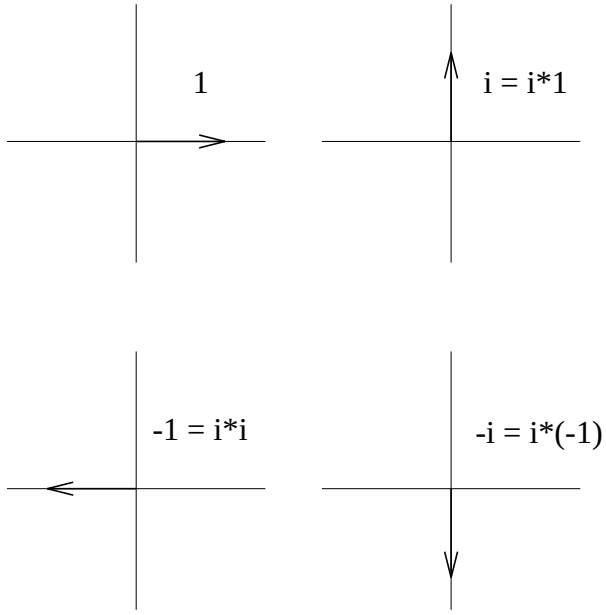


Figure 1.4: Multiplication by the pure imaginary i acts as a $\pi/4$ rotation in the complex plane.

And just as there is an equivalence between Cartesian and polar coordinates, so we can give a "polar" representation of every complex number. To understand this connection, consider the Maclauren series for e^x

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \frac{x^5}{120} + \frac{x^6}{720} + \cdots \tag{1.1.18}$$

Now replace x with ix. We get

$$\begin{aligned} e^{ix} &= 1 + ix + \frac{(ix)^2}{2} + \frac{(ix)^3}{6} + \frac{(ix)^4}{24} + \frac{(ix)^5}{120} + \frac{(ix)^6}{720} + \cdots \\ &= \left[1 - \frac{x^2}{2} + \frac{x^4}{24} + \cdots\right] + i\left[x - \frac{x^3}{6} + \frac{x^5}{120} + \cdots\right] \end{aligned} \tag{1.1.19}$$

In the limit of small x, this reduces to $e^{ix} \approx 1 + ix$, which is the small angle limit of $\cos(x) + i\sin(x)$. To see if this extends to large x, compute the Maclauren series for the sine and cosine:

$$\cos(x) = 1 - \frac{x^2}{2} + \frac{x^4}{24} + \cdots \tag{1.1.20}$$

$$\sin(x) = x - \frac{x^3}{6} + \frac{x^5}{120} + \cdots. \tag{1.1.21}$$

Thus we have proved what some call the most remarkable formula in mathematics:

$$\text{Euler's Formula } e^{ix} = \cos(x) + i\sin(x). \tag{1.1.22}$$

The geometry of the Cartesian and polar representations is summarized in Figure 1.5.

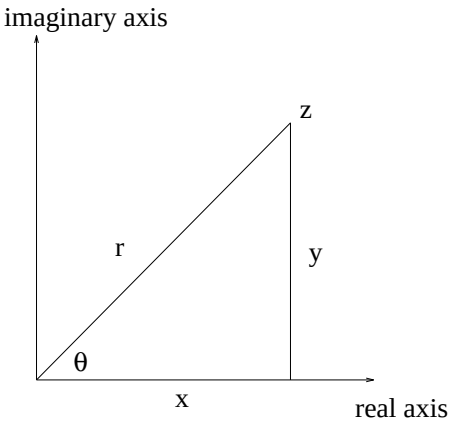


Figure 1.5: Every complex number z can be represented as a point in the complex plain. There are various ways or "coordinates" by which we can parameterize these points. The Cartesian parameterization is in terms of the real and imaginary parts. The polar parameterization is in terms of the length (or modulus) and angle (or phase). The connection between these two forms is the remarkable Euler formula: $re^{i\theta} = r\cos\ \theta + ir\sin\ \theta$. From Pythagoras' theorem we can see $r^2 = x^2 + y^2$ and the angle θ is just the arctangent of y/x.

1.1.4 Forced motion with damping

Now let's go ahead and do the fully general simple harmonic oscillator, including the effects of damping (i.e., dissipations). The causes of damping are extremely subtle. We will not go deeply into these effects here except to say that ultimately the physical processes which cause damping give rise to motion at the atomic and molecular level. If you calculate the characteristic frequencies of atoms, as we have done for the spring/mass system, you see that these frequencies are the same as electromagnetic radiation in the infrared. Heat in other words! But this heat is just the manifestation of a kind of oscillatory motion. When the energy of these oscillations is not too great, the atoms/molecules can be treated as simple harmonic oscillators.

However, it has long been observed empirically that to a reasonable approximation, the effect of damping or friction is to oppose the motion of the mass with a force that is proportional to the velocity. If the mass is at rest, there is no friction. As the velocity increases, the frictional force increases and this force opposes the motion. Try extending a damping piston of the sort used on doors. The faster you extend the piston, the greater the resistance. So as a first approximation, we can model the friction of our spring/mass system as

$$\ddot{x} + \gamma\dot{x} + \omega_0^2 x = \frac{F}{m} \tag{1.1.23}$$

where γ is a constant reflecting the strength of the damping. We can proceed just as before with the undamped, forced oscillations but the algebra is greatly simplified if we

use complex numbers. We have used $\cos(\omega t)$ to represent a oscillatory driving force. From Euler's formula we know that the cosine is the real part of $e^{i\omega t}$. So we are going to use a trick. We are going to use $e^{i\omega t}$ throughout the calculation and then take the real part when we're done. The reason for doing this is simply that exponentials are easier to work with than cosines. The fact that we get the right answer in the end depends critically on the equations being linear. This trick will not work for nonlinear equations. To prove this to yourself, assume for the moment that $x = x_r + ix_i$ and $F = F_r + iF_i$. Plug these into Equation 1.1.23 and show that

$$\ddot{x_r} + \gamma\dot{x_r} + \omega_0^2 x_r + i\left[\ddot{x_i} + \gamma\dot{x_i} + \omega_0^2 x_i\right] = \frac{F_r}{m} + i\frac{F_i}{m}. \tag{1.1.24}$$

In order for two complex numbers to be equal, their real and imaginary parts must be equal separately. Therefore the real part of the complex "displacement" must satisfy Equation 1.1.23, which is what was claimed.

So let's write the complex force as $F = \hat{F}e^{i\omega t}$ and the complex displacement as $x = \hat{x}e^{i\omega t}$. Plugging this into Equation 1.1.23 implies that

$$(-\omega^2 + i\gamma\omega + \omega_0^2)\hat{x}e^{i\omega t} = \frac{\hat{F}}{m}e^{i\omega t}. \tag{1.1.25}$$

Canceling the exponential gives

$$\hat{x} = \frac{\hat{F}/m}{-\omega^2 + i\gamma\omega + \omega_0^2}. \tag{1.1.26}$$

This equation requires a little analysis, but straight off we can see that the presence of the damping term γ has fixed the infinity we saw when we forced the oscillator at its resonant frequency; even when $\omega = \omega_0$ the amplitude is finite provided γ is not zero. For the moment let's just look at the denominator of the displacement.

$$-\omega^2 + i\gamma\omega + \omega_0^2 = \sqrt{(\omega_0^2 - \omega^2)^2 + \gamma^2\omega^2}\, e^{i\tan^{-1}\frac{\gamma\omega}{\omega_0^2 - \omega^2}}. \tag{1.1.27}$$

The function $\rho^2 = 1/((\omega_0^2 - \omega^2)^2 + \gamma^2\omega^2)$ has a characteristic shape seen in all resonance phenomena. It's peaked about the characteristic frequency ω_0 and has a full width of γ at half its maximum height as illustrated in Figure 1.6.

Finally, since γ has the dimension of inverse time, a useful dimensionless measure of damping can be obtained by taking the ratio of the characteristic frequency ω_0 and γ. This ratio is called the Q (for quality factor) of the peak. Typical values of Q at ultrasonic frequencies can range from 10-100 for sedimentary rocks, to a few thousand for aluminum, to nearly a million for monocrystalline quartz.

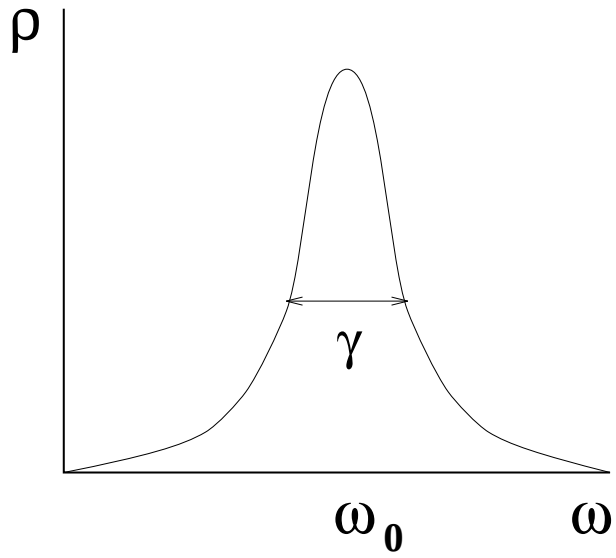


Figure 1.6: Square of the amplitude factor $\rho^2 = 1/((\omega_0^2 - \omega^2)^2 + \gamma^2\omega^2)$ for forced, damped motion near a resonance ω_0. The full width at half max of this curve is the damping factor γ, **provided γ is small!**

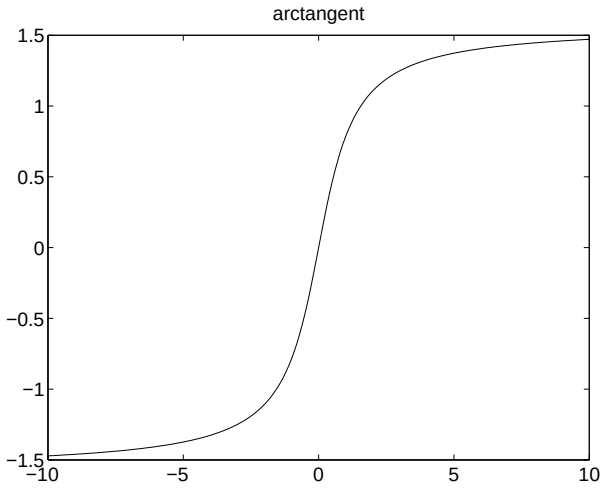


Figure 1.7: The arctangent function asymptotes at $\pm\pi/2$, so we should expect to see a phase shift of π when going through a resonance.

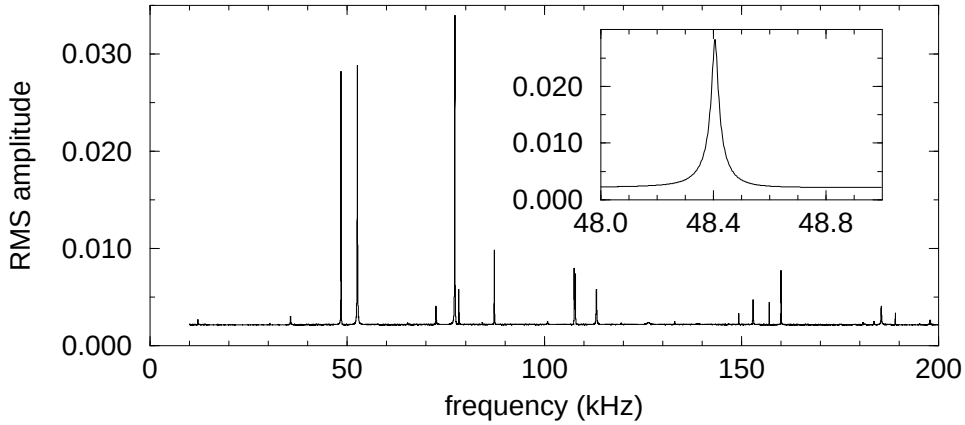


Figure 1.8: Here is a resonance spectrum for a piece of aluminum about the size shown in Figure 1.9. A swept sine wave is fed into the sample via a tiny transducer and recorded on another transducer. At a resonant frequency, there is a big jump in the amplitude. The DC level is shifted upward to make it easy to see the peaks. The inset shows a blow-up of one peak.

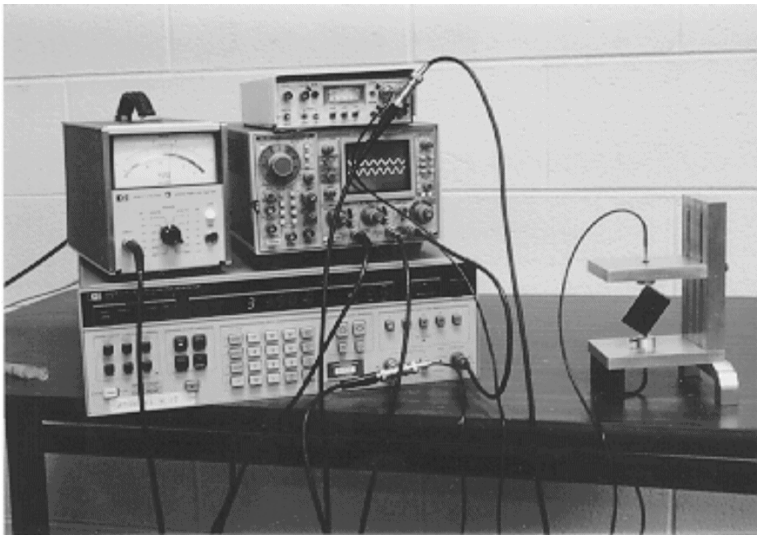


Figure 1.9: Resonance ultrasonic spectroscopy setup. The rock is held fast by two tiny transducers ("pinducers") which are used to transmit a signal and record the response. The two traces shown on the oscilloscope correspond to the transmitted and received signal. As the frequency is varied we see the characteristic resonance (cf Figure 1.8). To a first approximation, the frequency associated with the peak is one of the characteristic (eigen) frequencies of the sample.

1.1.5 Damped transient motion

Suppose we suspend our mass in a viscous fluid. Pull it down and let it go. The fluid will damp out the motion, more or less depending on whether it has the viscosity of water or honey. Mathematically this case is easy, all we have to do is set the right hand side of Equation 1.1.25 to zero. This leaves a simple quadratic for ω

$$\omega^2 - i\gamma\omega - \omega_0^2 = 0 \tag{1.1.28}$$

which has the two solutions

$$\begin{aligned} \omega &= \frac{i\gamma \pm \sqrt{4\omega_0^2 - \gamma^2}}{2} \\ &= i\frac{\gamma}{2} \pm \omega_0\sqrt{1 - \left(\frac{\gamma}{2\omega_0}\right)^2}. \end{aligned} \tag{1.1.29}$$

These give the following solutions for the motion (using $x(0) = x_0$)

$$x(t) = x_0 e^{-\frac{\gamma}{2}t} e^{\pm it\omega_0\sqrt{1-\left(\frac{\gamma}{2\omega_0}\right)^2}}. \tag{1.1.30}$$

This looks like the equation of a damped sinusoid. But the second term may or may not be a sinusoid, depending on whether the square root is positive. So we have to treat two special cases. First if $\frac{\gamma}{2\omega_0} < 1$, corresponding to small damping, then the argument of the square root is positive and indeed we have a damped sinusoid. On the other hand if $\frac{\gamma}{2\omega_0} > 1$, then we can rewrite the solution as

$$x(t) = x_0 e^{-\frac{\gamma}{2}t} e^{\pm t\omega_0\sqrt{(\frac{\gamma}{2\omega_0})^2-1}} \tag{1.1.31}$$

where, once again, we have arranged things so that the argument of the square root is positive. But now only the minus sign in the exponent makes sense, since otherwise the amplitude of the motion would increase with time. So, we have

$$x(t) = x_0 e^{-\frac{\gamma}{2}t} e^{-\omega_0 t\sqrt{(\frac{\gamma}{2\omega_0})^2-1}}. \tag{1.1.32}$$

In this case the motion is said to be "over-damped" since there is no oscillation. In a highly viscous fluid (high relative to ω_0) there is no oscillation at all, the motion is quickly damped to zero. The borderline case $\gamma = 2\omega_0$ is called critical damping, in which case $x(t) = x_0 e^{-\frac{\gamma}{2}t}$.

1.1.6 Another velocity-dependent force: the Zeeman effect

As a classical model for the radiation of light from excited atoms we can consider the electrons executing simple harmonic oscillations about their equilibrium positions under

the influence of a restoring force $F = -k\mathbf{r}$. Thus our picture is of an oscillating electric dipole. Remember, the restoring force $-k\mathbf{r}$ is just a linear approximation to the Coulomb force and therefore k, the "spring constant", is the first derivative of the Coulomb force evaluated at the equilibrium radius of the electron. So the vector differential equation governing this simple harmonic motion is:

$$m\ddot{\mathbf{r}} + k\mathbf{r} = 0. \tag{1.1.33}$$

Notice that there is no coupling between the different components of $\mathbf{r}$. In other words this one vector equation is equivalent to three completely separate scalar equations (using $\omega_0^2 = k/m$)

$$\begin{aligned} \ddot{x} + \omega_0^2 x &= 0 \\ \ddot{y} + \omega_0^2 y &= 0 \\ \ddot{z} + \omega_0^2 z &= 0 \end{aligned}$$

each of which has the same solution, a sinusoidal oscillation at frequency ω_0. Think of it this way: there are three equations and three frequencies of oscillation, but all the frequencies happen to be equal. This is called *degeneracy*. The equations are uncoupled in the sense that each unknown (x, y, z) occurs in only one equation; thus we can solve for x ignoring y and z.

Now let's suppose we apply a force that is not spherically symmetric. For instance, suppose we put the gas of atoms in a magnetic field pointed along, say, the z-axis. This results in another force on the electrons of the form $q\dot{\mathbf{r}} \times B\hat{\mathbf{z}}$ (from Lorentz's force law). Adding this force to the harmonic $(-k\mathbf{r})$ force gives[5]

$$\begin{aligned} \ddot{x} + \omega_0^2 x - \frac{qB}{m}\dot{y} &= 0 \\ \ddot{y} + \omega_0^2 y + \frac{qB}{m}\dot{x} &= 0 \\ \ddot{z} + \omega_0^2 z &= 0. \end{aligned}$$

The z equation hasn't changed so it's still true that $z(t) = Real(z_0 e^{i\omega_0 t})$. But now the x and y equations are coupled–we must solve for x and y simultaneously. Let us assume a solution of the form:

$$\begin{aligned} x(t) &= Real(x_0 e^{i\omega t}) \\ y(t) &= Real(y_0 e^{i\omega t}) \end{aligned}$$

[5]Remember the right-hand screw rule, so $\hat{\mathbf{y}} \times \hat{\mathbf{z}} = \hat{\mathbf{x}}$, $\hat{\mathbf{x}} \times \hat{\mathbf{z}} = -\hat{\mathbf{y}}$, and $\hat{\mathbf{z}} \times \hat{\mathbf{z}} = 0$

where x_0 and y_0 are constants to be determined. Plugging these into the equations for x and y gives the two amplitude equations

$$\begin{aligned}(\omega_0^2-\omega^2)x_0 &= \frac{qB}{m}i\omega y_0 \\ (\omega_0^2-\omega^2)y_0 &= -\frac{qB}{m}i\omega x_0.\end{aligned} \tag{1.1.34}$$

We can use the first equation to compute x_0 in terms of y_0 and then plug this into the second equation to get

$$(\omega_0^2-\omega^2)y_0 = -\frac{\left(\frac{qB}{m}i\omega\right)^2}{\omega_0^2-\omega^2}y_0.$$

Now we eliminate the y_0 altogether and get

$$\left(\omega_0^2-\omega^2\right)^2 = \left(\frac{qB}{m}\omega\right)^2 \tag{1.1.35}$$

Taking the square root we have

$$\omega_0^2-\omega^2 = \pm\frac{qB}{m}\omega. \tag{1.1.36}$$

This is a quadratic equation for the unknown frequency of motion ω. So we have

$$\omega = \frac{\pm\frac{qB}{m} \pm \sqrt{\left(\frac{qB}{m}\right)^2+4\omega_0^2}}{2}. \tag{1.1.37}$$

This is not too bad, but we can make a great simplification by assuming that the magnetic field is weak. Specifically, let's assume that $\frac{qB}{m} << \omega_0$ so the square root reduces to $2\omega_0$. We take the positive square root since otherwise we would have negative frequencies. Thus

$$\omega = \omega_0 \pm \frac{qB}{2m}. \tag{1.1.38}$$

As a result of the applied magnetic field, there are now three characteristic frequencies of oscillation:

$$\begin{aligned}\omega_1 &= \omega_0 + \frac{qB}{2m} \\ \omega_2 &= \omega_0 \\ \omega_3 &= \omega_0 - \frac{qB}{2m}.\end{aligned}$$

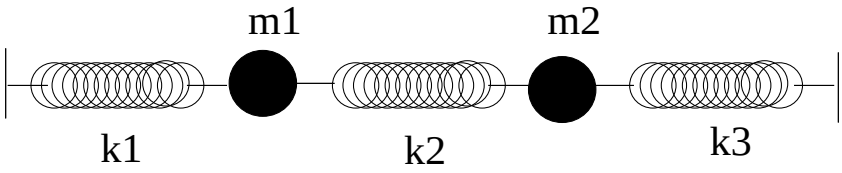


Figure 1.10: Two masses coupled by a spring and attached to walls.

This splitting of the degenerate frequency by an external magnetic field is called the Zeeman effect, after its discoverer Pieter Zeeman was born in May 1865, at Zonnemaire, a small village in the isle of Schouwen, Zeeland, The Netherlands. Zeeman was a student of the great physicists Onnes and Lorentz in Leyden. He was awarded the Nobel Prize in Physics in 1902. Zeeman succeeded Van der Waals (another Nobel prize winner) as professor and director of the Physics Laboratory in Amsterdam in 1908. In 1923 a new laboratory was built for Zeeman that included a quarter-million kilogram block of concrete for vibration free measurements.

We could continue the analysis by plugging these frequencies back into the amplitude equations 1.1.35. As an exercise, do this and show that the motion of the electron (and hence the electric field) is circularly polarized in the direction perpendicular to the magnetic field.

1.2 Two Coupled Masses

With only one mass and one spring, the range of motion is somewhat limited. There is only one characteristic frequency $\omega_0^2 = \frac{k}{m}$ so in the absence of damping, the transient (unforced) motions are all of the form $\cos(\omega_0 t + \Delta)$.

Now let us consider a slightly more general kind of oscillatory motion. Figure 1.10 shows two masses (m_1 and m_2) connected to fixed walls with springs k_1 and k_3 and connected to one another by a spring k_2. To derive the equations of motion, let's focus attention on one mass at a time. We know that for any given mass, say m_i (whose displacement from equilibrium we label x_i) it must be that

$$m_i \ddot{x}_i = F_i \tag{1.2.1}$$

where F_i is the total force acting on the ith mass. No matter how many springs and masses we have in the system, the force applied to a given mass must be transmitted by the two springs it is connected to. And the force each of these springs transmits is governed by the extent to which the spring is compressed or extended.

Referring to Figure 1.10, spring 1 can only be compressed or extended if mass 1 is displaced from its equilibrium. Therefore the force applied to m_1 from k_1 must be $-k_1 x_1$, just as before. Now, spring 2 is compressed or stretched depending on whether $x_1 - x_2$ is positive or not. For instance, suppose both masses are displaced to the right (positive x_i) with mass 1 being displaced more than mass 2. Then spring 2 is compressed relative to its equilibrium length and the force on mass 1 will in the negative x direction so as to restore the mass to its equilibrium position. Similarly, suppose both masses are displaced to the right, but now with mass 2 displaced more than mass 1, corresponding to spring 2 being stretched. This should result in a force on mass 1 in the positive x direction since the mass is being pulled away from its equilibrium position. So the proper expression of Hooke's law in any case is

$$m_1 \ddot{x}_1 = -k_1 x_1 - k_2(x_1 - x_2). \tag{1.2.2}$$

And similarly for mass 2

$$m_2 \ddot{x}_2 = -k_3 x_2 - k_2(x_2 - x_1). \tag{1.2.3}$$

These are the general equations of motion for a two mass/three spring system. Let us simplify the calculations by assuming that both masses and all three springs are the same. Then we have

$$\begin{aligned} \ddot{x}_1 &= -\frac{k}{m}x_1 - \frac{k}{m}(x_1 - x_2) \\ &= -\omega_0^2 x_1 - \omega_0^2(x_1 - x_2) \\ &= -2\omega_0^2 x_1 + \omega_0^2 x_2. \end{aligned} \tag{1.2.4}$$

and

$$\begin{aligned} \ddot{x}_2 &= -\frac{k}{m}x_2 - \frac{k}{m}(x_2 - x_1) \\ &= -\omega_0^2 x_2 - \omega_0^2(x_2 - x_1) \\ &= -2\omega_0^2 x_2 + \omega_0^2 x_1. \end{aligned} \tag{1.2.5}$$

Assuming trial solutions of the form

$$x_1 = Ae^{i\omega t} \tag{1.2.6}$$

$$x_2 = Be^{i\omega t} \tag{1.2.7}$$

we see that

$$(-\omega^2 + 2\omega_0^2)A = \omega_0^2 B \tag{1.2.8}$$

$$(-\omega^2 + 2\omega_0^2)B = \omega_0^2 A. \tag{1.2.9}$$

Substituting one into the other we get

$$A = \frac{\omega_0^2}{2\omega_0^2 - \omega^2}B, \tag{1.2.10}$$

and therefore

$$(2\omega_0^2 - \omega^2)B = \frac{\omega_0^4}{2\omega_0^2 - \omega^2}B. \tag{1.2.11}$$

This gives an equation for ω^2

$$(2\omega_0^2 - \omega^2)^2 = \omega_0^4. \tag{1.2.12}$$

There are two solutions of this equation, corresponding to $\pm\omega_0^2$ when we take the square root. If we choose the plus sign, then

$$2\omega_0^2 - \omega^2 = \omega_0^2 \Rightarrow \omega^2 = \omega_0^2. \tag{1.2.13}$$

On the other hand, if we choose the minus sign, then

$$2\omega_0^2 - \omega^2 = -\omega_0^2 \Rightarrow \omega^2 = 3\omega_0^2. \tag{1.2.14}$$

We have discovered an important fact: spring systems with two masses have two characteristic frequencies. We will refer to the frequency $\omega^2 = 3\omega_0^2$ as "fast" and $\omega^2 = \omega_0^2$ as "slow". Of course these are relative terms. Now that we have the frequencies we can investigate the amplitude. First, since

$$A = \frac{\omega_0^2}{2\omega_0^2 - \omega^2}B, \tag{1.2.15}$$

we have for the slow mode ($\omega = \omega_0$)

$$A = B, \tag{1.2.16}$$

which corresponds to the two masses moving in phase with the same amplitude. On the other hand, for the fast mode

$$A = -B. \tag{1.2.17}$$

For this mode, the amplitudes of the two mass' oscillation are the same, but they are out of phase. These two motions are easy to picture. The slow mode corresponds to both masses moving together, back and forth, as in Figure 1.11 (bottom). The fast mode corresponds to the two masses oscillating out of phase as in Figure 1.11 (top).

1.2.1 A Matrix Appears

There is a nice way to simplify the notation of the previous section and to introduce a powerful mathematical at the same time. Let's put the two displacements together into a vector. Define a vector $\mathbf{u}$ with two components, the displacements of the first and second mass:

$$\mathbf{u} = \begin{bmatrix} Ae^{i\omega t} \\ Be^{i\omega t} \end{bmatrix} = e^{i\omega t}\begin{bmatrix} A \\ B \end{bmatrix}. \tag{1.2.18}$$

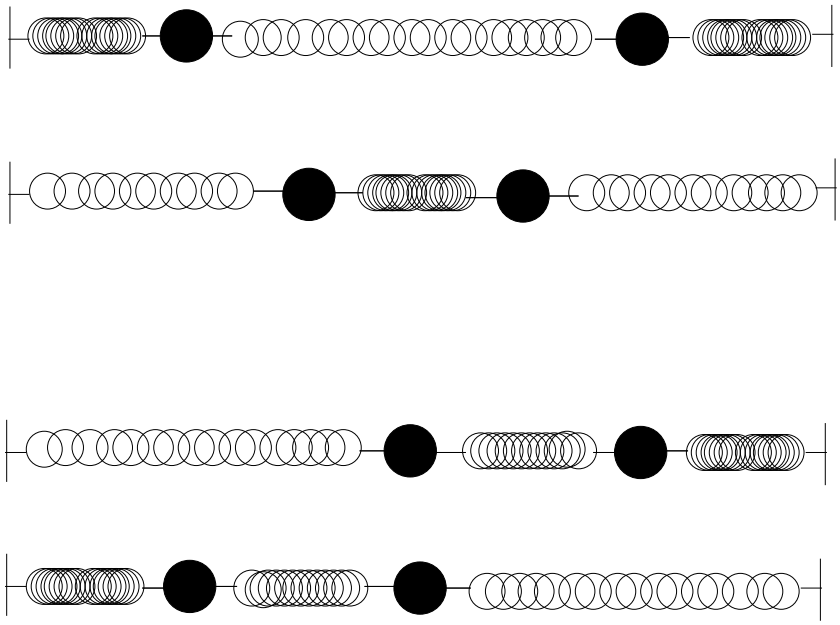


Figure 1.11: With two coupled masses there are two characteristic frequencies, one "slow" (bottom) and one "fast" (top).

We've already seen that we can multiply any solution by a constant and still get a solution, so we might as well take A and B to be equal to 1. So for the slow mode we have

$$\mathbf{u} = e^{i\omega_0 t} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \tag{1.2.19}$$

while for the fast mode we have

$$\mathbf{u} = e^{i\sqrt{3}\omega_0 t} \begin{bmatrix} 1 \\ -1 \end{bmatrix}. \tag{1.2.20}$$

Notice that the amplitude part of the two modes

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix} \text{ and } \begin{bmatrix} 1 \\ -1 \end{bmatrix} \tag{1.2.21}$$

are **orthogonal**. That means that the dot product of the two vectors is zero: $1 \times 1 + 1 \times (-1) = 0$.[6] As we will see in our discussion of linear algebra, this means that the two vectors point at right angles to one another. This orthogonality is an absolutely fundamental property of the natural modes of vibration of linear mechanical systems.

[6]
$$\begin{bmatrix} 1 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ -1 \end{bmatrix} \equiv [1,1] \begin{bmatrix} 1 \\ -1 \end{bmatrix} = 1 \cdot 1 - 1 \cdot 1 = 0.$$

1.2.2 Matrices for two degrees of freedom

The equations of motion are (see Figure 1.10):

$$m_1\ddot{x}_1 + k_1x_1 + k_2(x_1 - x_2) = 0 \tag{1.2.22}$$
$$m_2\ddot{x}_2 + k_3x_2 + k_2(x_2 - x_1) = 0. \tag{1.2.23}$$

We can write these in matrix form as follows.

$$\begin{bmatrix} m_1 & 0 \\ 0 & m_2 \end{bmatrix} \begin{bmatrix} \ddot{x}_1 \\ \ddot{x}_2 \end{bmatrix} + \begin{bmatrix} k_1 + k_2 & -k_2 \\ -k_2 & k_2 + k_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{1.2.24}$$

Or, defining a mass matrix

$$M = \begin{bmatrix} m_1 & 0 \\ 0 & m_2 \end{bmatrix} \tag{1.2.25}$$

and a "stiffness" matrix

$$K = \begin{bmatrix} k_1 + k_2 & -k_2 \\ -k_2 & k_2 + k_3 \end{bmatrix} \tag{1.2.26}$$

we can write the matrix equation as

$$M\ddot{\mathbf{u}} + K\mathbf{u} = \mathbf{0} \tag{1.2.27}$$

where

$$\mathbf{u} \equiv \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \tag{1.2.28}$$

This is much cleaner than writing out all the components and has the additional advantage that we can add more masses/springs without changing the equations, we just have to incorporate the additional terms into the definition of M and K.

Notice that the mass matrix is always invertible since it's diagonal and all the masses are presumably nonzero. Therefore

$$M^{-1} = \begin{bmatrix} {m_1}^{-1} & 0 \\ 0 & {m_2}^{-1} \end{bmatrix}. \tag{1.2.29}$$

So we can also write the equations of motion as

$$\ddot{\mathbf{u}} + M^{-1}K\mathbf{u} = \mathbf{0}. \tag{1.2.30}$$

And it is easy to see that

$$M^{-1}K = \begin{bmatrix} \frac{k_1+k_2}{m_1} & \frac{-k_2}{m_1} \\ \frac{-k_2}{m_2} & \frac{k_2+k_3}{m_2} \end{bmatrix}.$$

As another example, let's suppose that all the masses are the same and that $k_1 = k_3 = k$. Letting $\omega_0 = \sqrt{k/m}$ as usual and defining $\Omega = \sqrt{k_2/m}$, we have the following beautiful form for the matrix $M^{-1}K$:

$$M^{-1}K = \Omega^2 \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} + \omega_0^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{1.2.31}$$

In the limit that Ω goes to zero the coupling between the masses becomes progressively weaker. If $\Omega = 0$, then the equations of motion reduce to those for two uncoupled oscillators with the same characteristic frequency ω_0.

1.2.3 The energy method

In this example of two coupled masses, it's not entirely trivial to keep track of how the two masses interact. Unfortunately, we're forced into this by the Newtonian strategy of specifying forces explicitly. Fortunately this is not the only way to skin the cat. For systems in which energy conserved (no dissipation, also known as **conservative** systems), the force is the gradient of a potential energy function.[7]

Since energy is a scalar quantity it is almost always a lot easier to deal with than the force itself. In our 1-D system of masses and springs, that might not be apparent, but even so using energy simplifies life significantly. Think about it: the potential energy of the system must be the sum of the potential energies of the individual springs. And the potential energy of a spring is the spring constant times the square of amount the spring is compresses or extended. So the potential energy of the system is just $\frac{1}{2}\left[k_1 x_1^2 + k_2(x_2 - x_1)^2 + k_3 x_2^2\right]$. Unlike when dealing with the forces, it doesn't matter whether we write the second term as $x_2 - x_1$ or $x_1 - x_2$ since it gets squared.

The energy approach is easily extended to an arbitrary number of springs and masses. It's up to us to define just what the system will be. For instance do we connect the end springs to the wall, or do we connect the end masses? It doesn't matter much except in the labels we use and the limits of the summation. For now we will assume that we have n springs, the end springs being connected to rigid walls, and $n-1$ masses. So, $n-1$ masses $\{m_i\}_{i=1,n-1}$ and n spring constants $\{k_i\}_{i=1,n}$. Then the total energy is

$$E = \text{K.E.} + \text{P.E.} = \frac{1}{2}\sum_{i=1}^{n-1} m_i \dot{x}_i^2 + \frac{1}{2}\sum_{i=1}^{n} k_i (x_i - x_{i-1})^2. \tag{1.2.32}$$

[7] The work done by a force in displacing a system from a to b is $\int_a^b F\,dx$. If $F = -\frac{dU}{dx}$, then $\int_a^b F\,dx = -\int dU = -[U(b) - U(a)]$. In other words the work depends only on the endpoints, not the path taken. In particular, if the starting and ending point is the same, the work done is zero. This is true in 3 dimensions too where it is easier to visualize complicated paths.

