Applied Mathematics Lecture Notes

Peter J. Olver and Chehrzad Shakiban

Last Updated: May 1, 2006

OBSAH

Chapter 11

Boundary Value Problems in One Dimension

While its roots are firmly planted in the finite-dimensional world of matrices and vectors, the full scope of linear algebra is much broader. Its historical development and, hence, its structures, concepts, and methods, were strongly influenced by linear analysis — specifically, the need to solve linear differential equations, linear boundary value problems, linear integral equations, and the like. The time has come for us to fully transcend our finite dimensional limitations, and see linear algebra in action in infinite-dimensional function spaces.

In this chapter, we begin to analyze problems arising in continuum physics. The equilibrium equation of a one-dimensional continuum — an elastic bar, a bendable beam, and so on — is formulated as a boundary value problem for a scalar ordinary differential equation. The framework introduced for discrete mechanical systems in Chapter 6 will carry over, in its essence, to the infinite-dimensional setting appropriate to such problems. The underlying Euclidean vector space $\mathbb{R}^n$ is replaced by a function space. Vectors become functions, while matrices turn into linear differential operators. Physical boundary value problems are based on self-adjoint boundary value problems, founded on a suitable inner product on the function space. As in the discrete context, the positive definite cases are stable, and the equilibrium solution can be characterized by a minimization principle based on a quadratic energy functional.

Finite-dimensional linear algebra not only provides us with important insights into the underlying mathematical structure, but also motivates basic analytical and numerical solution schemes. In the function space framework, the general superposition principle is reformulated in terms of the effect of a combination of impulse forces concentrated at a single point of the continuum. However, constructing a function that represents a concentrated impulse turns out to be a highly non-trivial mathematical issue. Ordinary functions do not suffice, and we are led to develop a new calculus of generalized functions or distributions, including the remarkable delta function. The response of the system to a unit impulse force is known as the Green's function of the boundary value problem, in honor of the self-taught English mathematician (and miller) George Green. With the Green's function in hand, the general solution to the inhomogeneous system can be reconstructed by superimposing the effects of suitably scaled impulses on the entire domain. Understanding this construction will become increasingly important as we progress on to partial differential equations, where direct analytical solution techniques are far harder to come by.We begin with second order boundary value problems describing the equilibria of stretchable bars. We continue on to fourth order boundary value problems that govern the equilibrium of elastic beams, including piecewise cubic spline interpolants that play a key role in modern

computer graphics and numerical analysis, and to more general second order boundary value problems of Sturm–Liouville type, which arise in a host of physical applications that involve partial differential equations.

The simplest boundary value problems can be solved by direct integration. However, more complicated systems do not admit explicit formulae for their solutions, and one must rely on numerical approximations. In the final section, we introduce the powerful finite element method. The key idea is to restrict the infinite-dimensional minimization principle characterizing the exact solution to a suitably chosen finite-dimensional subspace of the function space. When properly formulated, the solution to the resulting finite-dimensional minimization problem approximates the true minimizer. As in Chapter 4, the finite-dimensional minimizer is found by solving the induced linear algebraic system, using either direct or iterative methods.

An alternative formulation of the finite element solution, that can be applied even in situations where there is no minimum principle available, is based on the idea of a weak solution to the boundary value problem, where one relaxes the classical differentiability requirements.

11.1. Elastic Bars.

A *bar* is a mathematical idealization of a one-dimensional linearly elastic continuum that can be stretched or contracted in the longitudinal direction, but is not allowed to bend in a transverse direction. (Materials that can bend are called beams, and will be analyzed in Section 11.4.) We will view the bar as the continuum limit of a one-dimensional chain of masses and springs, a system that we already analyzed in Section 6.1. Intuitively, the continuous bar consists of an infinite number of masses connected by infinitely short springs. The individual masses can be thought of as the "atoms" in the bar, although one should not try to read too much into the physics behind this interpretation.

We shall derive the basic equilibrium equations for the bar from first principles. Recall the three basic steps we already used to establish the corresponding equilibrium equations for a discrete mechanical system such as a mass–spring chain:

(*i*) First, use geometry to relate the displacement of the masses to the elongation in the connecting springs.

(*ii*) Second, use the constitutive assumptions such as Hooke's Law to relate the strain to the stress or internal force in the system.

(*iii*) Finally, impose a force balance between external and internal forces.

The remarkable fact, which will, when suitably formulated, carry over to the continuum, is that the force balance law is directly related to the geometrical displacement law by a transpose or, more correctly, adjoint operation.

Consider a bar of length ℓ hanging from a fixed support, with the bottom end left free, as illustrated in Figure 11.1. We use $0 \le x \le \ell$ to refer to the reference or unstressed configuration of the bar, so x measures the distance along the bar from the fixed end $x = 0$ to the free end $x = \ell$. Note that we are adopting the convention that the positive x axis points *down*. Let $u(x)$ denote the *displacement* of the bar from its reference configuration. This means that the "atom" that started at position x has moved to position $x + u(x)$.

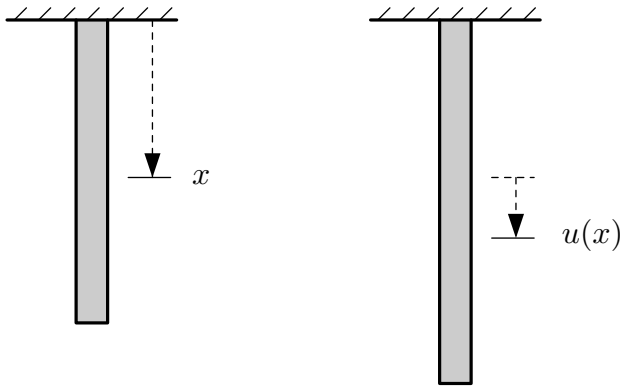


Figure 11.1. Bar with One Fixed Support.

With our convention, $u(x) > 0$ means that the atom has moved down, while if $u(x) < 0$, the atom has moved up. In particular,

$$u(0) = 0 \tag{11.1}$$

because we are assuming that the top end is fixed and cannot move.

The *strain* in the bar measures the relative amount of stretching. Two nearby atoms, at respective positions x and $x+\Delta x$, are moved to positions $x+u(x)$ and $x+\Delta x+u(x+\Delta x)$. The original, unstressed length of this small section of bar was Δx, while in the new configuration the same section has length

$$\big[\, x + \Delta x + u(x + \Delta x) \,\big] - \big[\, x + u(x) \,\big] = \Delta x + \big[\, u(x + \Delta x) - u(x) \,\big].$$

Therefore, this segment has been elongated by an amount $u(x + \Delta x) - u(x)$. The dimensionless strain measures the relative elongation, and so is obtained by dividing by the reference length: $[u(x + \Delta x) - u(x)]/\Delta x$. We now take the continuum limit by letting the interatomic spacing $\Delta x \to 0$. The result is the strain function

$$v(x) = \lim_{\Delta x \to 0} \frac{u(x + \Delta x) - u(x)}{\Delta x} = \frac{du}{dx} \tag{11.2}$$

that measures the local stretch in the bar at position x.

As noted above, we can approximate the bar by a chain of n masses connected by n springs, letting the bottom mass hang free. The mass–spring chain should also have total length ℓ, and so the individual springs have reference length

$$\Delta x = \frac{\ell}{n}\,.$$

The bar is to be viewed as the *continuum limit*, in which the number of masses $n \to \infty$ and the spring lengths $\Delta x \to 0$. The k^{th} mass starts out at position

$$x_k = k\,\Delta x = \frac{k\,\ell}{n}\,,$$

and, under forcing, experiences a displacement u_k. The relative elongation of the k^{th} spring[†] is

$$v_k = \frac{e_k}{\Delta x} = \frac{u_{k+1} - u_k}{\Delta x}. \tag{11.3}$$

In particular, since the fixed end cannot move, the first value $u_0 = 0$ is omitted from the subsequent equations.

The relation (11.3) between displacement and strain takes the familiar matrix form

$$\mathbf{v} = A\,\mathbf{u}, \qquad \mathbf{v} = \big(\, v_0, v_1, \dots, v_{n-1} \,\big)^T, \qquad \mathbf{u} = \big(\, u_1, u_2, \dots, u_n \,\big)^T,$$

where

$$A = \frac{1}{\Delta x} \begin{pmatrix} 1 & & & & & \\ -1 & 1 & & & & \\ & -1 & 1 & & & \\ & & -1 & 1 & & \\ & & & \ddots & \ddots & \\ & & & & -1 & 1 \end{pmatrix} \quad \longrightarrow \quad \frac{d}{dx} \tag{11.4}$$

is the *scaled incidence matrix* of the mass–spring chain. As indicated by the arrow, the derivative operator d/dx that relates displacement to strain in the bar equation (11.2) can be viewed as its continuum limit, as the number of masses $n \to \infty$ and the spring lengths $\Delta x \to 0$. Vice versa, the incidence matrix can be viewed as a discrete, numerical approximation to the derivative operator. Indeed, if we regard the discrete displacements and strains as approximations to the sample values of their continuous counterparts, so

$$u_k \approx u(x_k), \qquad v_k \approx v(x_k),$$

then (11.3) takes the form

$$v(x_k) \;=\; \frac{u(x_{k+1}) - u(x_k)}{\Delta x} \;=\; \frac{u(x_k + \Delta x) - u(x_k)}{\Delta x} \;\approx\; \frac{du}{dx}\,(x_k).$$

justifying the identification (11.4). The passage back and forth between the discrete and the continuous is the foundation of continuum mechanics — solids, fluids, gases, and plasmas. Discrete models both motivate and provide numerical approximations to continuum systems, which, in turn, simplify and give further insight into the discrete domain.

The next part of the mathematical framework is to use the constitutive relations to relate the strain to the *stress*, or internal force experienced by the bar. To keep matters simple, we shall only consider bars that have a linear relation between stress and strain. For a physical bar, this is a pretty good assumption as long as it is not stretched beyond its elastic limits. Let $w(x)$ denote the stress on the point of the bar that was at reference position x. Hooke's Law implies that

$$w(x) = c(x)\, v(x), \tag{11.5}$$

[†] We will find it helpful to label the springs from $k = 0$ to $k = n - 1$. This will facilitate comparisons with the bar, which, by convention, starts at position $x_0 = 0$.

where $c(x)$ measures the stiffness of the bar at position x. For a homogeneous bar, made out of a uniform material, $c(x) \equiv c$ is a constant function. The constitutive function $c(x)$ can be viewed as the continuum limit of the diagonal matrix

$$C = \begin{pmatrix} c_0 & & & \\ & c_1 & & \\ & & \ddots & \\ & & & c_{n-1} \end{pmatrix}$$

of individual spring constants c_k appearing in the discrete version

$$w_k = c_k\, v_k, \qquad \text{or} \qquad \mathbf{w} = C\,\mathbf{v}, \tag{11.6}$$

that relates stress to strain (internal force to elongation) in the individual springs. Indeed, (11.6) can be identified as the sampled version, $w(x_k) = c(x_k)\, v(x_k)$, of the continuum relation (11.5).