To derive the equations of motion, all we have to do is set $m_j\ddot{x}_j = -\frac{\partial U}{\partial x_j}$. Taking the derivative is slightly tricky. Since j is arbitrary (we want to be able to study any mass), there will be two nonzero terms in the derivative of U, corresponding to the two situations in which one of the terms in the sum is equal to x_j. This will happen when

- $i = j$, in which case the derivative is $k_j(x_j - x_{j-1})$.
- $i - 1 = j$, in which case $i = j + 1$ and the derivative is $-k_{j+1}(x_{j+1} - x_j)$.

Putting these two together we get

$$m_j\ddot{x}_j = -\frac{\partial U}{\partial x_j} = k_{j+1}(x_{j+1} - x_j) - k_j(x_j - x_{j-1}). \tag{1.2.33}$$

Once you get the hang of it, you'll see that in most cases the energy approach is a lot easier than dealing directly with the forces. After all, force is a vector, while energy is always a scalar. For now, let's simplify Equation 1.2.33 by taking all the masses to be the same m and all the spring constants to be the same k. Then, using $\omega_0^2 = k/m$ again, we have

$$\frac{1}{\omega_0^2}\ddot{x}_j = x_{j+1} - 2x_j + x_{j-1}. \tag{1.2.34}$$

1.2.4 Matrix form of the coupled spring/mass system

We can greatly simplify the notation of the coupled system using matrices. Let's consider the n mass case in Equation 1.2.34. We would like to be able to write this as

$$\frac{1}{\omega_0^2}\ddot{\mathbf{u}} \equiv \begin{bmatrix} \ddot{x}_1 \\ \ddot{x}_2 \\ \ddot{x}_3 \\ . \\ . \\ . \\ \ddot{x}_{n-1} \end{bmatrix} = \text{ some matrix dotted into } \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ . \\ . \\ . \\ x_{n-1} \end{bmatrix} \equiv \mathbf{u}. \tag{1.2.35}$$

The symbol $\equiv$ means the two things on either side are equal by definition.

Looking at Equation 1.2.34 we can see that this matrix must couple each mass to its nearest neighbors, with the middle mass getting a weight of -2 and the neighboring masses getting weights of 1. Thus the matrix must be

$$\begin{bmatrix} -2 & 1 & 0 & 0 & \dots \\ 1 & -2 & 1 & 0 & \dots \\ 0 & 1 & -2 & 1 & \dots \\ \vdots & & & \ddots & \\ 0 & \dots & 0 & 1 & -2 \end{bmatrix}. \tag{1.2.36}$$

So we have

$$\frac{1}{\omega_0^2}\ddot{\mathbf{u}} = \begin{bmatrix} \ddot{x}_1 \\ \ddot{x}_2 \\ \ddot{x}_3 \\ . \\ . \\ . \\ \ddot{x}_{n-1} \end{bmatrix} = \begin{bmatrix} -2 & 1 & 0 & 0 & \dots \\ 1 & -2 & 1 & 0 & \dots \\ 0 & 1 & -2 & 1 & \dots \\ \vdots & & & \ddots & \\ 0 & \dots & 0 & 1 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ . \\ . \\ . \\ x_{n-1} \end{bmatrix} = \mathbf{u}. \tag{1.2.37}$$

If we denote the matrix by K, then we collapse these n coupled second order differential equations to the following beautiful vector differential equation.

$$\frac{1}{\omega_0^2}\ddot{\mathbf{u}} = K \cdot \mathbf{u}. \tag{1.2.38}$$

We don't yet have the mathematical tools to analyze this equation properly, that is why we will spend a lot of time studying linear algebra. However we can proceed. Surprisingly enough if we add even more springs and masses to our system, we will get an equation we can solve analytically, but we need to an an infinite number of them! Let's see how we can do this.

First, let's be careful how we interpret the dependent and independent variables. If I write the vector of displacements from equilibrium as $\mathbf{u}$, then its components are $(\mathbf{u})_i \equiv x_i$. Let's forget about x and think only of displacements $\mathbf{u}$ or $(\mathbf{u})_i$. The reason is we want to be able to use x as a variable to denote the position along the spring/mass lattice at which we are measuring the displacement. Right now, with only a finite number of masses, we are using the index i for this purpose. But we want to let i go to infinity and have a continuous variable for this; this is what we will henceforth use x for. But before we do that, let's look at how we can approximate the derivative of a function. Suppose $f(x)$ is a differentiable function. Then, provided h is small

$$f'(x) \approx \frac{f(x+\frac{h}{2}) - f(x-\frac{h}{2})}{h}. \tag{1.2.39}$$

We can do this again for each of the two terms on the right hand side and achieve an approximation for the second derivative:

$$\begin{aligned} f''(x) &\approx \frac{f(x+h)-f(x)}{h^2} - \frac{f(x)-f(x-h)}{h^2} \\ &= \frac{1}{h^2}\left(f(x+h) - 2f(x) + f(x-h)\right). \end{aligned} \tag{1.2.40}$$

Now suppose that we want to look at this approximation to f'' at points x_i along the x-axis. For instance, suppose we want to know $f''(x_i)$ and suppose the distance between the x_i points is constant and equal to h. Then

$$f''(x_i) \approx \frac{1}{h^2}\left(f(x_{i+1}) - 2f(x_i) + f(x_{i-1})\right). \tag{1.2.41}$$

Or, if we denote $f(x_i)$ by f_i, then the approximate second derivative of the function at a given i location looks exactly like the ith row of the matrix above. In the limit that the number of mass points (and hence i locations) goes to infinity, the displacement $\mathbf{u}$ becomes a continuous function of the spatial location, which we now refer to as x, and K becomes a second derivative operator. To get the limit we have to introduce the lattice spacing h:

$$\frac{1}{\omega_0^2}\ddot{\mathbf{u}} = h^2\frac{1}{h^2}K\cdot\mathbf{u}. \tag{1.2.42}$$

We can identify each row of $\frac{1}{h^2}K\cdot\mathbf{u}$ as being the approximate second derivative of the corresponding displacement. But we can't quite take the limit yet, since ω_0 is defined in terms of the discrete mass and it's not clear what this would mean in the limit of a continuum. So let's write this as

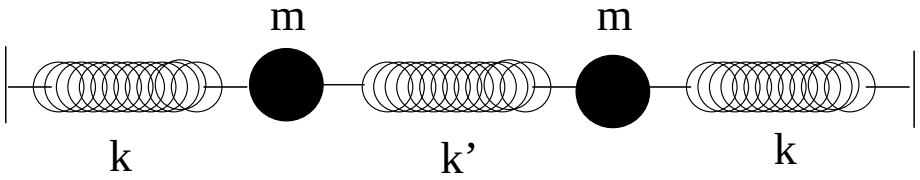
$$\ddot{\mathbf{u}} = \frac{k}{m}\frac{h^3}{h^3}K\cdot\mathbf{u} = \frac{k}{h}\frac{h^3}{m}\frac{1}{h^2}K\cdot\mathbf{u} \tag{1.2.43}$$

so that in the limit that the number of mass points goes to infinity, but the mass of each point goes to zero and the spacing h goes to zero, we can identify $\frac{m}{h^3}$ as the density and $\frac{k}{h}$ as the stiffness per unit length. Let's call the latter E. Now in this limit $\mathbf{u}$ is no longer a finite length vector, but a continuous function of the position x. Since it is also a function of time, these derivatives must become partial derivatives. So in this limit we end up with

$$\frac{\partial^2 u(x,t)}{\partial t^2} = \frac{E}{\rho}\frac{\partial^2 u(x,t)}{\partial x^2}. \tag{1.2.44}$$

This is called the wave equation.

Exercises



1.1 Write down the equations of motion for the system above in terms of the displacements of the two masses from their equilibrium positions. Call these displacements x_1 and x_2.

Answer: The equations of motion are

$$m\ddot{x_1} + kx_1 + k'(x_1 - x_2) = 0 \quad (1.2.45)$$
$$m\ddot{x_2} + kx_2 + k'(x_2 - x_1) = 0. \quad (1.2.46)$$

1.2 What are the two characteristic frequencies? (I.e., the frequencies of the fast and slow modes.)

Answer: Defining $\omega_0^2 = k/m$ and $\Omega^2 = k'/m$, we have $\omega_+^2 = \omega_0^2$ and $\omega_-^2 = \omega_0^2 + 2\Omega^2$. (The + and - sign denote sign of the square root.)This makes sense since if $k' = k$ we get the familiar result.

1.3 What is the difference in frequency between the fast mode and the slow mode in the limit that $k' \to 0$? What is the physical interpretation of this limit?

Answer: 0. This limit corresponds to the middle spring being cut or removed. So both masses oscilate at the same frequency ω_0.

1.4 Write down the equations of motion for a mass suspended from an undamped linear spring. Initially the spring has no mass attached to it. It's now at its relaxed length. You apply a 100 gram mass to the spring, pull the mass down and let it go. You measure its natural frequency as 2 Hz. What is the spring constant k? Don't forget that ω_0 is measured in radians/second.

Answer: $\sqrt{k/m} = \omega_0 = 4\pi \mathrm{s}^{-1}$. So $k = (4\pi)^2 .1\mathrm{kg/s}^2 = 15.8\mathrm{N/m}$.

1.5 Given that $g = 980cm/s^2$, what displacement from the spring's relaxed length would you expect the mass to cause.

Answer: $mg = kx$, so $x = mg/k = .1\mathrm{kg}\ 9.8\frac{\mathrm{m}}{\mathrm{s}^2}/15.8\frac{\mathrm{N}}{\mathrm{m}} = .06\mathrm{m} = 6\mathrm{cm}$.

1.6 Now suppose that the mass is sitting in a viscous fluid. What is the equation of motion?

Answer: $m\ddot{x} + m\gamma\dot{x} + kx = 0$.

1.7 For the characteristic frequency you estimated above, what is the minimum damping required to ensure that the mass does not oscillate if you pull it down and let it go.

Answer: $\gamma_{\mathrm{critical}} = 2\omega_0 = 2 \times 2\pi \times 2\mathrm{s}^{-1} \approx 25\mathrm{s}^{-1}$.

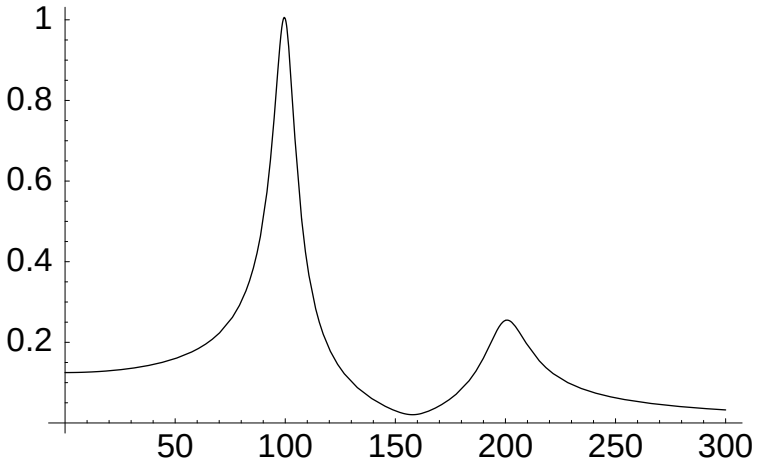
1.8 With this minimum (or "critical") damping, how long will it take for the mass to come to rest?

Answer: Strictly speaking $e^{-\gamma t/2}$ is never zero for any finite t. So the answer is infinite. But of course this is a non-physical result. It results from too simple a mathematical model. It would make more sense to ask: how long will it take before the displacement is, say, ϵ times the initial displacement, where ϵ is a small number. This you can easily answer since you simply need to find the t such that $e^{-\gamma t/2} = \epsilon$.

1.9 Prove that γ is the full width at half max of the ρ^2 curve provided γ is small. Hint: evaluate ρ^2 at $\omega_0 \pm \gamma/2$.

Answer: See question 13 below.

1.10 Estimate the Q of the two peaks in the curve below showing hypothetical amplitude versus circular frequency.



Answer: In both cases the Q is 10. That's 100/10 and 200/20. Unless the Q is very large it is not worth trying to be too precise about this measurement.

1.11 The equation describing the behavior of an RLC circuit are:

$$L\ddot{q} + R\dot{q} + \frac{1}{C}q = V(t)$$

where q is the charge on the capacitor, L is the inductance of the coil, R is the resistance, C the capacitance, and V is the applied voltage.

Reasoning by analogy with the spring/mass problem

(a) What is the damping constant (γ) for the circuit?
Answer: R/L.

(b) What is the characteristic frequency (ω_0)?
Answer: $\sqrt{\frac{1}{LC}}$

(c) What is the Q?
Answer: $\omega_0 \frac{L}{R} = \frac{1}{R}\sqrt{\frac{L}{C}}$

(d) If the inductance is 25×10^{-3} H (1 Henry = 1 volt per amp per second), what capacitance is required to have a characteristic period of 1 second?
Answer: $\sqrt{\frac{1}{LC}} = \omega_0 = 2\pi f_0 = \frac{2\pi}{1\text{sec}}$. Therefore $C = \frac{1}{L(2\pi)^2} \approx 1$Farad.

1.12 Solve the equations of motion for an undamped spring/spring mass system with three identical springs (fixed endpoints) and masses m_1 and m_2 where m_2 is much greater than m_1. In particular, what are the approximate frequencies of the slow and fast modes?

Answer: The equations of motion are

$$\begin{aligned} m_1\ddot{x_1} + kx_1 + k(x_1 - x_2) &= 0 \quad (1.2.47) \\ m_2\ddot{x_2} + kx_2 + k(x_2 - x_1) &= 0 \quad (1.2.48) \end{aligned}$$

or

$$\begin{aligned} \ddot{x_1} + 2\omega_1^2 x_1 - \omega_1^2 x_2 &= 0 \quad (1.2.49) \\ \ddot{x_2} + 2\omega_2^2 x_2 - \omega_2^2 x_1 &= 0. \quad (1.2.50) \end{aligned}$$

Now since $m_1 << m_2$, it follows that $\omega_1^2 = \frac{k}{m_1} >> \omega_2^2 = \frac{k}{m_2}$. Inserting the usual $x_1 = Ae^{i\omega t}$ and $x_2 = Be^{i\omega t}$ into the equations of motion we get the amplitude equations:

$$\begin{aligned} (-\omega^2 + 2\omega_1^2)A &= \omega_1^2 B \quad (1.2.51) \\ (-\omega^2 + 2\omega_2^2)B &= \omega_2^2 B. \quad (1.2.52) \end{aligned}$$

Eliminating A and B we get the equation for the frequency:

$$(-\omega^2 + 2\omega_1^2)(-\omega^2 + 2\omega_2^2) = \omega_1^2\omega_2^2.$$

This is a quadratic equation for ω^2 the roots of which are:

$$\omega_\pm^2 = \omega_1^2 + \omega_2^2 \pm \omega_1^2\sqrt{1 - \left(\frac{\omega_2}{\omega_1}\right)^2 + \left(\frac{\omega_2}{\omega_1}\right)^4}.$$

There are various levels of approximation we could pursue. First we could begin by dropping the quartic term in the square root in comparison to the quadratic. Or we could drop them both in comparison to 1. That would leave

$$\omega_\pm^2 = \omega_1^2 + \omega_2^2 \pm \omega_1^2$$

But of course if we are neglecting $\left(\frac{\omega_2}{\omega_1}\right)^2$ in comparison to 1, then we might as well neglect ω_2^2 in comparison to ω_1^2, in which case we have

$$\omega_\pm^2 = \omega_1^2 \pm \omega_1^2.$$

1.13 When we showed that γ was the full-width at half-max of the function $\rho^2 = ((\omega_0^2 - \omega^2)^2 + \gamma^2\omega^2)^{-1}$, we dropped certain terms by assuming that γ was small. So the question is: how small is small? To answer this look at $\rho^2(\omega_0 + \gamma/2)$, as we did in class, but don't drop any terms. Show that the error we incurred in dropping terms is on the order of $1/Q$.

Answer: We want to evaluate ρ^2 at $\omega_0 + \gamma/2$. It's easiest to begin by looking at $1/\rho^2(\omega)$:

$$(\omega^2 - \omega_0^2)^2 + \gamma^2\omega^2.$$

To evaluate this we need to evaluate $(\omega^2 - \omega_0^2)^2$ and $\gamma^2\omega^2$ for $\omega = \omega_0 + \gamma/2$: $\omega^2 = \omega_0^2 + \omega_0\gamma + \gamma^2/4$. So $(\omega^2 - \omega_0^2)^2 = \omega_0^2\gamma^2 + 1/2\omega_0\gamma^3 + \gamma^4/16$. Similarly $\gamma^2\omega^2 = \omega_0^2\gamma^2 + \omega_0\gamma^3 + \gamma^4/4$. So

$$\rho^{-2}(\omega_0 + \gamma/2) = 2\omega_0^2\gamma^2 + \frac{3}{2}\omega_0\gamma^3 + \frac{5}{16}\gamma^4.$$

Now if we neglect the terms that are order γ^3 and γ^4 we get our previous result that $\rho^2(\omega_0) = \frac{1}{2}\rho^2(\omega_0 + \gamma/2)$; i.e., that γ is the FWHM (full-width at half maximum). Now if we look at the ratio of the two highest powers of γ (neglecting constant numerical factors) we see that this ratio is

$$\frac{\gamma^3\omega_0}{\gamma^2\omega_0^2} = \frac{\gamma}{\omega_0} = \frac{1}{Q}.$$

So that if Q is roughly 100, then the error involved in taking γ to be the full-width at half-max is around 1%. But if Q is 10, this error is roughly 10%. One of the problems we face in geophysics, especially in the geophysics of the upper crust, is that our Qs are relatively small. Laboratory measurements of the Q of fluid saturated rocks might be less than 10!

1.3 More on coupled spring/mass lattices

The best tools for analyzing coupled linear systems involve Fourier analysis and eigenvector/eigenvalue analysis. We will develop as we go. But we can see the key ideas with a simple example. Let us return to the two mass/three spring case in which $m_1 = m_2 = m$ and $k_1 = k_2 = k_3 = k$, so there are two characteristic frequencies of vibration, the slow mode (in phase motion at ω_0) and the fast mode (out of phase motion at $\sqrt{3}\omega_0$). We can arrive at these results from a more abstract algebraic point of view that will serve us well for much more complicated situations. First let's look at the 2×2 matrix formulation of the problem:

$$\ddot{\mathbf{u}} + \omega_0^2 T\mathbf{u} = 0 \tag{1.3.1}$$

where $\mathbf{u}$ is a two-dimensional column vector with components x_1 and x_2 and

$$T = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}.$$

Let us make our usual guess at the solution of linear, constant-coefficient ODEs, but now in matrix notation:

$$\mathbf{u} = e^{i\omega t}\mathbf{z} \tag{1.3.2}$$

where $\mathbf{z}$ is a vector of amplitudes (i.e., it is independent of time). Inserting this guess into Equation 1.3.1 gives

$$\omega_0^2 T\mathbf{z} - \omega^2\mathbf{z} = 0. \tag{1.3.3}$$

We can write this slightly differently by introducing the 2×2 unit matrix

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

in which case we have

$$\left(\omega_0^2 T - \omega^2 I\right)\mathbf{z} = 0. \tag{1.3.4}$$

It is a basic fact in linear algebra, which we will prove soon, that any matrix equation of the form $A\mathbf{x} = 0$ can have a nonzero solution $\mathbf{x}$ if and only if the determinant of A is zero.[8] In our case the matrix whose determinant must be zero is $(\omega_0^2 T - \omega^2 I)$ which equals

$$\begin{bmatrix} 2\omega_0^2 - \omega^2 & -\omega_0^2 \\ -\omega_0^2 & 2\omega_0^2 - \omega^2 \end{bmatrix}.$$

The determinant of this matrix is $(2\omega_0^2 - \omega^2)^2 - \omega_0^4$. For this to be zero it must be that $2\omega_0^2 - \omega^2 = \pm\omega_0^2$, or $\omega_\pm^2 = \omega_0^2, 3\omega_0^2$. But these are just the slow and fast modes we found before. Further, if we substitute these frequencies back into Equation 1.3.4 we quickly discover that the $\mathbf{z}$ that works for $\omega_+^2 = \omega_0^2$ is

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

while the $\mathbf{z}$ that works for $\omega_-^2 = 3\omega_0^2$ is

$$\begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

[8] For a 2 by 2 matrix

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

the determinant is $ad - bc$.

So we have rediscovered what we know before but from a slightly different viewpoint. Equation 1.3.4 defines what are known as eigenvalues (the ω) and eigenvector (the $\mathbf{z}$ associated with the ω). Notice that the eigenvectors have the property that

$$\omega_0^2 T\mathbf{z} = \omega^2\mathbf{z}. \tag{1.3.5}$$

Remember, ω_0^2 is fixed, as are the elements of the matrix T. So this equation says that the matrix $\omega_0^2 T$ takes an unknown vector $\mathbf{z}$ and maps it into a constant of itself $\omega^2\mathbf{z}$. This is a rather unusual thing for a matrix to do.

The main advantage to introducing this new notation of eigenvalues and eigenvectors is that it lets us conveniently treat much more complicated problems. Such as the free oscillations of the entire planet! On the web page you will find a *Mathematica* notebook that solves the eigenvalue/eigenvector problem for a lattice of many coupled masses (you can select the number yourself).

A plot of some of these modes for a homogeneous (i.e., all the spring constants and masses are the same) lattice of 50 mass points is given in Figure 1.12. Notice especially that for the longer wavelengths, the modes are pure sinusoids, while for shorter wavelengths, the modes become modulated sinusoids. We will see later that this is an absolutely fundamental property of all discrete systems. The only way to make the high frequency modes be purely sinusoidal is to let there be a continuously infinite number of springs and masses.[9]

Now the equations of motion involve a second derivative with respect to time of the position (the "a" in $F = ma$). We have gotten around this so far by looking for solutions with an $e^{i\omega t}$ time dependence. (This is the definition of a normal mode in fact.) However, with computers available we can tackle the time derivatives by brute force. On the web page you will also find a *Mathematica* notebook that integrates the second derivatives for the coupled spring/mass system by a method known as finite differences.[10] Figure 1.13 shows some of the snapshots of the output of this program. In each column you see the time evolution of an initial displacement imposed on the lattice; on the left side the initial disturbance is smooth (a Gaussian in fact) on the right the initial disturbance is sharp (corresponding to displacing one mass point and letting it go). In addition, to make the problem slightly more interesting, I've put a bump in the middle to reflect the waves slightly. You can't see the bump since I'm not showing you the springs, but you can see the reflected pulse. Play with the *Mathematica* code and try putting different spring/mass distributions in, as well as initial conditions.

[9] This is known as the "continuum" approximation.

[10] The idea is very simple. We replace derivatives of the form $\dot{x}(t)$ with "finite differences" $\dot{x}(t) \approx x(t+h) - x(t)$, and $\ddot{x}(t) \approx \dot{x}(t+h) - \dot{x}(t) = x(t+2h) - 2x(t+h) + x(t)$, and so on. We just need to choose the step-size h small enough so that a) the equations are stable and b) the equations are reasonably accurate.

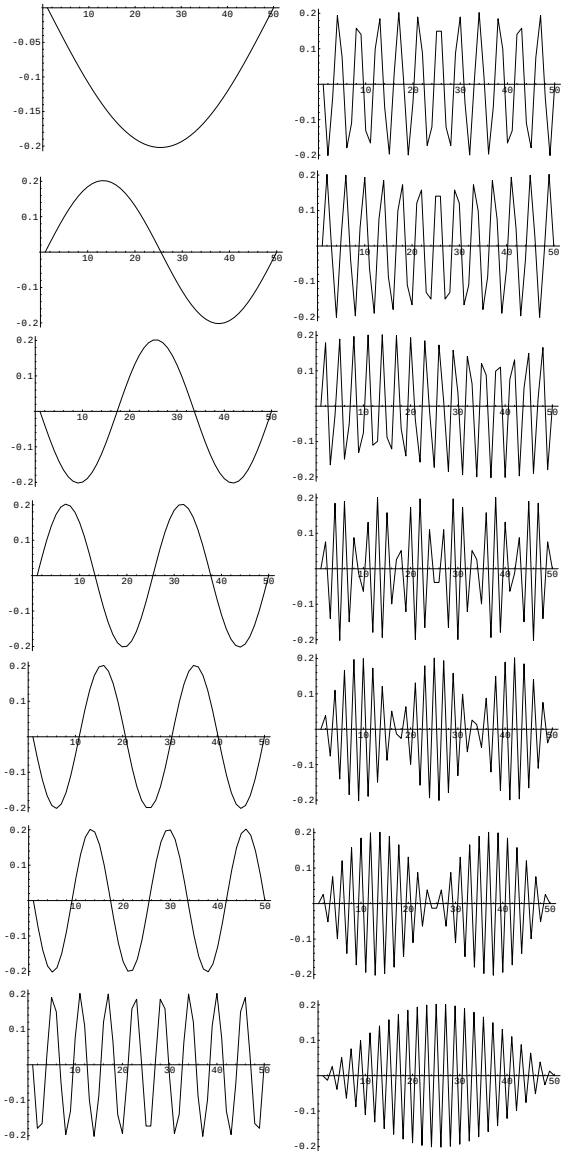


Figure 1.12: A sample of the normal modes (free oscillations) of a homogeneous 50 point lattice with fixed ends. The lower frequency modes are purely sinusoidal; the higher frequency modes become modulated sinusoids as a result of the dispersive effects of this being a discrete system.

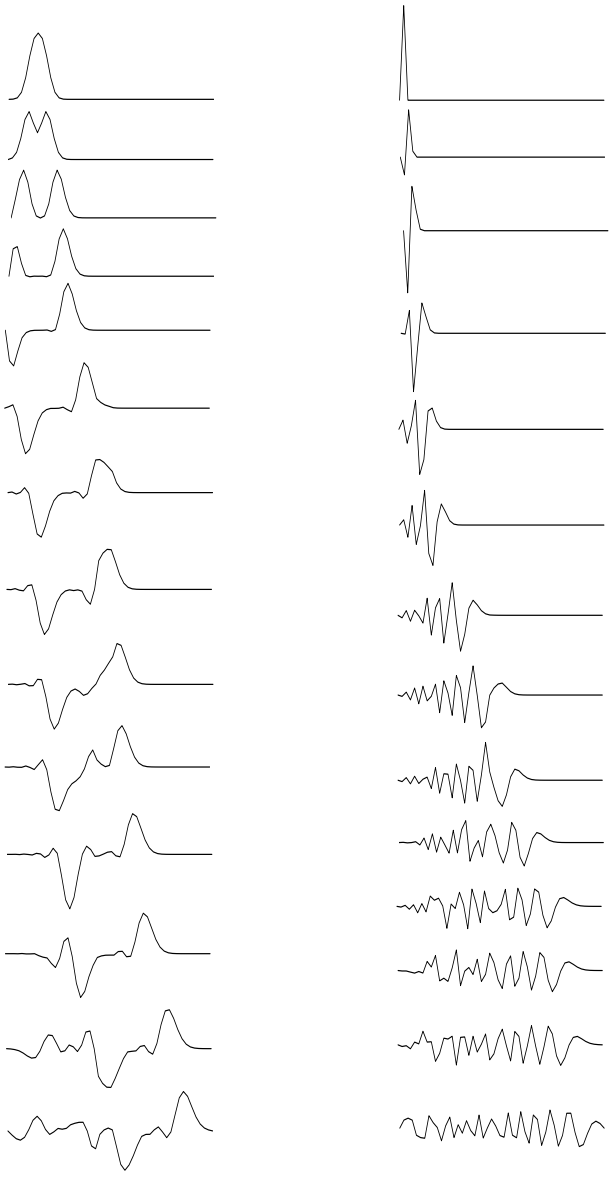


Figure 1.13: Waves on a lattice (discrete string). The two columns of figures are snapshots in time of waves propagating in a 1D medium. The only difference between the two is the initial conditions, shown at the top of each column. On the right, we are trying to propagate an inpulsive function (a displacement that is turned on only for one grid point in space and time). On the left, we are propagating a smoothed version of this. The medium is homogeneous except for an isolated reflecting layer in the middle. The dispersion seen in the right side simulation is the result of the discretness of the medium: waves whose wavelengths are comparable to the grid spacing sense the granularity of the medium and therefore propagate at a slightly different speed than longer wavelength disturbances.

Chapter 2

Waves and Modes in One and Two Spatial Dimensions

There are a number of different techniques for solving the 1-D wave equation:

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} \tag{2.0.1}$$

Perhaps the oldest is the method of traveling waves. In this method we look for solutions of the form $u(x,t) = f(x+ct)$ and $u(x,t) = f(x-ct)$. Using the chain rule you can see that

$$\frac{\partial u}{\partial t} = \pm c f'$$

where the prime denotes differentiation with respect to the argument ($x+ct$ for instance). Similarly

$$\frac{\partial^2 u}{\partial t^2} = c^2 f''$$

$$\frac{\partial^2 u}{\partial x^2} = f''.$$

As a result, **any** differentiable function evaluated at $x \pm ct$ is a solution of the 1-D wave equation. Think of the function f as representing some shape. As time increases x must increase at the rate ct in order for the shape to have the same value. This means that the shape, evaluated at $x - ct$, is actually moving to the right at a constant speed of c. Similarly $x + ct$ is moving to the left at speed ct.

2.1 1-D Separation of Variables: Summary of the Argument

Another approach to solving linear PDE's (not just the wave equation) is separation of variables. In certain coordinate systems we can find solutions which are factored. This means that the multivariate solution can be written as the product of univariate functions. The wave equation is separable in many coordinate systems, including Cartesian, spherical, and cylindrical.

Here is an overview of the argument for one spatial variable in Cartesian coordinates.

We want to solve

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} \tag{2.1.1}$$

such that, for example, $u(0,t) = u(l,t) = 0$ (clamped ends) and $u(x,0) = u_0(x)$ and $\partial u/\partial t(x,0) = v_0(x)$ where u_0 and v_0 represent the initial displacement and velocity.

Guess a solution of the form $u(x,t) = X(x)T(t)$. This doesn't always work. Plug this into Equation 2.1.1 and divide by XT.

This gives

$$c^2 \frac{X''}{X} = \frac{\ddot{T}}{T}. \tag{2.1.2}$$

This is an equation involving only x on the left and t on the right. The only way this can be true is if both sides are constant. Call this constant $-\omega^2$.

So $u = XT$ reduces (2.1.1) to two ODE's:

$$\ddot{T} + \omega^2 T = 0 \tag{2.1.3}$$

and

$$\ddot{X} + \frac{\omega^2}{c^2} X = 0. \tag{2.1.4}$$

Solve these as usual:

$$T(t) = A\cos(\omega t) + B\sin(\omega t) \tag{2.1.5}$$

$$X(x) = C\cos\left(\frac{\omega}{c}x\right) + D\sin\left(\frac{\omega}{c}x\right) \tag{2.1.6}$$

with A, B, C, D arbitrary constants.

The clamped end boundary conditions imply that $X(0) = X(l) = 0$. Therefore $C = 0$ and $\omega/c = \pi n/l$. Leaving:

$$T(t) = A\cos\left(\frac{\pi nc}{l}t\right) + B\sin\left(\frac{\pi nc}{l}t\right) \tag{2.1.7}$$

$$X(x) = D\sin\left(\frac{\pi n}{l}x\right) \tag{2.1.8}$$

Or,

$$u(x,t) = DA\sin\left(\frac{\pi n}{l}x\right)\cos\left(\frac{\pi nc}{l}t\right) + DB\sin\left(\frac{\pi n}{l}x\right)\sin\left(\frac{\pi nc}{l}t\right). \tag{2.1.9}$$

Let's relabel the constant DA, calling it A, and DB, calling it B. Then

$$u(x,t) = A\sin\left(\frac{\pi n}{l}x\right)\cos\left(\frac{\pi nc}{l}t\right) + B\sin\left(\frac{\pi n}{l}x\right)\sin\left(\frac{\pi nc}{l}t\right). \tag{2.1.10}$$

This solution obviously will not satisfy general initial conditions. However, linearity of the wave equation guarantees that if

$$A\sin\left(\frac{\pi n}{l}x\right)\cos\left(\frac{\pi nc}{l}t\right) + B\sin\left(\frac{\pi n}{l}x\right)\sin\left(\frac{\pi nc}{l}t\right) \tag{2.1.11}$$

is a solution, then so is

$$\sum_n A_n\sin\left(\frac{\pi n}{l}x\right)\cos\left(\frac{\pi nc}{l}t\right) + B_n\sin\left(\frac{\pi n}{l}x\right)\sin\left(\frac{\pi nc}{l}t\right) \tag{2.1.12}$$

where A_n and B_n are arbitrary constants.

Now we have some hope of satisfying the initial conditions. Let's see. If

$$u(x,t) = \sum_n A_n\sin\left(\frac{\pi n}{l}x\right)\cos\left(\frac{\pi nc}{l}t\right) + B_n\sin\left(\frac{\pi n}{l}x\right)\sin\left(\frac{\pi nc}{l}t\right) \tag{2.1.13}$$

then

$$u(x,0) = \sum_n A_n\sin\left(\frac{\pi n}{l}x\right) \tag{2.1.14}$$

and

$$\frac{\partial u(x,t)}{\partial t} = \sum_n B_n\frac{\pi nc}{l}\sin\left(\frac{\pi n}{l}x\right). \tag{2.1.15}$$

So this scheme will work if and only if we can choose the constants A_n and B_n such that

$$u_0(x) = \sum_n A_n\sin\left(\frac{\pi n}{l}x\right) \tag{2.1.16}$$

and

$$v_0(x) = \sum_n B_n \frac{\pi n c}{l} \sin\left(\frac{\pi n}{l} x\right). \tag{2.1.17}$$

This is our first example of a Fourier series. We will explore this last conjecture in detail. That this should be possible is not remotely obvious in my opinion and that it is true is one of the great triumphs of 19th century mathematical physics.

What's the simplest solution we could construct? We would displace the string into one of it's normal modes, initially at rest, and then let go. This corresponds to making all the B_n coefficients in 2.1.13 equal to zero (since the initial velocity is zero) and letting all but one of the A_n be zero. For instance, if we displace the string into its fundamental mode ($\sin(n\pi x/l)$, for $n = 1$) then the complete solution is

$$u(x,t) = \sin(\pi x/l) cos(\pi c t/l). \tag{2.1.18}$$

That's it. Notice that if you start the system out in one of its normal modes it stays there forever. In a linear system there is absolutely no way to transfer energy amongst the modes. Later on we'll be able to prove this directly: the energy of each mode is constant, so whatever energy a particular mode starts out with, it stays that way forever. (This is not too hard to prove. Why don't you give it a try. Just compute the energy (kinetic + potential) and integrate over one complete period of the motion.)

The symbol to the left indicates that on the WWW page you will find a *Mathematica* notebook; in this case one that solves the 1D problem for initial conditions corresponding to the string being pulled up in the middle and released at $t = 0$. We use *Mathematica*'s built-in Fourier series capability to represent a "hat" function as a 6 term sine-series. (Don't worry about the details of the Fourier analysis, we'll be covering that later.) But download this notebook and run it. You'll see a beautiful and realistic animation.

The results of running this code are shown in Figure 2.1.

2.2 2-D separation of variables

Separation of variables for the 2-D wave equation proceeds in the same way.

$$c^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = \frac{\partial^2 u}{\partial t^2} \tag{2.2.1}$$

We assume a solution of the form

$$u(x,y,t) = X(x)Y(y)T(t). \tag{2.2.2}$$

Equation 2.2.1 then becomes (after dividing by XYT)

$$c^2 \left(\frac{X''}{X} + \frac{Y''}{Y}\right) = \frac{\ddot{T}}{T}. \tag{2.2.3}$$

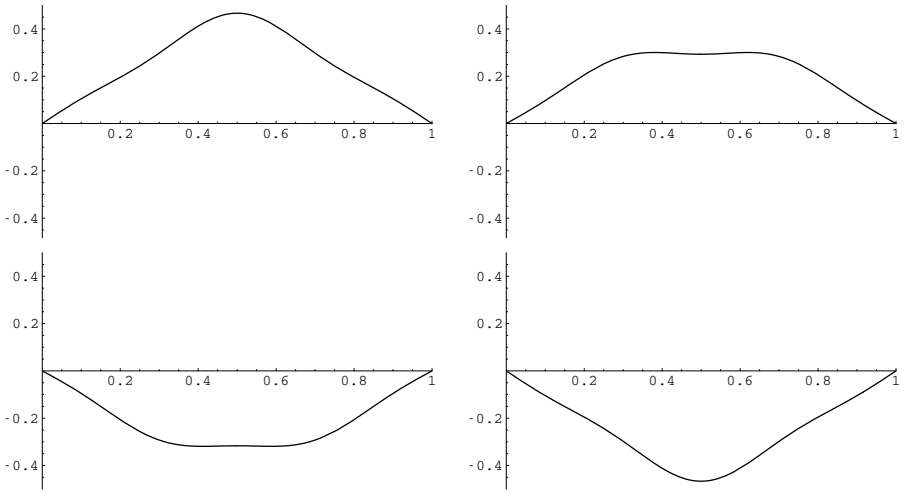


Figure 2.1: 4 of the 50 time snapshots of the plucked string. To compute this I approximated a triangular initial displacement by a 6 term Fourier series. This series is slowly convergent since the triangle function is not differentiable in the middle. But 6 terms are good enough to get a feel for what's happening.

As in 1-D, for this to be true, both sides of this equation must be constant. Let's call this constant $-\omega^2$. [1] So we have

$$\frac{\ddot{T}}{T} + \omega^2 = 0 \tag{2.2.4}$$

and

$$c^2 \left(\frac{X''}{X} + \frac{Y''}{Y} \right) = -\omega^2. \tag{2.2.5}$$

Let's rewrite this last equation as

$$\frac{X''}{X} + \frac{\omega^2}{c^2} = -\frac{Y''}{Y}. \tag{2.2.6}$$

We can apply the standard separation of variables argument again: an equation of x on the left and an equation of y on the right; this must mean that both sides equal yet another constant. Let's call this one k_y^2 (for reasons that will become apparent shortly):

$$\frac{X''}{X} + \frac{\omega^2}{c^2} = -\frac{Y''}{Y} = k_y^2. \tag{2.2.7}$$

So we have two de-coupled ODE's for the spatial variables

$$X'' + \left(\frac{\omega^2}{c^2} - k_y^2 \right) X = 0 \tag{2.2.8}$$

[1] You should convince yourself that it doesn't matter what we call this constant, plus, minus, with or without the square. It all works out the same in the end.

and

$$Y'' + k_y^2 Y = 0. \tag{2.2.9}$$

We can preserve the symmetry of these two equations by inventing a new label for $c^2 - k_y^2$. We'll call this k_x^2. Then we have the nice pair of equations

$$X'' + k_x^2 X = 0 \tag{2.2.10}$$
$$Y'' + k_y^2 Y = 0 \tag{2.2.11}$$

where, because of how we've defined k_x we have

$$\frac{\omega^2}{c^2} = k_x^2 + k_y^2. \tag{2.2.12}$$

The constants k_x and k_y have the dimensions of reciprocal length. $\frac{\omega}{c}$ is one over the wavelength, times 2π.

So we've successfully reduced the 2-D wave equation, which is a partial differential equation in two space variables and time, to three un-coupled ODE's. We already know how to solve these equations, so let's consider an interesting particular case. Let's consider a rectangular drum (a thin membrane, clamped on the sides) of lengths L_x and L_y. We'll put the origin of the coordinate system at $x = 0$, $y = 0$. Then in order for the displacement to be zero at $x = L_x$ and $y = L_y$, we must have

$$X(x) = A\sin(k_x x) \tag{2.2.13}$$
$$Y(y) = B\sin(k_y y) \tag{2.2.14}$$

where A and C are constants and $k_x = n\pi/L_x$ and $k_y = m\pi/L_y$ where m and n are arbitrary integers. So the spatial variation of the drum's vibration must be proportional to

$$\sin\left(\frac{n\pi x}{L_x}\right)\sin\left(\frac{m\pi y}{L_y}\right). \tag{2.2.15}$$

Now since

$$\frac{\omega^2}{c^2} = k_x^2 + k_y^2$$

we have

$$\frac{\omega^2}{c^2} = \pi^2\left(\frac{n^2}{L_x^2} + \frac{m^2}{L_y^2}\right) \tag{2.2.16}$$

As n and m vary over the integers, ω defines a doubly-infinite set of resonant frequencies. The same argument we made before about initial conditions applies here. To be able to solve a general initial value problem we need to be able to represent the initial conditions in a Fourier series. This will be a 2-D Fourier series in x and y but that's not a big deal.

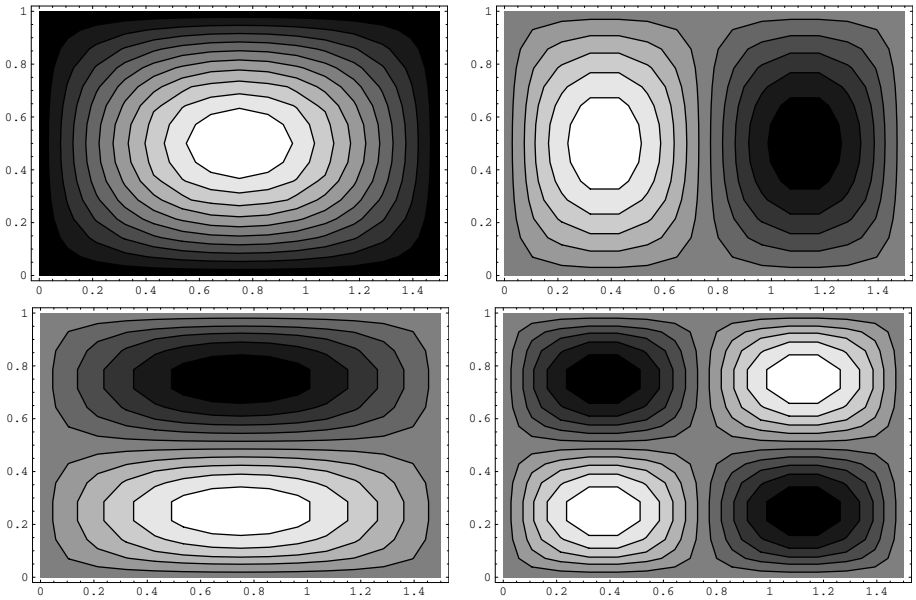


Figure 2.2: The first four modes of a rectangular drum. The aspect ratio of the drum is 1.5. The sides are clamped, so the displacement must be zero on the edge.

2.3 An Example

Here is a simple piece of *Mathematica* code that will draw the modes of a rectangular plate.

```
Lx = 1.5;
Ly = 1;
c = 1;
d[x_,y_,m_,n_]= Sin[m Pi x/Lx]Sin[n Pi y/Ly];
w[n_,m_] = c Sqrt[(m Pi /Lx)^2 + (n Pi/Ly)^2];


Do[
        Do[
                ContourPlot[d[x,y,m,n],{x,0,Lx},{y,0,Ly},
                                AspectRatio->Ly/Lx];
                ,{m,2}];
        ,{n,2}];
```

The results of running this code are shown in Figure 2.2.

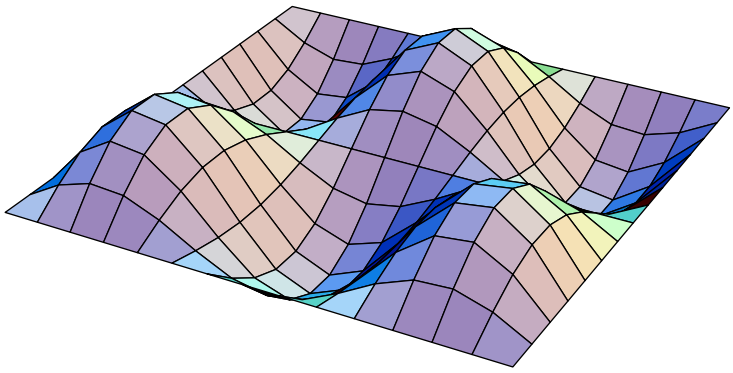


Figure 2.3: A perspective view of mode 3-2.

And in Figure 2.3 is a 3D perspective view of one of the modes. On the WWW page you'll find a *Mathematica* notebook that animates this.

2.4 Degeneracy

When we studied the Zeeman effect we saw that in the absence of a magnetic field, all three degrees of freedom oscillated with the same frequency. Applying a magnetic field splits this degenerate frequency into 3. The same thing happens with the drum. The expression we derived for the frequency of oscillation was

$$\omega_{n,m}^2 = c^2\pi^2\left(\frac{n^2}{L_x^2} + \frac{m^2}{L_y^2}\right).$$

Attaching the subscript to ω is a good reminder that it depends on the mode. Now, clearly if $L_x = L_y$, then $\omega_{i,j} = \omega_{j,i}$. This is degeneracy. If L_x is just slightly different than L_y, then the frequencies are different. But even if the frequencies are the same, the modes $n - m$ and $m - n$ are clearly different. For example, in Figure 2.4 you will see plots of the modes $n = 1 \;\; m = 2$ and $m = 1 \;\; n = 2$ for a drum for which $L_x = L_y = 1$. The two modes have different vibrational patterns, but the same frequency.

Suppose we excited the drum at a frequency $\omega_{12} = \omega_{21}$? What sort of pattern of nodal lines would we see? Like waves, modes will interfere constructively or destructively. This is a very interesting topic and we only touch upon it. But if the modes 12 and 21 were to constructively interfere, we would expect to see a vibrational pattern such as in Figure 2.5.

Finally we point out an interesting connection between number theory and normal modes. Let us rewrite our expression for the eigenfrequencies as

$$\omega_{n,m}^2 = \frac{c^2\pi^2}{L_x^2}\left(n^2 + m^2\frac{L_x^2}{L_y^2}\right).$$

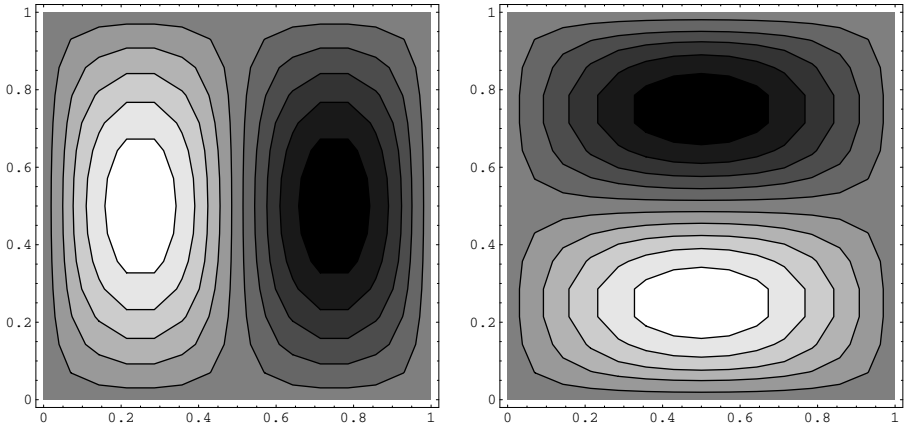


Figure 2.4: Modes 21 (left) and 12 (right) for a square drum.

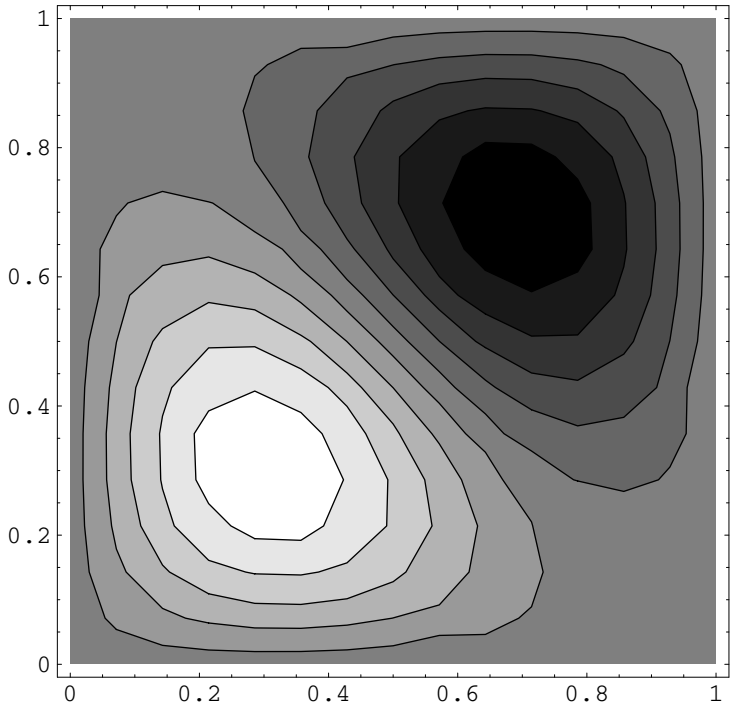


Figure 2.5: The sum of the two modes 12 and 21.

Let's suppose, just to keep life simple, that c is equal to L_x/π. And let's call the ratio $\frac{L_x^2}{L_y^2} = \xi$, so we have

$$\omega_{n,m}^2 = \left(n^2 + \xi m^2\right).$$

So the number-theoretic question I want to ask is: are there integers i, j and p, q such that the two frequencies $\omega_{i,j}$ and $\omega_{p,q}$ are equal? If they are equal then we have a degeneracy, if not, we don't. In other words, under what circumstances is it true that

$$p^2 + \xi q^2 = i^2 + \xi \text{J}^2?$$

Clearly this will be true if and only if

$$p^2 - i^2 = \xi(j^2 - q^2).$$

Assuming that $j \neq q$ of course, this implies that

$$\frac{p^2 - i^2}{j^2 - q^2} = \xi.$$

Since all of the numbers p, q, i, j are integers, this equation can only be true if ξ is a rational number. Therefore we have proved that if the ratio of the lengths of the sides of the drum is irrational, then there is no degeneracy. The Greeks, who knew all about harmonics of music, described sides whose ratio was irrational as being "incommensurate", a word that means not measurable. The Platonists had a theory that the universe was made of whole numbers. One would be a point. Two would be a line joining two points, three would be a triangle, and so on. They thought that everything could be built up out of these basic unit. It was a shock therefore to discover that the diagonal of a unit square was incommensurate: it could not be measured by any ruler made from the side of the square. No matter how finely you made the lines of the ruler, the diagonal would fall somewhere in between two lines. And not mid-way either, somewhere off to one side or the other.

2.5 Pictures of Modes

I'll conclude this chapter with some pictures of real modes. A "stadium" is a geometrical shape consisting of two circles on the end of a rectangle. The stadium shape has an important role in modern theories of chaos. Figure 2.6 shows two modes of such a shape visualized by dropping sand. The sand collects on node lines (i.e., places where the displacement is zero). These are called Chladni figures.

This particular stadium consists of an aluminum plate 194 mm long by 100 mm wide by about 3 mm thick. The plate is attached via a screw through a hole in the middle to a resonator. The whole system is driven by an HP function generator putting out a 10 volt

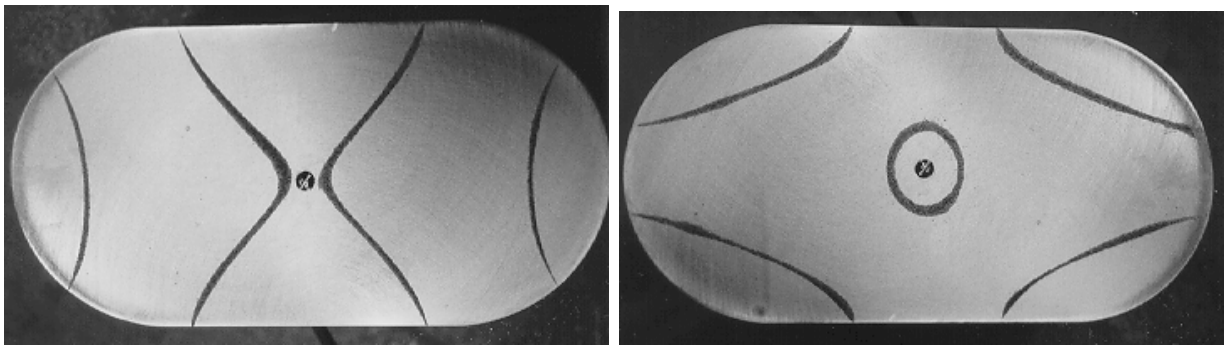


Figure 2.6: Chladni figures of two normal modes of a stadium plate. The mode on the left has a frequency of 1754 Hz and the one right 2116 Hz.

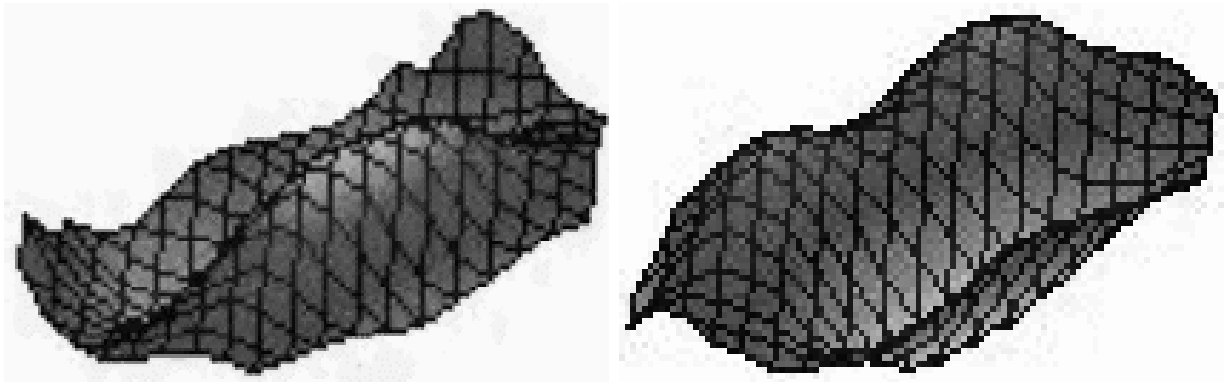


Figure 2.7: Displacement of the stadium plate at two different times when being driven in one of its normal modes. The measurement was performed with a laser-Doppler vibrometer.

RMS sine wave at the frequencies indicated in the caption. For more Chaldni figures of the stadium plate, see my web page.

Figure 2.7 shows a different visualization of a mode of the stadium. These are two snapshots of the instantaneous displacement of the plane when being driven in one of its modes. The measurements were made with a laser-Doppler vibrometer.

2.6 Spherical and Cylindrical Harmonics

In this section we will apply separation of variables to Laplace's equation in spherical and cylindrical coordinates. Laplace's equation is important in its own right as the cornerstone of potential theory, but the wave equation also involves the Laplacian derivative, so the ideas discussed in this section will be used to build solutions of the wave equation in spherical and cylindrical coordinates too. The treatment given here is completely

standard and I have nothing new to say on the subject. The results are given here for convenience; but many excellent textbooks cover the same ground. A particular favorite of mine, for undergraduates, is *Classical Electromagnetic Radiation* by Heald and Marion [5].

Spherical coordinates are important when treating problems with spherical or nearly-spherical symmetry. To a first approximation the earth is spherical and so is the hydrogen atom, with lots of other examples in-between. Before we treat the wave equation, let's look at the simpler problem of Laplace's equation:

$$\nabla^2\psi(x,y,z) = 0. \tag{2.6.1}$$

In Cartesian coordinates this is:

$$\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} + \frac{\partial^2\psi}{\partial z^2} \tag{2.6.2}$$

Laplace's equation is fundamental in geophysics since it describes the behavior of static electric and gravitational fields outside of the regions where this is charge or matter. For example, a point charge q at the origin creates an electric potential $\psi(r) = \frac{q}{r}$. As an exercise, carry out the differentiations of $\frac{1}{r} = (x^2+y^2+z^2)^{-1/2}$ and show that $\nabla^2\psi(x,y,z)$ is identically zero for $r > 0$, where the charge is located.

Joseph Louis Lagrange introduced the idea of potentials into his work on gravity. Lagrange is almost universally described as one of the great French mathematicians, but he was actually born in Turin (in what is now Italy) and baptized in the name of Giuseppe Lodovico Lagrangia. Lagrange, who worked for over 20 years in Berlin, made fundamental contributions in nearly all areas of mathematics and physics, in particular astronomy, the stability of the solar system, mechanics, dynamics, fluid mechanics, probability, and the foundations of the calculus as well as number theory. Lagrange died in Paris in April 1813.

This and other biographical material you will find in this book comes largely from the the St. Andrews University History of Mathematics WWW page:

`http://www-groups.dcs.st-andrews.ac.uk /~history/Mathematicians`.

Pierre-Simon Laplace really was French, having been born in Normandy in 1749. Laplace's mathematical talents were recognized early and he moved to Paris when he was 19 to further his studies. Laplace presented his first paper to the *Académie des Sciences* in Paris when he was 21 years old. He went on to make profound advances in differential equations and celestial mechanics. Laplace survived the reign of terror and was one of the first professors at the new *Ecole Normale* in Paris. Laplace propounded the nebular hypothesis for the origin of the solar system in his *Exposition du systeme du monde*. He also advanced the radical proposal that there could exist stars so massive that light could not escape them–we call these black holes now! And *Traité du Mécanique Céleste* is still print and widely read. Laplace also made fundamental contributions to mathematics, but I will mention only his book *Théorie Analytique des Probabilités*. He died on the third of March 1827 in Paris.

When solving boundary value problems for differential equations like Laplace's equation, it is extremely handy if the boundary on which you want to specify the boundary conditions can be represented by holding one of the coordinates constant. For instance, in Cartesian coordinates the surface of the unit cube can be represented by:

$$\begin{aligned} z &= \pm 1, \quad \text{for } -1 \le x \le 1 \text{ and } -1 \le y \le 1 \\ y &= \pm 1, \quad \text{for } -1 \le z \le 1 \text{ and } -1 \le x \le 1 \\ x &= \pm 1, \quad \text{for } -1 \le z \le 1 \text{ and } -1 \le y \le 1 \end{aligned}$$

On the other hand, if we tried to use Cartesian coordinates to solve a boundary value problem on a spherical domain, we couldn't represent this as a fixed value of any of the coordinates. Obviously this would be much simpler if we used spherical coordinates, since then we could specify boundary conditions on, for example, the surface $r =$ constant. The disadvantage to using coordinate systems other than Cartesian is that the differential operators are more complicated. To derive an expression for the Laplacian in spherical coordinates we have to change variables according to: $x = r\cos\phi\sin\theta$, $y = r\sin\phi\sin\theta$, $z = r\cos\theta$. The angle θ runs from 0 to π, while the angle ϕ runs from 0 to 2π.

Here is the result, the Laplacian in spherical coordinates:

$$\nabla^2\psi(x,y,z) = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\psi}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\psi}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2\psi}{\partial\phi^2} \tag{2.6.3}$$

Physical interpretation of the Laplacian

In 1-dimension, Laplace's equation says: $\phi''(x) = 0$. This equation can be integrated to give: $\phi(x) = ax+b$. So in 1-D any linear function (or a constant) satisfies Laplace's equation. The Laplacian operator itself measures (more or less) the curvature of a function of space. So since Laplace's equation says that this must be zero, it stands to reason the harmonic functions would be relatively smooth.

2.6.1 separation of variables

Look for solutions of the form: $\psi(r,\theta,\phi) = R(r)P(\theta)Q(\phi)$. So,

$$\begin{aligned}\nabla^2\psi(r,\theta,\phi) &= \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\psi}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\psi}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2\psi}{\partial\phi^2} \qquad (2.6.4)\\ &= \frac{PQ}{r^2}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right) + \frac{RQ}{r^2\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{dP}{d\theta}\right) + \frac{RP}{r^2\sin^2\theta}\frac{d^2Q}{d\phi^2} = 0.\end{aligned}$$

Dividing, as usual, by RPQ we have:

$$\frac{1}{Rr^2}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right) + \frac{1}{Pr^2\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{dP}{d\theta}\right) + \frac{1}{Qr^2\sin^2\theta}\frac{d^2Q}{d\phi^2} = 0. \qquad (2.6.5)$$

This looks more complicated than what we had with rectangular coordinates. In fact it looks like we're stuck since all three terms involve both r and θ. However, multiplying by $r^2\sin^2\theta$ makes the third term just $\frac{1}{Q}\frac{d^2Q}{d\phi^2}$. So,

$$\frac{\sin^2\theta}{R}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right) + \frac{\sin\theta}{P}\frac{d}{d\theta}\left(\sin\theta\frac{dP}{d\theta}\right) = -\frac{1}{Q}\frac{d^2Q}{d\phi^2}. \qquad (2.6.6)$$

This we can make some progress with since we have a function of r and θ on the left side and a function of ϕ on the right; therefore both sides must be equal to a constant, which we'll call m^2. Thus

$$\frac{d^2Q}{d\phi^2} + m^2Q = 0 \qquad (2.6.7)$$

and so Q must be proportional to $e^{im\phi}$. In order that the solution be continuous, we must require that $Q(\phi) = Q(\phi+2\pi)$ so m must be an integer. Of course, it may happen that one is interested in solving a boundary on a subset of a sphere, in which case it may not be true that Q is continuous; in that case m need not be an integer.

Next, for the r and θ part of Equation 2.6.6.

$$\frac{\sin^2\theta}{R}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right) = -\frac{\sin\theta}{P}\frac{d}{d\theta}\left(\sin\theta\frac{dP}{d\theta}\right) + m^2. \qquad (2.6.8)$$

Again, to separate the r and θ we divide by $\sin^2\theta$:

$$\frac{1}{R}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right) = -\frac{1}{P\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{dP}{d\theta}\right) + \frac{m^2}{\sin^2\theta}. \tag{2.6.9}$$

Now we can introduce another separation constant, call it k^2. With this we get the radial equation:

$$\frac{1}{R}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right) = k^2. \tag{2.6.10}$$

This turns out to be easy to solve if we guess a trial solution of the form $R = Ar^\alpha$. Plugging this into the radial equation we get:

$$\alpha(\alpha+1)Ar^\alpha - k^2Ar^\alpha = 0 \tag{2.6.11}$$

which implies that $k^2 = \alpha(\alpha+1)$. Now if we were to write the separation constant k^2 as $k^2 = \ell(\ell+1)$, then it would be easy to see that

$$\alpha(\alpha+1) = \ell(\ell+1) \tag{2.6.12}$$

is the same as

$$(\alpha-\ell)(\alpha+(\ell+1) = 0. \tag{2.6.13}$$

This equation is satisfied for $\alpha = \ell$ and $\alpha = -(\ell+1)$. In other words, the solution to

$$\frac{1}{R}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right) = \ell(\ell+1) \tag{2.6.14}$$

is

$$R(r) = A_\ell r^\ell + B_\ell r^{-(\ell+1)}. \tag{2.6.15}$$

Lastly, we must solve the θ equation for P (i.e., the right side of Equation 2.6.9 set equal to $\ell(\ell+1)$):

$$\frac{1}{\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{dP}{d\theta}\right) + \left[\ell(\ell+1) - \frac{m^2}{\sin^2\theta}\right]P = 0. \tag{2.6.16}$$

This is called Legendre's equation and is sometimes written in terms of the variable $x = \cos\theta$ since then, $\frac{1}{\sin\theta}\frac{d}{d\theta} = -\frac{d}{dx}$ which leads to

$$\frac{d}{dx}\left[(1-x^2)\frac{dP}{dx}\right] + \left[\ell(\ell+1) - \frac{m^2}{1-x^2}\right]P. \tag{2.6.17}$$

Remember that the angle θ runs from 0 to π, so the same interval in x corresponds to $[-1, 1]$.

The solutions to Equations 2.6.16 or 2.6.17 are called *associated Legendre* functions and are named after Adrien-Marie Legendre, one of the towering figures of 18th and 19th century mathematics. Legendre was born (Sept 1752) and died (Jan 1833) in Paris. He produced major works on number theory, elliptic functions, geometry and celestial mechanics.

The standard way to proceed with the solution of Legendre's equation is by power series. The solution P is expanded in a power series in x (or $\cos\theta$) of the form:

$$P(x) = (1-x^2)^{m/2} \sum_{n=0}^{\infty} a_n x^n \tag{2.6.18}$$

Since the solutions must be defined on the interval $[-1, 1]$, we do not include any negative powers of x. So, to find the coefficients a_n, we insert the power series into Equation 2.6.17 and derive a recursion relation. I will skip the details, which you can find in many applied mathematics books, but the essential idea is that the power series solution diverges at the end-points $x = \pm 1$ unless $\ell \geq |m|$. And in this case the power series actually terminates and becomes a polynomial in x: the coefficients a_n are zero when $n > \ell - |m|$. This is why the solutions are called Legendre polynomials; they are written $P_{\ell m}(x)$, with ℓ and m integers and $\ell \geq |m|$. Strictly speaking $P_{\ell m}(x)$ are called associated Legendre polynomials. The term Legendre polynomial is reserved to the special case of $m = 0$.

The case of axial symmetry: $m = 0$

The separation constant m appears in both the ϕ (i.e., Q) and θ (i.e., P) equations. However, if $m = 0$ then $Q(\phi)$ is just a constant. So, for problems which are symmetrical about the z axis (independent of ϕ) the θ equation reduces to

$$\frac{1}{\sin\theta}\frac{d}{d\theta}\left(\sin\theta\frac{dP}{d\theta}\right) + \ell(\ell+1)P = 0 \tag{2.6.19}$$

while the x equation reduces to

$$(1-x^2)\frac{d^2P}{dx^2} - 2x\frac{dP}{dx} + \ell(\ell+1)P = 0. \tag{2.6.20}$$

The solution depends on only one index now, ℓ, and is written $P_l(x)$. By examining the recursion relation for the coefficients of the power series one can derive the following two formulae for the Legendre polynomials:

$$P_\ell(x) = \frac{1}{2^\ell \ell!} \frac{d^\ell}{dx^\ell} \left(x^2 - 1\right)^\ell . \tag{2.6.21}$$

$$P_{\ell m}(x) = (-1)^m (1 - x^2)^{m/2} \frac{d^m}{dx^m} P_\ell(x). \tag{2.6.22}$$

Of course P_ℓ^m reduces to P_ℓ when $m = 0$. These expression for the Legendre polynomials are referred to as *Rodrigues' formulae.*[2]

So the separable solutions to Laplace's equation involves multiplying the r solutions by the θ solutions by the ϕ solutions:

$$\psi(r, \theta, \phi) = \left\{ \begin{array}{c} r^\ell \\ r^{-(\ell+1)} \end{array} \right\} P_{\ell m}(\cos\theta) e^{im\phi} \tag{2.6.23}$$

which reduces in the axi-symmetric case to

$$\psi(r, \theta, \phi) = \left\{ \begin{array}{c} r^\ell \\ r^{-(\ell+1)} \end{array} \right\} P_\ell(\cos\theta). \tag{2.6.24}$$

This is the final result of separation of variables. You will have to take it on faith for now that **any** solution of Laplace's equation can be built up by superposition out of these basic separation of variables solutions. In other words, **any potential function** (solution to Laplace's equation) can be written as:

$$\psi(r, \theta, \phi) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} \left(A_{\ell m} r^\ell + B_{\ell m} r^{-(\ell+1)} \right) Y_{\ell m}(\theta, \phi). \tag{2.6.25}$$

Shortly, when we solve boundary value problems for Laplace's equation, all the work will boil down to computing the A and B coefficients given the boundary values.

Here are the first few Legendre polynomials:

$$P_1(x) = 1 \tag{2.6.26}$$

$$P_2(x) = x \tag{2.6.27}$$

$$P_3(x) = \frac{1}{2}(3x^2 - 1) \tag{2.6.28}$$

It is standard to put the θ and ϕ dependence of the solutions to Laplace's equations together into a single set of functions called *spherical harmonics.*[3] The spherical harmonics

[2]Rodriques was actually Benjamin Olinde, born in Bordeaux in 1794, the son of a wealthy Jewish banker. Olinde studied mathematics at the *Ecole Normale* in Paris, taking his doctors degree in 1816 with a thesis containing the famous formulae for Legendre polynomials.

[3]A harmonic is any solution of Laplace's equation.

are defined as:

$$Y_{\ell m}(\theta,\phi) = \sqrt{\frac{2\ell+1}{4\pi}\frac{(\ell-m)!}{(\ell+m)!}} P_{\ell m}(\theta,\phi) e^{im\phi}. \tag{2.6.29}$$

The first few spherical harmonics are:

$$\begin{aligned}
Y_{00}(\theta,\phi) &= \sqrt{\frac{1}{4\pi}} && (2.6.30)\\
Y_{10}(\theta,\phi) &= \sqrt{\frac{3}{4\pi}}\cos\theta && (2.6.31)\\
Y_{1\pm1}(\theta,\phi) &= \mp\sqrt{\frac{3}{8\pi}}\sin\theta e^{\pm i\phi} && (2.6.32)\\
Y_{20}(\theta,\phi) &= \sqrt{\frac{5}{16\pi}}(2\cos^2\theta - \sin^2\theta) && (2.6.33)\\
Y_{2\pm1}(\theta,\phi) &= \mp\sqrt{\frac{15}{8\pi}}\cos\theta\sin\theta e^{\pm i\phi} && (2.6.34)\\
Y_{2\pm2}(\theta,\phi) &= \sqrt{\frac{15}{32\pi}}\sin^2\theta e^{\pm 2i\phi} && (2.6.35)
\end{aligned}$$

2.6.2 Properties of Spherical Harmonics and Legendre Polynomials

The Legendre polynomials and the spherical harmonics satisfy the following "orthogonality" relations. We will see shortly that these properties are the analogs for functions of the usual orthogonality relations you already know for vectors.

$$\begin{aligned}
\int_{-1}^{-1} P_{\ell'}(x)P_\ell(x)dx &= \frac{2}{2\ell+1}\delta_{\ell\ell'} && (2.6.36)\\
\int_{-1}^{-1} P_{\ell' m}(x)P_{\ell m}(x)dx &= \frac{2}{2\ell+1}\frac{(\ell+m)!}{(\ell-m)!}\delta_{\ell\ell'} && (2.6.37)\\
\int_{4\pi} Y_{\ell m}(\theta,\phi)\bar{Y}_{\ell' m'}(\theta,\phi)d\Omega &= && \\
\int_0^{2\pi}\int_0^{\pi} Y_{\ell m}(\theta,\phi)\bar{Y}_{\ell' m'}(\theta,\phi)\sin\theta d\theta d\phi &= \delta_{\ell\ell'}\delta_{mm'} && (2.6.38)
\end{aligned}$$

where the over-bar denotes complex conjugation and Ω represents solid angle: $d\Omega \equiv \sin\theta d\theta d\phi$. Using 4π as the limit of integration is symbolic of the fact that if you integrate $d\Omega$ over the sphere (θ going from 0 to π and ϕ going from 0 to 2π) you get 4π. Notice that the second relation is slightly different than the others; it says that for any given value of m, the polynomials $P_{\ell m}$ and $P_{\ell' m}$ are orthogonal.