Finally, we need to impose a force balance at each point of the bar. Suppose $f(x)$ is an external force at position x on the bar, where $f(x) > 0$ means the force is acting downwards. Physical examples include mechanical, gravitational, or magnetic forces acting solely in the vertical direction. In equilibrium[†], the bar will deform so as to balance the external force with its own internal force resulting from stretching. Now, the internal force per unit length on the section of the bar lying between nearby positions x and $x + \Delta x$ is the relative difference in stress at the two ends, namely $[w(x + \Delta x) - w(x)]/\Delta x$. The force balance law requires that, in the limit,

$$0 = f(x) + \lim_{\Delta x \to 0} \frac{w(x + \Delta x) - w(x)}{\Delta x} = f(x) + \frac{dw}{dx}\,,$$

or

$$f = -\,\frac{dw}{dx}\,. \tag{11.7}$$

This can be viewed as the continuum limit of the mass–spring chain force balance equations

$$f_k = \frac{w_{k-1} - w_k}{\Delta x}, \qquad\qquad w_n = 0, \tag{11.8}$$

where the final condition ensures the correct formula for the force on the free-hanging bottom mass. (Remember that the springs are numbered from 0 to $n - 1$.) This indicates that we should also impose an analogous boundary condition

$$w(\ell) = 0 \tag{11.9}$$

at the bottom end of the bar, which is hanging freely and so is unable to support any internal stress. The matrix form of the discrete system (11.8) is

$$\mathbf{f} = A^T \mathbf{w},$$

[†] The dynamical processes leading to equilibrium will be discussed in Chapter 14.

where the transposed scaled incidence matrix

$$A^T = \frac{1}{\Delta x} \begin{pmatrix} 1 & -1 & & & & & \\ & 1 & -1 & & & & \\ & & 1 & -1 & & & \\ & & & 1 & -1 & & \\ & & & & \ddots & \ddots & \end{pmatrix} \longrightarrow \; - \, \frac{d}{dx} \tag{11.10}$$

should approximate the differential operator $-d/dx$ that appears in the continuum force balance law (11.7). Thus, we should somehow interpret $-d/dx$ as the "transpose" or "adjoint" of the differential operator d/dx. This important point will be developed properly in Section 11.3. But before trying to push the theory any further, we will pause to analyze the mathematical equations governing some simple configurations.

But first, let us summarize our progress so far. The three basic equilibrium equations (11.2, 5, 7) are

$$v(x) = \frac{du}{dx}\,, \qquad\qquad w(x) = c(x)\,v(x), \qquad\qquad f(x) = -\,\frac{dw}{dx}\,. \tag{11.11}$$

Substituting the first into the second, and then the resulting formula into the last equation, leads to the equilibrium equation

$$K[u] = -\,\frac{d}{dx}\left(c(x)\,\frac{du}{dx}\right) = f(x), \qquad\qquad 0 < x < \ell. \tag{11.12}$$

Thus, the displacement $u(x)$ of the bar is obtained as the solution to a second order ordinary differential equation. As such, it will depend on two arbitrary constants, which will be uniquely determined by the boundary conditions† (11.1, 9) at the two ends:

$$u(0) = 0, \qquad\qquad w(\ell) = c(\ell)\,u'(\ell) = 0. \tag{11.13}$$

Usually $c(\ell) > 0$, in which case it can be omitted from the second boundary condition, which simply becomes $u'(\ell) = 0$. The resulting boundary value problem is to be viewed as the continuum limit of the linear system

$$K\,\mathbf{u} = A^T C\,A\,\mathbf{u} = \mathbf{f} \tag{11.14}$$

modeling a mass-spring chain with one free end, cf. (6.11), in which

$$A \;\longrightarrow\; \frac{d}{dx}\,, \qquad C \;\longrightarrow\; c(x), \qquad A^T \;\longrightarrow\; -\,\frac{d}{dx}\,, \qquad \mathbf{u} \;\longrightarrow\; u(x), \qquad \mathbf{f} \;\longrightarrow\; f(x).$$

And, as we will see, most features of the finite-dimensional linear algebraic system have, when suitably interpreted, direct counterparts in the continuous boundary value problem.

† We will sometimes use primes, as in $u' = du/dx$, to denote derivatives with respect to x.

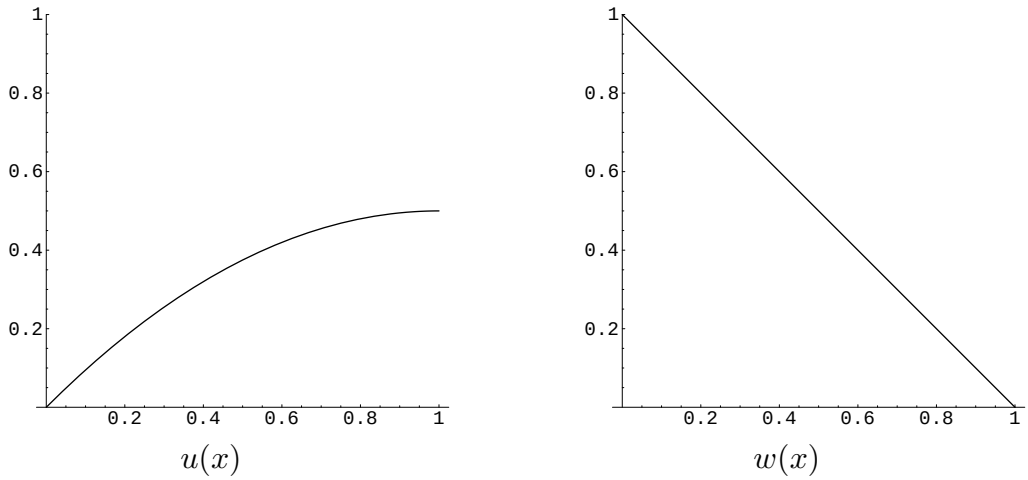


Figure 11.2. Displacement and Stress of Bar with One Fixed End.

Example 11.1. Consider the simplest case of a uniform bar of unit length $\ell = 1$ subjected to a uniform force, e.g., gravity. The equilibrium equation (11.12) is

$$-c\,\frac{d^2u}{dx^2} = f, \tag{11.15}$$

where we are assuming that the force f is constant. This elementary second order ordinary differential equation can be immediately integrated:

$$u(x) = -\tfrac{1}{2}\,\alpha\,x^2 + a\,x + b, \qquad \text{where} \qquad \alpha = \frac{f}{c} \tag{11.16}$$

is the ratio of the force to the stiffness of the bar. The values of the integration constants a and b are fixed by the boundary conditions (11.13), so

$$u(0) = b = 0, \qquad u'(1) = -\alpha + a = 0.$$

Therefore, there is a unique solution to the boundary value problem, yielding the displacement

$$u(x) = \alpha\big(\,x - \tfrac{1}{2}\,x^2\,\big), \tag{11.17}$$

which is graphed in Figure 11.2 for $\alpha = 1$. Note the parabolic shape, with zero derivative, indicating no strain, at the free end. The displacement reaches its maximum, $u(1) = \frac{1}{2}\,\alpha$, at the free end of the bar, which is the point which moves downwards the farthest. The stronger the force or the weaker the bar, the farther the overall displacement.

Remark: This example illustrates the simplest way to solve boundary value problems, which is adapted from the usual solution technique for initial value problems. First, solve the differential equation by standard methods (if possible). For a second order equation, the general solution will involve two arbitrary constants. The values of the constants are found by substituting the solution formula into the two boundary conditions. Unlike initial value problems, the existence and/or uniqueness of the solution to a general boundary value problem is not guaranteed, and you may encounter situations where you are unable to complete the solution; see, for instance, Example 7.42. A more sophisticated method, based on the Green's function, will be presented in the following section.

As in the discrete situation, this particular mechanical configuration is *statically determinate*, meaning that we can solve directly for the stress $w(x)$ in terms of the external force $f(x)$ without having to compute the displacement $u(x)$ first. In this particular example, we need to solve the first order boundary value problem

$$-\frac{dw}{dx} = f, \qquad w(1) = 0,$$

arising from the force balance law (11.7). Since f is constant,

$$w(x) = f\,(1-x), \qquad \text{and} \qquad v(x) = \frac{w(x)}{c} = \alpha\,(1-x).$$

Note that the boundary condition uniquely determines the integration constant. We can then find the displacement $u(x)$ by solving another boundary value problem

$$\frac{du}{dx} = v(x) = \alpha\,(1-x), \qquad u(0) = 0,$$

resulting from (11.2), which again leads to (11.17). As before, the appearance of one boundary condition implies that we can find a unique solution to the differential equation.

Remark: We motivated the boundary value problem for the bar by taking the continuum limit of the mass–spring chain. Let us see to what extent this limiting procedure can be justified. To compare the solutions, we keep the reference length of the chain fixed at $\ell = 1$. So, if we have n identical masses, each spring has length $\Delta x = 1/n$. The k^{th} mass will start out at reference position $x_k = k/n$. Using static determinacy, we can solve the system (11.8), which reads

$$w_{k-1} = w_k + \frac{f}{n}\,, \qquad w_n = 0,$$

directly for the stresses:

$$w_k = f\,\left(1 - \frac{k}{n}\right) = f\,(1 - x_k)\,, \qquad k = 0, \dots, n-1.$$

Thus, in this particular case, the continuous bar and the discrete chain have equal stresses at the sample points: $w(x_k) = w_k$. The strains are also in agreement:

$$v_k = \frac{1}{c}\,w_k = \alpha\left(1 - \frac{k}{n}\right) = \alpha\,(1 - x_k) = v(x_k)\,,$$

where $\alpha = f/c$, as before. We then obtain the displacements by solving

$$u_{k+1} = u_k + \frac{v_k}{n} = u_k + \frac{\alpha}{n}\left(1 - \frac{k}{n}\right).$$

Since $u_0 = 0$, the solution is

$$u_k = \frac{\alpha}{n}\sum_{i=0}^{k-1}\left(1 - \frac{i}{n}\right) = \alpha\left(\frac{k}{n} - \frac{k(k-1)}{2\,n^2}\right) = \alpha\left(\,x_k - \tfrac{1}{2}\,x_k^2\,\right) + \frac{\alpha\,x_k}{2\,n} = u(x_k) + \frac{\alpha\,x_k}{2\,n}\,. \tag{11.18}$$

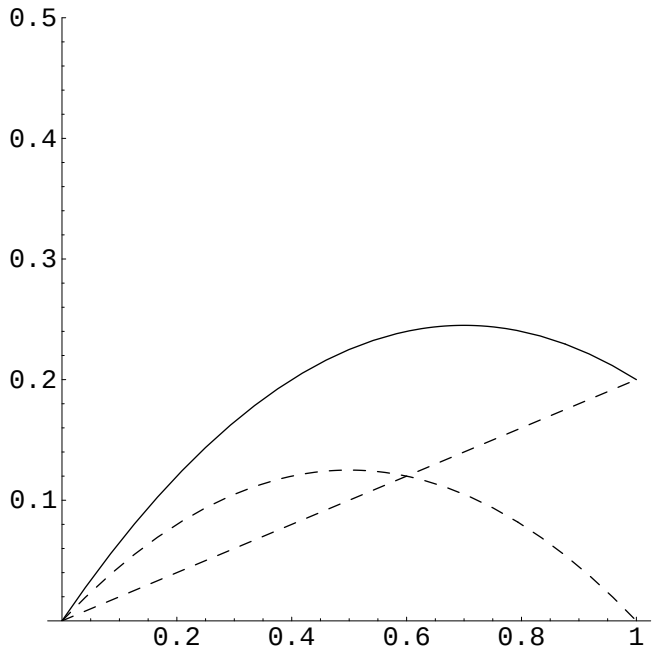


Figure 11.3. Displacements of a Bar with Two Fixed Ends.

The sampled displacement $u(x_k)$ is not exactly equal to u_k, but their difference tends to zero as the number of masses $n \to \infty$. In this way, we have completely justified our limiting interpretation.

Example 11.2. Consider the same uniform, unit length bar as in the previous example, again subject to a uniform constant force, but now with two fixed ends. We impose inhomogeneous boundary conditions

$$u(0) = 0, \qquad u(1) = d,$$

so the top end is fixed, while the bottom end is displaced an amount d. (Note that $d > 0$ means the bar is stretched, while $d < 0$ means it is compressed.) The general solution to the equilibrium equation (11.15) is, as before, given by (11.16). The values of the arbitrary constants a, b are determined by plugging into the boundary conditions, so

$$u(0) = b = 0, \qquad u(1) = -\tfrac{1}{2}\,\alpha + d = 0.$$

Thus

$$u(x) = \tfrac{1}{2}\,\alpha\,(x - x^2) + d\,x \tag{11.19}$$

is the unique solution to the boundary value problem, The displacement is a linear superposition of two functions; the first is induced by the external force f, while the second represents a uniform stretch induced by the boundary condition. In Figure 11.3, the dotted curves represent the two constituents, and the solid graph is their sum, which is the actual displacement.