There is also the following "parity" property:

$$Y_{\ell m}(\pi - \theta, \phi + \pi) = (-1)^{\ell} Y_{\ell m}(\theta, \phi). \tag{2.6.39}$$

orthogonal function expansions

The functions $P_\ell(x)$ have a very special property. They are *complete* in the set of functions on $[-1, 1]$. This means that *any* (reasonable) function defined on $[-1, 1]$ can be represented as a superposition of the Legendre polynomials:

$$f(x) = \sum_{\ell=0}^{\infty} A_\ell P_\ell(x). \tag{2.6.40}$$

To compute the coefficients of this expansion we use the orthogonality relation exactly as you would with an ordinary vector. For example, suppose you want to know the $x-$component of a vector $\mathbf{T}$. All you have to do is take the inner product of $\mathbf{T}$ with $\hat{\mathbf{x}}$. This is because

$$\mathbf{T} = T_x\hat{\mathbf{x}} + T_y\hat{\mathbf{y}} + T_z\hat{\mathbf{z}}$$

so

$$\hat{\mathbf{x}} \cdot \mathbf{T} = T_x\hat{\mathbf{x}} \cdot \hat{\mathbf{x}} + T_y\hat{\mathbf{x}} \cdot \hat{\mathbf{y}} + T_z\hat{\mathbf{x}} \cdot \hat{\mathbf{z}} = T_x$$

since $\hat{\mathbf{x}} \cdot \hat{\mathbf{z}} = \hat{\mathbf{x}} \cdot \hat{\mathbf{y}} = 0$ and $\hat{\mathbf{x}} \cdot \hat{\mathbf{x}} = 1$. When you take the inner product of two vectors you sum the product of their components. The analog of this for functions is to sum the product of the values of the function at each point in their domains. Since the variables are continuous, we use an integration instead of a summation. So the "dot" or inner product of two functions $f(x)$ and $g(x)$ defined on $[-1, 1]$ is:

$$(f, g) = \int_{-1}^{1} f(x)g(x)dx. \tag{2.6.41}$$

So, to find the expansion coefficients of a function $f(x)$ we take the inner product of f with each of the Legendre "basis vectors" $P_\ell(x)$:

$$\int_{-1}^{1} f(x)P_{\ell'}(x)dx = \sum_{\ell=0}^{\infty} A_\ell \int_{-1}^{1} P_\ell(x)P_{\ell'}(x)dx = \sum_{\ell=0}^{\infty} A_\ell \frac{2}{2\ell+1}\delta_{\ell\ell'} = \frac{2A_{\ell'}}{2\ell'+1}. \tag{2.6.42}$$

So, the ℓ-th coefficient of the expansion of a function $f(x)$ is

$$A_\ell = \frac{2\ell+1}{2} \int_{-1}^{1} f(x)P_\ell(x)dx \tag{2.6.43}$$

Similarly, we can expand any function defined on the surface of the unit sphere in terms of the $Y_{\ell m}(\theta, \phi)$:

$$\psi(\theta, \phi) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} A_{\ell m} Y_{\ell m}(\theta, \phi) \tag{2.6.44}$$

with expansion coefficients

$$A_{\ell m} = \int_{4\pi} \psi(\theta, \phi) \bar{Y}_{\ell m}(\theta, \phi) d\Omega. \tag{2.6.45}$$

For example, what is the expansion in spherical harmonics of 1? Only the $\ell = 0, m = 0$ spherical harmonic is constant, so

$$1 = \sqrt{4\pi} Y_{00}.$$

In other words, $A_{\ell m} = \sqrt{4\pi}\delta_{0,0}$.

What is a field?

The term "field" is used to refer to any function of space. This could be a scalar function or it could be a vector or even tensor function. Examples of scalar fields include: temperature, acoustic pressure and mass density. Examples of vector fields include the electric and magnetic fields, gravity, elastic displacement. Examples of tensor fields include the stress and strain inside continuous bodies.

2.7 Exercises

1. Apply separation of variables to Laplace's equation in cylindrical coordinates:

$$\nabla^2 \psi(r, \theta, z) = \frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial \psi}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 \psi}{\partial \theta^2} + \frac{\partial^2 \psi}{\partial z^2} = 0.$$

answer: We make the, by now, standard assumption that we can write the solution in the form $\psi(r, \theta, z) = R(r)Q(\theta)Z(z)$. Plugging this into Laplace's equation and dividing by RQZ we have:

$$\frac{1}{Rr} \frac{d}{dr} \left(r \frac{dR}{dr} \right) + \frac{1}{r^2 Q} \frac{d^2 Q}{d\theta^2} + \frac{1}{Z} \frac{d^2 Z}{dz^2}. \tag{2.7.1}$$

At this point we have a choice as to the order of the solution. We could first isolate the z equation or we could isolate the θ equation. Also, in choosing the sign of the separation constant, we are in effect choosing whether we want an exponentially decaying solution or a sinusoidal one. I suggest isolating the θ equation first since

we almost always want our solutions to be continuous in angle. That means we expect the fundamental θ dependence to be sinusoidal, so we write

$$\frac{r}{R}\frac{d}{dr}\left(r\frac{dR}{dr}\right)+\frac{r^2}{Z}\frac{d^2Z}{dz^2}=-\frac{1}{Q}\frac{d^2Q}{d\theta^2}=m^2. \tag{2.7.2}$$

This gives us

$$\frac{d^2Q}{d\theta^2}+m^2Q=0 \tag{2.7.3}$$

which has solutions proportional to $e^{\pm im\theta}$. Now if we had chosen the separation constant to be $-m^2$, then we would have gotten

$$\frac{d^2Q}{d\theta^2}-m^2Q=0 \tag{2.7.4}$$

the solutions of which are proportional to $e^{\pm m\theta}$. Since we usually don't expect exponential decay with angle, we choose the plus sign for the separation constant. As we will see shortly, the choice is less clear cut for the other variables. In any case we now have for the r,θ dependence:

$$\frac{r}{R}\frac{d}{dr}\left(r\frac{dR}{dr}\right)-m^2=\frac{r^2}{Z}\frac{d^2Z}{dz^2} \tag{2.7.5}$$

or

$$\frac{1}{Rr}\frac{d}{dr}\left(r\frac{dR}{dr}\right)-\frac{m^2}{r^2}=-\frac{1}{Z}\frac{d^2Z}{dz^2}. \tag{2.7.6}$$

Once again we must decide on the sign of the separation constant. Looking at the z equation we could imaging either sinusoidal or exponential dependence. So let's do both cases. First let's look for exponential z dependence. That means we'll need a negative separation constant, say, $-k^2$:

$$\frac{1}{Rr}\frac{d}{dr}\left(r\frac{dR}{dr}\right)-\frac{m^2}{r^2}=-\frac{1}{Z}\frac{d^2Z}{dz^2}=-k^2, \tag{2.7.7}$$

which implies

$$\frac{r}{R}\frac{d}{dr}\left(r\frac{dR}{dr}\right)+r^2k^2-m^2=0 \tag{2.7.8}$$

and

$$\frac{d^2Z}{dz^2}-k^2Z=0. \tag{2.7.9}$$

The Z solutions are now proportional to $e^{\pm iz}$.

Now for the R solutions. These satisfy

$$r\frac{d}{dr}\left(r\frac{dR}{dr}\right)+\left(r^2k^2-m^2\right)R=0$$

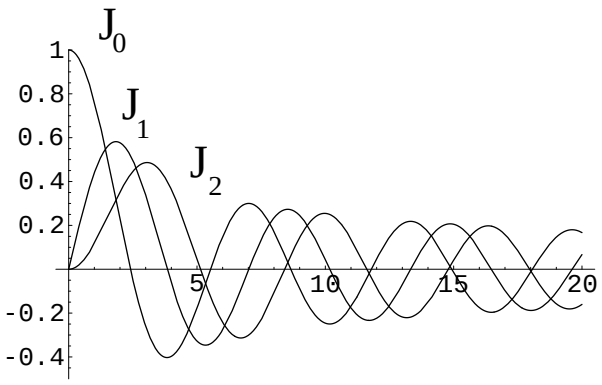


Figure 2.8: The first 3 cylindrical Bessel functions.

The constant k arose from $z-$separation, so if we set $k = 0$ this corresponds to no $z-$dependence. This is useful, for example, in studying the potentials of long wires, where we can neglect the z-dependence. It is easy to show, similarly to what we did with the axi-symmetric spherical harmonics, that in the case $k = 0$, the radial solutions are of the form r^m and r^{-m} (for $m > 0$). To see this just make a trial solution of the form $R(r) = Ar^\alpha$, then show that this satisfies the radial equation if and only if $\alpha^2 = m^2$.

The radial equation above is almost in the standard form of Bessel's equation.

Friedrich Wilhelm Bessel (born 22 July 1784 in Minden, Westphalia, died 17 March 1846 in Konigsberg, Prussia) was a leading figure in 19th century astronomy. Bessel made fundamental advances in the calculation of planetary orbits and is also well-known for his work as a teacher and educational reformer.

To get it in the standard form we make the substitution: $u = kr$, then

$$\frac{1}{u}\frac{d}{du}\left(u\frac{dR}{du}\right) + \left(1 - \frac{m^2}{u^2}\right)R = 0 \tag{2.7.10}$$

or after multiplication by u^2:

$$u^2\frac{d^2R}{du^2} + u\frac{dR}{du} + \left(u^2 - m^2\right)R = 0 \tag{2.7.11}$$

Solutions of this last equation are called cylindrical Bessel functions and are denoted by $J_m(u)$.

The power series solution to Bessel's equation can be found in many textbooks on differential equations and electricity and magnetism such as [5]. Here I will just

quote the result:

$$J_m(u) = \frac{u^m}{2^m m!} \sum_{\ell=0}^{\infty} \frac{(-1)^\ell}{2^{2\ell}\ell!(m+1)(m+2)\cdots(m+\ell)} u^{2\ell} \tag{2.7.12}$$

This is hard to deal with analytically. However for small and large values of $u = kr$, there are nice analytic approximations:

$$J_m(kr) \approx \frac{1}{(m+1)!}\left(\frac{kr}{2}\right)^m \quad kr << 1 \tag{2.7.13}$$

and

$$J_m(kr) \approx \sqrt{\frac{2}{\pi kr}} \cos\left(kr - \frac{m\pi}{2} - \frac{\pi}{4}\right) \quad kr >> 1 \tag{2.7.14}$$

sinusoidal z-dependence and modified Bessel functions

On the other hand, if instead of choosing the separation constant to be $-k^2$ we had chosen k^2, then

$$\frac{1}{Rr}\frac{d}{dr}\left(r\frac{dR}{dr}\right) - \frac{m^2}{r^2} = -\frac{1}{Z}\frac{d^2Z}{dz^2} = k^2, \tag{2.7.15}$$

and we would have gotten Z solutions proportional to $e^{\pm ikz}$ and the radial equation would have been:

$$r\frac{d}{dr}\left(r\frac{dR}{dr}\right) - (k^2r^2 + m^2)R = 0 \tag{2.7.16}$$

or equivalently

$$r^2\frac{d^2R}{dr^2} + r\frac{dR}{dr} - (k^2r^2 + m^2)R = 0. \tag{2.7.17}$$

The solutions of this equation are called modified Bessel functions (of the first kind).

2. Expand $f(x) = e^{-|x|}$ on $[-1, 1]$ in terms of $\ell = 0, 1, 2$ Legendre polynomials.
   answer: $A_0 = 1 - \frac{1}{e}$. $A_1 = 0$. $A_2 = \frac{25}{2} - \frac{35}{e}$.

3. A grounded conducting sphere of radius a is placed in a plane parallel electric field $\mathbf{E} = E_0\hat{\mathbf{z}}$. What is the electric field outside the conductor?

   answer: First we will compute the potential $V(r, \theta, \phi)$ then we will take the gradient of this to get the electric field. Since the electric field is axisymmetric about the $z-$axis, in fact the potential does not depend on ϕ. So we can be sure that we can write the unknown potential as:

$$V(r, \theta) = \sum_{\ell=0}^{\infty} \left(A_\ell r^\ell + B_\ell r^{-(\ell+1)}\right) P_l(\cos\theta).$$

We have two boundary conditions that we can apply. First the potential on the surface is zero, so

$$V(r=a,\theta)=\sum_{\ell=0}^{\infty}\left(A_\ell a^\ell+B_\ell a^{-(\ell+1)}\right)P_l(\cos\theta)=0.$$

The second boundary condition is that as $r\to\infty$, the potential must approach that of the original, unperturbed E-field:

$$\lim_{r\to\infty}V(r,\theta)=-E_0 z$$

where $\mathbf{E}=-\nabla(-E_0 z)$. If we apply the orthogonality condition to the $r=a$ boundary condition, we can see that

$$\begin{aligned}0 &= \sum_{\ell=0}^{\infty}\left(A_\ell a^\ell+B_\ell a^{-(\ell+1)}\right)\int_{-1}^{1}P_l(x)P_{\ell'}(x)dx\\ &= \sum_{\ell=0}^{\infty}\left(A_\ell a^\ell+B_\ell a^{-(\ell+1)}\right)\frac{2}{2\ell+1}\delta_{\ell\ell'}\\ &= A_\ell a^\ell+B_\ell a^{-(\ell+1)}. \qquad (2.7.18)\end{aligned}$$

So we end up with the constraint that: $B_\ell=-a^{2\ell+1}A_\ell$.

Next we apply the large-r condition. In the limit of large r, our boundary condition only constrains terms involving positive power of r, since the negative powers of r go to zero. So we must have

$$\lim_{r\to\infty}-E_0 r\cos\theta\equiv\lim_{r\to\infty}-E_0 rP_1(\cos\theta)=\sum_{\ell=0}^{\infty}A_\ell r^\ell P_\ell(\cos\theta).$$

It is clear from this that we can satisfy the boundary condition at infinity only if all the of A coefficients are zero expect the $\ell=1$ term. So $A_\ell=0$ for all ℓ except 1, and $A_1=-E_0$. We combine this with the constraint we found on the A and B coefficients above to get: $B_1=-A_1a^3=E_0a^3$. With the result that the potential everywhere outside the sphere is:

$$V(r,\theta)=-E_0 r\cos\theta+E_0a^3\frac{\cos\theta}{r^2}=-E_0\left(1-\left(\frac{a}{r}\right)^3\right)r\cos\theta.$$

From this it follows by taking the gradient in spherical coordinates that:

$$E_r=E_0\left(1+2\left(\frac{a}{r}\right)^3\right)\cos\theta$$

and

$$E_\theta=-E_0\left(1-\left(\frac{a}{r}\right)^3\right)\sin\theta.$$

4. A grounded, spherical conductor of radius a is placed in an electric field such that far away from the sphere the potential is $V(r,\theta,\phi) = r^2 \sin 2\theta \cos\phi$. Find the potential everywhere outside the sphere.

   answer: We can write any solution of Laplace's equation as:

$$\psi(r,\theta,\phi) = \sum_{\ell,m} \left(A_{\ell m} r^{\ell} + B_{\ell m} r^{-(\ell+1)} \right) Y_{\ell m}(\theta,\phi). \tag{2.7.19}$$

   In this case we are told that far away from the conductor the potential is: $r^2 \sin 2\theta \cos\phi$. OK, for large r we can only say something about the A coefficients since the terms involving the B coefficients decay at least as fast as $1/r$. Of the A coefficients it is clear that since the field must be proportional to r^2 for large r, only the $\ell = 2$ terms can be nonzero. So straight away we can see that for large r the field must be of the form

$$\psi(r \to \infty, \theta, \phi) = r^2 \left(A_{22} Y_{22} + A_{21} Y_{21} + A_{20} Y_{20} + A_{2-1} Y_{2-1} + A_{2-1} Y_{2-2} \right). \tag{2.7.20}$$

   If you look at the $\ell = 2$ spherical harmonics you will see that only the $m = 1$ terms are needed:

$$Y_{21} = -\sqrt{\frac{15}{8\pi}} \sin\theta \cos\theta e^{i\phi}$$

   and

$$Y_{2-1} = \sqrt{\frac{15}{8\pi}} \sin\theta \cos\theta e^{-i\phi}$$

   so

$$\sqrt{\frac{8\pi}{15}} \left(Y_{2-1} - Y_{21} \right) = 2 \sin\theta \cos\theta \cos\phi.$$

   In fact, since $\sin 2\theta = 2 \sin\theta \cos\theta$, it follows that:

$$r^2 \sin 2\theta \cos\phi = \sqrt{\frac{8\pi}{15}} r^2 \left(-Y_{21} + Y_{2-1} \right). \tag{2.7.21}$$

   Therefore $Y_{21} = -A_{2-1} = -\sqrt{\frac{8\pi}{15}}$.

   Now just as we did in the previous problem, we can apply the boundary condition that $\psi(r = a, \theta, \phi) = 0$ to give a constraint on the A and B coefficients:

$$B_{\ell m} = -a^{2\ell+1} A_{\ell m}. \tag{2.7.22}$$

   Hence only $B_{21} = a^5 \sqrt{\frac{8\pi}{15}}$ and $B_{2-1} = -a^5 \sqrt{\frac{8\pi}{15}}$ are nonzero. So now we have all four nonzero coefficients in the spherical harmonic expansion of ψ:

$$\begin{aligned}\psi(r,\theta,\phi) &= \sqrt{\frac{8\pi}{15}}\left[\left(-r^2+\frac{a^5}{r^3}\right)Y_{21}+\left(r^2-\frac{a^5}{r^3}\right)Y_{2-1}\right] \\ &= \sqrt{\frac{8\pi}{15}}\left(r^2-\frac{a^5}{r^3}\right)(-Y_{21}+Y_{2-1}) \qquad (2.7.23) \\ &= \left(1-\left(\frac{a}{r}\right)^5\right)r^2\sin 2\theta\cos\phi. \qquad (2.7.24)\end{aligned}$$

Notice that this agrees with the boundary condition when $r = a$. Always check your results against what you know.

5. Suppose the potential is constant on a sphere of radius a: $\psi(r = a, \theta, \phi) = V_0$. Use the spherical harmonic expansion of Laplace's equation to find the potential everywhere on the exterior of the sphere.

   answer: On the surface of the sphere, the potential is a constant. Only the $\ell = m = 0$ spherical harmonic is constant so

   $$\psi(r = a, \theta, \phi) = V_0 = V_0\sqrt{4\pi}Y_{00}. \qquad (2.7.25)$$

   This means that only the $\ell = m = 0$ term in the expansion of the field is present. This tells us immediately that

   $$\psi(r,\theta,\phi) = \left(A_{00} + B_{00}r^{-1}\right)Y_{00}. \qquad (2.7.26)$$

   Usually we don't care about constant potentials since they don't contribute to the electric or gravitational fields (the gradient of a constant is zero). So we can always shift the potential by a constant amount without changing physics; this means that we can ignore the A_{00} term. At $r = a$ we have:

   $$\psi(r = 0, \theta, \phi) = B_{00}a^{-1}Y_{00} = V_0\sqrt{4\pi}Y_{00}, \qquad (2.7.27)$$

   so $B_{00} = V_0 a^{-1}\sqrt{4\pi}$ and the complete potential outside the sphere is

   $$\psi(r,\theta,\phi) = \left(V_0 a\sqrt{4\pi}\right)r^{-1}Y_{00} = \frac{aV_0}{r}. \qquad (2.7.28)$$

6. Consider the gravitational potential on the Earth's surface. The Earth is not exactly a sphere. A better approximation is:

   $$\psi(r = R, \theta, \phi) = V_0\left(1 - J_2P_2(\cos\theta)\right)$$

   where J_2 and V_0 are constants. This is a bit of a trick actually since we're still assuming the surface is a sphere. What is the potential for $r > R$?

answer: On the surface $r = R$ the potential depends only on θ, so

$$\psi(r = R, \theta) = V_0 \left(1 - J_2 P_2(\cos\theta)\right). \tag{2.7.29}$$

Since this problem is axi-symmetric (no ϕ-dependence), the complete solution of Laplace's equation is

$$\psi(r, \theta) = \sum_{\ell=0}^{\infty} \left(A_\ell r^\ell + B_\ell r^{-(\ell+1)}\right) P_\ell(\cos\theta). \tag{2.7.30}$$

Applying the boundary condition we have

$$V_0 \left(1 - J_2 P_2(\cos\theta)\right) = \sum_{\ell=0}^{\infty} \left(A_\ell R^\ell + B_\ell R^{-(\ell+1)}\right) P_\ell(\cos\theta). \tag{2.7.31}$$

Using the orthogonality of the Legendre polynomials this equation implies two constraints on the A and B coefficients:

$$V_0 = A_0 + B_0 R^{-1}$$

and

$$-J_2 V_0 = \left(A_2 R^2 + B_2 R^{-3}\right).$$

The gravitational potential of the Earth cannot grow as you go farther away from the surface, so the A_2 term must be zero. And as before we can set any constant potential term to zero. So we're left with: $B_0 = RV_0$ and $B_2 = -J_2 V_0 R^3$. Which gives for the final solution:

$$\psi(r, \theta, \phi) = \frac{V_0 R}{r}\left[1 - J_2 \left(\frac{R}{r}\right)^2 P_2(\cos\theta)\right]. \tag{2.7.32}$$

The term J_2 corresponds to the flattening of the Earth. If this term is zero, we're left with the usual $1/r$ potential which equals V_0 on the surface. In any case, the effects of the J_2 term decay like $1/r^2$ as you recede from the Earth's surface.

7. Consider two concentric, conducting spheres of radius r_0 and r_1 respectively. The inner sphere is grounded while the outer sphere is held at a fixed potential V_0. Find the potential between the spheres by directly integrating Laplace's equation in spherical coordinates. Hint: this problem is spherically symmetric.

   answer: The spherical symmetry implies that the solution depends only on r. Therefore Laplace's equation reduces to

$$\nabla^2 \phi(r) = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial \phi}{\partial r}\right) = 0.$$

   This implies that $r^2 \frac{\partial \phi}{\partial r}$ is a constant; call it c. Integrating once more we have

$$\phi(r) = d - \frac{c}{r}$$

where r is the second integration constant. Applying the two boundary conditions we can see that $c = \frac{V_0(r_1 r_0)}{r_1 - r_0}$ and $d = \frac{c}{r}$. The final result is that

$$\phi(r) = V_0 \frac{r_1}{r} \frac{r - r_0}{r_1 - r_0} = V_0 \left[\frac{1 - \frac{r_0}{r}}{1 - \frac{r_0}{r_1}} \right].$$

8. Two functions $f(x)$ and $g(x)$ defined on an interval $[-1, 1]$ are said to be orthogonal on that interval if $(f, g) \equiv \int_{-1}^{1} f(x)g(x)dx = 0$. Similarly we can define the squared "length" of a function on the interval by: $(f, f) = \int_{-1}^{1} f^2(x)dx$. Here are two functions (polynomials of zeroth and first order) $Q_0(x) = \sqrt{\frac{1}{2}}$ and $Q_1(x) = \sqrt{\frac{3}{2}}x$, that are orthogonal and have unit length. Compute the unique quadratic function $Q_2(x)$ by using the three conditions:

$$\begin{aligned} (Q_0, Q_2) &= 0 \\ (Q_1, Q_2) &= 0 \\ (Q_2, Q_2) &= 1 \end{aligned}$$

answer: Since $Q2$ is a quadratic, it can be written $Q2(x) = ax^2 + bx + c$. The conditions $(Q_0, Q_2) = 0$ and $(Q_1, Q_2) = 0$ force $b = 0$ and $c = -1/3a$. The normalization condition (Q_2, Q_2) gives $c = \frac{3}{2}\sqrt{\frac{5}{2}}$. So,

$$Q2(x) = \frac{3}{2}\sqrt{\frac{5}{2}} \left(x^2 - \frac{1}{3} \right).$$

9. - Give the spherical harmonic expansion of $\sin\theta\cos\phi$.
   - If this is the potential on a conducting sphere of radius 1, what is the potential for $r > 1$?.

answer: In the absence of any other fields, for solutions on the exterior of bodies we want potentials that decay with r. So the general solution must be of the form

$$\psi(r, \theta, \phi) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} A_{\ell m} r^{-(\ell+1)} Y_{\ell m}(\theta, \phi)$$

The boundary condition is:

$$\psi(r = 1, \theta, \phi) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} A_{\ell m} Y_{\ell m}(\theta, \phi) = \sin\theta\cos\phi.$$

But the right-hand side is

$$\sin\theta\cos\phi = \frac{1}{2}\sqrt{\frac{8\pi}{3}} \left(Y_{11} - Y_{1-1} \right).$$

So clearly only the $\ell = 1$ $A-$coefficients are nonzero: $A_{11} = -A_{1-1} = \frac{1}{2}\sqrt{\frac{8\pi}{3}}$. Hence, the potential outside the sphere is

$$\psi(r, \theta, \phi) = \frac{1}{2}\sqrt{\frac{8\pi}{3}} r^{-2} \left(Y_{11} - Y_{1-1}\right) = \frac{\sin\theta\cos\phi}{r^2}.$$

10. The 2D Laplace's equation in Cartesian coordinates is:

$$\nabla^2\psi(x, y) = \frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} = 0.$$

Apply separation of variables to this equation to get two ordinary differential equations. Solve these equations. Explain how the choice of sign of the separation constant influences the solutions.

answer: By the now standard argument, we look for solutions of the form $X(x)Y(y)$, in which case Laplace's equation reduces to:

$$\frac{X''}{X} = -\frac{Y''}{Y}.$$

So we can choose the separation constant to be either k^2 or $-k^2$. If we choose the plus sign, then the solution will be oscillatory in the y direction and exponential in the x direction. If we choose the negative sign, the solution will be oscillatory in the x direction and exponential in the y direction. E.g., with the positive sign we get

$$X'' - k^2 X = 0$$

and

$$Y'' + k^2 Y = -0.$$

So the basic solutions are of the form

$$\psi(x, y) = \sum_k e^{ikx} e^{-ky}$$

or

$$\psi(x, y) = \sum_k e^{-kx} e^{iky}$$

2.8 More on vectors

In the next chapter we will study vectors systematically, but you already know quite a lot about them. You were taught that a vector is something that has both a magnitude (length) and direction. Examples include gravity and the electric field. You also know

that any vector can be resolved into components. For example a vector $\mathbf{T}$ in three dimensions can be resolved as

$$\mathbf{T} = T_x\hat{\mathbf{x}} + T_y\hat{\mathbf{y}} + T_z\hat{\mathbf{z}} \tag{2.8.1}$$

We will refer to the set of **all** three dimensional vectors like this as $\mathbf{R}^3$ since it's really the real line $\mathbf{R}^1$ in each of the three dimensions. The $x-y$ plane is $\mathbf{R}^2$.

To find the components of the vector in the various directions we "dot" the basis vectors (i.e., $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$ and $\hat{\mathbf{z}}$) into $\mathbf{T}$. For example

$$\hat{\mathbf{x}} \cdot \mathbf{T} = T_x\hat{\mathbf{x}} \cdot \hat{\mathbf{x}} + T_y\hat{\mathbf{x}} \cdot \hat{\mathbf{y}} + T_z\hat{\mathbf{x}} \cdot \hat{\mathbf{z}}. \tag{2.8.2}$$

But since the basis vectors are mutually perpendicular (orthogonal) $\hat{\mathbf{x}} \cdot \hat{\mathbf{y}} = \hat{\mathbf{x}} \cdot \hat{\mathbf{z}} = 0$ and $\hat{\mathbf{x}} \cdot \hat{\mathbf{x}} = 1$. So $T_x = \hat{\mathbf{x}} \cdot \mathbf{T}$ and similarly for the other components.

You can see that what really counts are the components of the vector in the mutually orthogonal directions. It doesn't really matter what we call these directions so we could also write

$$\mathbf{T} = T_x\hat{\mathbf{x}} + T_y\hat{\mathbf{y}} + T_z\hat{\mathbf{z}} = T_1\hat{\mathbf{e}}_1 + T_2\hat{\mathbf{e}}_2 + T_3\hat{\mathbf{e}}_3 = \sum_{i=1}^{3} T_i\hat{\mathbf{e}}_i \tag{2.8.3}$$

The advantage of labeling the directions by numbers is that it frees us from the constraints of ordinary three-dimensional geometry. Consider a time-series. Suppose I record the noon temperature for three days. Here are the data: $(15.3, 8.5, 11.0)$. I can pretend that these are three components of a vector in $\mathbf{R}^3$. The fact is there is no physical "daily temperature vector" but I can treat these three numbers as if they were components of a three-dimensional vector. Or they could be the first three samples of a seismic trace. And if three, why not four? Or five? It turns out that a set of numbers such as $(15.3, 8.5, 11.0, 12.1, 14.3)$ has exactly the same sort geometrical properties in $\mathbf{R}^5$ as $(15.3, 8.5, 11.0)$ does in $\mathbf{R}^3$, it's just that I can't make a plot of this vector on a piece of paper. I could extend this approach to quite long vectors, such as all the samples in a typical seismic trace, which might number in the thousands. Suppose that

$$\mathbf{A} = (a_1, a_2, a_3 ... a_{1000})$$

is a one thousand sample seismic trace and

$$\mathbf{B} = (b_1, b_2, b_3 ... b_{1000})$$

is another. Anything you can do with a three-dimensional vector you can do with a thousand-dimensional vector such as $\mathbf{A}$ and $\mathbf{B}$, except plot them. We can add two seismic traces component-wise just as you would add two force vectors:

$$\mathbf{A} + \mathbf{B} = \sum_{i=1}^{1000} A_i + B_i.$$

We can take take the length of a seismic trace:

$$\|\mathbf{A}\|^2 = \sum_{i=1}^{1000} A_i^2,$$

which is just the Pythagorean theorem in a 1000-dimensional space.

We can even compute the "angle" between two traces: since $\mathbf{A} \cdot \mathbf{B} = \|\mathbf{A}\| \, \|\mathbf{B}\| \cos\theta$ works for ordinary vectors, there is no reason we cannot extend the idea to these abstract vectors. Then, the angle between the two traces is *naturally*:

$$\begin{aligned} \cos\theta &= \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \, \|\mathbf{B}\|} \\ &= \frac{\sum_{i=1}^{1000} A_i B_i}{\sum_{i=1}^{1000} A_i^2 \sum_{i=1}^{1000} B_i^2} \end{aligned}$$

Don't be put off by this apparently abstract notion. The point is we can manipulate things like seismic traces as vectors and gain tremendous advantage from our prior geometrical understanding in $\mathbf{R}^3$, even though we're not in $\mathbf{R}^3$ any more. And for the final stroke, you should not be surprised if I tell you that we need not limit ourselves to vectors of finite length. Consider for example a power series:

$$f(x) = \sum_{i=0}^{\infty} a_i x^i.$$

Think of the powers of x here as being like our basis vectors $\hat{\mathbf{e}}_i$ in $\mathbf{R}^3$. Then the coefficients a_i are just like the coefficients A_i of our length-1000 time series above; there just happens to be an infinite number of them! Well, OK, here we **do** run into a minor spot of difficulty. For a finite length series $\sum_{i=0}^{N} a_i x^i$ we don't have to worry about convergence, but for infinite series like $\sum_{i=0}^{\infty} a_i x^i$, we do. Apart from that, we can **still** use our geometrical intuition even in spaces of infinite dimension.

Look again at the dot or inner product of two finite length vectors:

$$\mathbf{A} \cdot \mathbf{B} = \sum_{i=1}^{1000} A_i B_i$$

exactly as in $\mathbf{R}^3$. We can certainly use the same formula for the dot product of two infinite dimensional vectors:

$$\mathbf{A} \cdot \mathbf{B} = \sum_{i=1}^{\infty} A_i B_i$$

provided the sum converges. The dot product for functions is just like this except that we can't use a summation, we must use an integration:

$$f(x) \cdot g(x) = \int f(x) g(x) dx$$

where the integration is over whatever interval the functions are defined. It would be unconventional to use the "dot" for the inner product of functions, although we could. The standard notation for the dot product of functions is (f, g), thus

$$(f, g) \equiv \int f(x)g(x)dx.$$

So, when we said a few pages ago that

$$\int_{-1}^{1} P_{\ell'}(x)P_{\ell}(x)dx = \frac{2}{2\ell+1}\delta_{\ell\ell'}$$

this means that the dot product of any P_ℓ with any other is zero. So, $(P_0, P_1) = 0$ and $(P_1, P_{23}) = 0$. It really does make sense to visualize P_0, P_1 and so on, as orthogonal vectors in some infinite dimensional space. There is nothing super-complicated about this idea; it really is quite natural when you get used to it. And not only does it save you a lot of time in the long run, it also allows you to apply your geometrical insight to solve problems that would be very difficult otherwise to solve.

One last thing. If I say that

$$a_x\hat{\mathbf{x}} + a_y\hat{\mathbf{y}} + a_z\hat{\mathbf{z}} = b_x\hat{\mathbf{x}} + b_y\hat{\mathbf{y}} + b_z\hat{\mathbf{z}}$$

it will come as no surprise to you that this implies that the coefficients of each basis vector must be equal $a_x = b_x$, $a_y = b_y$, and $a_z = b_z$. This is easy to prove just by dotting the above equation with each basis vector in succession. But now suppose I say that

$$a_0 + a_1x = b_0 + b_1x?$$

Does this imply that $a_0 = b_0$ and $a_1 = b_1$? Yes and for exactly the same reason. The basis vectors are $1 = P_0$ and $x = P_1$ are mutually orthogonal (at least on $[-1, 1]$)

$$\int_{-1}^{1} 1 \cdot xdx = 0.$$

Another way to see this is to re-arrange the equation $a_0 + a_1x + = b_0 + b_1x$ as:

$$(a_0 - b_0) = (b_1 - a_1)x.$$

But x is completely arbitrary here, so the only way this equation can possibly be true for **all** x is if it really says $0 = 0x$, or $a_0 = b_0$, and $a_1 = b_1$.

This orthogonality of the basis vectors is why we could say in the HW problem on rotational elipticity of the Earth that an equation such as

$$V_0\left(1 - J_2P_2(\cos\theta)\right) = \sum_{\ell=0}^{\infty}\left(A_\ell R^\ell + B_\ell R^{-(\ell+1)}\right)P_\ell(\cos\theta). \tag{2.8.4}$$

forces us to conclude that

$$V_0 = A_0 + B_0 R^{-1}$$

and

$$-J_2 V_0 = \left(A_2 R^2 + B_2 R^{-3} \right).$$

The first equation comes from equating terms proportional P_0 (which equals 1) on both sides of equation 2.8.4 and the second comes from equating terms proportional to P_2. All the other P_ℓ terms are zero on the left side of equation 2.8.4 so they must be zero on the right.