Unlike a bar with a free end, this configuration is *statically indeterminate*. There is

no boundary condition on the force balance equation

$$-\frac{dw}{dx} = f,$$

and so the integration constant a in the stress $w(x) = a - f\,x$ *cannot* be determined without first figuring out the displacement (11.19):

$$w(x) = c\,\frac{du}{dx} = f\left(\tfrac{1}{2} - x\right) + c\,d.$$

Example 11.3. Finally, consider the case when both ends of the bar are left free. The boundary value problem

$$-u'' = f(x), \qquad u'(0) = 0, \qquad u'(\ell) = 0, \tag{11.20}$$

represents the continuum limit of a mass–spring chain with two free ends and corresponds to a bar floating in outer space, subject to a nonconstant external force. Based on our finite-dimensional experience, we expect the solution to manifest an underlying instability of the physical problem. Solving the differential equation, we find that

$$u(x) = a\,x + b - \int_0^x \left(\int_0^y f(z)\,dz \right) dy,$$

where the constants a, b are to be determined by the boundary conditions. Since

$$u'(x) = a - \int_0^x f(z)\,dz,$$

the first boundary condition $u'(0) = 0$ requires $a = 0$. The second boundary condition requires

$$u'(\ell) = \int_0^\ell f(x)\,dx = 0, \tag{11.21}$$

which is not automatically valid! The integral represents the total force per unit length exerted on the bar. As in the case of a mass-spring chain with two free ends, if there is a non-zero net force, the bar cannot remain in equilibrium, but will move off in space and the equilibrium boundary value problem has no solution. On the other hand, if the forcing satisfies the constraint (11.21), then the resulting solution of the boundary value problem has the form

$$u(x) = b - \int_0^x \left(\int_0^y f(z)\,dz \right) dy, \tag{11.22}$$

where the constant b is arbitrary. Thus, when it exists, the solution to the boundary value problem is not unique. The constant b solves the corresponding homogeneous problem, and represents a rigid translation of the entire bar by a distance b.

Physically, the free boundary value problem corresponds to an unstable structure: there is a translational instability in which the bar moves off rigidly in the longitudinal direction. Only balanced forces of mean zero can maintain equilibrium. Furthermore,

when it does exist, the equilibrium solution is not unique since there is nothing to tie the bar down to any particular spatial position.

This dichotomy should remind you of our earlier study of linear algebraic systems. An inhomogeneous system $K\mathbf{u} = \mathbf{f}$ consisting of n equations in n unknowns either admits a unique solution for all possible right hand sides $\mathbf{f}$, or, when K is singular, either no solution exists or the solution is not unique. In the latter case, the constraints on the right hand side are prescribed by the Fredholm alternative (5.78), which requires that $\mathbf{f}$ be orthogonal, with respect to the Euclidean inner product, to all elements of $\operatorname{coker} K$. In physical equilibrium systems K is symmetric, and so $\operatorname{coker} K = \ker K$. Thus, the Fredholm alternative requires that the forcing be orthogonal to all the unstable modes. In the function space for the bar, the finite-dimensional dot product is replaced by the L^2 inner product

$$\langle\, f\,, g\,\rangle = \int_0^\ell f(x)\, g(x)\, dx.$$

Since the kernel or solution space to the homogeneous boundary value problem is spanned by the constant function 1, the Fredholm alternative† requires that the forcing function be orthogonal to it, $\langle\, f\,, 1\,\rangle = \displaystyle\int_0^\ell f(x)\, dx = 0$. This is precisely the condition (11.21) required for existence of a (non-unique) equilibrium solution, and so the analogy between the finite and infinite dimensional categories is complete.

Remark: The boundary value problems that govern the mechanical equilibria of a simple bar arise in many other physical systems. For example, the equation for the thermal equilibrium of a bar under an external heat source is modeled by the same boundary value problem (11.12); in this case, $u(x)$ represents the temperature of the bar, $c(x)$ represents the *diffusivity* or *thermal conductivity* of the material at position x, while $f(x)$ represents an external heat source. A fixed boundary condition $u(\ell) = a$ corresponds to an end that is held at a fixed temperature a, while a free boundary condition $u'(\ell) = 0$ represents an insulated end that does not allow heat energy to enter or leave the bar. Details of the physical derivation can be found in Section 14.1.

11.2. Generalized Functions and the Green's Function.

The general superposition principle for inhomogeneous linear systems inspires an alternative, powerful approach to the solution of boundary value problems. This method relies on the solution to a particular type of inhomogeneity, namely a concentrated unit impulse. The resulting solutions are collectively known as the Green's function for the boundary value problem. Once the Green's function is known, the response of the system to any other external forcing can be constructed through a continuous superposition of these fundamental solutions. However, rigorously formulating a concentrated impulse force turns out to be a serious mathematical challenge.

† In fact, historically, Fredholm first made his discovery while studying such infinite-dimensional systems. Only later was it realized that the same underlying ideas are equally valid in finite-dimensional linear algebraic systems.

To motivate the construction, let us return briefly to the case of a mass–spring chain. Given the equilibrium equations

$$K\,\mathbf{u} = \mathbf{f}, \tag{11.23}$$

let us decompose the external forcing $\mathbf{f} = (\,f_1, f_2, \ldots, f_n\,)^T \in \mathbb{R}^n$ into a linear combination

$$\mathbf{f} = f_1\,\mathbf{e}_1 + f_2\,\mathbf{e}_2 + \,\cdots\, + f_n\,\mathbf{e}_n \tag{11.24}$$

of the standard basis vectors of $\mathbb{R}^n$. Suppose we know how to solve each of the individual systems

$$K\,\mathbf{u}_i = \mathbf{e}_i, \qquad\qquad i = 1, \ldots, n. \tag{11.25}$$

The vector $\mathbf{e}_i$ represents a unit force or, more precisely, a *unit impulse*, which is applied solely to the i^{th} mass in the chain; the solution $\mathbf{u}_i$ represents the response of the chain. Since we can decompose any other force vector as a superposition of impulse forces, as in (11.24), the superposition principle tells us that the solution to the inhomogeneous system (11.23) is the self-same linear combination of the individual responses, so

$$\mathbf{u} = f_1\,\mathbf{u}_1 + f_2\,\mathbf{u}_2 + \,\cdots\, + f_n\,\mathbf{u}_n. \tag{11.26}$$

Remark: The alert reader will recognize that $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are the columns of the inverse matrix, K^{-1}, and so we are, in fact, reconstructing the solution to the linear system (11.23) by inverting the coefficient matrix K. Thus, this observation, while noteworthy, does not lead to an efficient solution technique for discrete systems. In contrast, in the case of continuous boundary value problems, this approach leads to one of the most valuable solution paradigms in both practice and theory.

The Delta Function

Our aim is to extend this algebraic technique to boundary value problems. The key question is how to characterize an impulse force that is concentrated on a single atom[†] of the bar. In general, a *unit impulse* at position $x = y$ will be described by something called the *delta function*, and denoted by $\delta_y(x)$. Since the impulse is supposed to be concentrated solely at $x = y$, we should have

$$\delta_y(x) = 0 \qquad \text{for} \qquad x \neq y. \tag{11.27}$$

Moreover, since it is a *unit* impulse, we want the total amount of force exerted on the bar to be equal to one. Since we are dealing with a continuum, the total force is represented by an integral over the length of the bar, and so we also require that the delta function to satisfy

$$\int_0^\ell \delta_y(x)\,dx = 1, \qquad \text{provided that} \qquad 0 < y < \ell. \tag{11.28}$$

Alas, there is no *bona fide* function that enjoys both of the required properties! Indeed, according to the basic facts of Riemann (or even Lebesgue) integration, two functions

[†] As before, "atom" is used in a figurative sense.

which are the same everywhere except at one single point have exactly the same integral, [**45**, **141**]. Thus, since δ_y is zero except at one point, its integral should be 0, not 1. The mathematical conclusion is that the two requirements, (11.27, 28) are inconsistent!

This unfortunate fact stopped mathematicians dead in their tracks. It took the imagination of a British engineer, with the unlikely name Oliver Heaviside, who was not deterred by the lack of rigorous justification, to start utilizing delta functions in practical applications — with remarkable effect. Despite his success, Heaviside was ridiculed by the pure mathematicians of his day, and eventually succumbed to mental illness. But, some thirty years later, the great theoretical physicist Paul Dirac resurrected the delta function for quantum mechanical applications, and this finally made theoreticians sit up and take notice. (Indeed, the term "Dirac delta function" is quite common.) In 1944, the French mathematician Laurent Schwartz finally established a rigorous theory of *distributions* that incorporated such useful, but non-standard objects, [**104**, **140**]. (Thus, to be more accurate, we should really refer to the *delta distribution*; however, we will retain the more common, intuitive designation "delta function" throughout.) It is beyond the scope of this introductory text to develop a fully rigorous theory of distributions. Rather, in the spirit of Heaviside, we shall concentrate on learning, through practice with computations and applications, how to tame these wild mathematical beasts.

There are two distinct ways to introduce the delta function. Both are important and both worth knowing.

Method #1. Limits: The first approach is to regard the delta function $\delta_y(x)$ as a limit of a sequence of ordinary smooth functions[†] $g_n(x)$. These functions will represent more and more concentrated unit forces, which, in the limit, converge to the desired unit impulse concentrated at a single point, $x = y$. Thus, we require

$$\lim_{n \to \infty} g_n(x) = 0, \qquad\qquad x \neq y, \tag{11.29}$$

while the total amount of force remains fixed at

$$\int_0^\ell g_n(x)\, dx = 1. \tag{11.30}$$

On a formal level, the limit "function"

$$\delta_y(x) = \lim_{n \to \infty} g_n(x)$$

will satisfy the key properties (11.27–28).

An explicit example of such a sequence is provided by the rational functions

$$g_n(x) = \frac{n}{\pi\big(1 + n^2 x^2\big)}\,. \tag{11.31}$$

[†] To keep the notation compact, we suppress the dependence of the functions g_n on the point y where the limiting delta function is concentrated.

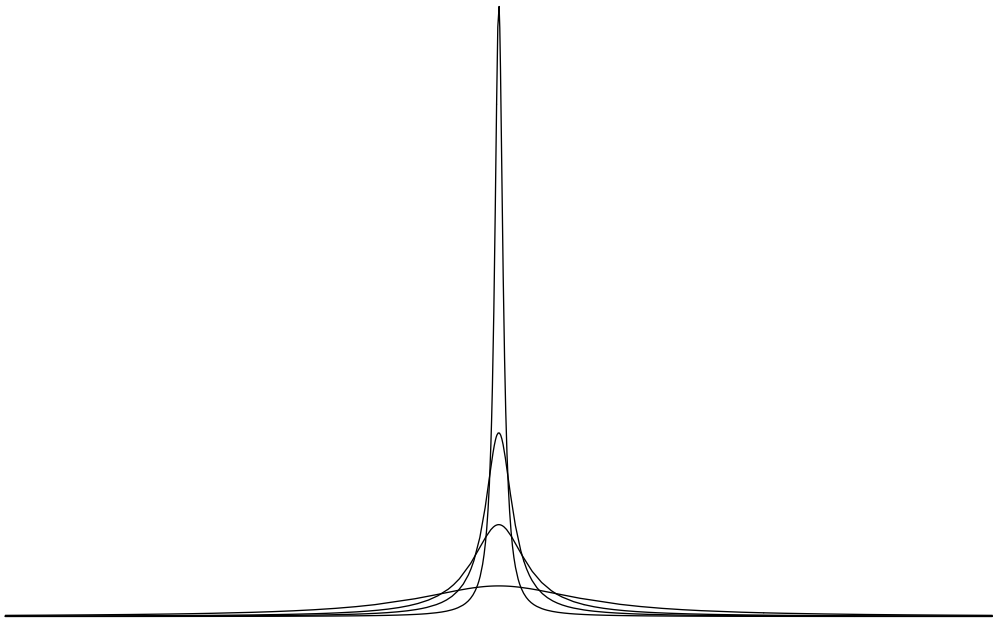


Figure 11.4. Delta Function as Limit.