Chapter 3

A Little More Linear Algebra

3.1 Linear Vector Spaces

In the last chapter we introduced a couple of ideas about vectors that no doubt seemed very strange at first glance. For example, that the ordinary geometry and algebra of vectors in three-dimensional physical space could be extended to spaces of any dimension, and that objects that at first bore no resemblance to vectors you know about, seismic traces for instance, had all the properties of physical vectors, you just couldn't draw pictures of them. We will be a little more systematic in this chapter; not much, just a little. The main ideas that we need to understand are linear dependence, orthogonal projection, orthogonal subspaces, and eigenvectors/eigenvalues. If you want to consult a standard reference on linear algebra, you could do no better than Strang's beautiful book [6]

You are already familiar with concrete examples of the algebra and geometry of vectors, at least in the case of vectors in three-dimensional space. We can add any two, say, force vectors and get another force vector. For example, consider a mass hanging from a spring. The total force on the mass is the sum of the gravitational force (which is pointed down) and the restoring force of the spring (which can be either up or down depending on where the mass is relative to its equilibrium position). We can add the individual forces to get the total force because they are vectors. We can also scale any such vector by a numerical quantity and still have a legitimate vector.

However, we also use vectors to encapsulate discrete information. If we record one seismogram one second in length at a sample rate of 1000 samples per second, then we can put these 1000 bytes of information in a 1000-tuple

$$(s_1, s_2, s_3, \cdots, s_{1000}) \tag{3.1.1}$$

where s_i is the i-th sample, and treat it just as we would a 3-component physical vec-

tor. That is, we can add any two such vectors together, scale them, and so on. When we "stack" seismic traces, we're just adding these n-dimensional vectors component by component

$$s + t = (s_1 + t_1, s_2 + t_2, s_3 + t_3, \cdots, s_{1000} + t_{1000}). \tag{3.1.2}$$

Now, the physical vectors have a life independent of the particular 3-tuple we use to represent them. We will get a different 3-tuple depending on whether we use Cartesian or spherical coordinates, for example, but the force vector itself is independent of these considerations. Whereas our use of vector spaces is purely abstract. There is no physical seismogram vector; all we have is the n-tuple.

In fact, the mathematical definition of a *vector space* is sufficiently general to incorporate objects that you might not consider as vectors at first glance–such as functions and matrices. The definition of such a space actually requires two sets of objects: a set of vectors V and a one of scalars F. For our purposes the scalars will always be either the real numbers $\mathbf{R}$ or the complex numbers $\mathcal{C}$.

Definition 1 Linear Vector Space *A linear vector space over a field F of scalars is a set of elements V together with a function called addition from $V \times V$ into V[1] and a function called scalar multiplication from $F \times V$ into V satisfying the following conditions for all $x, y, z \in V$ and all $\alpha, \beta \in F$:*

V1: $(x + y) + z = x + (y + z)$

V2: $x + y = y + x$

V3: There is an element 0 *in* V *such that* $x + 0 = x$ *for all* $x \in V$.

V4: For each $x \in V$ *there is an element* $-x \in V$ *such that* $x + (-x) = 0$.

V5: $\alpha(x + y) = \alpha x + \alpha y$

V6: $(\alpha + \beta)x = \alpha x + \beta x$

V7: $\alpha(\beta x) = (\alpha\beta)x$

V8: $1 \cdot x = x$

The simplest example of a vector space is $\mathbf{R}^n$, whose vectors are n-tuples of real numbers. Addition and scalar multiplication are defined component-wise:

$$(x_1, x_2, \cdots, x_n) + (y_1, y_2, \cdots, y_n) = (x_1 + y_1, x_2 + y_2, \cdots, x_n + y_n) \tag{3.1.3}$$

[1]The Cartesian product $A \times B$ of two sets A and B is the set of all ordered pairs (a, b) where $a \in A$ and $b \in B$.

and

$$\alpha(x_1, x_2, \cdots, x_n) = (\alpha x_1, \alpha x_2, \cdots, \alpha x_n). \tag{3.1.4}$$

In the case of $n = 1$ the vector space V and the field F are the same. So trivially, F is a vector space over F.

A few observations: first, by adding $-x$ to both sides of $x + y = x$, you can show that $x + y = x$ if and only if $y = 0$. This implies the uniqueness of the zero element and also that $\alpha \cdot 0 = 0$ for all scalars α.

Functions themselves are vectors according to this definition. Consider the space of functions mapping some nonempty set onto the scalars, with addition and multiplication defined by:

$$[f + g](t) = f(t) + g(t) \tag{3.1.5}$$

and

$$[\alpha f](t) = \alpha f(t). \tag{3.1.6}$$

We use the square brackets to separate the function from its arguments. In this case, the zero element is the function whose value is zero everywhere. And the minus element is inherited from the scalars: $[-f](t) = -f(t)$.

3.2 Matrices

The set of all $n \times m$ matrices with scalar entries is a linear vector space with addition and scalar multiplication defined component-wise. We denote this space by $\mathbf{R}^{n\times m}$. Two matrices have the same dimensions if they have the same number of rows and columns. We use upper case roman letters to denote matrices, lower case roman[2] to denote ordinary vectors and greek letters to denote scalars. For example, let

$$A = \begin{bmatrix} 2 & 5 \\ 3 & 8 \\ 1 & 0 \end{bmatrix}. \tag{3.2.1}$$

Then the components of A are denoted by A_{ij}. The *transpose* of a matrix, denoted by A^T, is achieved by exchanging the columns and rows. In this example

$$A^T = \begin{bmatrix} 2 & 3 & 1 \\ 5 & 8 & 0 \end{bmatrix}. \tag{3.2.2}$$

Thus $A_{21} = 3 = A^T_{12}$.

[2]For emphasis, and to avoid any possible confusion, we will henceforth also use bold type for ordinary vectors.

You can prove for yourself that

$$(AB)^T = B^T A^T. \tag{3.2.3}$$

A matrix which equals its transpose ($A^T = A$) is said to be symmetric. If $A^T = -A$ the matrix is said to be skew-symmetric. We can split any square matrix A into a sum of a symmetric and a skew-symmetric part via

$$A = \frac{1}{2}(A + A^T) + \frac{1}{2}(A - A^T). \tag{3.2.4}$$

The Hermitian transpose of a matrix is the complex conjugate of its transpose. Thus if

$$A = \begin{bmatrix} 4-i & 8 & 12+i \\ -12 & -8 & -4-i \end{bmatrix} \tag{3.2.5}$$

then

$$\bar{A^T} \equiv A^H = \begin{bmatrix} 4+i & -12 \\ 8 & -8 \\ 12-i & -4+i \end{bmatrix}. \tag{3.2.6}$$

Sometimes it is useful to have a special notation for the columns of a matrix. So if

$$A = \begin{bmatrix} 2 & 5 \\ 3 & 8 \\ 1 & 0 \end{bmatrix} \tag{3.2.7}$$

then we write

$$A = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 \end{bmatrix} \tag{3.2.8}$$

where

$$\mathbf{a}_1 = \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}. \tag{3.2.9}$$

Addition of two matrices A and B only makes sense if they have the same number of rows and columns. In which case we can add them component-wise

$$(A + B)_{ij} = [A_{ij} + B_{ij}]. \tag{3.2.10}$$

For example if

$$A = \begin{bmatrix} 1 & 2 & 3 \\ -3 & -2 & -1 \end{bmatrix} \tag{3.2.11}$$

and

$$B = \begin{bmatrix} 0 & 6 & 2 \\ 1 & 1 & 1 \end{bmatrix} \tag{3.2.12}$$

Then

$$A + B = \begin{bmatrix} 1 & 8 & 5 \\ -2 & -1 & 0 \end{bmatrix}. \tag{3.2.13}$$

Scalar multiplication, once again, is done component-wise. If

$$A = \begin{bmatrix} 1 & 2 & 3 \\ -3 & -2 & -1 \end{bmatrix} \tag{3.2.14}$$

and $\alpha = 4$ then

$$\alpha A = \begin{bmatrix} 4 & 8 & 12 \\ -12 & -8 & -4 \end{bmatrix}. \tag{3.2.15}$$

So both matrices and vectors can be thought of as vectors in the abstract sense. Matrices can also be thought of as operators acting on vectors in $\mathbf{R}^n$ via the matrix-vector inner (or "dot") product. If $A \in \mathbf{R}^{n\times m}$ and $\mathbf{x} \in \mathbf{R}^m$, then $A \cdot \mathbf{x} = \mathbf{y} \in \mathbf{R}^n$ is defined by

$$y_i = \sum_{j=1}^{m} A_{ij} x_j. \tag{3.2.16}$$

This is an algebraic definition of the inner product. We can also think of it geometrically. Namely, the inner product is a linear combination of the columns of the matrix. For example,

$$A \cdot \mathbf{x} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_1 \begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \end{bmatrix} + x_2 \begin{bmatrix} a_{12} \\ a_{22} \\ a_{32} \end{bmatrix}. \tag{3.2.17}$$

A special case of this occurs when A is just an ordinary vector. We can think of this as $A \in \mathbf{R}^{n\times m}$ with $n = 1$. Then $\mathbf{y} \in \mathbf{R}^1$ is just a scalar. A vector $\mathbf{z}$ in $\mathbf{R}^{1\times m}$ looks like

$$(z_1, z_2, z_3, \cdots, z_m) \tag{3.2.18}$$

so the inner product of two vectors $\mathbf{z}$ and $\mathbf{x}$ is just

$$[z_1, z_2, z_3, \cdots, z_n] \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = [z_1 x_1 + z_2 x_2 + z_3 x_3 + \cdots + z_n x_n]. \tag{3.2.19}$$

By default, a vector $\mathbf{x}$ is regarded as a column vector. So this vector-vector inner product is also written as $\mathbf{z}^T\mathbf{x}$ or as $(\mathbf{z}, \mathbf{x})$. Similarly if $A \in \mathbf{R}^{n\times m}$ and $B \in \mathbf{R}^{m\times p}$, then the matrix-matrix AB product is defined to be a matrix in $\mathbf{R}^{n\times p}$ with components

$$(AB)_{ij} = \sum_{k=1}^{m} a_{ik} b_{kj}. \tag{3.2.20}$$

For example,

$$AB = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 2 & 3 \end{bmatrix} = \begin{bmatrix} 4 & 7 \\ 8 & 15 \end{bmatrix}. \tag{3.2.21}$$

On the other hand, note well that

$$BA = \begin{bmatrix} 0 & 1 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 3 & 4 \\ 11 & 16 \end{bmatrix} \neq AB. \tag{3.2.22}$$

This definition of matrix-matrix product even extends to the case in which both matrices are vectors. If $\mathbf{x} \in \mathbf{R}^m$ and $\mathbf{y} \in \mathbf{R}^n$, then $\mathbf{xy}$ (called the "outer" product and usually written as $\mathbf{xy}^T$) is

$$(xy)_{ij} = x_i y_j. \tag{3.2.23}$$

So if

$$\mathbf{x} = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \tag{3.2.24}$$

and

$$\mathbf{y} = \begin{bmatrix} 1 \\ 3 \\ 0 \end{bmatrix} \tag{3.2.25}$$

then

$$\mathbf{xy}^T = \begin{bmatrix} -1 & -3 & 0 \\ 1 & 3 & 0 \end{bmatrix}. \tag{3.2.26}$$

Here is a brief summary of the notation for inner products:

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^T \mathbf{y} = (\mathbf{x}, \mathbf{y}) = \sum_i x_i y_i = x_i y_i \text{ summation convention}$$

3.3 Some Special Matrices

The identity element in the space of square $n \times n$ matrices is a matrix with ones on the main diagonal and zeros everywhere else

$$I_n = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots \\ 0 & 1 & 0 & 0 & \dots \\ 0 & 0 & 1 & 0 & \dots \\ \vdots & & & \ddots & \\ 0 & \dots & 0 & 0 & 1 \end{bmatrix}. \tag{3.3.1}$$

As an exercise, show that $AI_n = I_n A = A$ for any $n \times x$ matrix A.

Even if the matrix is not square, there is still a *main diagonal* of elements given by A_{ii} where i runs from 1 to the smaller of the number of rows and columns. We can take any vector in $\mathbf{R}^n$ and make a diagonal matrix out of it just by putting it on the main diagonal and filling in the rest of the elements of the matrix with zeros. There is a special notation for this:

$$\mathrm{diag}(x_1, x_2, \cdots, x_n) = \begin{bmatrix} x_1 & 0 & 0 & 0 & \dots \\ 0 & x_2 & 0 & 0 & \dots \\ 0 & 0 & x_3 & 0 & \dots \\ \vdots & & & \ddots & \\ 0 & \dots & 0 & 0 & x_n \end{bmatrix}. \tag{3.3.2}$$

A matrix $Q \in \mathbf{R}^{n \times n}$ is said to be orthogonal if $Q^T Q = I_n$. In this case, each column of Q is an orthonormal vector: $\mathbf{q}_i \cdot \mathbf{q}_i = 1$. So why are these matrices called orthogonal? No good reason. As an example

$$Q = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}. \tag{3.3.3}$$

Now convince yourself that $Q^T Q = I_n$ implies that $QQ^T = I_n$ as well. In which case the rows of Q must be orthonormal vectors too.

Another interpretation of the matrix-vector inner product is as a mapping from one vector space to another. Suppose $A \in \mathbf{R}^{n \times m}$, then A maps vectors in $\mathbf{R}^m$ into vectors in $\mathbf{R}^n$. An orthogonal matrix has an especially nice geometrical interpretation. To see this first notice that for any matrix A, the inner product $(A \cdot \mathbf{x}) \cdot \mathbf{y}$, which we write as $(A\mathbf{x}, \mathbf{y})$, is equal to $(\mathbf{x}, A^T \mathbf{y})$, as you can readily verify. Similarly

$$(A^T \mathbf{x}, \mathbf{y}) = (\mathbf{x}, A\mathbf{y}). \tag{3.3.4}$$

As a result, for an orthogonal matrix Q

$$(Q\mathbf{x}, Q\mathbf{x}) = (Q^T Q\mathbf{x}, \mathbf{x}) = (\mathbf{x}, \mathbf{x}). \tag{3.3.5}$$

Now, as you already know, and we will discuss shortly, the inner product of a vector with itself is related to the length, or norm, of that vector. Therefore an orthogonal matrix maps a vector into another vector of the same norm. In other words it does a rotation.

3.4 Matrix and Vector Norms

We need some way of comparing the relative "size" of vectors and matrices. For scalars, the obvious answer is the absolute value. The absolute value of a scalar has the property

that it is never negative and it is zero if and only if the scalar itself is zero. For vectors and matrices both we can define a generalization of this concept of length called a *norm*. A norm is a function from the space of vectors onto the scalars, denoted by $\|\cdot\|$ satisfying the following properties for any two vectors v and u and any scalar α:

Definition 2 *Norms*

N1: $\|v\| > 0$ *for any* $v \neq 0$ *and* $\|v\| = 0 \Leftrightarrow v = 0$

N2: $\|\alpha v\| = |\alpha|\|v\|$

N3: $\|v + u\| \leq \|v\| + \|u\|$

Property $N3$ is called the *triangle inequality*.

The most useful class of norms for vectors in $\mathbf{R}^n$ is the ℓ_p norm defined for $p \geq 1$ by

$$\|\mathbf{x}\|_{\ell_p} = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}. \tag{3.4.1}$$

For $p = 2$ this is just the ordinary Euclidean norm: $\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^T\mathbf{x}}$. A finite limit of the ℓ_p norm exists as $p \to \infty$ called the ℓ_∞ norm:

$$\|x\|_{\ell_\infty} = \max_{1 \leq i \leq n} |x_i| \tag{3.4.2}$$

We won't need matrix norms in this class, but in case you're interested any norm on vectors in $\mathbf{R}^n$ induces a norm on matrices via

$$\|A\| = \max_{\mathbf{x} \neq 0} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}. \tag{3.4.3}$$

E.g., Let $\mathbf{x} = (1, 1)$, then $\|\mathbf{x}\| = \sqrt{1 \cdot 1 + 1 \cdot 1} = \sqrt{2}$.

3.5 Projecting Vectors Onto Other Vectors

Figure 3.1 illustrates the basic idea of projecting one vector onto another. We can always represent one, say $\mathbf{b}$, in terms of its components parallel and perpendicular to the other. The length of the component of $\mathbf{b}$ along $\mathbf{a}$ is $\|\mathbf{b}\| \cos\theta$ which is also $\mathbf{b}^T\mathbf{a}/\|\mathbf{a}\|$

Now suppose we want to construct a vector in the direction of $\mathbf{a}$ but whose length is the component of $\mathbf{b}$ along $\|\mathbf{b}\|$. We did this, in effect, when we computed the tangential force

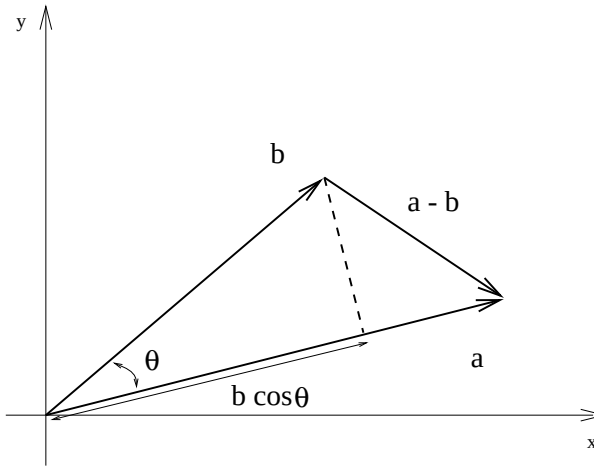


Figure 3.1: Let $\mathbf{a}$ and $\mathbf{b}$ be any two vectors. We can always represent one, say $\mathbf{b}$, in terms of its components parallel and perpendicular to the other. The length of the component of $\mathbf{b}$ along $\mathbf{a}$ is $\|\mathbf{b}\|\cos\theta$ which is also $\mathbf{b}^T\mathbf{a}/\|\mathbf{a}\|$.

of gravity on a simple pendulum. What we need to do is multiply $\|\mathbf{b}\|\cos\theta$ by a unit vector in the $\mathbf{a}$ direction. Obviously a convenient unit vector in the $\mathbf{a}$ direction is $\mathbf{a}/\|\mathbf{a}\|$, which equals

$$\frac{\mathbf{a}}{\sqrt{\mathbf{a}^T\mathbf{a}}}.$$

So a vector in the $\mathbf{a}$ with length $\|\mathbf{b}\|\cos\theta$ is given by

$$\|\mathbf{b}\|\cos\theta\hat{\mathbf{a}} = \frac{\mathbf{a}^T\mathbf{b}}{\|\mathbf{a}\|}\frac{\mathbf{a}}{\|\mathbf{a}\|} \tag{3.5.1}$$

$$= \frac{\mathbf{a}}{\|\mathbf{a}\|}\frac{\mathbf{a}^T\mathbf{b}}{\|\mathbf{a}\|} = \frac{\mathbf{a}\mathbf{a}^T\mathbf{b}}{\mathbf{a}^T\mathbf{a}} = \frac{\mathbf{a}\mathbf{a}^T}{\mathbf{a}^T\mathbf{a}}\mathbf{b} \tag{3.5.2}$$

As an exercise verify that in general $\mathbf{a}(\mathbf{a}^T\mathbf{b}) = (\mathbf{a}\mathbf{a}^T)\mathbf{b}$. This is not completely obvious since in one expression there is an inner product in the parenthesis and in the other there is an outer product.

What we've managed to show is that the projection of the vector $\mathbf{b}$ into the direction of $\mathbf{a}$ can be achieved with the following matrix (operator)

$$\frac{\mathbf{a}\mathbf{a}^T}{\mathbf{a}^T\mathbf{a}}.$$

This is our first example of a projection operator.

3.6 Linear Dependence and Independence

Suppose we have n nonzero vectors

$$\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\} \tag{3.6.1}$$

of the same dimension. The question is, under what circumstances can the linear combination of these vectors be zero:

$$\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \cdots \alpha_n\mathbf{x}_n = 0. \tag{3.6.2}$$

If this is true with at least one of the coefficients α_i nonzero, then we could isolate a particular vector on the right hand side, expressing it as a linear combination of the other vectors. In this case the original set of n vectors are said to be *linearly dependent*. On the other hand, if the only way for this sum of vectors to be zero is for all the coefficients themselves to be zero, then we say that the vectors are *linearly independent*.

Now, this linear combination of vectors can also be written as a matrix-vector inner product. With $\mathbf{a} = (\alpha_1, \alpha_2, \cdots, \alpha_n)$, and $X = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$ we have the condition for linear dependence being

$$X\mathbf{a} = 0 \tag{3.6.3}$$

for some nonzero vector $\mathbf{a}$, and the condition for linear independence being

$$X\mathbf{a} = 0 \Rightarrow \mathbf{a} = 0. \tag{3.6.4}$$

As a result, if we are faced with a linear system of equations to solve

$$A\mathbf{x} = \mathbf{b} \tag{3.6.5}$$

we can think in two different ways. On the one hand, we can investigate the equation in terms of the *existence* of a vector $\mathbf{x}$ satisfying the equation. On the other hand, we can think in terms of the *compatibility* of the right hand side with the columns of the matrix.

Linear independence is also central to the notion of how big a vector space is–its *dimension*. It's intuitively clear that no two linearly independent vectors are adequate to represent an arbitrary vector in $\mathbf{R}^3$. For example, $(1, 0, 0)$ and $(0, 1, 0)$ are linearly independent, but there are no scalar coefficients that will let us write $(1, 1, 1)$ as a linear combination of the first two. Conversely, since any vector in $\mathbf{R}^3$ can be written as a combination of the three vectors $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$, it is impossible to have more than three linearly independent vectors in $\mathbf{R}^3$. So the dimension of a space is the number of linearly independent vectors required to represent an arbitrary element.

3.7 The Four Fundamental Spaces

Now if we take two basis vectors $\mathbf{R}^2$ $(1,0)$ and $(0,1)$,[3] and consider all possible linear combinations of them–this is called the *span* of the two vectors–we will incorporate all the elements in $\mathbf{R}^2$. On the other hand, if we consider these two vectors as being in $\mathbf{R}^3$, so that we write them as $(1,0,0)$ and $(0,1,0)$, then their span clearly doesn't fill up all of $\mathbf{R}^3$. It does, however, fill up a subspace of $\mathbf{R}^3$, the $x-y$ plane. The technical definition of a subspace is that it is a subset *closed* under addition and scalar multiplication:

Definition 3 *Subspaces A subspace of a vector space is a nonempty subset S that satisfies*

S1: The sum of any two elements from S is in S, and

S2: The scalar multiple of any element from S is in S.

If we take a general matrix $A \in \mathbf{R}^{n\times m}$, then the span of the columns must be a subspace of $\mathbf{R}^n$. Whether this subspace amounts to the whole of $\mathbf{R}^n$ obviously depends on whether the columns are linearly independent or not. This subspace is called the *column space* of the matrix and is usually denoted by $R(A)$, for "range". The dimension of the column space is called the *rank* of the matrix.

Another fundamental subspace associated with any matrix A is composed by the solutions of the homogeneous equation $A\mathbf{x} = 0$. Why is this a subspace? Well, take any two such solutions, say $\mathbf{x}$ and $\mathbf{y}$ and we have

$$A(\mathbf{x}+\mathbf{y}) = A\mathbf{x} + A\mathbf{y} = 0. \tag{3.7.1}$$

Similarly,

$$A(\alpha\mathbf{x}) = \alpha A\mathbf{x}. \tag{3.7.2}$$

This subspace is called the *nullspace* or *kernel* and is extremely important from the point of view of inverse theory. As we shall see, in an inverse calculation the right hand side of a matrix equations is usually associated with perturbations to the data. Vectors in the nullspace have no effect on the data and are therefore unresolved in an experiment. Figuring out what features of a model are unresolved is a major goal of inversion.

3.7.1 Spaces associated with a linear system $A\mathbf{x} = \mathbf{y}$

The span of the columns is a subset of $\mathbf{R}^n$ and the span of the rows is a subset of $\mathbf{R}^m$. In other words the rows of A have m components while the columns of A have n components. This is easiest to visual if you keep in mind a picture of a generic n by m matrix (Figure 3.2).

[3] Any other pair of linearly independent vectors, such as $(2,0)$ and $(1,15)$ would also work.

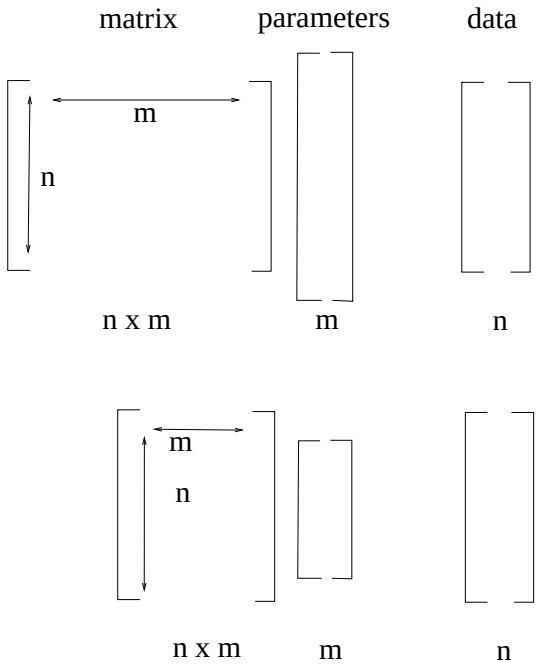


Figure 3.2: A generic $n \times m$ matrix can have more columns than rows (top), more rows than columns (bottom), or it could be square.

Now the column space and the nullspace are generated by A. What about the column space and the null space of A^T? These are, respectively, the row space and the left nullspace of A. The nullspace and row space are subspaces of $\mathbf{R}^m$, while the column space and the left nullspace are subspaces of $\mathbf{R}^n$.

Here is probably the most important result in linear algebra: For any matrix whatsoever, the number of linearly independent rows equals the number of linearly independent columns. We summarize this by saying that **row rank = column rank**. For a generic $n \times m$ matrix, this is not an obvious result. We can summarize these spaces as follows:

Theorem 1 Fundamental Theorem of Linear Algebra *Let $A \in \mathbf{R}^{n \times m}$. Then*

1: Dimension of column space equals r, the rank.

2: Dimension of nullspace equals $m - r$.

3: Dimension of row space equals r.

4: Dimension of left nullspace equals $n - r$.

3.7.2 A Geometrical Picture

Any vector in the null space of a matrix, must be orthogonal to all the rows (since each component of the matrix dotted into the vector is zero). Therefore all the elements in the null space are orthogonal to all the elements in the row space. In mathematical terminology, the null space and the row space are *orthogonal complements* of one another. Or, to say the same thing, they are *orthogonal subspaces* of $\mathbf{R}^m$. Similarly, vectors in the left null space of a matrix are orthogonal to all the columns of this matrix. This means that the left null space of a matrix is the orthogonal complement of the column space; they are orthogonal subspaces of $\mathbf{R}^n$.

3.8 Matrix Inverses

A *left inverse* of a matrix $A \in \mathbf{R}^{n\times m}$ is defined to be a matrix B such that

$$BA = I. \tag{3.8.1}$$

A *right inverse* C therefore must satisfy

$$AC = I. \tag{3.8.2}$$

If there exists a left and a right inverse of A then they must be equal since matrix multiplication is associative:

$$AC = I \Rightarrow B(AC) = B \Rightarrow (BA)C = B \Rightarrow C = B. \tag{3.8.3}$$

Now if we have more equations than unknowns then the columns cannot possibly span all of $\mathbf{R}^n$. Certainly the rank r must be less than or equal to n, but it can only equal n if we have at least as many unknowns as equations. The basic existence result is then:

Theorem 2 Existence of solutions to $A\mathbf{x} = \mathbf{y}$ *The system $A\mathbf{x} = \mathbf{y}$ has at least one solution $\mathbf{x}$ for every $\mathbf{y}$ (there might be infinitely many solutions) if and only if the columns span $\mathbf{R}^n$ ($r = n$), in which case there exists an $m \times n$ right inverse C such that $AC = I_n$. This is only possible if $n \leq m$.*

Don't be mislead by the picture above into neglecting the important special case when $m = n$. The point is that the basic issues of existence and, next, uniqueness, depend on whether there are more or fewer rows than equations. The statement of uniqueness is:

Theorem 3 Uniqueness of solutions to $A\mathbf{x} = \mathbf{y}$ *There is at most one solution to $A\mathbf{x} = \mathbf{y}$ (there might be none) if and only if the columns of A are linearly independent ($r = m$), in which case there exists an $m \times n$ left inverse B such that $BA = I_m$. This is only possible if $n \geq m$.*

Clearly then, in order to have both existence and uniqueness, we must have that $r = m = n$. This precludes having existence and uniqueness for rectangular matrices. For square matrices $m = n$, so **existence implies uniqueness and uniqueness implies existence.**

Using the left and right inverses we can find solutions to $A\mathbf{x} = \mathbf{y}$: if they exist. For example, given a right inverse A, then since $AC = I$, we have $AC\mathbf{y} = \mathbf{y}$. But since $A\mathbf{x} = \mathbf{y}$ it follows that $\mathbf{x} = C\mathbf{y}$. But C is not necessarily unique. On the other hand, if there exists a left inverse $BA = I$, then $BA\mathbf{x} = B\mathbf{y}$, which implies that $\mathbf{x} = B\mathbf{y}$.

Some examples. Consider first the case of more equations than unknowns. Let

$$A = \begin{bmatrix} -1 & 0 \\ 0 & 3 \\ 0 & 0 \end{bmatrix} \tag{3.8.4}$$

Since the columns are linearly independent and there are more rows than columns, there can be at most one solution. You can readily verify that any matrix of the form

$$\begin{bmatrix} -1 & 0 & \gamma \\ 0 & 1/3 & \iota \end{bmatrix} \tag{3.8.5}$$

is a left inverse. The particular left inverse given by the formula $(A^TA)^{-1}A^T$ (cf. the exercise at the end of this chapter) is the one for which γ and ι are zero. But there are infinitely many other left inverses. As for solutions of $A\mathbf{x} = \mathbf{y}$, if we take the inner product of A with the vector $(x_1, x_2)^T$ we get

$$\begin{bmatrix} -x_1 \\ 3x_2 \\ 0 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \tag{3.8.6}$$

So, clearly, we must have $x_1 = -y_1$ and $x_2 = 1/3y_2$. But, there will not be any solution unless $y_3 = 0$.

Next, let's consider the case of more columns (unknowns) than rows (equations). Let

$$A = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 3 & 0 \end{bmatrix} \tag{3.8.7}$$

Here you can readily verify that any matrix of the form

$$\begin{bmatrix} -1 & 0 \\ 0 & 1/3 \\ \gamma & \iota \end{bmatrix} \tag{3.8.8}$$

is a right inverse. The particular right inverse (shown in the exercise at the end of this chapter) $A^T(AA^T)^{-1}$ corresponds to $\gamma = \iota = 0$.

Now if we look at solutions of the linear system $A\mathbf{x} = \mathbf{y}$ with $\mathbf{x} \in \mathbf{R}^3$ and $\mathbf{y} \in \mathbf{R}^2$ we find that $x_1 = -y_1$, $x_2 = 1/3y_2$, and that x_3 is completely undetermined. So there is an infinite set of solutions corresponding to the different values of x_3.

3.9 Elementary operations and Gaussian Elimination

I am assuming that you've seen this before, so this is a very terse review. If not, see the book by Strang in the bibliography.

Elementary matrix operations consist of:

- Interchanging two rows (or columns)
- Multiplying a row (or column) by a nonzero constant
- Adding a multiple of one row (or column) to another row (or column)

If you have a matrix that can be derived from another matrix by a sequence of elementary operations, then the two matrices are said to be row or column equivalent. For example

$$A = \begin{pmatrix} 1 & 2 & 4 & 3 \\ 2 & 1 & 3 & 2 \\ 1 & -1 & 2 & 3 \end{pmatrix}$$

is row equivalent to

$$B = \begin{pmatrix} 2 & 4 & 8 & 6 \\ 1 & -1 & 2 & 3 \\ 4 & -1 & 7 & 8 \end{pmatrix}$$

because we can add 2 times row 3 of A to row 2 of A; then interchange rows 2 and 3; finally multiply row 1 by 2.

Gaussian elimination consists of two phases. The first is the application of elementary operations to try to put the matrix in row-reduced form; i.e., making zero all the elements below the main diagonal (and normalizing the diagonal elements to 1). The second phase is back-substitution. Unless the matrix is very simple, calculating any of the four fundamental subspaces is probably easiest if you put the matrix in row-reduced form first.

3.9.1 Examples

1. Find the row-reduced form and the null-space of

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$$

Answer A row-reduced form of the matrix is

$$\begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 2 \end{pmatrix}$$

Now, some people reserve the term row-reduced (or row-reduced echelon) form for the matrix that also has zeros above the ones. We can get this form in one more step:

$$\begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \end{pmatrix}$$

The null space of A can be obtained by solving the system

$$\begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

So we must have $x_1 = x_3$ and $x_2 = -2x_3$. So the null space is is the line spanned by

$$(1, -2, 1)$$

.

2. Solve the linear system $A\mathbf{x} = \mathbf{y}$ with $\mathbf{y} = (1, 1)$:

   Answer

   Any vector of the form $(z - 1, 1 - 2z, z)$ will do. For instance, $(-1, 1, 0)$.

3. Solve the linear system $A\mathbf{x} = \mathbf{y}$ with $\mathbf{y} = (0, -1)$:

   Answer One example is

   $$\left(-\frac{2}{3}, \frac{1}{3}, 0\right)$$

4. Find the row-reduced form and the null space of the matrix

   $$B = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$$

   Answer The row-reduced matrix is

   $$\begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \\ 0 & 0 & 0 \end{pmatrix}$$

   The null space is spanned by

   $$(1, -2, 1)$$

   .

5. Find the row-reduced form and the null space of the matrix

$$C = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 1 & 0 & 1 \end{pmatrix}$$

Answer The row-reduced matrix is

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

The only element in the null space is the zero vector.

6. Find the null space of the matrix

$$D = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 2 \end{pmatrix}$$

Answer You can solve the linear system $D\mathbf{x} = \mathbf{y}$ with $\mathbf{y} = (0, 0, 0)$ and discover that $x_1 = -2x_3 = -2x_2$. This means that the null space is spanned $(-2, 1, 1)$. The row-reduced form of the matrix is

$$\begin{pmatrix} 1 & 0 & 2 \\ 0 & 1 & -1 \end{pmatrix}$$

7. Are the following vectors in R^3 linearly independent or dependent? If they are dependent express one as a linear combination of the others.

$$\left\{ \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 3 \\ 6 \\ 6 \end{pmatrix} \right\}$$

Answer The vectors are obviously dependent since you cannot have four linearly independent vectors in a three dimensional space. If you put the matrix in row-reduced form you will get

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

The first three vectors are indeed linearly independent. Note that the determinant of

$$\begin{pmatrix} 1 & 0 & 1 \\ 1 & 2 & 2 \\ 0 & 3 & 3 \end{pmatrix}$$

is equal to 3.