These functions satisfy

$$\lim_{n \to \infty} g_n(x) = \begin{cases} 0, & x \neq 0, \\ \infty, & x = 0, \end{cases} \tag{11.32}$$

while[‡]

$$\int_{-\infty}^{\infty} g_n(x)\, dx = \frac{1}{\pi} \tan^{-1} n\, x \,\Big|_{x=-\infty}^{\infty} = 1. \tag{11.33}$$

Therefore, formally, we identify the limiting function

$$\lim_{n \to \infty} g_n(x) = \delta(x) = \delta_0(x),$$

with the unit impulse delta function concentrated at $x = 0$. As sketched in Figure 11.4, as n gets larger and larger, each successive function $g_n(x)$ forms a more and more concentrated spike, while maintaining a unit total area under its graph. The limiting delta function can be thought of as an infinitely tall spike of zero width, entirely concentrated at the origin.

Remark: There are many other possible choices for the limiting functions $g_n(x)$. See Exercise ■ for another useful example.

Remark: This construction of the delta function highlights the perils of interchanging limits and integrals without proper justification. In any standard theory of integration,

[‡] For the moment, it will be slightly simpler here to consider the entire real line — corresponding to a bar of infinite length. Exercise ■ discusses how to modify the construction for a finite interval.

the limit of the functions g_n would be indistinguishable from the zero function, so the limit of their integrals (11.33) would *not* equal the integral of their limit:

$$1 = \lim_{n \to \infty} \int_{-\infty}^{\infty} g_n(x)\, dx \;\neq\; \int_{-\infty}^{\infty} \lim_{n \to \infty} g_n(x)\, dx = 0.$$

The delta function is, in a sense, a means of sidestepping this analytic inconvenience. The full ramifications and theoretical constructions underlying such limits must, however, be deferred to a rigorous course in real analysis, [**45**, **141**].

Once we have found the basic delta function $\delta(x) = \delta_0(x)$, which is concentrated at the origin, we can obtain a delta function concentrated at any other position y by a simple translation:

$$\delta_y(x) = \delta(x - y). \tag{11.34}$$

Thus, $\delta_y(x)$ can be realized as the limit of the translated functions

$$\widehat{g}_n(x) = g_n(x - y) = \frac{n}{\pi \left(1 + n^2 (x - y)^2 \right)}\,. \tag{11.35}$$

Method #2. Duality: The second approach is a bit more abstract, but much closer to the proper rigorous formulation of the theory of distributions like the delta function. The critical observation is that if $u(x)$ is any continuous function, then

$$\int_0^{\ell} \delta_y(x)\, u(x)\, dx = u(y), \qquad \text{for} \qquad 0 < y < \ell. \tag{11.36}$$

Indeed, since $\delta_y(x) = 0$ for $x \neq y$, the integrand only depends on the value of u at the point $x = y$, and so

$$\int_0^{\ell} \delta_y(x)\, u(x)\, dx = \int_0^{\ell} \delta_y(x)\, u(y)\, dx = u(y) \int_0^{\ell} \delta_y(x)\, dx = u(y).$$

Equation (11.36) serves to define a linear functional[†] $L_y \colon \mathrm{C}^0[0, \ell] \to \mathbb{R}$ that maps a continuous function $u \in \mathrm{C}^0[0, \ell]$ to its value at the point $x = y$:

$$L_y[u] = u(y) \in \mathbb{R}.$$

In the dual approach to generalized functions, the delta function is, in fact, *defined* as this particular linear functional. The function $u(x)$ is sometimes referred to as a *test function*, since it serves to "test" the form of the linear functional L.

[†] Linearity, which requires that $L_y[c\, f + d\, g] = c\, L_y[f] + d\, L_y[g]$ for all functions f, g and all scalars (constants) $c, d \in \mathbb{R}$, is easily established; see also Example 7.7.

Remark: If the impulse point y lies outside the integration domain, then

$$\int_0^\ell \delta_y(x)\, u(x)\, dx = 0, \qquad \text{when} \qquad y < 0 \qquad \text{or} \qquad y > \ell, \tag{11.37}$$

because the integrand is identically zero on the entire interval. For technical reasons, we will not attempt to define the integral (11.37) if the impulse point $y = 0$ or $y = \ell$ lies on the boundary of the interval of integration.

The interpretation of the linear functional L_y as representing a kind of function $\delta_y(x)$ is based on the following line of thought. According to Theorem 7.10, every scalar-valued linear function $L\colon V \to \mathbb{R}$ on a finite-dimensional inner product space is given by an inner product with a fixed element $\mathbf{a} \in V$, so

$$L[\mathbf{u}] = \langle\, \mathbf{a}\,, \mathbf{u}\,\rangle.$$

In this sense, linear functions on $\mathbb{R}^n$ are the "same" as vectors. (But bear in mind that the identification does depend upon the choice of inner product.) Similarly, on the infinite-dimensional function space $\mathrm{C}^0[0,\ell]$, the L^2 inner product

$$L_g[u] = \langle\, g\,, u\,\rangle = \int_0^\ell g(x)\, u(x)\, dx \tag{11.38}$$

taken with a fixed function $g \in \mathrm{C}^0[0,\ell]$ defines a real-valued linear functional $L_g\colon \mathrm{C}^0[0,\ell] \to \mathbb{R}$. However, unlike the finite-dimensional situation, *not* every real-valued linear functional has this form! In particular, there is no actual function $\delta_y(x)$ such that the identity

$$\langle\, \delta_y\,, u\,\rangle = \int_0^\ell \delta_y(x)\, u(x)\, dx = u(y) \tag{11.39}$$

holds for every continuous function $u(x)$. Every (continuous) function defines a linear functional, but not conversely. Or, stated another way, while the dual space to a finite-dimensional vector space like $\mathbb{R}^n$ can be identified, via an inner product, with the space itself, this is not the case in infinite-dimensional function space; the dual is an entirely different creature. This disconcerting fact highlights yet another of the profound differences between finite- and infinite-dimensional vector spaces!

But the dual interpretation of generalized functions acts as if this were true. *Generalized functions are real-valued linear functionals on function space, but viewed as a kind of function via the inner product*. Although the identification is not to be taken literally, one can, with a little care, manipulate generalized functions as if they were actual functions, but always keeping in mind that a rigorous justification of such computations must ultimately rely on their true characterization as linear functionals.

The two approaches — limits and duality — are completely compatible. Indeed, with a little extra work, one can justify the dual formula (11.36) as the limit

$$u(y) = \lim_{n \to \infty} \int_0^\ell g_n(x)\, u(x)\, dx = \int_0^\ell \delta_y(x)\, u(x)\, dx \tag{11.40}$$

of the inner products of the function u with the approximating concentrated impulse functions $g_n(x)$ satisfying (11.29–30). In this manner, the linear functional $L[u] = u(y)$ represented by the delta function is the limit, $L_y = \lim_{n \to \infty} L_n$, of the approximating linear functionals

$$L_n[u] = \int_0^\ell g_n(x)\, u(x)\, dx.$$

Thus, the choice of interpretation of the generalized delta function is, on an operational level, a matter of taste. For the novice, the limit interpretation of the delta function is perhaps the easier to digest at first. However, the dual, linear functional interpretation has stronger connections with the rigorous theory and, even in applications, offers some significant advantages.

Although on the surface, the delta function might look a little bizarre, its utility in modern applied mathematics and mathematical physics more than justifies including it in your analytical toolbox. Even though you are probably not yet comfortable with either definition, you are advised to press on and familiarize yourself with its basic properties, to be discussed next. With a little care, you usually won't go far wrong by treating it as if it were a genuine function. After you gain more practical experience, you can, if desired, return to contemplate just exactly what kind of object the delta function really is.

Remark: If you are familiar with basic measure theory, [**141**], there is yet a third interpretation of the delta function as a point mass or atomic measure. However, the measure-theoretic approach has definite limitations, and does not cover the full gamut of generalized functions.

Calculus of Generalized Functions

In order to develop a working relationship with the delta function, we need to understand how it behaves under the basic operations of linear algebra and calculus. First, we can take linear combinations of delta functions. For example,

$$f(x) = 2\,\delta(x) + 3\,\delta(x-1)$$

represents a combination of an impulse of magnitude 2 concentrated at $x = 0$ and one of magnitude 3 concentrated at $x = 1$. Since $\delta_y(x) = 0$ for any $x \neq y$, multiplying the delta function by an ordinary function is the same as multiplying by a constant:

$$g(x)\,\delta_y(x) = g(y)\,\delta_y(x), \tag{11.41}$$

provided that $g(x)$ is continuous at $x = y$. For example, $x\,\delta(x) \equiv 0$ is the same as the constant zero function.

Warning: It is *not* permissible to multiply delta functions together, or to use more complicated algebraic operations. Expressions like $\delta(x)^2$, $1/\delta(x)$, $e^{\delta(x)}$, etc., are *not* well defined in the theory of generalized functions. This makes their application to nonlinear systems much more problematic .

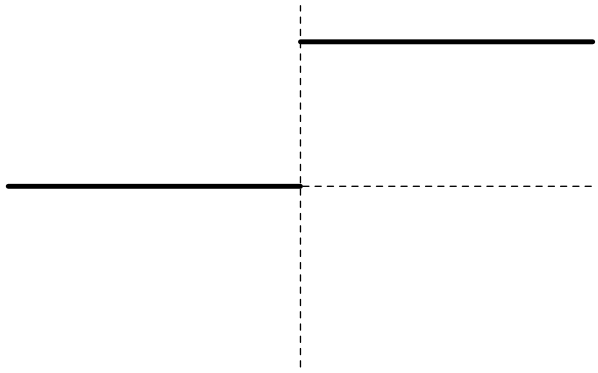


Figure 11.5. The Step Function.

The integral of the delta function is known as a *step function*. More specifically, the basic formulae (11.36, 37) imply that

$$\int_a^x \delta_y(t)\,dt = \sigma_y(x) = \sigma(x-y) = \begin{cases} 0, & a < x < y, \\ 1, & x > y > a. \end{cases} \tag{11.42}$$

Figure 11.5 shows the graph of $\sigma(x) = \sigma_0(x)$. Unlike the delta function, the step function $\sigma_y(x)$ is an ordinary function. It is continuous — indeed constant — except at $x = y$. The value of the step function at the discontinuity $x = y$ is left unspecified, although a popular choice, motivated by Fourier theory, cf. Chapter 12 , is to set $\sigma_y(y) = \frac{1}{2}$, the average of its left and right hand limits.

We note that the integration formula (11.42) is compatible with our characterization of the delta function as the limit of highly concentrated forces. If we integrate the approximating functions (11.31), we obtain

$$f_n(x) = \int_{-\infty}^x g_n(t)\,dt = \frac{1}{\pi}\,\tan^{-1} n\,x + \frac{1}{2}\,.$$

Since

$$\lim_{y \to \infty} \tan^{-1} y = \tfrac{1}{2}\,\pi, \qquad \text{while} \qquad \lim_{y \to -\infty} \tan^{-1} y = -\tfrac{1}{2}\,\pi,$$

these functions converge to the step function:

$$\lim_{n \to \infty} f_n(x) = \sigma(x) = \begin{cases} 1, & x > 0, \\ \frac{1}{2}, & x = 0, \\ 0, & x < 0. \end{cases} \tag{11.43}$$

A graphical illustration of this limiting process appears in Figure 11.6.

The integral of the discontinuous step function (11.42) is the continuous *ramp function*

$$\int_a^x \sigma_y(z)\,dz = \rho_y(x) = \rho(x-y) = \begin{cases} 0, & a < x < y, \\ x - y, & x > y > a, \end{cases} \tag{11.44}$$

which is graphed in Figure 11.7. Note that $\rho(x-y)$ has a corner at $x = y$, and so is not differentiable there; indeed, its derivative $\dfrac{d\rho}{dx} = \sigma$ has a jump discontinuity, and its second

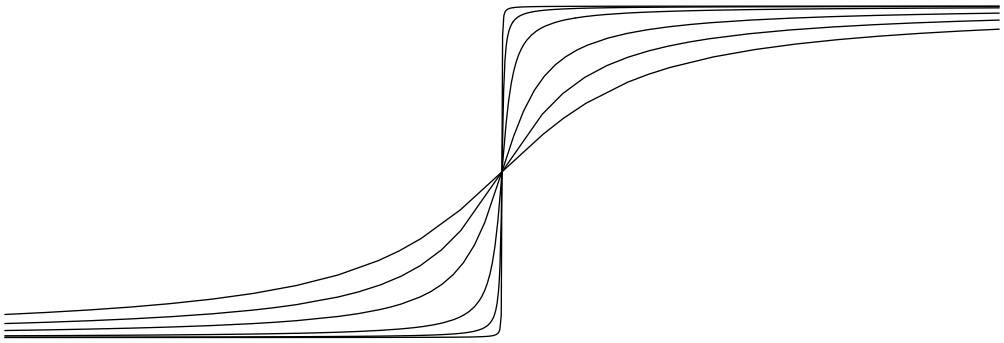


Figure 11.6. Step Function as Limit.

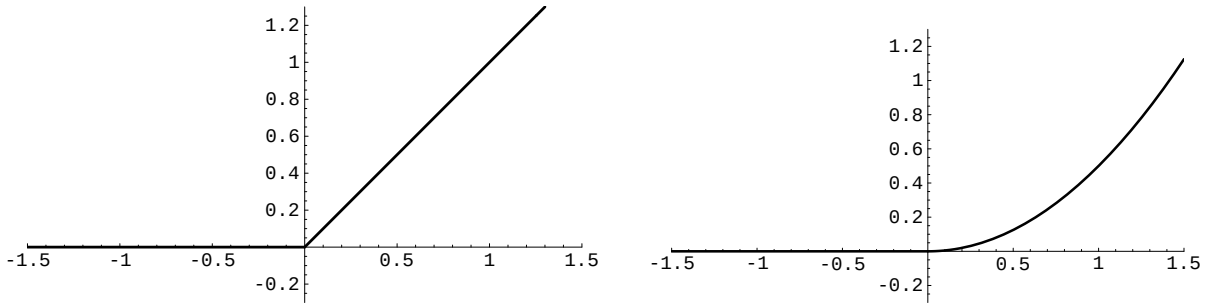


Figure 11.7. First and Second Order Ramp Functions.

derivative $\dfrac{d^2\rho}{dx^2} = \delta$ is no longer an ordinary function. We can continue to integrate; the n^{th} integral of the delta function is the n^{th} *order ramp function*

$$\rho_n(x-y) = \begin{cases} \dfrac{(x-y)^n}{n!}\,, & x > y, \\ 0, & x < y. \end{cases} \tag{11.45}$$

What about differentiation? Motivated by the Fundamental Theorem of Calculus, we shall use formula (11.42) to identify the derivative of the step function with the delta function

$$\frac{d\sigma}{dx} = \delta\,. \tag{11.46}$$

This fact is highly significant. In basic calculus, one is not allowed to differentiate a discontinuous function. Here, we discover that the derivative can be defined, not as an ordinary function, but rather as a generalized delta function.

This basic identity is a particular instance of a general rule for differentiating functions with discontinuities. We use

$$f(y^-) = \lim_{x \to y^-} f(x), \qquad f(y^+) = \lim_{x \to y^+} f(x), \tag{11.47}$$

to denote, respectively, the left and right sided limits of a function at a point y. The function $f(x)$ is *continuous* at the point y if and only if its one-sided limits exist and are equal to its value: $f(y) = f(y^-) = f(y^+)$. If the one-sided limits are the same, but not

equal to $f(y)$, then the function is said to have a *removable discontinuity*, since redefining $f(y) = f(y^-) = f(y^+)$ serves to make f continuous at the point in question. An example is the function $f(x)$ that is equal to 0 for all $x \neq 0$, but has[†] $f(0) = 1$. Removing the discontinuity by setting $f(0) = 0$ makes $f(x) \equiv 0$ equal to a continuous constant function. Since removable discontinuities play no role in our theory or applications, they will always be removed without penalty.

Warning: Although $\delta(0^+) = 0 = \delta(0^-)$, we will emphatically *not* call 0 a removable discontinuity of the delta function. Only standard functions have removable discontinuities.

Finally, if both the left and right limits exist, but are not equal, then f is said to have a *jump discontinuity* at the point y. The *magnitude* of the jump is the difference

$$\beta = f(y^+) - f(y^-) = \lim_{x \to y^+} f(x) \; - \lim_{x \to y^-} f(x) \tag{11.48}$$

between the right and left limits. The magnitude of the jump is positive if the function jumps up, when moving from left to right, and negative if it jumps down. For example, the step function $\sigma(x)$ has a unit, i.e., magnitude 1, jump discontinuity at the origin:

$$\sigma(0^+) - \sigma(0^-) = 1 - 0 = 1,$$

and is continuous everywhere else. Note the value of the function at the point, namely $f(y)$, which may not even be defined, plays no role in the specification of the jump.

In general, the derivative of a function with jump discontinuities is a generalized function that includes delta functions concentrated at each discontinuity. More explicitly, suppose that $f(x)$ is differentiable, in the usual calculus sense, everywhere except at the point y where it has a jump discontinuity of magnitude β. We can re-express the function in the convenient form

$$f(x) = g(x) + \beta\, \sigma(x - y), \tag{11.49}$$

where $g(x)$ is continuous everywhere, with a removable discontinuity at $x = y$, and differentiable except possibly at the jump. Differentiating (11.49), we find that

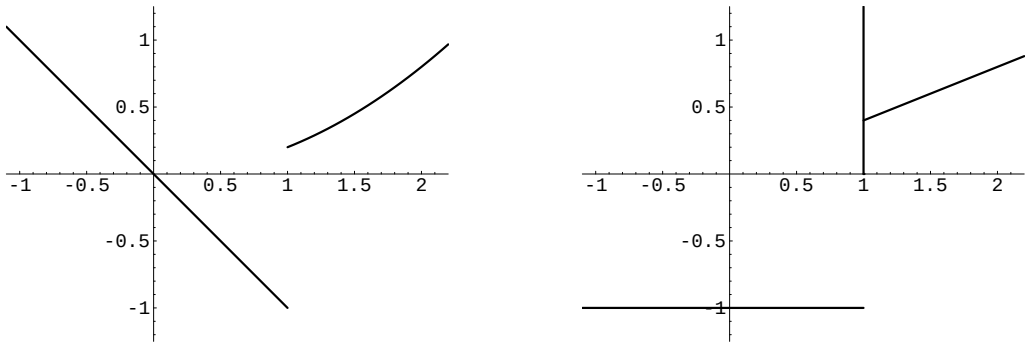
$$f'(x) = g'(x) + \beta\, \delta(x - y), \tag{11.50}$$

has a delta spike of magnitude β at the discontinuity. Thus, the derivatives of f and g coincide everywhere except at the discontinuity.

Example 11.4. Consider the function

$$f(x) = \begin{cases} -x, & x < 1, \\ \frac{1}{5}\, x^2, & x > 1, \end{cases} \tag{11.51}$$

[†] This function is *not* a version of the delta function. It is an ordinary function, and its integral is 0, not 1.



$f(x)$ $\qquad\qquad$ $f'(x)$

Figure 11.8. The Derivative of a Discontinuous Function.

which we graph in Figure 11.8. We note that f has a single jump discontinuity of magnitude $\frac{6}{5}$ at $x = 1$. This means that

$$f(x) = g(x) + \tfrac{6}{5}\,\sigma(x-1), \qquad \text{where} \qquad g(x) = \begin{cases} -x, & x < 1, \\ \frac{1}{5}\,x^2 - \frac{6}{5}, & x > 1, \end{cases}$$

is continuous everywhere, since its right and left hand limits at the original discontinuity are equal: $g(1^+) = g(1^-) = -1$. Therefore,

$$f'(x) = g'(x) + \tfrac{6}{5}\,\delta(x-1), \qquad \text{where} \qquad g'(x) = \begin{cases} -1, & x < 1, \\ \frac{2}{5}\,x, & x > 1, \end{cases}$$

while $g'(1)$, and hence $f'(1)$, is not defined. In Figure 11.8, the delta spike in the derivative of f is symbolized by a vertical line — although this pictorial device fails to indicate its magnitude of $\frac{6}{5}$.

Since $g'(x)$ can be found by directly differentiating the formula for $f(x)$, once we determine the magnitude and location of the jump discontinuities of $f(x)$, we can compute its derivative directly without introducing to the auxiliary function $g(x)$.

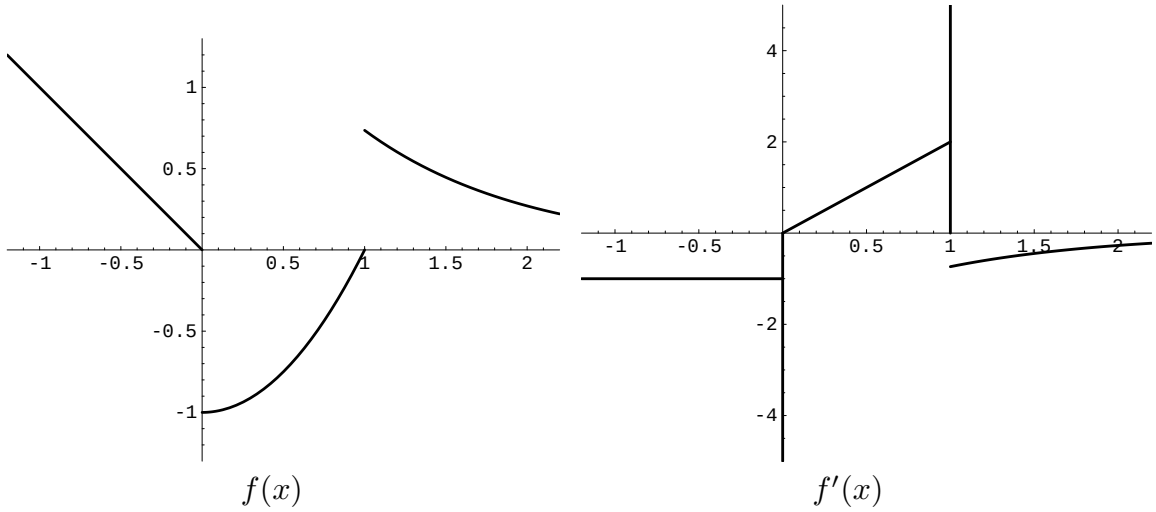
Example 11.5. As a second, more streamlined example, consider the function

$$f(x) = \begin{cases} -x, & x < 0, \\ x^2 - 1, & 0 < x < 1, \\ 2\,e^{-x}, & x > 1, \end{cases}$$

which is plotted in Figure 11.9. This function has jump discontinuities of magnitude -1 at $x = 0$, and of magnitude $2/e$ at $x = 1$. Therefore, in light of the preceding remark,

$$f'(x) = -\,\delta(x) + \frac{2}{e}\,\delta(x-1) + \begin{cases} -1, & x < 0, \\ 2\,x, & 0 < x < 1, \\ -2\,e^{-x}, & x > 1, \end{cases}$$

where the final terms are obtained by directly differentiating $f(x)$.



$f(x)$ $\qquad\qquad$ $f'(x)$

Figure 11.9. The Derivative of a Discontinuous Function.

Example 11.6. The derivative of the absolute value function

$$a(x) = |\, x\,| = \begin{cases} x, & x > 0, \\ -x, & x < 0, \end{cases}$$

is the *sign function*

$$s(x) = a'(x) = \begin{cases} +1, & x > 0, \\ -1, & x < 0. \end{cases} \tag{11.52}$$

Note that there is no delta function in $a'(x)$ because $a(x)$ is continuous everywhere. Since $s(x)$ has a jump of magnitude 2 at the origin and is otherwise constant, its derivative $s'(x) = a''(x) = 2\,\delta(x)$ is twice the delta function.