To find the desired linear combination we need to solve:

$$x\begin{pmatrix}1\\1\\0\end{pmatrix}+y\begin{pmatrix}0\\2\\3\end{pmatrix}+z\begin{pmatrix}1\\2\\3\end{pmatrix}=\begin{pmatrix}3\\6\\6\end{pmatrix}$$

or

$$\begin{pmatrix}1&0&1\\1&2&2\\0&3&3\end{pmatrix}\begin{pmatrix}x\\y\\z\end{pmatrix}=\begin{pmatrix}3\\6\\6\end{pmatrix}$$

Gaussian elimination could proceed as follows (the sequence of steps is not unique of course): first divide the third row by 3

$$\begin{matrix}1&0&1&3\\1&2&2&6\\0&1&1&2\end{matrix}$$

$$\begin{matrix}1&0&1&3\\0&2&1&3\\0&1&1&2\end{matrix}$$

$$\begin{matrix}1&0&1&3\\0&0&-1&-1\\0&1&1&2\end{matrix}$$

$$\begin{matrix}1&0&1&3\\0&0&1&1\\0&1&0&1\end{matrix}$$

$$\begin{matrix}1&0&1&3\\0&1&0&1\\0&0&1&1\end{matrix}$$

Thus we have $z = y = 1$ and $x + z = 3$, which implies that $x = 2$. So, the solution is $(2, 1, 1)$ and you can verify that

$$2\begin{pmatrix}1\\1\\0\end{pmatrix}+1\begin{pmatrix}0\\2\\3\end{pmatrix}+1\begin{pmatrix}1\\2\\3\end{pmatrix}=\begin{pmatrix}3\\6\\6\end{pmatrix}$$

3.10 Least Squares

In this section we will consider the problem of solving $A\mathbf{x} = \mathbf{y}$ when no solution exists! I.e., we consider what happens when there is no vector that satisfies the equations exactly. This sort of situation occurs all the time in science and engineering. Often we

make repeated measurements which, because of noise, for example, are not exactly consistent. Suppose we make n measurements of some quantity x. Let x_i denote the i-th measurement. You can think of this as n equations with 1 unknown:

$$\begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix}$$

Obviously unless all the x_i are the same, there cannot be a value of x which satisfies all the equations simultaneously. Being practical people we could, at least for this simple problem, ignore all the linear algebra and simply assert that we want to find the value of x which minimizes the sum of squared errors:

$$\min_x \sum_{i=1}^{n} (x - x_i)^2 .$$

Differentiating this equation with respect to x and setting the result equal to zero gives:

$$x_{\text{ls}} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

where we have used x_{ls} to denote the *least squares* value of x. In other words the value of x that minimizes the sum of squares of the errors is just the mean of the data.

In more complicated situations (with n equations and m unknowns) it's not quite so obvious how to proceed. Let's return to the basic problem of solving $A\mathbf{x} = \mathbf{y}$. If $\mathbf{y}$ were in the column space of A, then there would exist a vector $\mathbf{x}$ such that $A\mathbf{x} = \mathbf{y}$. On the other hand, if $\mathbf{y}$ is not in the column space of A a reasonable strategy is to try to find an approximate solution from within the column space. In other words, find a linear combination of the columns of A that is as close as possible in a least squares sense to the data. Let's call this approximate solution $\mathbf{x_{ls}}$. Since $A\mathbf{x_{ls}}$ is, by definition, confined to the column space of A then $A\mathbf{x_{ls}} - \mathbf{y}$ (the error in fitting the data) must be in the orthogonal complement of the column space. The orthogonal complement of the column space is the left null space, so $A\mathbf{x_{ls}} - \mathbf{y}$ must get mapped into zero by A^T:

$$A^T \left(A\mathbf{x_{ls}} - \mathbf{y} \right) = 0$$

or

$$A^T A\mathbf{x_{ls}} = A^T \mathbf{y}$$

These are called the *normal equations*. Now we saw in the last chapter that the outer product of a vector or matrix with itself defined a projection operator onto the subspace spanned by the vector (or columns of the matrix). If we look again at the normal

equations and assume for the moment that the matrix A^TA is invertible, then the least squares solution is:

$$\mathbf{x_{ls}} = (A^TA)^{-1}A^T\mathbf{y}$$

The matrix $(A^TA)^{-1}A^T$ is an example of what is called a *generalized inverse* of A. In the even that A is not invertible in the usual sense, this provides a reasonable generalization (not the only one) of the ordinary inverse.

Now A applied to the least squares solution is the approximation to the data from within the column space. So $A\mathbf{x_{ls}}$ is precisely the projection of the data $\mathbf{y}$ onto the column space:

$$A\mathbf{x_{ls}} = A(A^TA)^{-1}A^T\mathbf{y}.$$

Before when we did orthogonal projections, the projecting vectors/matrices were orthogonal, so A^TA term would have been the identity, but the outer product structure in $A\mathbf{x_{ls}}$ is evident.

The generalized inverse projects the data onto the column space of A.

A few observations:

- When A is invertible (square, full rank) $A(A^TA)^{-1}A^T = AA^{-1}(A^T)^{-1}A^T = I$, so every vector projects to itself.
- A^TA has the same null space as A. Proof: clearly if $A\mathbf{x} = 0$, then $A^TA\mathbf{x} = 0$. Going the other way, suppose $A^TA\mathbf{x} = 0$. Then $\mathbf{x}^TA^TA\mathbf{x} = 0$. But this can also be written as $(A\mathbf{x}, A\mathbf{x}) = \|A\mathbf{x}\|^2 = 0$. By the properties of the norm, $\|A\mathbf{x}\|^2 = 0 \Rightarrow A\mathbf{x} = 0$.
- As a corollary of this, if A has linearly independent columns (i.e., the rank $r = m$) then A^TA is invertible.

Finally, it's not too difficult to show that the normal equations can also be derived by directly minimizing the following function:

$$\|A\mathbf{x} - \mathbf{y}\|^2 = (A\mathbf{x} - \mathbf{y}, A\mathbf{x} - \mathbf{y}).$$

This is just the sum of the squared errors, but for n simultaneous equations in m unknowns. You can either write this vector function out explicitly in terms of its components and use ordinary calculus, or you can actually differentiate the expression with respect to the vector $\mathbf{x}$ and set the result equal to zero. So for instance, since

$$(A\mathbf{x}, A\mathbf{x}) = (A^TA\mathbf{x}, \mathbf{x}) = (\mathbf{x}, A^TA\mathbf{x})$$

differentiating $(A\mathbf{x}, A\mathbf{x})$ with respect to $\mathbf{x}$ yields $2A^TAx$, one factor coming from each factor of $\mathbf{x}$. The details will be left as an exercise.

3.10.1 Examples of Least Squares

Let us return to the problem we started above:

$$\begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix}$$

Ignoring linear algebra and just going for a least squares value of the parameter x we came up with:

$$x_{\text{ls}} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Let's make sure we get the same thing using the generalized inverse approach. Now, $A^T A$ is just

$$(1, 1, 1, ..., 1) \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = n.$$

So the generalized inverse of A is

$$(A^T A)^{-1} A^T = \frac{1}{n} (1, 1, 1, ..., 1).$$

Hence the generalized inverse solution is:

$$\frac{1}{n} (1, 1, 1, ..., 1) \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

as we knew already.

Consider a more interesting example

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}$$

Thus $x + y = \alpha$, $y = \beta$ and $2y = \gamma$. So, for example, if $\alpha = 1$, and $\beta = \gamma = 0$, then $x = 1$, $y = 0$ is a solution. In that case the right hand side is in the column space of A.

But now suppose the right hand side is $\alpha = \beta = 0$ and $\gamma = 1$. It is not hard to see that the column vector $(0, 1, 1)^T$ is not in the column space of A. (Show this as an exercise.) So what do we do? We solve the normal equations. Here are the steps. We want to solve (in the least squares sense) the following system:

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

So first compute

$$A^T A = \begin{pmatrix} 1 & 1 \\ 1 & 6 \end{pmatrix}.$$

The inverse of this matrix is

$$\left(A^T A\right)^{-1} = \frac{1}{5} \begin{pmatrix} 6 & -1 \\ -1 & 1 \end{pmatrix}.$$

So the generalized inverse solution (i.e., the least squares solution) is

$$\mathbf{x_{ls}} = \begin{pmatrix} 1 & -1/5 & -2/5 \\ 1 & 1/5 & 2/5 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} -2/5 \\ 2/5 \end{pmatrix}.$$

The interpretation of this solution is that it satisfies the first equation exactly (since $x + y = 0$) and it does an *average* job of satisfying the second and third equations. Least squares tends to average inconsistent information.

3.11 Eigenvalues and Eigenvectors

Recall that in Chapter 1 we showed that the equations of motion for two coupled masses are

$$m_1 \ddot{x}_1 = -k_1 x_1 - k_2(x_1 - x_2).$$

$$m_2 \ddot{x}_2 = -k_3 x_2 - k_2(x_2 - x_1).$$

or, restricting ourselves to the case in which $m_1 = m_2 = m$ and $k_1 = k_2 = k$

$$\begin{aligned} \ddot{x}_1 &= -\frac{k}{m} x_1 - \frac{k}{m}(x_1 - x_2) \\ &= -\omega_0^2 x_1 - \omega_0^2 (x_1 - x_2) \\ &= -2\omega_0^2 x_1 + \omega_0^2 x_2. \end{aligned} \tag{3.11.1}$$

and

$$\begin{aligned} \ddot{x}_2 &= -\frac{k}{m}x_2 - \frac{k}{m}(x_2 - x_1) \\ &= -\omega_0^2 x_2 - \omega_0^2(x_2 - x_1) \\ &= -2\omega_0^2 x_2 + \omega_0^2 x_1. \end{aligned} \tag{3.11.2}$$

If we look for the usual suspect solutions

$$x_1 = Ae^{i\omega t} \tag{3.11.3}$$

$$x_2 = Be^{i\omega t} \tag{3.11.4}$$

we see that the relationship between the displacement amplitudes A and B and ω can be written as the following matrix equation:

$$\begin{pmatrix} 2\omega_0^2 & -\omega_0^2 \\ -\omega_0^2 & 2\omega_0^2 \end{pmatrix} \begin{pmatrix} A \\ B \end{pmatrix} = \omega^2 \begin{pmatrix} A \\ B \end{pmatrix}. \tag{3.11.5}$$

This equation has the form of a matrix times a vector is equal to a scalar times the same vector:

$$K\mathbf{u} = \omega^2 \mathbf{u}. \tag{3.11.6}$$

In other words, the action of the matrix is to map the vector $((A, B)^T)$ into a scalar multiple of itself. This is a very special thing for a matrix to do.

Without using any linear algebra we showed way back on page 22 that the solutions of the equations of motion had two characteristic frequencies ($\omega = \omega_0$ and $\omega = \sqrt{3}\omega_0$), while the vector $(A, B)^T$ was either $(1, 1)^T$ for the slow mode ($\omega = \omega_0$) or $(1, -1)^T$ for the fast mode ($\omega = \sqrt{3}\omega_0$). You can quickly verify that these two sets of vectors/frequencies do indeed satisfy the matrix equation 3.11.5.

Now we will look at equations of the general form of 3.11.6 more systematically. We will see that finding the eigenvectors of a matrix gives us fundamental information about the system which the matrix models. Usually when a matrix operates on a vector, it changes the direction of the vector as well as its length. But for a special class of vectors, *eigenvectors*, the action of the matrix is to simply scale the vector:

$$A\mathbf{x} = \lambda \mathbf{x}. \tag{3.11.7}$$

If this is true, then $\mathbf{x}$ is an eigenvector of the matrix A associated with the eigenvalue λ. Now, $\lambda\mathbf{x}$ equals $\lambda I\mathbf{x}$ so we can rearrange this equation and write

$$(A - \lambda I)\mathbf{x} = 0. \tag{3.11.8}$$

Clearly in order that $\mathbf{x}$ be an eigenvector we must choose λ so that $(A - \lambda I)$ has a nullspace and we must choose $\mathbf{x}$ so that it lies in that nullspace. That means we must

choose λ so that $\mathrm{Det}(A - \lambda I) = 0$. This determinant is a polynomial in λ, called the characteristic polynomial. Let's look at a simple example:

$$A = \begin{pmatrix} 5 & 1 \\ 1 & 5 \end{pmatrix}. \tag{3.11.9}$$

The characteristic polynomial equation is

$$(5 - \lambda)^2 - 1 = 0 \tag{3.11.10}$$

the roots of which are

$$\lambda = 5 \pm 1.$$

So now all we have to do is solve the linear systems

$$\begin{pmatrix} 5 & 1 \\ 1 & 5 \end{pmatrix} \begin{pmatrix} y \\ y \end{pmatrix} = 6 \begin{pmatrix} y \\ y \end{pmatrix} \tag{3.11.11}$$

and

$$\begin{pmatrix} 5 & 1 \\ 1 & 5 \end{pmatrix} \begin{pmatrix} y \\ y \end{pmatrix} = 4 \begin{pmatrix} y \\ y \end{pmatrix} \tag{3.11.12}$$

You can easily see that in the first case $(1, 1)^T$ is a solution, while in the second $(1, -1)^T$ is. So $(1, 1)^T$ is an eigevector associated with the eigenvalue 6 and $(1, -1)^T$ an is eigenvector associated with the eigenvalue 4.

Here is another example Let

$$A = \begin{pmatrix} 5 & 3 \\ 4 & 5 \end{pmatrix}. \tag{3.11.13}$$

The characteristic polynomial is

$$\lambda^2 - 10\lambda + 13. \tag{3.11.14}$$

The roots of this polynomial are

$$\lambda = 5 + 2\sqrt{3}, \text{ and } \lambda = 5 - 2\sqrt{3}. \tag{3.11.15}$$

Now all we have to do is solve the two homogeneous systems:

$$\begin{bmatrix} 2\sqrt{3} & 3 \\ 4 & 2\sqrt{3} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0 \tag{3.11.16}$$

and

$$\begin{bmatrix} -2\sqrt{3} & 3 \\ 4 & -2\sqrt{3} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0 \tag{3.11.17}$$

from which we arrive at the two eigenvectors

$$\begin{bmatrix} \frac{\sqrt{3}}{2} \\ 1 \end{bmatrix}, \begin{bmatrix} -\frac{\sqrt{3}}{2} \\ 1 \end{bmatrix} \tag{3.11.18}$$

But note well, that these eigenvectors are not unique. Because they solve a homogeneous system, we can multiply them by any scalar we like and not change the fact that they are eigenvectors.

This exercise was straightforward. But imagine what would have happened if we had needed to compute the eigenvectors/eigenvalues of a 10×10 matrix. Can you imagine having to compute the roots of a 10-th order polynomial? In fact, once you get past order 4, there is no algebraic formula for the roots of a polynomial. The eigenvalue problem is much harder than solving $A\mathbf{x} = \mathbf{y}$.

The following theorem gives us the essential computational tool for using eigenvectors.

Theorem 4 Matrix diagonalization *Let A be an $n \times n$ matrix with n linearly independent eigenvectors. Let S be a matrix whose columns are these eigenvectors. Then $S^{-1}AS$ is a diagonal matrix Λ whose elements are the eigenvalues of A.*

The proof is easy. The elements in the first column of the product matrix AS are precisely the elements of the vector which is the inner product of A with the first column of S. The first column of S, say $\mathbf{s}_1$, is, by definition, an eigenvector of A. Therefore the first column of AS is $\lambda_1\mathbf{s}_1$. Since this is true for all the columns, it follows that AS is a matrix whose columns are $\lambda_i\mathbf{s}_i$. But now we're in business since

$$[\lambda_1\mathbf{s}_1 \ \lambda_2\mathbf{s}_2 \ \cdots \lambda_n\mathbf{s}_n] = [\mathbf{s}_1 \ \mathbf{s}_2 \ \cdots \mathbf{s}_n] \operatorname{diag}(\lambda_1, \lambda_2, \cdots, \lambda_n) \equiv S\Lambda. \tag{3.11.19}$$

Therefore $AS = S\Lambda$ which means that $S^{-1}AS = \Lambda$. S must be invertible since we've assumed that all it's columns are linearly independent.

Some points to keep in mind:

- Any matrix in $\mathbf{R}^{n\times n}$ with n distinct eigenvalues can be diagonalized.
- Because the eigenvectors themselves are not unique, the diagonalizing matrix S is not unique.
- Not all square matrices possess n linearly independent eigenvectors. For example, what are the eigenvectors of

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}? \tag{3.11.20}$$

  Well the characteristic polynomial of this matrix is simply λ^2. So the roots are both 0. So the eigenvectors of the matrix are any vectors in the null space. Now a vector $(x, y)^T$ gets mapped into zero by $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ if an only if $y = 0$. So any vector of the form $(x, 0)^T$ is an eigenvector.

- A matrix can be invertible without being diagonalizable. For example,

$$\begin{pmatrix} 3 & 1 \\ 0 & 3 \end{pmatrix}. \tag{3.11.21}$$

  Its two eigenvalues are both equal to 3 and its eigenvectors cannot be linearly independent. However the inverse of this matrix is straightforward

$$\begin{pmatrix} 1/3 & -1/9 \\ 0 & 1/3 \end{pmatrix}. \tag{3.11.22}$$

We can summarize these ideas with a theorem whose proof can be found in linear algebra books.

Theorem 5 Linear independence of eigenvectors *If n eigenvectors of an $n \times n$ matrix correspond to n different eigenvalues, then the eigenvectors are linearly independent.*

An important class of matrices for inverse theory are the real symmetric matrices. The reason is that since we have to deal with rectangular matrices, we often end up treating the matrices $A^T A$ and AA^T instead. And these two matrices are manifestly symmetric. In the case of real symmetric matrices, the eigenvector/eigenvalue decomposition is especially nice, since in this case the diagonalizing matrix S can be chosen to be an orthogonal matrix Q.

Theorem 6 Orthogonal decomposition of a real symmetric matrix *A real symmetric matrix A can be factored into*

$$A = Q\Lambda Q^T \tag{3.11.23}$$

with orthonormal eigenvectors in Q and real eigenvalues in Λ.

3.12 Orthogonal decomposition of rectangular matrices

[4] For dimensional reasons there is clearly no hope of the kind of eigenvector decomposition discussed above being applied to rectangular matrices. However, there is an amazingly useful generalization that pertains if we allow a different orthogonal matrix on each side of A. It is called the *Singular Value Decomposition* and **works for any matrix whatsoever.** Essentially the singular value decomposition generates orthogonal bases of $\mathbf{R}^m$ and $\mathbf{R}^n$ simultaneously.

[4] This section can be skipped on first reading.

Theorem 7 Singular value decomposition *Any matrix $A \in \mathbf{R}^{n\times m}$ can be factored as*

$$A = U\Lambda V^T \tag{3.12.1}$$

where the columns of $U \in \mathbf{R}^{n\times n}$ are eigenvectors of AA^T and the columns of $V \in \mathbf{R}^{m\times m}$ are the eigenvectors of A^TA. $\Lambda \in \mathbf{R}^{n\times m}$ is a rectangular matrix with the singular values on its main diagonal and zero elsewhere. The singular values are the square roots of the eigenvalues of A^TA, which are the same as the nonzero eigenvalues of AA^T. Further, there are exactly r nonzero singular values, where r is the rank of A.

The columns of U and V span the four fundamental subspaces. The column space of A is spanned by the first r columns of U. The row space is spanned by the first r columns of V. The left nullspace of A is spanned by the last $n-r$ columns of U. And the nullspace of A is spanned by the last $m-r$ columns of V.

A direct approach to the SVD, attributed to the physicist Lanczos, is to make a symmetric matrix out of the rectangular matrix A as follows: Let

$$S = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}. \tag{3.12.2}$$

Since A is in $\mathbf{R}^{n\times m}$, S must be in $\mathbf{R}^{(n+m)\times(n+m)}$. And since S is symmetric it has orthogonal eigenvectors $\mathbf{w}_i$ with real eigenvalues λ_i

$$S\mathbf{w}_i = \lambda_i \mathbf{w}_i. \tag{3.12.3}$$

If we split up the eigenvector $\mathbf{w}_i$, which is in $\mathbf{R}^{n+m}$, into an n-dimensional data part and an m-dimensional model part

$$\mathbf{w}_i = \begin{bmatrix} \mathbf{u}_i \\ \mathbf{v}_i \end{bmatrix} \tag{3.12.4}$$

then the eigenvalue problem for S reduces to two coupled eigenvalue problems, one for A and one for A^T

$$A^T\mathbf{u}_i = \lambda_i \mathbf{v}_i \tag{3.12.5}$$

$$A\mathbf{v}_i = \lambda_i \mathbf{u}_i. \tag{3.12.6}$$

We can multiply the first of these equations by A and the second by A^T to get

$$A^TA\mathbf{v}_i = {\lambda_i}^2 \mathbf{v}_i \tag{3.12.7}$$

$$AA^T\mathbf{u}_i = {\lambda_i}^2 \mathbf{u}_i. \tag{3.12.8}$$

So we see, once again, that the model eigenvectors $\mathbf{u}_i$ are eigenvectors of AA^T and the data eigenvectors $\mathbf{v}_i$ are eigenvectors of A^TA. Also note that if we change sign of the eigenvalue we see that $(-\mathbf{u}_i, \mathbf{v}_i)$ is an eigenvector too. So if there are r pairs of nonzero

eigenvalues $\pm\lambda_i$ then there are r eigenvectors of the form $(\mathbf{u}_i, \mathbf{v}_i)$ for the positive λ_i and r of the form $(-\mathbf{u}_i, \mathbf{v}_i)$ for the negative λ_i.

Keep in mind that the matrices U and V whose columns are the model and data eigenvectors are square (respectively $n \times n$ and $m \times m$) and orthogonal. Therefore we have $U^TU = UU^T = I_n$ and $V^TV = VV^T = I_m$. But it is important to distinguish between the eigenvectors associated with zero and nonzero eigenvalues. Let U_r and V_r be the matrices whose columns are the r model and data eigenvectors associated with the r nonzero eigenvalues and U_0 and V_0 be the matrices whose columns are the eigenvectors associated with the zero eigenvalues, and let Λ_r be the diagonal matrix containing the r nonzero eigenvalues. Then we have the following eigenvalue problem

$$AV_r = U_r\Lambda_r \tag{3.12.9}$$

$$A^TU_r = V_r\Lambda_r \tag{3.12.10}$$

$$AV_0 = 0 \tag{3.12.11}$$

$$A^TU_0 = 0. \tag{3.12.12}$$

Since the full matrices U and V satisfy $U^TU = UU^T = I_n$ and $V^TV = VV^T = I_m$ it can be readily seen that $AV = U\Lambda$ implies $A = U\Lambda V^T$ and therefore

$$A = [U_r, U_0]\begin{bmatrix} \Lambda_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\begin{bmatrix} V_r^T \\ V_0^T \end{bmatrix} = U_r\Lambda_r V_r^T, \tag{3.12.13}$$

This is the singular value decomposition. Notice that $\mathbf{0}$ represent rectangular matrices of zeros. Since Λ_r is $r \times r$ and Λ is $n \times m$ then the lower left block of zeros must be $n - r \times r$, the upper right must be $r \times m - r$ and the lower right must be $n - r \times m - r$.

It is important to keep the subscript r in mind since the fact that A can be reconstructed from the eigenvectors associated with the nonzero eigenvalues means that the experiment is unable to see the contribution due to the eigenvectors associated with zero eigenvalues.

3.13 Eigenvectors and Orthogonal Projections

Above we said that the matrices V and U were orthogonal so that $V^TV = VV^T = I_m$ and $U^TU = UU^T = I_n$. There is a nice geometrical picture we can draw for these equations having to do with projections onto lines or subspaces. Let $\mathbf{v}_i$ denote the ith column of the matrix V. (The same argument applies to U of course.) The outer product $\mathbf{v}_i\mathbf{v}_i^T$ is an $m \times m$ matrix. It is easy to see that the action of this matrix on a vector is to project that vector onto the one-dimensional subspace spanned by $\mathbf{v}_i$:

$$\left(\mathbf{v}_i\mathbf{v}_i^T\right)\mathbf{x} = (\mathbf{v}_i^T\mathbf{x})\mathbf{v}_i.$$

A "projection" operator is defined by the property that once you've applied it to a vector, applying it again doesn't change the result: $P(P\mathbf{x}) = P\mathbf{x}$, in other words. For the operator $\mathbf{v}_i\mathbf{v}_i^T$ this is obviously true since $\mathbf{v}_i^T\mathbf{v}_i = 1$.

Now suppose we consider the sum of two of these projection operators: $\mathbf{v}_i\mathbf{v}_i^T + \mathbf{v}_j\mathbf{v}_j^T$. This will project any vector in $\mathbf{R}^m$ onto the plane spanned by $\mathbf{v}_i$ and $\mathbf{v}_j$. We can continue this procedure and define a projection operator onto the subspace spanned by any number p of the model eigenvectors:

$$\sum_{i=1}^{p} \mathbf{v}_i\mathbf{v}_i^T.$$

If we let $p = m$ then we get a projection onto all of $\mathbf{R}^m$. But this must be the identity operator. In effect we've just proved the following identity:

$$\sum_{i=1}^{m} \mathbf{v}_i\mathbf{v}_i^T = VV^T = I.$$

On the other hand, if we only include the terms in the sum associated with the r nonzero singular values, then we get a projection operator onto the non-null space (which is the row space). So

$$\sum_{i=1}^{r} \mathbf{v}_i\mathbf{v}_i^T = V_rV_r^T$$

is a projection operator onto the row space. By the same reasoning

$$\sum_{i=r+1}^{m} \mathbf{v}_i\mathbf{v}_i^T = V_0V_0^T$$

is a projection operator onto the null space. Putting this all together we can say that

$$V_rV_r^T + V_0V_0^T = I.$$

This says that any vector in $\mathbf{R}^m$ can be written in terms of its component in the null space and its component in the row space of A. Let $\mathbf{x} \in \mathbf{R}^m$, then

$$\mathbf{x} = I\mathbf{x} = \left(V_rV_r^T + V_0V_0^T\right)\mathbf{x} = (\mathbf{x})_{\text{row}} + (\mathbf{x})_{\text{null}}. \tag{3.13.1}$$

3.14 A few examples

This example shows that often matrices with repeated eigenvalues cannot be diagonalized. But symmetric matrices can **always** be diagonalized.

$$A = \begin{bmatrix} 3 & 1 \\ 0 & 3 \end{bmatrix} \tag{3.14.1}$$

The eigenvalues of this matrix are obviously 3 and 3. This matrix has a one-dimensional family of eigenvectors; any vector of the form $(x, 0)^T$ will do. So it cannot be diagonalized, it doesn't have enough eigenvectors.

Now consider

$$A = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix} \tag{3.14.2}$$

The eigenvalues of this matrix are still 3 and 3. But it will be diagonalized **by any invertible matrix**! So, of course, to make our lives simple we will choose an orthogonal matrix. How about

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}? \tag{3.14.3}$$

That will do. But so will

$$\frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}. \tag{3.14.4}$$

So, as you can see, repeated eigenvalues give us choice. And for symmetric matrices we nearly always choose to diagonalize with orthogonal matrices.

Exercises

3.1 Solve the following linear system for a, b and c.

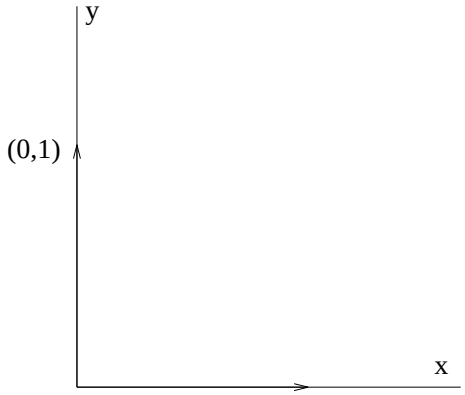
$$\begin{bmatrix} -2 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -2 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

3.2 Consider the linear system

$$\begin{bmatrix} a & b \\ b & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Assume x and y are nonzero. Try to solve this system for x and y and thereby show what conditions must be put on the elements of the matrix such that there is a nonzero solution of these equations.

3.3 Here is a box generated by two unit vectors, one in the x direction and one in the y direction.



If we take a two by two matrix

$$A = \begin{bmatrix} a & b \\ b & d \end{bmatrix}$$

and apply it to the two unit vectors, we get two new vectors that form a different box. (I.e., take the dot product of A with the two column vectors $(1,0)^T$ and $(0,1)^T$.) Draw the resulting boxes for the following matrices and say in words what the transformation is doing.

(a)

$$\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$$

(b)

$$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

(c)

$$\begin{bmatrix} 2 & 0 \\ 0 & 1/2 \end{bmatrix}$$

(d)

$$\begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$$

(e)

$$\begin{bmatrix} -2 & 1 \\ 1 & -2 \end{bmatrix}$$

3.4 For the matrices

$$A = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix}$$

and

$$B = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$$

compute A^{-1}, B^{-1}, $(BA)^{-1}$, and $(AB)^{-1}$

3.5 The next 5 questions concern a particular linear system. Let

$$A = \begin{pmatrix} 0 & 2 & 4 & -6 \\ 1 & -2 & 4 & 3 \\ 2 & 2 & -4 & 0 \end{pmatrix}$$

Compute the row-reduced form of A and A^T. Clearly label the pivots for each case.

3.6 Write down basis vectors for the row and column spaces of A. What is the rank of the matrix?

3.7 Write down basis vectors for the left and right null spaces of A.

3.8 What are the free variable(s) of the linear system $A\mathbf{r} = \mathbf{b}$ where

$$\mathbf{r} = \begin{pmatrix} w \\ x \\ y \\ z \end{pmatrix} \text{ and } \mathbf{b} = \begin{pmatrix} 0 \\ 6 \\ 0 \end{pmatrix}.$$

Compute the particular solution of this system by setting the free variable(s) equal to zero. Show for this system the general solution is equal to this particular solution plus an element of the null space.

3.9 How many of the columns are linearly independent?

How many of the rows are linearly independent?

3.10 Let

$$A = \begin{pmatrix} 3/2 & -5/2 \\ -5/2 & 3/2 \end{pmatrix}$$

Compute the eigenvalues and eigenvectors of this matrix. Are the eigenvectors orthogonal?

3.11 Let Q be the matrix of eigenvectors from the previous question and L be the diagonal matrix of eigenvalues. Show by direct calculation that Q diagonalizes A, i.e., $QAQ^T = L$.

3.12 Give an example of a real, nondiagonal 2×2 matrix whose eigenvalues are complex.

3.13 In terms of its eigenvalues, what does it mean for a matrix to be invertible? Are diagonalizable matrices always invertible?

3.14 Give specific (nonzero) examples of 2 by 2 matrices satisfying the following properties:

$$A^2 = 0, A^2 = -I_2, \text{and } AB = -BA \tag{3.14.5}$$

3.15 Let A be an upper triangular matrix. Suppose that all the diagonal elements are nonzero. Show that the columns must be linearly independent and that the nullspace contains only the zero vector.

3.16 Figure out the column space and null space of the following two matrices:

$$\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} \text{ and } \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{3.14.6}$$

3.17 Which of the following two are subspaces of $\mathbf{R}^n$: the plane of all vectors whose first component is zero; the plane of all vectors whose first component is 1.

3.18 Let P be a plane in $\mathbf{R}^3$ defined by $x_1 - 6x_2 + 13x_3 = -3$. What is the equation of the plane P_0 parallel to P but passing through the origin? Is either P or P_0 a subspace of $\mathbf{R}^3$?

3.19 Let

$$\mathbf{x} = \begin{bmatrix} 9 \\ -12 \end{bmatrix}. \tag{3.14.7}$$

Compute $\|x\|_1$, $\|x\|_2$, and $\|x\|_\infty$.

3.20 Show that $B = (A^TA)^{-1}A^T$ is a left inverse and $C = A^T(AA^T)^{-1}$ is a right inverse of a matrix A, provided that AA^T and A^TA are invertible. It turns out that A^TA is invertible if the rank of A is equal to n, the number of columns; and AA^T is invertible if the rank is equal to m, the number of rows.

3.21 Consider the matrix

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \tag{3.14.8}$$

The trace of this matrix is $a + d$ and the determinant is $ad - cb$. Show by direct calculation that the product of the eigenvalues is equal to the determinant and the sum of the eigenvalues is equal to the trace.

3.22 As we have seen, an orthogonal matrix corresponds to a rotation. Consider the eigenvalue problem for a simple orthogonal matrix such as

$$Q = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \tag{3.14.9}$$

How can a rotation map a vector into a multiple of itself?

3.23 Show that the eigenvalues of A^j are the j-th powers of the eigenvalues of A.

3.24 Compute the SVD of the matrix

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & -1 \end{bmatrix} \tag{3.14.10}$$

directly by computing the eigenvectors of $A^T A$ and AA^T.

Chapter 4

Fourier Analysis

4.1 Motivation

At the beginning of this course, we saw that superposition of functions in terms of sines and cosines was extremely useful for solving problems involving linear systems. For instance, when we studied the forced harmonic oscillator, we first solved the problem by assuming the forcing function was a sinusoid (or complex exponential). This turned out to be easy. We then argued that since the equations were linear this was enough to let us build the solution for an arbitrary forcing function if only we could represent this forcing function as a sum of sinusoids. Later, when we derived the continuum limit of the coupled spring/mass system we saw that separation of variables led us to a solution, but only **if** we could somehow represent general initial conditions as a sum of sinusoids. The representation of arbitrary functions in terms of sines and cosines is called *Fourier analysis.*

Jean Baptiste Joseph Fourier. Born: 21 March 1768 in Auxerre. Died: 16 May 1830 in Paris. Fourier trained as a priest and nearly lost his head (literally) in the French revolution. He is best known for his work on heat conduction. Fourier established the equation governing diffusion and used infinite series of trigonometric functions to solve it. Fourier was also a scientific adviser to Napoleon's army in Egypt.

4.2 The Fourier Series

So, the motivation for further study of such a Fourier superposition is clear. But there are other important reasons as well. For instance, consider the data shown in Figure 4.1.

These are borehole tiltmeter measurements. A tiltmeter is a device that measures the local tilt relative to the earth's gravitational field. The range of tilts shown here is between -40 and 40 nanoradians! (There are 2 π radians in 360 degrees, so this range corresponds to about 8 millionths of a degree.) With this sensitivity, you would expect that the dominant signal would be due to earth tides. So buried in the time-series on the top you would expect to see two dominant frequencies, one that was diurnal (1 cycle per day) and one that was semi-diurnal (2 cycles per day). If we somehow had an automatic way of representing these data as a superposition of sinusoids of various frequencies, then might we not expect these characteristic frequencies to manifest themselves in the size of the coefficients of this superposition? The answer is yes, and this is one of the principle aims of Fourier analysis. In fact, the power present in the data at each frequency is called the power spectrum. Later we will see how to estimate the power spectrum using a Fourier transform.