We are even allowed to differentiate the delta function. Its first derivative

$$\delta_y'(x) = \delta'(x - y)$$

can be interpreted in two ways. First, we may view $\delta'(x)$ as the limit of the derivatives of the approximating functions (11.31):

$$\frac{d\delta}{dx} = \lim_{n \to \infty} \frac{dg_n}{dx} = \lim_{n \to \infty} \frac{-2\,n^3\,x}{\pi(1 + n^2\,x^2)^2}\,. \tag{11.53}$$

The graphs of these rational functions take the form of more and more concentrated spiked "doublets", as illustrated in Figure 11.10. To determine the effect of the derivative on a test function $u(x)$, we compute the limiting integral

$$\begin{aligned} \langle\, \delta'\,, u\,\rangle &= \int_{-\infty}^{\infty} \delta'(x)\, u(x)\, dx = \lim_{n \to \infty} \int_{-\infty}^{\infty} g_n'(x)\, u(x)\, dx \\ &= -\lim_{n \to \infty} \int_{-\infty}^{\infty} g_n(x)\, u'(x)\, dx = -\int_{-\infty}^{\infty} \delta(x)\, u'(x)\, dx = -\,u'(0). \end{aligned} \tag{11.54}$$

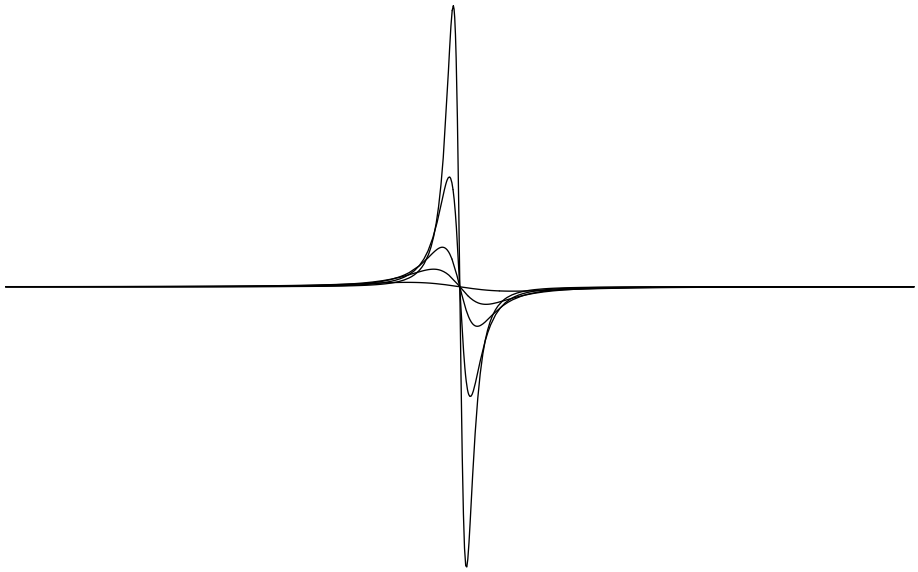


Figure 11.10. Derivative of Delta Function as Limit of Doublets.

In the middle step, we used an integration by parts, noting that the boundary terms at $\pm\infty$ vanish, provided that $u(x)$ is continuously differentiable and bounded as $|\,x\,| \to \infty$. Pay attention to the minus sign in the final answer.

In the dual interpretation, the generalized function $\delta_y'(x)$ corresponds to the linear functional

$$L_y'[\,u\,] = -\,u'(y) = \langle\,\delta_y'\,,u\,\rangle = \int_0^\ell \delta_y'(x)\,u(x)\,dx, \qquad \text{where} \qquad 0 < y < \ell, \tag{11.55}$$

that maps a continuously differentiable function $u(x)$ to *minus* its derivative at the point y. We note that (11.55) is compatible with a formal integration by parts

$$\int_0^\ell \delta\,'(x-y)\,u(x)\,dx = \,\delta(x-y)\,u(x)\,\Big|_{x=0}^{\ell} - \int_0^\ell \delta(x-y)\,u'(x)\,dx = -\,u'(y).$$

The boundary terms at $x = 0$ and $x = \ell$ automatically vanish since $\delta(x-y) = 0$ for $x \neq y$.

Warning: The functions $\widetilde{g}_n(x) = g_n(x) + g_n'(x)$ satisfy $\lim\limits_{n\to\infty} \widetilde{g}_n(x) = 0$ for all $x \neq y$, while $\displaystyle\int_{-\infty}^{\infty} \widetilde{g}_n(x)\,dx = 1$. However, $\lim\limits_{n\to\infty} \widetilde{g}_n = \lim\limits_{n\to\infty} g_n + \lim\limits_{n\to\infty} g_n' = \delta + \delta\,'$. Thus, our original conditions (11.29–30) are *not* in fact sufficient to characterize whether a sequence of functions has the delta function as a limit. To be absolutely sure, one must, in fact, verify the more comprehensive limiting formula (11.40).

The Green's Function

To further cement our new-found friendship, we now put the delta function to work to solve inhomogeneous boundary value problems. Consider a bar of length ℓ subject to a unit impulse force $\delta_y(x) = \delta(x-y)$ concentrated at position $0 < y < \ell$. The underlying

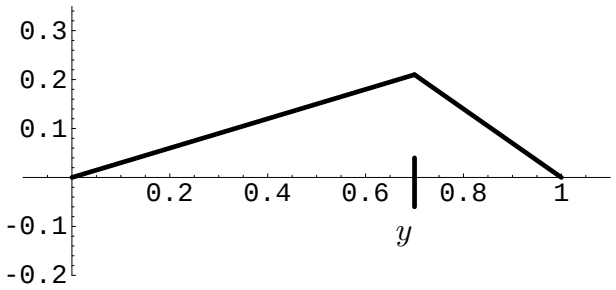


Figure 11.11. Green's function for a Bar with Fixed Ends.

differential equation (11.12) takes the special form

$$-\frac{d}{dx}\left(c(x)\,\frac{du}{dx}\right) = \delta(x-y), \qquad 0 < x < \ell, \tag{11.56}$$

which we supplement with homogeneous boundary conditions that lead to a unique solution. The solution is known as the *Green's function* for the boundary value problem, and will be denoted by $G_y(x) = G(x,y)$.

Example 11.7. Let us look at the simple case of a homogeneous bar, of unit length $\ell = 1$, with constant stiffness c, and fixed at both ends. The boundary value problem for the Green's function $G(x,y)$ takes the form

$$-c\,u'' = \delta(x-y), \qquad u(0) = 0 = u(1), \tag{11.57}$$

where $0 < y < 1$ indicates the point at which we apply the impulse force. The solution to the differential equation is obtained by direct integration. First, by (11.42),

$$u'(x) = -\frac{\sigma(x-y)}{c} + a,$$

where a is a constant of integration. A second integration leads to

$$u(x) = -\frac{\rho(x-y)}{c} + a\,x + b, \tag{11.58}$$

where ρ is the ramp function (11.44). The integration constants a, b are fixed by the boundary conditions; since $0 < y < 1$, we have

$$u(0) = b = 0, \qquad u(1) = -\frac{1-y}{c} + a + b = 0, \qquad \text{and so} \qquad a = \frac{1-y}{c}\,.$$

Therefore, the Green's function for the problem is

$$G(x,y) = -\rho(x-y) + (1-y)x = \begin{cases} x(1-y)/c, & x \le y, \\ y(1-x)/c, & x \ge y, \end{cases} \tag{11.59}$$

Figure 11.11 sketches a graph of $G(x,y)$ when $c = 1$. Note that, for each fixed y, it is a continuous and piecewise affine function of x — meaning that its graph consists of connected straight line segments, with a corner where the unit impulse force is being applied.

Once we have determined the Green's function, we are able to solve the general inhomogeneous boundary value problem

$$-u'' = f(x), \qquad u(0) = 0 = u(1), \tag{11.60}$$

The solution formula is a consequence of linear superposition. We first express the forcing function $f(x)$ as a linear combination of impulses concentrated at points along the bar. Since there is a continuum of possible positions $0 \le y \le 1$ at which impulse forces may be applied, we will use an integral to sum them up, thereby writing the external force as

$$f(x) = \int_0^1 f(y)\,\delta(x-y)\,dy. \tag{11.61}$$

In the continuous context, sums are replaced by integrals, and we will interpret (11.61) as the (continuous) superposition of an infinite collection of impulses $f(y)\,\delta(x-y)$, of magnitude $f(y)$ and concentrated at position y.

The superposition principle states that, for linear systems, linear combinations of inhomogeneities produce linear combinations of solutions. Again, we adapt this principle to the continuum by replacing the sums by integrals. Thus, we claim that the solution to the boundary value problem is the self-same linear superposition

$$u(x) = \int_0^1 f(y)\,G(x,y)\,dy \tag{11.62}$$

of the Green's function solutions to the individual unit impulse problems.

For the particular boundary value problem (11.60), we use the explicit formula (11.59) for the Green's function. Breaking the integral (11.62) into two parts, for $y < x$ and $y > x$, we arrive at the explicit solution formula

$$u(x) = \frac{1}{c}\int_0^x (1-x)\,y\,f(y)\,dy \;+\; \frac{1}{c}\int_x^1 x\,(1-y)\,f(y)\,dy. \tag{11.63}$$

For example, under a constant unit force f, (11.63) reduces to

$$u(x) = \frac{f}{c}\int_0^x (1-x)\,y\,dy + \frac{f}{c}\int_x^1 x\,(1-y)\,dy = \frac{f}{2\,c}\,(1-x)\,x^2 + \frac{f}{2\,c}\,x\,(1-x)^2 = \frac{f}{2\,c}\,(x-x^2),$$

in agreement with our earlier solution (11.19) in the special case $d = 0$. Although this relatively simple problem was perhaps easier to solve directly, the Green's function approach helps crystallize our understanding, and provides a unified framework that covers the full range of linear boundary value problems arising in applications, including those governed by partial differential equations, [**104**, **152**, **170**].

Let us, finally, convince ourselves that the superposition formula (11.63) does indeed give the correct answer. First,

$$\begin{aligned}\frac{du}{dx} &= (1-x)\,x\,f(x) + \int_0^x [-y\,f(y)]\,dy - x\,(1-x)\,f(x) + \int_x^1 (1-y)\,f(y)\,dy\\ &= -\int_0^1 y\,f(y)\,dy + \int_x^1 f(y)\,dy.\end{aligned}$$

Differentiating again, we conclude that $\dfrac{d^2u}{dx^2} = -f(x)$, as claimed.

Remark: In computing the derivatives of u, we made use of the calculus formula

$$\frac{d}{dx}\int_{\alpha(x)}^{\beta(x)} F(x,y)\,dy = F(x,\beta(x))\,\frac{d\beta}{dx} - F(x,\alpha(x))\,\frac{d\alpha}{dx} + \int_{\alpha(x)}^{\beta(x)} \frac{\partial F}{\partial x}\,(x,y)\,dy \qquad (11.64)$$

for the derivative of an integral with variable limits, which is a straightforward consequence of the Fundamental Theorem of Calculus and the chain rule, [**9**, **151**]. As with all limiting processes, one must always be careful when interchanging the order of differentiation and integration.

We note the following fundamental properties, that serve to uniquely characterize the Green's function. First, since the delta forcing vanishes except at the point $x = y$, the Green's function satisfies the homogeneous differential equation†

$$\frac{\partial^2 G}{\partial x^2}\,(x,y) = 0 \qquad \text{for all} \qquad x \neq y. \qquad (11.65)$$

Secondly, by construction, it must satisfy the boundary conditions,

$$G(0,y) = 0 = G(1,y).$$

Thirdly, for each fixed y, $G(x,y)$ is a continuous function of x, but its derivative $\partial G/\partial x$ has a jump discontinuity of magnitude $-1/c$ at the impulse point $x = y$. The second derivative $\partial^2 G/\partial x^2$ has a delta function discontinuity there, and thereby solves the original impulse boundary value problem (11.57).

Finally, we cannot help but notice that the Green's function is a symmetric function of its two arguments: $G(x,y) = G(y,x)$. Symmetry has the interesting physical consequence that the displacement of the bar at position x due to an impulse force concentrated at position y is exactly the same as the displacement of the bar at y due to an impulse of the same magnitude being applied at x. This turns out to be a rather general, although perhaps unanticipated phenomenon. (For the finite-dimensional counterpart for mass-spring chains, circuits, and structures see Exercises ■, ■ and ■.) Symmetry is a consequence of the underlying symmetry or "self-adjointness" of the boundary value problem, to be developed properly in the following section.

Remark: The Green's function $G(x,y)$ should be be viewed as the continuum limit of the inverse of the stiffness matrix, $G = K^{-1}$, appearing in the discrete equilibrium equations $K\mathbf{u} = \mathbf{f}$. Indeed, the entries G_{ij} of the inverse matrix are approximations to the sampled values $G(x_i, x_j)$. In particular, symmetry of the Green's function, whereby $G(x_i, x_j) = G(x_j, x_i)$, corresponds to symmetry, $G_{ij} = G_{ji}$, of the inverse of the symmetric stiffness matrix. In Exercise ■, you are asked to study this limiting procedure in some detail.

† Since $G(x,y)$ is a function of two variables, we switch to partial derivative notation to indicate its derivatives.

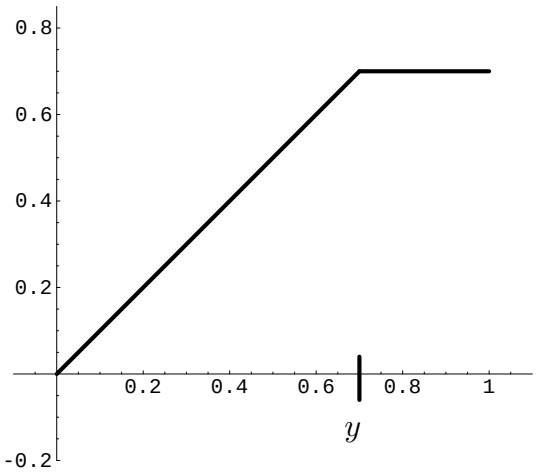


Figure 11.12. Green's Function for Bar with One Fixed and One Free End.

Let us summarize the fundamental properties that serve to characterize the Green's function, in a form that applies to general second order boundary value problems.

Basic Properties of the Green's Function

(i) Solves the homogeneous differential equation:

$$-\frac{\partial}{\partial x}\left(c(x)\,\frac{\partial}{\partial x}\,G(x,y)\right) = 0, \qquad \text{for all} \qquad x \neq y. \tag{11.66}$$

(ii) Satisfies the homogeneous boundary conditions.

(iii) Is a continuous function of its arguments.

(iv) As a function of x, its derivative $\dfrac{\partial G}{\partial x}$ has a jump discontinuity of magnitude $-\dfrac{1}{c(y)}$ at $x = y$.

(v) Is a symmetric function of its arguments:

$$G(x,y) = G(y,x). \tag{11.67}$$

(vi) Generates a superposition principle for the solution under general forcing functions:

$$u(x) = \int_0^\ell G(x,y)\,f(y)\,dy. \tag{11.68}$$

.

Example 11.8. Consider a uniform bar of length $\ell = 1$ with one fixed and one free end, subject to an external force. The displacement $u(x)$ satisfies the boundary value problem

$$-c\,u'' = f(x), \qquad u(0) = 0, \qquad u'(1) = 0, \tag{11.69}$$

where c is the elastic constant of the bar. To determine the Green's function, we appeal to its characterizing properties, although one could equally well use direct integration as in the preceding Example 11.7.

First, since, as a function of x, it must satisfy the homogeneous differential equation $-c\,u'' = 0$ for all $x \neq y$, the Green's function must be of the form

$$G(x,y) = \begin{cases} p\,x + q, & x \leq y, \\ r\,x + s, & x \geq y, \end{cases}$$

for certain constants p, q, r, s. Second, the boundary conditions require

$$q = G(0,y) = 0, \qquad r = \frac{\partial G}{\partial x}\,(1,y) = 0.$$

Continuity of the Green's function at $x = y$ imposes the further constraint

$$p\,y = G(y^-, y) = G(y^+, y) = s.$$

Finally, the derivative $\partial G/\partial x$ must have a jump discontinuity of magnitude $-1/c$ at $x = y$, and so

$$-\frac{1}{c} = \frac{\partial G}{\partial x}\,(y^+, y) - \frac{\partial G}{\partial x}\,(y^-, y) = 0 - p, \qquad \text{and so} \qquad p = s = \frac{1}{c}.$$

We conclude that the Green's function for this problem is

$$G(x,y) = \begin{cases} x/c, & x \leq y, \\ y/c, & x \geq y, \end{cases} \tag{11.70}$$

which, for $c = 1$, is graphed in Figure 11.12. Note that $G(x,y) = G(y,x)$ is indeed symmetric, which helps check the correctness of our computation. Finally, the superposition principle (11.68) implies that the solution to the boundary value problem (11.69) can be written as a single integral, namely

$$u(x) = \int_0^1 G(x,y) f(y)\,dy \;=\; \frac{1}{c}\int_0^x y\,f(y)\,dy + \frac{1}{c}\int_x^1 x\,f(y)\,dy. \tag{11.71}$$

The reader may wish to verify this directly, as we did in the previous example.

11.3. Adjoints and Minimum Principles.

One of the profound messages of this text is that the linear algebraic structures that were initially† designed for finite-dimensional problems all have direct counterparts in the infinite-dimensional function spaces. To further develop this theme, let us now discuss how the boundary value problems for continuous elastic bars fit into our general equilibrium framework of positive (semi-)definite linear systems. As we will see, the associated energy minimization principle not only leads to a new mathematical characterization of the equilibrium solution, it also, through the finite element method, underlies the most important class of numerical approximation algorithms for such boundary value problems.

† Sometimes, the order is reversed, and, at least historically, basic linear algebra concepts make their first appearance in function space. Examples include the Cauchy–Schwarz inequality, the Fredholm alternative, and the Fourier transform.

Adjoints of Differential Operators

In discrete systems, a key step was the recognition that the matrix appearing in the force balance law is the transpose of the incidence matrix relating displacements and elongations. In the continuum limit, the discrete incidence matrix has turned into a differential operator. But how do you take the "transpose" of a differential operator? The abstract answer to this quandary can be found in Section 7.5. The transpose of a matrix is a particular instance of the general notion of the *adjoint* of a linear function, which relies on the specification of inner products on its domain and target spaces. In the case of the matrix transpose, the adjoint is taken with respect to the standard dot product on Euclidean space. Thus, the correct interpretation of the "transpose" of a differential operator is as the adjoint linear operator with respect to suitable inner products on function space.

For bars and similar one-dimensional media, the role of the incidence matrix is played by the derivative $v = D[u] = du/dx$, which defines a linear operator $D\colon U \to V$ from the vector space of possible displacements $u(x)$, denoted by U, to the vector space of possible strains $v(x)$, denoted by V. In order to compute its adjoint, we need to impose inner products on both the displacement space U and the strain space V. The simplest is to adopt the same standard L^2 inner product

$$\langle\, u\,, \widetilde{u}\,\rangle = \int_0^\ell u(x)\, \widetilde{u}(x)\, dx, \qquad\qquad \langle\!\langle\, v\,, \widetilde{v}\,\rangle\!\rangle = \int_0^\ell v(x)\, \widetilde{v}(x)\, dx, \tag{11.72}$$

on both vector spaces. These are the continuum analogs of the Euclidean dot product, and, as we shall see, will be appropriate when dealing with homogeneous bars. According to the defining equation (7.72), the adjoint D^* of the derivative operator must satisfy the inner product identity

$$\langle\!\langle\, D[u]\,, v\,\rangle\!\rangle = \langle\, u\,, D^*[v]\,\rangle \qquad \text{for all} \qquad u \in U, \quad v \in V. \tag{11.73}$$

First, we compute the left hand side:

$$\langle\!\langle\, D[u]\,, v\,\rangle\!\rangle = \left\langle\!\!\left\langle\, \frac{du}{dx}\,;\, v\,\right\rangle\!\!\right\rangle = \int_0^\ell \frac{du}{dx}\; v\; dx. \tag{11.74}$$

On the other hand, the right hand side should equal

$$\langle\, u\,, D^*[v]\,\rangle = \int_0^\ell u\, D^*[v]\, dx. \tag{11.75}$$

Now, in the latter integral, we see u multiplying the result of applying the linear operator D^* to v. To identify this integrand with that in the previous integral (11.74), we need to somehow remove the derivative from u. The secret is integration by parts, which allows us to rewrite the first integral in the form

$$\int_0^\ell \frac{du}{dx}\; v\, dx = \big[\, u(\ell)\, v(\ell) - u(0)\, v(0)\,\big] - \int_0^\ell u\; \frac{dv}{dx}\, dx. \tag{11.76}$$

Ignoring the two boundary terms for a moment, the remaining integral has the form of an inner product

$$-\int_0^\ell u\,\frac{dv}{dx}\,dx = \int_0^\ell u\left[-\frac{dv}{dx}\right]dx = \left\langle\, u\,,\,-\frac{dv}{dx}\,\right\rangle = \langle\, u\,,-D[\,v\,]\,\rangle. \tag{11.77}$$

Equating (11.74) and (11.77), we deduce that

$$\langle\!\langle\, D[\,u\,]\,,v\,\rangle\!\rangle = \left\langle\!\!\left\langle\, \frac{du}{dx}\,;v\,\right\rangle\!\!\right\rangle = \left\langle\, u\,,\,-\frac{dv}{dx}\,\right\rangle = \langle\, u\,,-D[\,v\,]\,\rangle.$$

Thus, to satisfy the adjoint equation (11.73), we must have

$$\langle\, u\,,D^*[\,v\,]\,\rangle = \langle\, u\,,-D[\,v\,]\,\rangle \qquad \text{for all} \qquad u\in U,\quad v\in V,$$

and so

$$\left(\frac{d}{dx}\right)^* = D^* = -\,D = -\,\frac{d}{dx}\,. \tag{11.78}$$

The final equation confirms our earlier identification of the derivative operator D as the continuum limit of the incidence matrix A, and its negative $-D = D^*$ as the limit of the transposed (or adjoint) incidence matrix $A^T = A^*$.

However, the preceding argument is valid *only* if the boundary terms in the integration by parts formula (11.76) vanish:

$$u(\ell)\,v(\ell) - u(0)\,v(0) = 0, \tag{11.79}$$

which necessitates imposing suitable boundary conditions on the functions u and v. For example, in the case of a bar with both ends fixed, the boundary conditions

$$u(0) = 0, \qquad\qquad u(\ell) = 0, \tag{11.80}$$

will ensure that (11.79) holds, and therefore validate (11.78). The homogeneous boundary conditions serve to define the vector space

$$U = \left\{\, u(x)\in \mathrm{C}^2[0,\ell] \;\;\middle|\;\; u(0) = u(\ell) = 0 \,\right\}$$

of allowable displacements, consisting of all twice continuously differentiable functions that vanish at the ends of the bar.

The fixed boundary conditions (11.80) are not the only possibilities that ensure the vanishing of the boundary terms (11.79). An evident alternative is to require that the strain vanish at both endpoints, $v(0) = v(\ell) = 0$. In this case, the strain space

$$V = \left\{\, v(x)\in \mathrm{C}^1[0,\ell] \;\;\middle|\;\; v(0) = v(\ell) = 0 \,\right\}$$

consists of all functions that vanish at the endpoints. Since the derivative $D\colon U\to V$ must map a displacement $u(x)$ to an *allowable* strain $v(x)$, the vector space of possible displacements takes the form

$$U = \left\{\, u(x)\in \mathrm{C}^2[0,\ell] \;\;\middle|\;\; u'(0) = u'(\ell) = 0 \,\right\}.$$

Thus, this case corresponds to the free boundary conditions of Example 11.3. Again, restricting $D\colon U \to V$ to these particular vector spaces ensures that the boundary terms (11.79) vanish, and so (11.78) holds in this situation too.

Let us list the most important combinations of boundary conditions that imply the vanishing of the boundary terms (11.79), and so ensure the validity of the adjoint equation (11.78).