You'll notice in the tiltmeter spectrum that the two peaks (diurnal and semi-diurnal seem to be split; i.e., there are actually two peaks centered on 1 cycle/day and two peaks centered on 2 cycles/day. Consider the superposition of two sinusoids of nearly the same frequency:

$$\sin((\omega - \epsilon)t) + \sin((\omega + \epsilon)t).$$

Show that this is equal to

$$2\cos(\epsilon t)\sin(\omega t).$$

Interpret this result physically, keeping in mind that the way we've set the problem up, ϵ is a small number compared to ω. It might help to make some plots. Once you've figured out the interpretation of this last equation, do you see evidence of the same effect in the tiltmeter data?

There is also a *drift* in the tiltmeter. Instead of the tides fluctuating about 0 tilt, they slowly drift upwards over the course of 50 days. This is likely a drift in the instrument and not associated with any tidal effect. Think of how you might *correct* the data for this drift.

As another example Figure 4.2 shows 50 milliseconds of sound (a low C) made by a soprano saxophone and recorded on a digital oscilloscope. Next to this is the estimated power spectrum of the same sound. Notice that the peaks in the power occur at integer multiples of the frequency of the first peak (the nominal frequency of a low C).

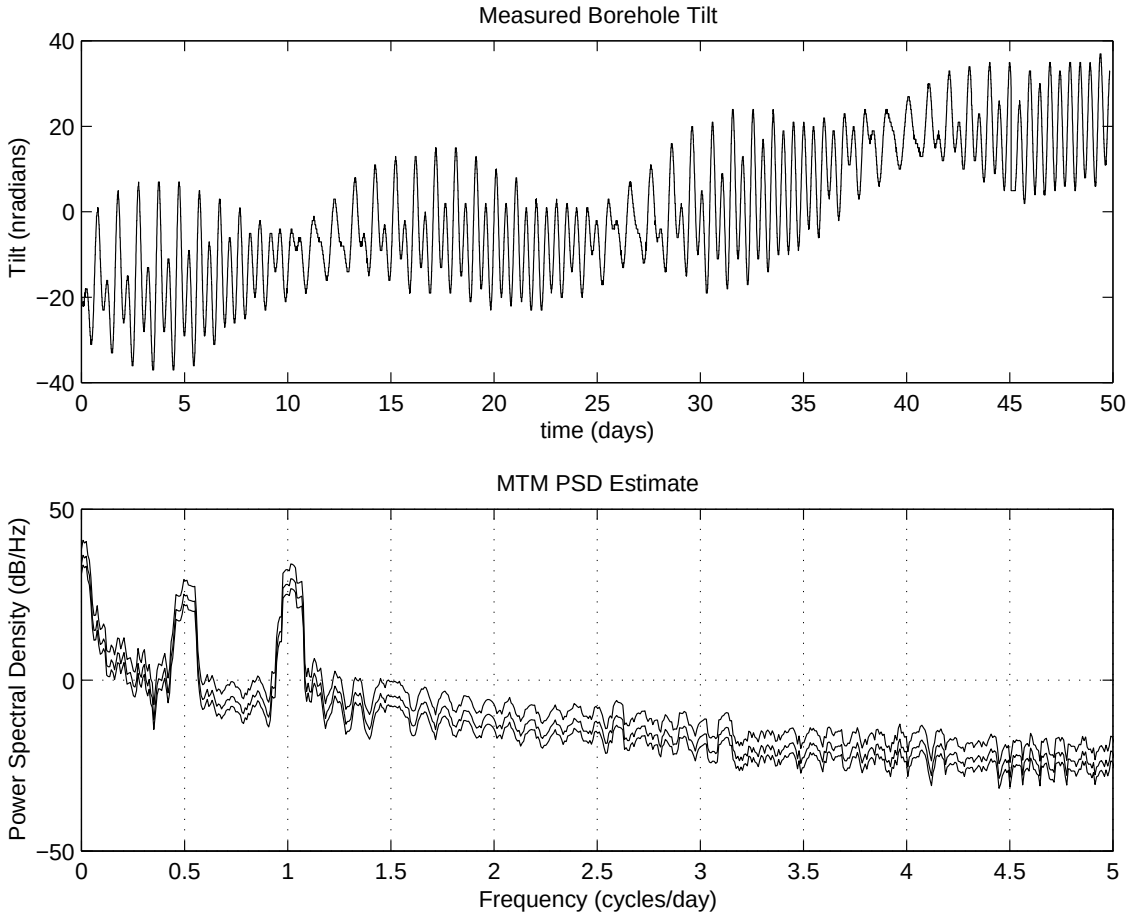


Figure 4.1: Borehole tiltmeter measurements. Data courtesy of Dr. Judah Levine (see [**?**] for more details). The plot on the top shows a 50 day time series of measurements. The figure on the bottom shows the estimated power in the data at each frequency over some range of frequencies. This is known as an estimate of the *power spectrum* of the data. Later we will learn how to compute estimates of the power spectrum of time series using the *Fourier transform*. Given what we know about the physics of tilt, we should expect that the diurnal tide (once per day) should peak at 1 cycle per day, while the semi-diurnal tide (twice per day) should peak at 2 cycles per day. This sort of analysis is one of the central goals of Fourier theory.

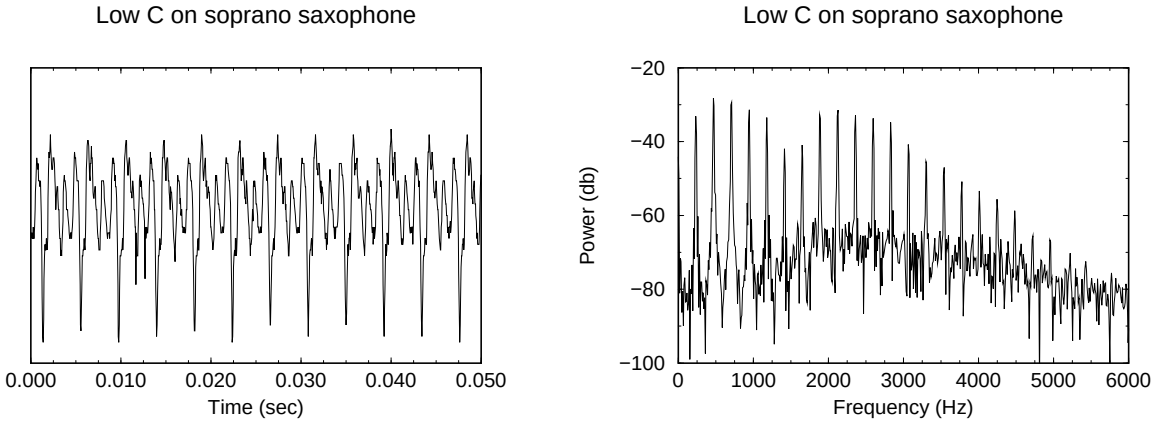


Figure 4.2: On the left is .05 seconds of someone playing low C on a soprano saxophone. On the right is the power spectrum of these data. We'll discuss later how this computation is made, but essentially what you're seeing the power as a function of frequency. The first peak on the right occurs at the nominal frequency of low C. Notice that all the higher peaks occur at integer multiples of the frequency of the first (fundamental) peak.

Definition of the Fourier Series

For a function periodic on the interval $[-l, l]$, the Fourier series is defined to be:

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos(n\pi x/l) + b_n \sin(n\pi x/l). \tag{4.2.1}$$

or equivalently,

$$f(x) = \sum_{n=-\infty}^{\infty} c_n e^{in\pi x/l}. \tag{4.2.2}$$

We will see shortly how to compute these coefficients. The connection between the real and complex coefficients is:

$$c_k = \frac{1}{2}(a_k - ib_k) \quad c_{-k} = \frac{1}{2}(a_k + ib_k). \tag{4.2.3}$$

In particular notice that the sine/cosine series has only positive frequencies, while the exponential series has both positive and negative. The reason is that in the former case each frequency has two functions associated with it. If we introduce a single complex function (the exponential) we avoid this by using negative frequencies. In other words, any physical vibration always involves two frequencies, one positive and one negative.

Later on you will be given two of the basic convergence theorems for Fourier series. Now let's look at some examples.

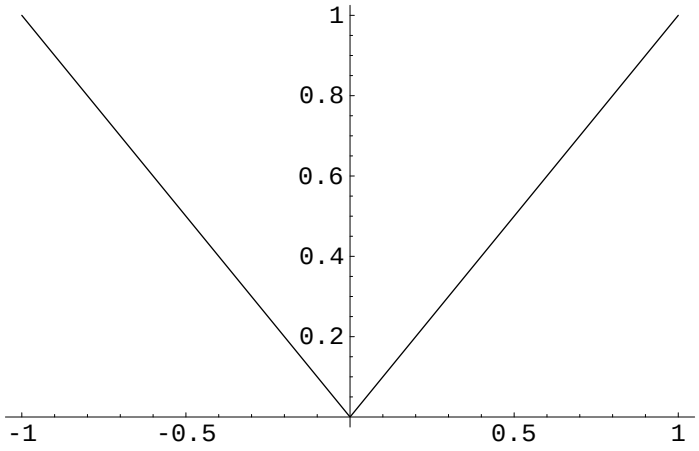


Figure 4.3: Absolute value function.

4.2.1 Examples

Let $f(x) = \text{abs}(x)$, as shown in Figure 4.3. The first few terms of the Fourier series are:

$$\frac{1}{2} - \frac{4\,\cos(\pi\,x)}{\pi^2} - \frac{4\,\cos(3\,\pi\,x)}{9\,\pi^2} - \frac{4\,\cos(5\,\pi\,x)}{25\,\pi^2} \tag{4.2.4}$$

This approximation is plotted in Figure 4.3.

Observations

Note well that the convergence is slowest at the origin, where the absolute value function is not differentiable. (At the origin, the slope changes abruptly from -1 to +1. So the left derivative and the right derivative both exist, but they are not the same.) Also, as for any even function (i.e., $f(x) = f(-x)$) only the cosine terms of the Fourier series are nonzero.

Suppose now we consider an odd function (i.e., $f(x) = -f(-x)$), such as $f(x) = x$. The first four terms of the Fourier series are

$$\frac{2\,\sin(\pi\,x)}{\pi} - \frac{\sin(2\,\pi\,x)}{\pi} + \frac{2\,\sin(3\,\pi\,x)}{3\,\pi} - \frac{\sin(4\,\pi\,x)}{2\,\pi} \tag{4.2.5}$$

Here you can see that only the sine terms appear, and no constant (zero-frequency) term. A plot of this approximation is shown in Figure 4.4.

So why the odd behavior at the endpoints? It's because we've assume the function is periodic on the interval $[-1, 1]$. The *periodic extension* of $f(x) = x$ must therefore have

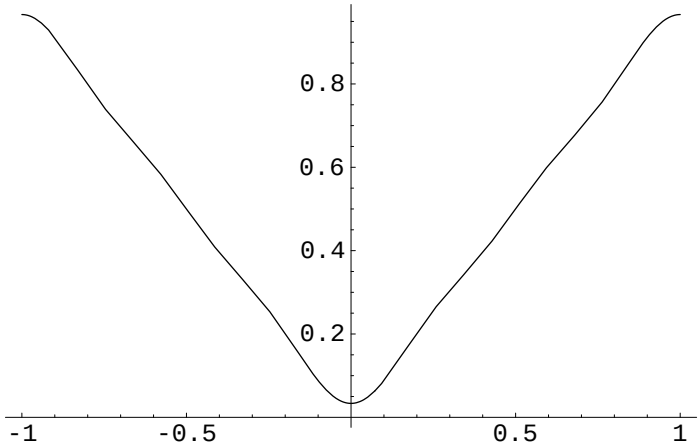


Figure 4.4: First four nonzero terms of the Fourier series of the function $f(x) = \text{abs}(x)$.

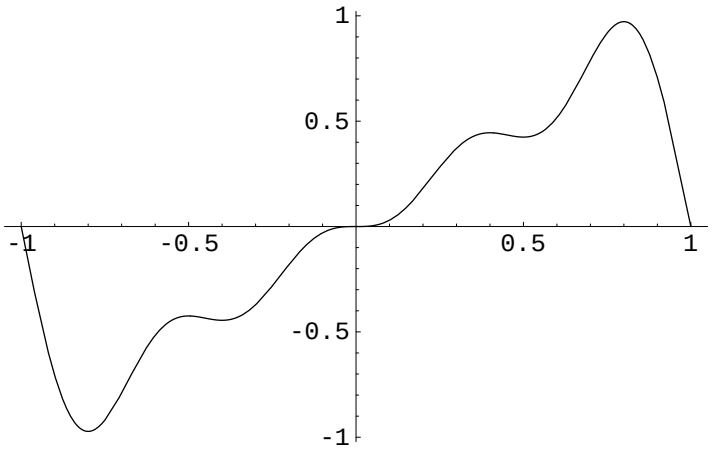


Figure 4.5: First four nonzero terms of the Fourier series of the function f(x) = x.

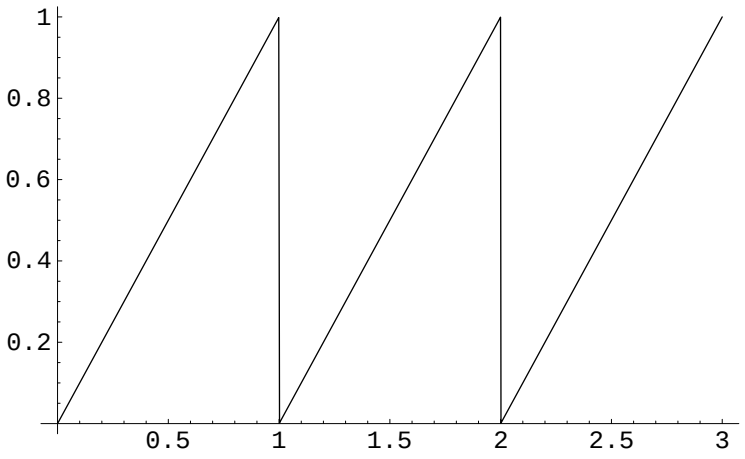


Figure 4.6: Periodic extension of the function $f(x) = x$ relative to the interval $[0, 1]$.

a sort of sawtooth appearance. In other words any non-periodic function defined on a finite interval can be used to generate a periodic function just by cloning the function over and over again. Figure 4.6 shows the periodic extension of the function $f(x) = x$ relative to the interval $[0, 1]$. It's a potentially confusing fact that the same function will give rise to different periodic extensions on different intervals. What would the periodic extension of $f(x) = x$ look like relative to the interval $[-.5, .5]$?

4.3 Superposition and orthogonal projection

Now, recall that for any set of N linearly independent vectors $\mathbf{x}_i$ in R^N, we can represent an arbitrary vector $\mathbf{z}$ in R^N as a superposition

$$\mathbf{z} = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots + c_N\mathbf{x}_N, \tag{4.3.1}$$

which is equivalent to the linear system

$$\mathbf{z} = X \cdot \mathbf{c} \tag{4.3.2}$$

where X is the matrix whose columns are the $\mathbf{x}_i$ vectors and $\mathbf{c}$ is the vector of unknown expansion coefficients. As you well know, matrix equation has a unique solution $\mathbf{c}$ if and only if the $\mathbf{x}_i$ are linearly independent. But the solution is especially simple if the $\mathbf{x}_i$ are orthogonal. Suppose we are trying to find the coefficients of

$$\mathbf{z} = c_1\mathbf{q}_1 + c_2\mathbf{q}_2 + \cdots + \mathbf{q}_N, \tag{4.3.3}$$

when $\mathbf{q}_i \cdot \mathbf{q}_j = \delta_{ij}$. In this case we can find the coefficients easily by projecting onto the orthogonal directions:

$$c_i = \mathbf{q}_i \cdot \mathbf{z}, \tag{4.3.4}$$

or, in the more general case where the $\mathbf{q}$ vectors are orthogonal but not necessarily normalized

$$c_i = \frac{\mathbf{q}_i \cdot \mathbf{z}}{\mathbf{q}_i \cdot \mathbf{q}_i}. \tag{4.3.5}$$

We have emphasized throughout this course that functions are vectors too, they just happen to live in an infinite dimensional vector space (for instance, the space of square integrable functions). So it should come as no surprise that we would want to consider a formula just like 4.3.3, but with functions instead of finite dimensional vectors; e.g.,

$$f(x) = c_1 q_1(x) + c_2 q_2(x) + \cdots + c_n q_n(x) + \cdots. \tag{4.3.6}$$

In general, the sum will require an infinite number of coefficients c_i, since a function has an infinite amount of information. (Think of representing $f(x)$ by its value at each point x in some interval.) Equation 4.3.6 is nothing other than a Fourier series if the $q(x)$ happen to be sinusoids. Of course, you can easily think of functions for which all but a finite number of the coefficients will be zero; for instance, the sum of a finite number of sinusoids.

Now you know exactly what is coming. If the basis functions $q_i(x)$ are "orthogonal", then we should be able to compute the Fourier coefficients by simply projecting the function $f(x)$ onto each of the orthogonal "vectors" $q_i(x)$. So, let us define a dot (or inner) product for functions on an interval $[-l, l]$ (this could be an infinite interval)

$$(u, v) \equiv \int_{-l}^{l} u(x)v(x)dx. \tag{4.3.7}$$

Then we will say that two functions are orthogonal if their inner product is zero.

Now we simply need to show that the sines and cosines (or complex exponentials) are orthogonal. Here is the theorem. Let $\phi_k(x) = \sin(k\pi x/l)$ and $\psi_k(x) = \cos(k\pi x/l)$. Then

$$(\phi_i, \phi_j) = (\psi_i, \psi_j) = l\delta_{ij} \tag{4.3.8}$$

$$(\phi_i, \psi_j) = 0. \tag{4.3.9}$$

The proof, which is left as an exercise, makes use of the addition formulae for sines and cosines. (If you get stuck, the proof can be found in [2], Chapter 10.) A similar result holds for the complex exponential, where we define the basis functions as $\xi_k(x) = e^{ik\pi x/l}$.

Using Equations 4.3.8 and 4.3.9 we can compute the Fourier coefficients by simply projecting $f(x)$ onto each orthogonal basis vector:

$$a_n = \frac{1}{l}\int_{-l}^{l} f(x)\cos(n\pi x/l)dx = \frac{1}{l}(f, \psi_n), \tag{4.3.10}$$

and

$$b_n = \frac{1}{l}\int_{-l}^{l} f(x)\sin(n\pi x/l)dx = \frac{1}{l}(f, \phi_n). \tag{4.3.11}$$

Or, in terms of complex exponentials

$$c_n = \frac{1}{2l}\int_{-l}^{l} f(x)e^{-in\pi x/l}dx. \tag{4.3.12}$$

4.4 The Fourier Integral

For a function defined on any finite interval, we can use the Fourier series, as above. For functions that are periodic on some other interval than $[-l, l]$ all we have to do to use the above formulae is to make a linear change of variables so that in the new coordinate the function is defined on $[-l, l]$. And for functions that are not periodic at all, but still defined on a finite interval, we can fake the periodicity by replicating the function over and over again. This is called periodic extension.

OK, so we have a function that is periodic on an interval $[-l, l]$. Looking at its Fourier series (either Equation 4.2.1 or 4.2.2) we see straight away that the frequencies present in the Fourier synthesis are

$$f_1 = \frac{1}{2l},\ \ f_2 = \frac{2}{2l},\ \ f_3 = \frac{3}{2l}, \cdots,\ \ f_k = \frac{k}{2l}\cdots \tag{4.4.1}$$

Suppose we were to increase the range of the function to a larger interval $[-L, L]$ trivially by defining it to be zero on $[-L, -l]$ and $[l, L]$. To keep the argument simple, let us suppose that $L = 2l$. Then we notice two things straight away. First, the frequencies appearing in the Fourier synthesis are now

$$f_1 = \frac{1}{2L},\ \ f_2 = \frac{2}{2L},\ \ f_3 = \frac{3}{2L}, \cdots,\ \ f_k = \frac{k}{2L}\cdots \tag{4.4.2}$$

So the frequency interval is half what it was before. And secondly, we notice that half of the Fourier coefficients are the same as before, with the new coefficients appearing mid-way between the old ones. Imagine continuing this process indefinitely. The Fourier coefficients become more and more densely distributed, until, in the limit that $L \to \infty$, the coefficient sequence c_n becomes a continuous function. We call this function the Fourier transform of $f(x)$ and denote it by $F(k)$. In this case, our Fourier series

$$f(x) = \sum_{n=-\infty}^{\infty} c_n e^{in\pi x/l}$$

becomes

$$f(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} F(k)e^{ikx}dk \tag{4.4.3}$$

with the "coefficient" function $F(k)$ being determined, once again, by orthogonal projection:

$$F(k) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(x)e^{-ikx}dx \tag{4.4.4}$$

Normalization

A function $f(t)$ is related to its Fourier transform $f(\omega)$ via:

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(\omega) e^{i\omega t} \, d\omega \tag{4.4.5}$$

and

$$F(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) e^{-i\omega t} \, dt \tag{4.4.6}$$

It doesn't matter how we split up the 2π normalization. For example, in the interest of symmetry we have defined both the forward and inverse transform with a $1/\sqrt{2\pi}$ out front. Another common normalization is

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) e^{i\omega t} \, d\omega \tag{4.4.7}$$

and

$$F(\omega) = \int_{-\infty}^{\infty} f(t) e^{-i\omega t} \, dt. \tag{4.4.8}$$

It doesn't matter how we do this as long as we're consistent. We could get rid of the normalization altogether if we stop using circular frequencies ω in favor of f measured in hertz or cycles per second. Then we have

$$g(t) = \int_{-\infty}^{\infty} G(f) e^{2\pi i f t} \, df \tag{4.4.9}$$

and

$$G(f) = \int_{-\infty}^{\infty} g(t) e^{-2\pi i f t} \, dt \tag{4.4.10}$$

Here, using time and frequency as variables, we are thinking in terms of time series, but we could just as well use a distance coordinate such as x and a wavenumber k:

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(k) e^{ikx} \, dk \tag{4.4.11}$$

with the inverse transformation being

$$F(k) = \int_{-\infty}^{\infty} f(x) e^{-ikx} \, dx. \tag{4.4.12}$$

Invertibility: the Dirichlet Kernel

These transformations from time to frequency or space to wavenumber are invertible in the sense that if we apply one after the other we recover the original function. To see this plug Equation (4.4.12) into Equation (4.4.11):

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \, dk \int_{-\infty}^{\infty} f(x') e^{-ik(x'-x)} \, dx'. \tag{4.4.13}$$

If we define the kernel function $K(x - x', \mu)$ such that

$$K(x' - x, \mu) = \frac{1}{2\pi} \int_{-\mu}^{\mu} e^{-ik(x'-x)}\, dk = \frac{\sin \mu(x' - x)}{\pi(x' - x)} \tag{4.4.14}$$

then we have

$$f(x) = \int_{-\infty}^{\infty} f(x')K(x' - x)dx' \tag{4.4.15}$$

where $K(x' - x)$ is the limit (assuming that it exists) of $K(x' - x, \mu)$ as $\mu \to \infty$. In order for this to be true $K(x' - x)$ will have to turn out to be a *Dirac delta function.*

In one space dimension, the Dirac delta function is defined by the property that for any interval I, $f(x) = \int_I f(y)\delta(y - x)dy$ if x is in I and zero otherwise. (We can also write this as $f(0) = \int_I f(y)\delta(y)dy$.) No ordinary function can have this property since it implies that $\delta(y - x)$ must be zero except when $x = y$. If you try integrating any function which is finite at only one point (and zero everywhere else), then you always get zero. This means that $\int_I f(y)\delta(y - x)dy$ would always be zero if $\delta(0)$ is finite. So $\delta(0)$ must be infinite. And yet the function $\delta(x)$ itself must integrate to 1 since if we let $f(x) = 1$, then the basic property of the delta function says that: $1 = \int \delta(y)dy$. So we are forced to conclude that $\delta(x)$ has the strange properties that it is zero, except when $x = 0$, it is infinite when $x = 0$ and that it integrates to 1. This is no ordinary function.

The Dirac delta function is named after the Nobel prize-winning English physicist Paul A.M. Dirac (born August 1902, Bristol, England; died October 1984, Tallahassee, Florida). Dirac was legendary for making inspired physical predictions based on abstract arguments. His book *Principals of Quantum Mechanics* was one of the most influential scientific books of the 20th century. He got the Nobel Prize in Physics in 1933. Amazingly, Dirac had published 11 significant papers before his completed his PhD work. Along with Newton, Dirac is buried in Westminster Abbey.

We won't attempt to prove that the kernel function converges to a delta function and hence that the Fourier transform is invertible; you can look it up in most books on analysis. But Figure 4.7 provides graphical evidence. We show plots of this kernel function for $x = 0$ and four different values of μ, 10, 100, 1000, and 10000. It seems pretty clear that in the limit that $\mu \to \infty$, the function K becomes a Dirac delta function.

4.4.1 Examples

Let's start with an easy but interesting example. Suppose we want the compute the Fourier transform of a box-shaped function. Let $f(x)$ be equal to 1 for x in the interval

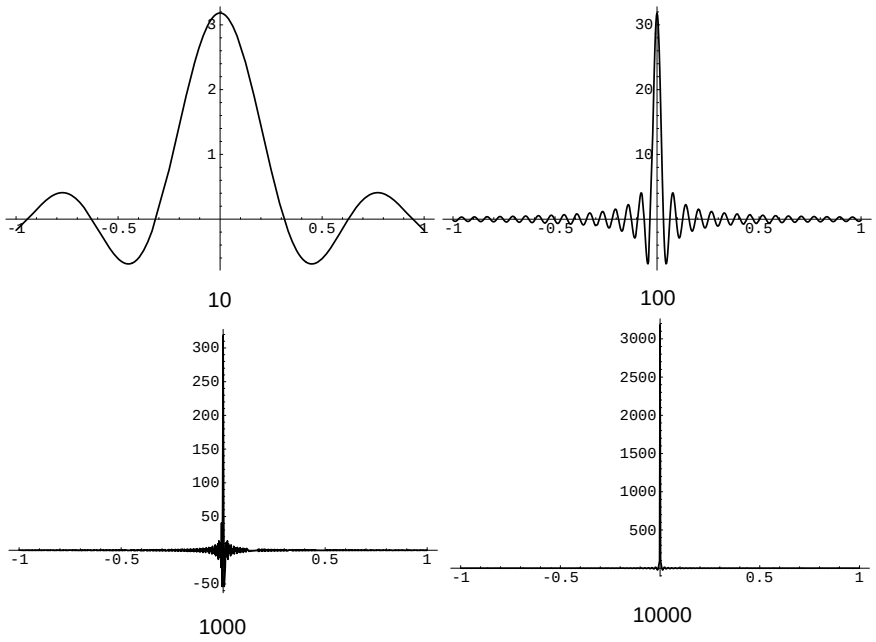


Figure 4.7: The kernel $\sin \mu x/\pi x$ for $\mu = 10$, 100, 1000, and 10000.

$[-1, 1]$ and 0 otherwise. So we need to compute

$$\int_{-1}^{1} e^{-ikx} dx = \frac{2 \sin\ k}{k}.$$

This function is shown in Figure 4.8 and is just the Dirichlet kernel for $\mu = 1$, centered about the origin.[1]

Here is a result which is a special case of a more general theorem telling us how the Fourier transform scales. Let $f(x) = e^{-x^2/a^2}$. Here a is a parameter which corresponds to the width of the bell-shaped curve. Make a plot of this curve. When a is small, the curve is relatively sharply peaked. When a is large, the curve is broadly peaked. Now compute the Fourier transform of f:

$$F(k) \propto \int_{-\infty}^{\infty} e^{-x^2/a^2} e^{-ikx} dx.$$

The trick to doing integrals of this form is to complete the square on the exponentials. You want to write the whole thing as an integral of the form

$$\int_{-\infty}^{\infty} e^{-z^2} dz.$$

As you'll see shortly, this integral can be done analytically. The details will be left as an exercise, here we will just focus on the essential feature, the exponential.

$$e^{-x^2/a^2} e^{-ikx} = e^{-1/a^2\left[(x+ika^2/2)^2+(ka^2/2)^2\right]}.$$

[1] To evaluate the limit of this function at $k = 0$, use L'Hôpital's rule.

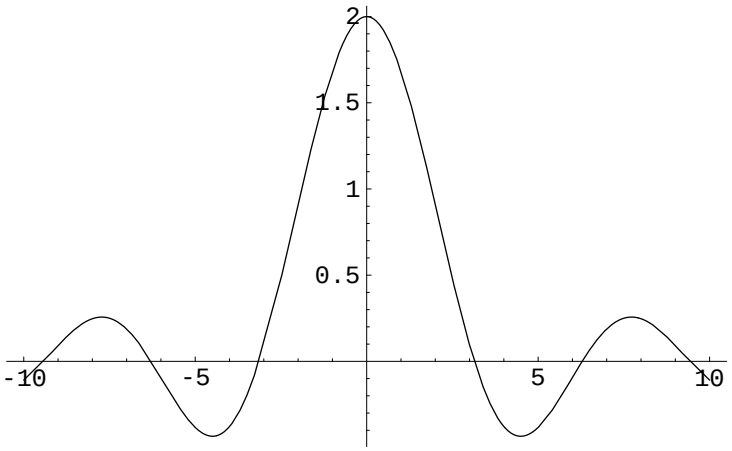


Figure 4.8: The Fourier transform of the box function.

So the integral reduces to

$$ae^{-k^2a^2/4}\int_{-\infty}^{\infty} e^{-z^2}dz = \sqrt{\pi}ae^{-k^2a^2/4}.$$

(The $\sqrt{\pi}$ will come next.) So we see that in the Fourier domain the factor of a^2 appears in the numerator of the exponential, whereas in the original domain, it appeared in the denominator. Thus, making the function more peaked in the space/time domain makes the Fourier transform more broad; while making the function more broad in the space/time domain, makes it more peaked in the Fourier domain. This is a very important idea.

Now the trick to doing the Gaussian integral. Since

$$H = \int_{-\infty}^{\infty} e^{-x^2}\ dx$$

$$H^2 = \left[\int_{-\infty}^{\infty} e^{-x^2}\ dx\right]\left[\int_{-\infty}^{\infty} e^{-y^2}\ dy\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)}\ dx\ dy.$$

Therefore

$$H^2 = \int_0^{\infty}\int_0^{2\pi} e^{-r^2} r\ dr\ d\theta = \frac{1}{2}\int_0^{\infty}\int_0^{2\pi} e^{-\rho}\ d\rho\ d\theta = \pi$$

So $H = \sqrt{\pi}$.

4.4.2 Some Basic Theorems for the Fourier Transform

It is very useful to be able think of the Fourier transform as an operator acting on functions. Let us define an operator Φ via

$$\Phi[f] = F \tag{4.4.16}$$

where

$$F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-i\omega t}\ dt. \tag{4.4.17}$$

Then it is easy to see that Φ is a linear operator

$$\Phi[c_1 f_1 + c_2 f_2] = c_1\Phi[f_1] + c_2\Phi[f_2]. \tag{4.4.18}$$

Next, if $f^{(k)}$ denotes the k-th derivative of f, then

$$\Phi[f^{(k)}] = (i\omega)^k\Phi[f] \quad k = 1, 2, \ldots \tag{4.4.19}$$

This result is crucial in using Fourier analysis to study differential equations. Next, suppose c is a real constant, then

$$\Phi[f(t-c)] = e^{-icw}\Phi[f] \tag{4.4.20}$$

and

$$\Phi[e^{ict}f(t)] = F(t-c) \tag{4.4.21}$$

where $F = \Phi(f)$. And finally, we have the convolution theorem. For any two functions $f(t)$ and $g(t)$ with $F = \Phi(f)$ and $G = \Phi(g)$, we have

$$\Phi(f)\Phi(g) = \Phi[f * g] \tag{4.4.22}$$

where "*" denotes convolution:

$$[f * g](t) = \int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau. \tag{4.4.23}$$

The convolution theorem is one of the most important in time series analysis. Convolutions are done often and by going to the frequency domain we can take advantage of the algorithmic improvements of the fast Fourier transform algorithm (FFT).

The proofs of all these but the last will be left as an exercise. The convolution theorem is worth proving. Start by multiplying the two Fourier transforms. We will throw caution to the wind and freely exchange the orders of integration. Also, let's ignore the normalization for the moment:

$$\begin{aligned} F(\omega)G(\omega) &= \int_{-\infty}^{\infty} f(t)e^{-i\omega t}dt \int_{-\infty}^{\infty} g(t')e^{-i\omega t'}dt' && (4.4.24)\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-i\omega(t+t')}f(t)g(t')dt\ dt' && (4.4.25)\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-i\omega\tau}f(t)g(\tau-t)dt\ d\tau && (4.4.26)\\ &= \int_{-\infty}^{\infty} e^{-i\omega\tau}\left[\int_{-\infty}^{\infty} f(t)g(\tau-t)dt\right]d\tau. && (4.4.27) \end{aligned}$$

This completes the proof, but now what about the normalization? If we put the symmetric $1/\sqrt{2\pi}$ normalization in front of both transforms, we end up with a left-over factor of

$1/\sqrt{2\pi}$ because we started out with two Fourier transforms and we ended up with only one and a convolution. On the other hand, if we had used an asymmetric normalization, then the result would be different depending on whether we put the $1/(2\pi)$ on the forward or inverse transform. This is a fundamental ambiguity since we can divide up the normalization anyway we want as long as the net effect is $1/(2\pi)$. This probably the best argument for using f instead of ω since then the 2πs are in the exponent and the problem goes away.

4.5 The Sampling Theorem

Now returning to the Fourier transform, suppose the spectrum of our time series $f(t)$ is zero outside of some symmetric interval $[-2\pi f_s, 2\pi f_s]$ about the origin.[2] In other words, the signal does not contain any frequencies higher than f_s hertz. Such a function is said to be *band limited*; it contains frequencies only in the band $[-2\pi f_s, 2\pi f_s]$. Clearly a band limited function has a finite inverse Fourier transform

$$f(t) = \frac{1}{2\pi} \int_{-2\pi f_s}^{2\pi f_s} F(\omega) e^{-i\omega t} \, d\omega. \tag{4.5.1}$$

sampling frequencies and periods

f_s is called the *sampling* frequency. Hence the sampling period is $T_s \equiv 1/f_s$. It is sometimes convenient to normalize frequencies by the sampling frequency. Then the maximum normalized frequency is 1:

$$\hat{f} = \frac{\hat{\omega}}{2\pi} = fT_s = f/f_s.$$

Since we are now dealing with a function on a finite interval we can represent it as a Fourier series:

$$F(\omega) = \sum_{n=-\infty}^{\infty} \phi_n e^{i\omega n/2f_s} \tag{4.5.2}$$

where the Fourier coefficients ϕ_n are to be determined by

$$\phi_n = \frac{1}{4\pi f_s} \int_{-2\pi f_s}^{2\pi f_s} F(\omega) e^{-i\omega n/2f_s} \, d\omega. \tag{4.5.3}$$

[2] In fact the assumption that the interval is symmetric about the origin is made without loss of generality, since we can always introduce a change of variables which maps an arbitrary interval into a symmetric one centered on 0.

Comparing this result with our previous work we can see that

$$\phi_n = \frac{f(n/2f_s)}{2f_s} \tag{4.5.4}$$

where $f(n/2f_s)$ are the samples of the original continuous time series $f(t)$. Putting all this together, one can show that the band limited function $f(t)$ is completely specified by its values at the countable set of points spaced $1/2f_s$ apart:

$$\begin{aligned} f(t) &= \frac{1}{4\pi f_s} \sum_{n=-\infty}^{\infty} f(n/2f_s) \int_{-2\pi f_s}^{2\pi f_s} e^{i(\omega n/2f_s - \omega t)}\, d\omega \\ &= \sum_{n=-\infty}^{\infty} f(n/2f_s) \frac{\sin(\pi(2f_s t - n))}{\pi(2f_s t - n)}. \end{aligned} \tag{4.5.5}$$

The last equation is known as the **Sampling Theorem**. Notice that the function $\sin x/x$ appears here too. Since this function appears frequently it is given a special name, it is called the sinc function:

$$\mathrm{sinc}(x) = \frac{sinx}{x}.$$

And we know that the sinc function is also the Fourier transform of a box-shaped function. So the sampling theorem says take the value of the function, sampled every $1/2f_s$, multiply it by a sinc function centered on that point, and then sum these up for all the samples.

It is worth repeating for emphasis: any band limited function is completely determined by its samples chosen $1/2f_s$ apart, where f_s is the maximum frequency contained in the signal. This means that in particular, a time series of finite duration (i.e., any real time series) is completely specified by a finite number of samples. It also means that in a sense, the information content of a band limited signal is infinitely smaller than that of a general continuous function. So if our band-limited signal $f(t)$ has a maximum frequency of f_s hertz, and the length of the signal is T, then the total number of samples required to describe f is $2f_sT$.

A sampling exercise

Consider the continuous sinusoidal signal:

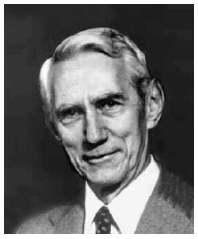
$$x(t) = A\cos(2\pi f t + \phi)$$

Suppose we sample this signal at a sampling period of T_s. Let us denote the discrete samples of the signal with square brackets:

$$x[n] \equiv x(nT_s) = A\cos(2\pi f n T_s + \phi).$$

Now consider a different sinusoid of the same amplitude and phase, but sampled at a frequency of $f + \ell f_s$, where ℓ is an integer and $f_s = 1/T_s$. Let the samples of this second sinusoid be denoted by $y[n]$. Show that $x[n] = y[n]$. This is an example of *aliasing*. These two sinusoids have exactly the same samples, so the frequency of one appears to be the same.

The sampling theorem is due to Harry Nyquist, a researcher at Bell Labs in New Jersey. In a 1928 paper Nyquist laid the foundations for the sampling of continuous signals and set forth the sampling theorem. Nyquist was born on February 7, 1889 in Nilsby, Sweden and emigrated to the US in 1907. He got his PhD in Physics from Yale in 1917. Much of Nyquist's work in the 1920's was inspired by the telegraph. In addition to his work in sampling, Nyquist also made an important theoretical analysis of thermal noise in electrical systems. In fact this sort of noise is sometimes called Nyquist noise. Nyquist died on April 4, 1976 in Harlingen, Texas.



A generation after Nyquist's pioneering work Claude Shannon, also at Bell Labs, laid the broad foundations of modern communication theory and signal processing. Shannon (Born: April 1916 in Gaylord, Michigan; Died: Feb 2001 in Medford, Massachusetts) was the founder of modern information theory. After beginning his studies in electrical engineering, Shannon took his PhD in mathematics from MIT in 1940. Shannon's *A Mathematical Theory of Communication* published in 1948 in the Bell System Technical Journal, is one of the profoundly influential scientific works of the 20th century. In it he introduced many ideas that became the basis for electronic communication, such as breaking down information into sequences of 0's and 1's (this is where the term *bit* first appeared), adding extra bits to automatically correct for errors and measuring the information or variability of signals. Shannon's paper and many other influential papers on communication are compiled in the book *Key papers in the development of information theory* [**?**].

4.5.1 Aliasing

As we have seen, if a time-dependent function contains frequencies up to f_s hertz, then discrete samples taken at an interval of $1/2f_s$ seconds completely determine the signal. Looked at from another point of view, for any sampling interval Δ, there is a special frequency (called the Nyquist frequency), given by $f_s = \frac{1}{2\Delta}$. The extrema (peaks and troughs) of a sinusoid of frequency f_s will lie exactly $1/2f_s$ apart. This is equivalent to saying that critical sampling of a sine wave is 2 samples per wavelength.

We can sample at a finer interval without introducing any error; the samples will be

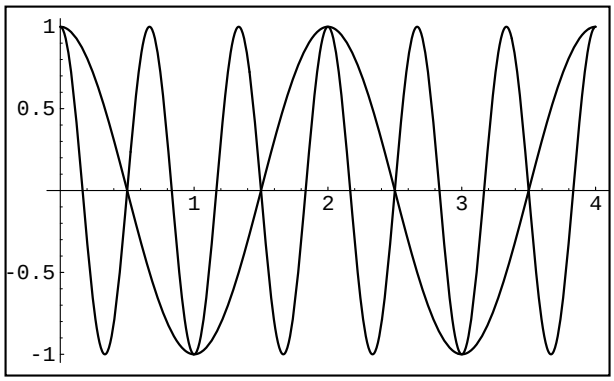


Figure 4.9: A sinusoid sampled at less than the Nyquist frequency gives rise to spurious periodicities.

redundant, of course. However, if we sample at a coarser interval a very serious kind of error is introduced called aliasing. Figure 4.9 shows a cosine function sampled at an interval longer than $1/2f_s$; this sampling produces an apparent frequency of $1/3$ the true frequency. This means that any frequency component in the signal lying outside the interval $(-f_s, f_s)$ will be spuriously shifted into this interval. Aliasing is produced by under-sampling the data: once that happens there is little that can be done to correct the problem. The way to prevent aliasing is to know the true band-width of the signal (or band-limit the signal by analog filtering) and then sample appropriately so as to give at least 2 samples per cycle at the highest frequency present.

4.6 The Discrete Fourier Transform

Now we consider the third major use of the Fourier superposition. Suppose we have discrete data, not a continuous function. In particular, suppose we have data f_k recorded at locations x_k. To keep life simple, let us suppose that the data are recorded at N evenly spaced locations $x_k = 2\pi k/N$, $k = 0, 1, \ldots N-1$. Think of f_k as being samples of an unknown function, which we want to approximate. Now we write down a Fourier approximation for the unknown function (i.e., a Fourier series with coefficients to be determined):

$$p(x) = \sum_{n=0}^{N-1} c_n e^{inx}. \tag{4.6.1}$$

Now we will compute the coefficients in such a way that p interpolates (i.e., fits exactly) the data at each x_k:

$$f_k = p(x_k) = \sum_{n=0}^{N-1} c_n e^{in2\pi k/N}. \tag{4.6.2}$$

In theory we could do this for any linearly independent set of basis functions by solving a linear system of equations for the coefficients. But since sines/cosines are orthogonal, the c_n coefficients can be computed directly:

$$c_k = \frac{1}{N} \sum_{n=0}^{N-1} f_n e^{-in2\pi k/N}. \tag{4.6.3}$$

This is the discrete version of the Fourier transform (DFT). f_n are the data and c_k are the harmonic coefficients of a trigonometric function that interpolates the data. Now, of course, there are many ways to interpolate data, but it is a theorem that the only way to interpolate with powers of $e^{i2\pi x}$ is Equation 4.6.3.

Optional Exercise In the handout you will see some Mathematica code for computing and displaying discrete Fourier transforms. Implement the previous formula and compare the results with Mathematica's built in *Fourier* function. You should get the same result, but it will take dramatically longer than Mathematica would for 100 data points. The reason is that Mathematica uses a special algorithm called the FFT (Fast Fourier Transform). See Strang for an extremely clear derivation of the FFT algorithm.

4.7 The Linear Algebra of the DFT

Take a close look at Equation 4.6.3. Think of the DFT coefficients c_k and the data points f_n as being elements of vectors $\mathbf{c}$ and $\mathbf{f}$. There are N coefficients and N data so both $\mathbf{c}$ and $\mathbf{f}$ are elements of R^N. The summation in the Fourier interpolation is therefore a matrix-vector inner product. Let's identify the coefficients of the matrix. Define a matrix Q such that

$$Q_{nk} = e^{in2\pi k/N}. \tag{4.7.1}$$

N is fixed, that's just the number of data points. The matrix appearing in Equation 4.6.3 is the complex conjugate of Q; i.e., Q^*. We can write Equation 4.6.2 as

$$\mathbf{f} = Q \cdot \mathbf{c}. \tag{4.7.2}$$

The matrix Q is almost orthogonal. We have said that a matrix A is orthogonal if $AA^T = A^T A = I$, where I is the N-dimensional identity matrix. For complex matrices we need to generalize this definition slightly; for complex matrices we will say that A is

orthogonal if $(A^T)^* A = A(A^T)^* = I$.[3] In our case, since Q is obviously symmetric, we have:

$$Q^* Q = Q Q^* = I. \tag{4.7.3}$$

Once again, orthogonality saves us from having to solve a linear system of equations: since $Q^* = Q^{-1}$, we have

$$\mathbf{c} = Q^* \cdot \mathbf{f}. \tag{4.7.4}$$

Now you may well ask: what happens if we use fewer Fourier coefficients than we have data? That corresponds to having fewer unknowns (the coefficients) than data. So you wouldn't expect there to be an exact solution as we found with the DFT, but how about a least squares solution? Let's try getting an approximation function of the form

$$p(x) = \sum_{n=0}^{m} c_n e^{inx} \tag{4.7.5}$$

where now we sum only up to $m < N - 1$. Our N equations in m unknowns is now:

$$f_k = \sum_{n=0}^{m} c_n e^{in2\pi k/N}. \tag{4.7.6}$$

So to minimize the mean square error we set the derivative of

$$||\mathbf{f} - Q \cdot \mathbf{c}||^2 \tag{4.7.7}$$

with respect to an arbitrary coefficient, say c_j, equal to zero. But this is just an ordinary least squares problem.

4.8 The DFT from the Fourier Integral

In this section we will use the f (cycles per second) notation rather than the ω (radians per second), because there are slightly fewer factors of 2π floating around. You should be comfortable with both styles, but mind those 2πs! Also, up to now, we have avoided any special notation for the Fourier transform of a function, simply observing whether it was a function of space-time or wavenumber-frequency. Now that we are considering discrete transforms and real data, we need to make this distinction since we will generally have both the sampled data and its transform stored in arrays on the computer. So for this section we will follow the convention that if $h = h(t)$ then $H = H(f)$ is its Fourier transform.

[3] Technically such a matrix is called Hermitian or self-adjoint–the operation of taking the complex conjugate transpose being known at the adjoint–but we needn't bother with this distinction here.

We suppose that our data are samples of a function and that the samples are taken at equal intervals, so that we can write

$$h_k \equiv h(t_k), \qquad t_k \equiv k\Delta, \qquad k = 0, 1, 2, \ldots, N-1, \tag{4.8.1}$$

where N is an even number. In our case, the underlying function $h(t)$ is unknown; all we have are the digitally recorded time series. But in either case we can estimate the Fourier transform $H(f)$ at at most N discrete points chosen in the range $-f_s$ to f_s where f_s is the Nyquist frequency:[4]

$$f_n \equiv \frac{n}{\Delta N}, \qquad n = \frac{-N}{2}, \ldots, \frac{N}{2}. \tag{4.8.2}$$

The two extreme values of frequency $f_{-N/2}$ and $f_{-N/2}$ are not independent ($f_{-N/2} = -f_{N/2}$), so there are actually only N independent frequencies specified above.

A sensible numerical approximation for the Fourier transform integral is thus:

$$H(f_n) = \int_{-\infty}^{\infty} h(t) e^{-2\pi i f_n t}\, dt \approx \sum_{k=0}^{N-1} h_k e^{-2\pi i f_n t_k} \Delta. \tag{4.8.3}$$

Therefore

$$H(f_n) \approx \Delta \sum_{k=0}^{N-1} h_k e^{-2\pi i k n/N}. \tag{4.8.4}$$

Defining the *Discrete Fourier Transform* (DFT) by

$$H_n = \sum_{k=0}^{N-1} h_k e^{-2\pi i k n/N} \tag{4.8.5}$$

we then have

$$H(f_n) \approx \Delta H_n \tag{4.8.6}$$

where f_n are given by Equation (4.8.2).

Now, the numbering convention implied by Equation (4.8.2) has ± Nyquist at the extreme ends of the range and zero frequency in the middle. However it is clear that the DFT is periodic with period N:

$$H_{-n} = H_{N-n}. \tag{4.8.7}$$

As a result, it is standard practice to let the index n in H_n vary from 0 to $N-1$, with n and k varying over the same range. In this convention 0 frequency occurs at

[4]The highest frequency f_s in the Fourier representation of a time series sampled at a time interval of Δ is $\frac{1}{2\Delta}$. This maximum frequency is called the Nyquist frequency. You'll study this in detail in the digital course.

$n = 0$; positive frequencies from from $1 \leq n \leq N/2 - 1$; negative frequencies run from $N/2 + 1 \leq n \leq N - 1$. Nyquist sits in the middle at $n = N/2$. The inverse transform is:

$$h_k = \frac{1}{N} \sum_{n=0}^{N-1} H_n e^{2\pi i k n/N} \tag{4.8.8}$$

Mathematica, on the other hand, uses different conventions. It uses the symmetric normalization ($1/\sqrt{N}$ in front of both the forward and inverse transform), and defines arrays running from 1 to N in Fortran fashion. So in *Mathematica*, the forward and inverse transforms are, respectively:

$$H_n = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} h_k e^{-2\pi i (k-1)(n-1)/N} \tag{4.8.9}$$

and

$$h_k = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} H_n e^{2\pi i (k-1)(n-1)/N}. \tag{4.8.10}$$

If you are using canned software, make sure you know what conventions are being used.

4.8.1 Discrete Fourier Transform Examples

Here we show a few examples of the use of the DFT. What we will do is construct an unknown time series' DFT by hand and inverse transform to see what the resulting time series looks like. In all cases the time series h_k is 64 samples long. Figures 4.10 and 4.11 show the real (left) and imaginary (right) parts of six time series that resulted from inverse DFT'ing an array H_n which was zero except at a single point (i.e., it's a Kronecker delta: $H_i = \delta_{i,j} = 1$ if $i = j$ and zero otherwise; here a different j is chosen for each plot). Starting from the top and working down, we choose j to be the following samples: the first, the second, Nyquist-1, Nyquist, Nyquist+1, the last. We can see that the first sample in frequency domain is associated with the zero-frequency or DC component of a signal and that the frequency increases until we reach Nyquist, which is in the middle of the array. Next, in Figure 4.12, we show at the top an input time series consisting of a pure sinusoid (left) and the real part of its DFT. Next we add some random noise to this signal. On the left in the middle plot is the real part of the noisy signals DFT. Finally, at the bottom, we show a Gaussian which we convolve with the noisy signal in order to attenuate the frequency components in the signal. The real part of the inverse DFT of this convolved signal is shown in the lower right plot.

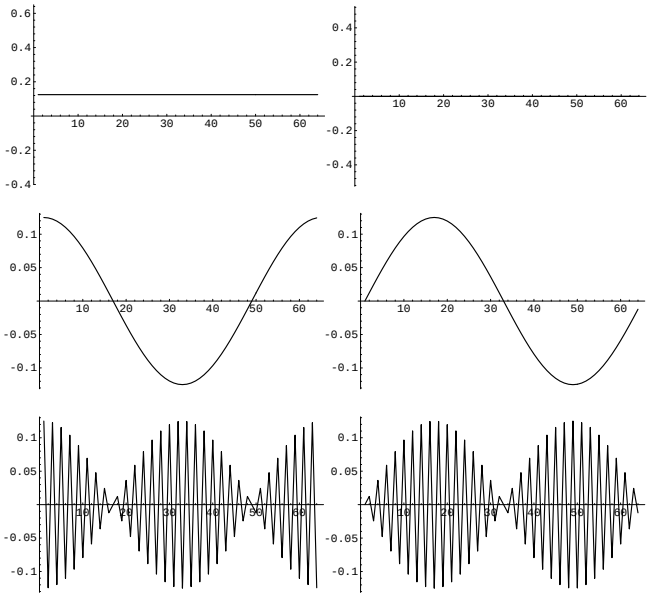


Figure 4.10: The real (left) and imaginary (right) parts of three length 64 time series, each associated with a Kronecker delta frequency spectrum. These time series are reconstructed from the spectra by inverse DFT. At the top the input spectrum is $\delta_{i,0}$, in the middle $\delta_{i,1}$, and at the bottom, $\delta_{i,64/2-1}$.

4.9 Convergence Theorems

One has to be a little careful about saying that a particular function is equal to its Fourier series since there exist piecewise continuous functions whose Fourier series diverge everywhere! However, here are two basic results about the convergence of such series.

Point-wise Convergence Theorem: If f is piecewise continuous and has left and right derivatives at a point c[5] then the Fourier series for f converges converges to

$$\frac{1}{2}\left(f(c-)+f(c+)\right) \tag{4.9.1}$$

where the + and - denote the limits when approached from greater than or less than c.

Another basic result is the **Uniform Convergence Theorem**: If f is continuous with period 2π and f' is piecewise continuous, then the Fourier series for f converges uniformly to f. For more details, consult a book on analysis such as *The Elements of Real Analysis* by Bartle [1] or *Real Analysis* by Haaser and Sullivan [**?**].

[5] A right derivative would be: $\lim_{t\to 0}(f(c+t)-f(c))/t$, $t>0$. Similarly for a left derivative.

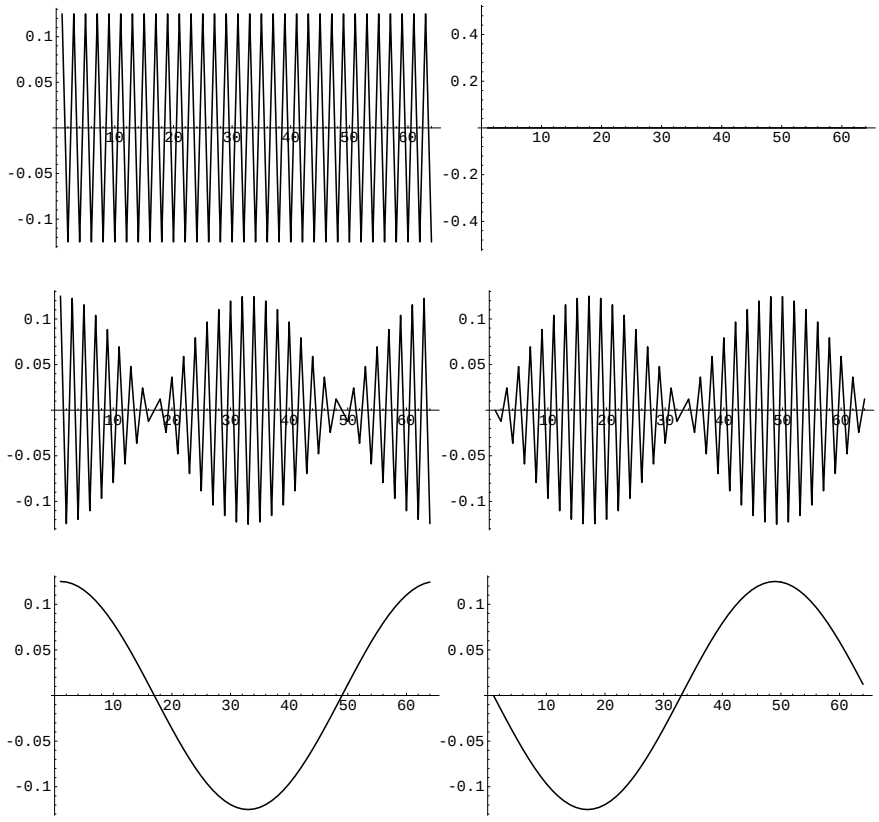


Figure 4.11: The real (left) and imaginary (right) parts of three time series of length 64, each associated with a Kronecker delta frequency spectrum. These time series are reconstructed from the spectra by inverse DFT. At the top the input spectrum is $\delta_{i,64/2}$, in the middle $\delta_{i,64/2+1}$, and at the bottom $\delta_{i,64}$.

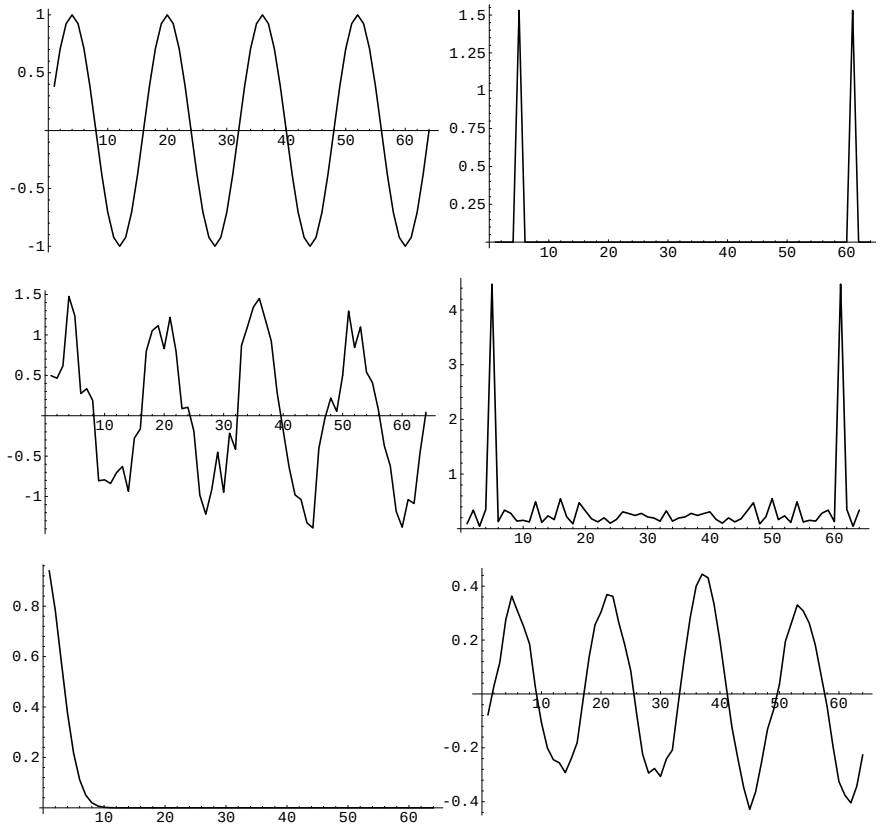


Figure 4.12: The top left plot shows an input time series consisting of a single sinusoid. In the top right we see the real part of its DFT. Note well the wrap-around at negative frequencies. In the middle we show the same input sinusoid contaminated with some uniformly distributed pseudo-random noise and its DFT. At the bottom left, we show a Gaussian time series that we will use to smooth the noisy time series by convolving it with the DFT of the noisy signal. When we inverse DFT to get back into the "time" domain we get the smoothed signal shown in the lower right.

4.10 Basic Properties of Delta Functions

Another representation of the delta function is in terms of Gaussian functions:

$$\delta(x) = \lim_{\mu\to\infty} \frac{\mu}{\sqrt{\pi}} e^{-\mu^2 x^2}. \tag{4.10.1}$$

You can verify for yourself that the area under any of the Gaussian curves associated with finite μ is one.

The spectrum of a delta function is completely flat since

$$\int_{-\infty}^{\infty} e^{-ikx}\delta(x)\ dx = 1. \tag{4.10.2}$$

For delta functions in higher dimensions we need to add an extra $1/2\pi$ normalization for each dimension. Thus

$$\delta(x,y,z) = \left(\frac{1}{2\pi}\right)^3 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{i(k_x x + k_y y + k_z z)}\ dk_x\ dk_y\ dk_z. \tag{4.10.3}$$

The other main properties of delta functions are the following:

$$\begin{aligned}
\delta(x) &= \delta(-x) && (4.10.4)\\
\delta(ax) &= \frac{1}{|a|}\delta(x) && (4.10.5)\\
x\delta(x) &= 0 && (4.10.6)\\
f(x)\delta(x-a) &= f(a)\delta(x-a) && (4.10.7)\\
\int \delta(x-y)\delta(y-a)\ dy &= \delta(x-a) && (4.10.8)\\
\int_{-\infty}^{\infty} \delta^{(m)} f(x)\ dx &= (-1)^m f^{(m)}(0) && (4.10.9)\\
\int \delta'(x-y)\delta(y-a)\ dy &= \delta'(x-a) && (4.10.10)\\
x\delta'(x) &= -\delta(x) && (4.10.11)\\
\delta(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{ikx}\ dk && (4.10.12)\\
\delta'(x) &= \frac{i}{2\pi}\int_{-\infty}^{\infty} k e^{ikx}\ dk && (4.10.13)
\end{aligned}$$

Exercises

4.1 Prove Equations 4.2.4 and 4.2.5.

4.2 Compute the Fourier transform of the following function. $f(x)$ is equal to 0 for $x < 0$, x for $0 \leq x \leq 1$ and 0 for $x > 1$.

4.3 Prove Equations 4.4.18, 4.4.19, 4.4.20, 4.4.22.

4.4 Compute the Fourier transform of $f(x) = e^{-x^2/a^2}$. If a is small, this bell-shaped curve is sharply peaked about the origin. If a is large, it is broad. What can you say about the Fourier transform of f in these two cases?

4.5 Let $f(x)$ be the function which is equal to -1 for $x < 0$ and +1 for $x > 0$. Assuming that

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos(k\pi x/l) + \sum_{k=1}^{\infty} b_k \sin(k\pi x/l),$$

compute a_0, a_1, a_2, b_1 and b_2 by hand, taking the interval of periodicity to be $[-1, 1]$.

4.6 For an odd function, only the sine or cosine terms appear in the Fourier series. Which is it?

4.7 Consider the complex exponential form of the Fourier series of a real function

$$f(x) = \sum_{n=-\infty}^{\infty} c_n e^{in\pi x/l}.$$

Take the complex conjugate of both sides. Then use the fact that since f is real, it equals its complex conjugate. What does this tell you about the coefficients c_n?

.

Chapter 5

Linear Systems

This chapter is under construction. The goal is to build a bridge to the material covered in GP 404. This will involve basic ideas of FIR filters, the frequency response of a filter, impulse responses, and z-transforms.

> Can you hear the shape of a bell?
>
> –Marc Kaz

When you strike a bell with a hammer it rings. The result of impulsively exciting a system, such as striking a bell with a hammer, is called the impulse response. In this chapter one of the key ideas we will study is how to characterize systems from measurements of their impulse responses, sort of hearing the shape of the bell. It is useful to have the a mental picture of the bell as a black box. Into the left side of the black box we put the input signal–in this case the hammer strike. The system acts on the input and produces an output–the ringing sound. This is illustrated in Figure 5.1. Of course the system need not be linear: if we strike the hammer hard enough we might dent or even crack the bell. But for now we will limit the discussions to linear systems.

The way to represent an impulsive function mathematically is with a Dirac delta. So the official definition of the impulse response of a system is the result of applying a delta function excitation. Now in practice we cannot actually measure the true impulse response since it is impossible to have a true delta function in nature–they're infinite after all. But we can come pretty close.

Figure 5.2 show an example of such an measurement. An impulsive voltage is applied to an ultrasonic transducer–a sort of high-frequency microphone. The transducer converts voltages into force, so the force generated by the transducer would be appoximately impulsive. The transducer was coupled to a rock sample. So the rock is our linear system, the impulsive force generated by the transducer is the input to the system and

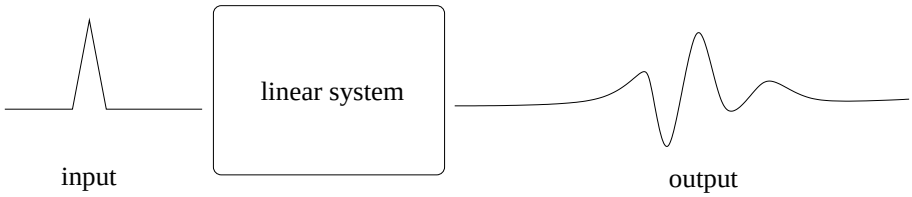


Figure 5.1: For a linear system with one input channel and one ouput channel, the impulse response is the output associated with an impulsive excitation of the system. Theoretically the excitation should be a delta function, but in practice we approximate this as best we can. It is possible to excite the system and measure its response at many different spatial locations, then the impulse response becomes a function of both the source and detector coordinates. In this case we call it a Green function.

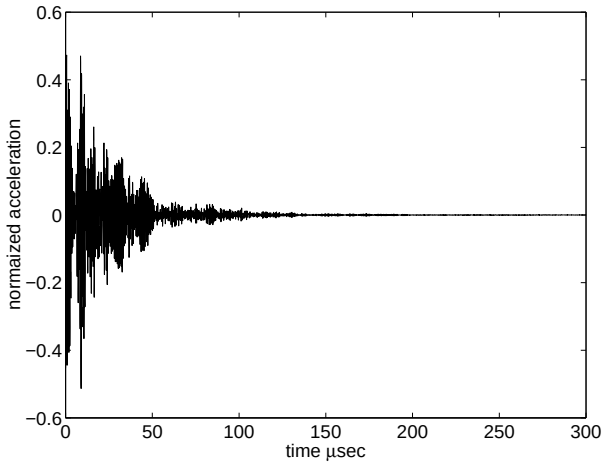


Figure 5.2: Ultrasonic impulse response measured in a cylindrical core of sandstone. An impulsive voltage is applied to a piezoelectric transducer and the response is measured with the same transducer.

Figure 5.2 shows the signal measured by the same transducer for a period of 300 μs after the excitation. What you're seening is the ringing of the rock sample to which the impulsive force was applied. The time-series associated with a bell would look very similar.

You can probably think of lots of examples of impulse response measurements. We could also set off an exposive charge in the soil; this would generate an impulsive stress wave in the earth. We could strike an object with a hammer, as is done in near-surface seismic exploration. In laboratory acoustics it is common to break capilary tubes or pencil lead on surfaces to create impulsive acoustic sources. Active sonar is a pulse-echo technique: set of an audible ping and record the echo.

One detail we've ignored up to this point is the spatial dependence of the input and output functions. For a spatially extended system, the force could be applied or the

response measured at any point, or over a region or volume of points. So for these sorts of systems, where the input and output can are functions of space, the impulse responce is actually a function of both the source and detector coordinates. In this case it is usually called a Green function. We would have to write it as something like the following: $g(\mathbf{r}, t; \mathbf{r}', t')$. This is complicated but just keep in mind that the two sets of coordinates, space and time, are associated with the location of the source and the location of the detector.

In Chapter 4, the Dirac delta function appeared somewhat mysteriously from our desire to see if the Fourier transformation was invertible. However, there is a much more intuitive explanation for the importance of this unusual function. Suppose we have a physical system that produces some output $u(t)$ when a input $f(t)$ is applied. For example u could be the acoustic pressure and f could be a force. In this particular case we know that the ouput (pressure) is related to the input (force) via the wave equation:

$$\nabla^2 u(t) - \frac{1}{c^2}\frac{\partial^2 u(t)}{\partial t^2} = f(t). \tag{5.0.1}$$

Now, u actually depends on spatial coordinates as well as time (e.g., $u(\mathbf{r}, t)$), but for now let's not explicitly show these. The left-hand side of this equation can be thought of as an operator acting on the unknown output u. Let's call this operator L, for linear:

$$L(u) = f(t). \tag{5.0.2}$$

OK, at this point the forcing funtion f could be anything. But we have already seen that because of the linearity of L, if we can solve Equation 5.0.2 when $f(t) = e^{i\omega t}$, then we can use Fourier synthesis to build the solution for an arbitrary forcing function. There is another special kind of forcing function that will work in much the same way. If we make the forcing function impulsive (i.e., we *ping* the sysem), then provided we can solve Equation 5.0.2, we can solve it for any f. Since our mathematical model of an impulsive force is a Dirac delta function, the claim is that if we can solve:

$$L(g) = \delta(t) \tag{5.0.3}$$

then we can use this solution to solve Equation 5.0.2 for *any* right-hand side. Now, I'm not trying to pull any tricks by using g instead of u here. The symbols g or u are just labels. The interpretation of these functions depends on the structure of L. If L is the wave equation operator then the solutions of Equation 5.0.2 are waves excited by the right-hand; we can call the solution whatever we want. But since we are claiming that the solution for a delta function force is special, it helps to have a special symbol for the solution in this case. So g is the response of our system to an impulsive force: this is why it is known as the *impulse response*.

So $g(t)$ will denote the solution of Equation 5.0.2 for a delta function right-hand side. Now I have to show you that this is useful. Let us summarize the situation as follows:

$$\begin{aligned} L(u(t)) &= f(t) \\ L(g(t)) &= \delta(t). \end{aligned} \tag{5.0.4}$$

Now suppose we convolve the force f with the impulse response. Let

$$h \equiv f \star g.$$

I claim that h is what we're after, namely a solution to the general problem Equation 5.0.2. The proof is easy, but subtle; all we have to do is operate on h with L and we must end up with f.

$$L[h(t)] = L\left[\int f(t-\tau)g(\tau)d\tau\right]. \tag{5.0.5}$$

Here is the trick. I know what happens when I apply L to g; I get a delta function. But on the left side of this equation L is operating on a function of t. So it must do this on the right side too. τ is just the dummy integration variable. So if I change variables of integration I can switch the t dependence to g:

$$\begin{aligned} L\left[h(t)\right] &= L\left[\int_{-\infty}^{\infty} f(t-\tau)g(\tau)\right]d\tau & (5.0.6)\\ &= L\left[\int_{-\infty}^{\infty} f(t')g(t-t')dt'\right] & (5.0.7)\\ &= \int_{-\infty}^{\infty} f(t')L\left[g(t-t')\right]dt' & (5.0.8)\\ &= \int_{-\infty}^{\infty} f(t')\delta(t-t')dt' & (5.0.9)\\ &= f(t). & (5.0.10) \end{aligned}$$

Going from the first equation to the second is just a change of variables. Going from the second to the third depends on our understanding that L is operating only on functions of t. Going from the third to the fourth equation is simply using the fact that since g is the impulse response of the sytem, L acting on g must produce a delta function. Finally we use the basic property of the delta function and end up with the statement that $L[h(t)] = f(t)$. This means that h is a solution to our basic Equation 5.0.2, which is what we set out to prove. So if we know the impulse reponse of *any* linear system (not just for the wave equation), we can obtain the solution for an arbitrary forcing function by simply convolving the forcing function with the impulse response.

Bibliography

[1] R. G. Bartle. *The elements of real analysis*. Wiley, 1976.

[2] W. E. Boyce and R. C. DiPrima. *Elementary differential equations and boundary value problems*. Wiley, 1977.

[3] R. Feynman, M. Sands, and R. Leighton. *Lectures on Physics*. Addison Wesley, 1994.

[4] N. B. Haaser and J. A. Sullivan. *Real analysis*. Van Nostrand Reinhold, 1971.

[5] M. A. Heald and J. B. Marion. *Classical electromagnetic radiation*. Saunders College Publishing, 1995.

[6] G. Strang. *Linear algebra and its application*. Saunders College Publishing, Fort Worth, 1988.