Self-Adjoint Boundary Conditions for a Bar

(a)	Both ends fixed:	$u(0) = u(\ell) = 0.$		
(b)	One free and one fixed end:	$u(0) = u'(\ell) = 0$	or	$u'(0) = u(\ell) = 0.$
(c)	Both ends free:	$u'(0) = u'(\ell) = 0.$		
(d)	Periodic bar or ring:	$u(0) = u(\ell), \quad u'(0) = u'(\ell).$		

In all cases, the boundary conditions impose restrictions on the displacement space U and, in cases (b–d) when identifying $v(x) = u'(x)$, the strain space V also.

In mathematics, a fixed boundary condition, $u(a) = 0$, is commonly referred to as a *Dirichlet boundary condition*, to honor the nineteenth century French analyst Lejeune Dirichlet. A free boundary condition, $u'(a) = 0$, is known as a *Neumann boundary condition*, after his German contemporary Carl Gottfried Neumann. The *Dirichlet boundary value problem* (a) has both ends fixed, while the *Neumann boundary value problem* (c) has both ends free. The intermediate case (b) is known as a *mixed boundary value problem*. The periodic boundary conditions (d) represent a bar that has its ends joined together to form a circular[†] elastic ring, and represents the continuum limit of the periodic mass–spring chain discussed in Exercise ■.

Summarizing, for a homogeneous bar with unit stiffness $c(x) \equiv 1$, the displacement, strain, and external force are related by the adjoint formulae

$$v = D[u] = u', \qquad\qquad f = D^*[v] = -v',$$

provided that we impose a suitable pair of homogeneous boundary conditions. The equilibrium equation has the self-adjoint form

$$K[u] = f, \qquad \text{where} \qquad K = D^* \circ D = -D^2. \tag{11.81}$$

We note that

$$K^* = (D^* \circ D)^* = D^* \circ (D^*)^* = D^* \circ D = K, \tag{11.82}$$

which proves self-adjointness of the differential operator. In gory detail,

$$\langle\, K[u]\,, \widetilde{u}\,\rangle = \int_0^\ell \big[\, -u''(x)\, \widetilde{u}(x)\,\big]\, dx = \int_0^\ell \big[\, -u(x)\, \widetilde{u}''(x)\,\big]\, dx = \langle\, u\,, K[\widetilde{u}]\,\rangle \tag{11.83}$$

for all displacements $u, \widetilde{u} \in U$. A direct verification of this formula relies on two integration by parts, employing the selected boundary conditions to eliminate the boundary contributions.

[†] The circle is sufficiently large so that we can safely ignore any curvature effects.

To deal with nonuniform materials, we must modify the inner products. Let us retain the ordinary L^2 inner product

$$\langle\, u\,,\widetilde{u}\,\rangle = \int_0^\ell u(x)\,\widetilde{u}(x)\,dx, \qquad\qquad u,\widetilde{u}\in U, \tag{11.84}$$

on the vector space of possible displacements, but adopt a weighted inner product

$$\langle\!\langle\, v\,,\widetilde{v}\,\rangle\!\rangle = \int_0^\ell v(x)\,\widetilde{v}(x)\,c(x)\,dx, \qquad\qquad v,\widetilde{v}\in V, \tag{11.85}$$

on the space of strain functions. The weight function $c(x) > 0$ coincides with the stiffness of the bar; its positivity, which is required for (11.85) to define a *bona fide* inner product, is in accordance with the underlying physical assumptions.

Let us recompute the adjoint of the derivative operator $D\colon U \to V$, this time with respect to the inner products (11.84–85). Now we need to compare

$$\langle\!\langle\, D[u]\,,v\,\rangle\!\rangle = \int_0^\ell \frac{du}{dx}\; v(x)\,c(x)\,dx, \qquad \text{with} \qquad \langle\, u\,,D^*[v]\,\rangle = \int_0^\ell u(x)\,D^*[v]\,dx.$$

Integrating the first expression by parts, we find

$$\int_0^\ell \frac{du}{dx}\; c\,v\,dx = \big[\,u(\ell)\,c(\ell)\,v(\ell) - u(0)\,c(0)\,v(0)\,\big] - \int_0^\ell u\,\frac{d(c\,v)}{dx}\,dx = \int_0^\ell u\left[-\frac{d(c\,v)}{dx}\right]dx, \tag{11.86}$$

provided that we choose our boundary conditions so that

$$u(\ell)\,c(\ell)\,v(\ell) - u(0)\,c(0)\,v(0) = 0. \tag{11.87}$$

As you can check, this follows from any of the listed boundary conditions: Dirichlet, Neumann, or mixed, as well as the periodic case, assuming $c(0) = c(\ell)$. Therefore, in such situations, the weighted adjoint of the derivative operator is

$$D^*[v] = -\,\frac{d(c\,v)}{dx} = -\,c\,\frac{dv}{dx} - c'\,v. \tag{11.88}$$

The self-adjoint combination $K = D^*\circ D$ is

$$K[u] = -\,\frac{d}{dx}\left(\,c(x)\,\frac{du}{dx}\,\right), \tag{11.89}$$

and hence we have formulated the original equation (11.12) for a nonuniform bar in the same abstract self-adjoint form.

As an application, let us show how the self-adjoint formulation leads directly to the symmetry of the Green's function $G(x,y)$. As a function of x, the Green's function satisfies

$$K[\,G(x,y)\,] = \delta(x-y).$$

Thus, by the definition of the delta function and the self-adjointness identity (11.83),

$$G(z,y) = \int_0^\ell G(x,y)\,\delta(x-z)\,dx = \langle\, G(x,y)\,, \delta(x-z)\,\rangle = \langle\, G(x,y)\,, K[G(x,z)]\,\rangle \quad (11.90)$$
$$= \langle\, K[G(x,y)]\,, G(x,z)\,\rangle = \langle\, \delta(x-y)\,, G(x,z)\,\rangle = \int_0^\ell G(x,z)\,\delta(x-y)\,dx = G(y,z),$$

for any $0 < y, z < \ell$, which validates[†] the symmetry equation (11.67).

Positivity and Minimum Principles

We are now able to characterize the solution to a stable self-adjoint boundary value problem by a quadratic minimization principle. Again, the development shadows the finite-dimensional case presented in Chapter 6. So the first step is to understand how a differential operator defining a boundary value problem can be positive definite.

According to the abstract Definition 7.59, a linear operator $K\colon U \to U$ on an inner product space U is *positive definite*, provided that it is

(a) self-adjoint, so $K^* = K$, and

(b) satisfies the positivity criterion $\langle\, K[u]\,, u\,\rangle > 0$ for all $0 \neq u \in U$.

Self-adjointness of the product operator $K = D^* \circ D$ was established in (11.82). Furthermore, Theorem 7.62 tells us that K is positive definite if and only if $\ker D = \{0\}$. Indeed, by the definition of the adjoint,

$$\langle\, K[u]\,, u\,\rangle = \langle\, D^*[D[u]]\,, u\,\rangle = \langle\!\langle\, D[u]\,, D[u]\,\rangle\!\rangle = \|\, D[u]\,\|^2 \geq 0, \quad (11.91)$$

so $K = D^* \circ D$ is automatically positive semi-definite. Moreover, $\langle\, K[u]\,, u\,\rangle = 0$ if and only if $D[u] = 0$, i.e., $u \in \ker K$. Thus $\ker D = \{0\}$ is both necessary and sufficient for the positivity criterion to hold.

Now, in the absence of constraints, the kernel of the derivative operator D is *not* trivial. Indeed, $D[u] = u' = 0$ if and only if $u(x) \equiv c$ is constant, and hence $\ker D$ is the one-dimensional subspace of $\mathrm{C}^1[0,\ell]$ consisting of all constant functions. However, we are viewing D as a linear operator on the vector space U of allowable displacements, and so the elements of $\ker D \subset U$ must also be allowable, meaning that they must satisfy the boundary conditions. Thus, positivity reduces, in the present situation, to the question of whether or not there are any nontrivial constant functions that satisfy the prescribed homogeneous boundary conditions.

Clearly, the only constant function that satisfies a homogeneous Dirichlet boundary condition is the zero function. Therefore, when restricted to the Dirichlet displacement space $U = \{u(0) = u(\ell) = 0\}$, the derivative operator has trivial kernel, $\ker D = \{0\}$, so $K = D^* \circ D$ defines a positive definite linear operator on U. A similar argument applies to the mixed boundary value problem, which is also positive definite. On the other hand, any constant function satisfies the homogeneous Neumann boundary conditions, and so $\ker D \subset \widetilde{U} = \{u'(0) = u'(\ell) = 0\}$ is a one-dimensional subspace. Therefore,

[†] Symmetry at the endpoints follows from continuity.

the Neumann boundary value problem is only positive semi-definite. A similar argument shows that the periodic problem is also positive semi-definite. Observe that, just as in the finite-dimensional version, the positive definite cases are stable, and the boundary value problem admits a unique equilibrium solution under arbitrary external forcing, whereas the semi-definite cases are unstable, and have either no solution or infinitely many equilibrium solutions, depending on the nature of the external forcing.

In the positive definite, stable cases, we can characterize the equilibrium solution as the unique function $u \in U$ that minimizes the quadratic functional

$$\mathcal{P}[u] = \tfrac{1}{2} \| D[u] \|^2 - \langle u, f \rangle = \int_0^\ell \big[\tfrac{1}{2} c(x) u'(x)^2 - f(x) u(x) \big] \, dx. \tag{11.92}$$

A proof of this general fact appears following Theorem 7.61. Pay attention: the norm in (11.92) refers to the strain space V, and so is associated with the weighted inner product (11.85), whereas the inner product term refers to the displacement space U, which has been given the L^2 inner product. Physically, the first term measures the internal energy due to the stress in the bar, while the second term is the potential energy induced by the external forcing. Thus, as always, the equilibrium solution seeks to minimize the total energy in the system.

Example 11.9. Consider the homogeneous Dirichlet boundary value problem

$$-u'' = f(x), \qquad u(0) = 0, \qquad u(\ell) = 0. \tag{11.93}$$

for a uniform bar with two fixed ends. This is a stable case, and so the underlying differential operator $K = D^* \circ D = -D^2$, when acting on the space of displacements satisfying the boundary conditions, is positive definite. Explicitly, positive definiteness requires

$$\langle K[u], u \rangle = \int_0^\ell [-u''(x) u(x)] \, dx = \int_0^\ell [u'(x)]^2 \, dx > 0 \tag{11.94}$$

for all nonzero $u(x) \not\equiv 0$ with $u(0) = u(\ell) = 0$. Notice how we used an integration by parts, invoking the boundary conditions to eliminate the boundary contributions, to expose the positivity of the integral. The corresponding energy functional is

$$\mathcal{P}[u] = \tfrac{1}{2} \| u' \|^2 - \langle u, f \rangle = \int_0^\ell \big[\tfrac{1}{2} u'(x)^2 - f(x) u(x) \big] \, dx.$$

Its minimum value, taken over all possible displacement functions that satisfy the boundary conditions, occurs precisely when $u = u_\star$ is the solution to the boundary value problem.

A direct verification of the latter fact may be instructive. As in our derivation of the adjoint operator, it relies on an integration by parts. Since $-u_\star'' = f$,

$$\begin{aligned} \mathcal{P}[u] &= \int_0^\ell \big[\tfrac{1}{2} (u')^2 + u_\star'' u \big] \, dx = u_\star'(\ell) u(\ell) - u_\star'(0) u(0) + \int_0^\ell \big[\tfrac{1}{2} (u')^2 - u_\star' u' \big] \, dx \\ &= \int_0^\ell \tfrac{1}{2} (u' - u_\star')^2 \, dx \; - \int_0^\ell \tfrac{1}{2} (u_\star')^2 \, dx, \end{aligned} \tag{11.95}$$

where the boundary terms vanish owing to the boundary conditions on $u_\star$ and u. In the final expression for $\mathcal{P}[u]$, the first integral is always ≥ 0, and is actually equal to 0 if and only if $u'(x) = u'_\star(x)$ for all $0 \leq x \leq \ell$. On the other hand, the second integral does not depend upon u at all. Thus, for $\mathcal{P}[u]$ to achieve a minimum, $u(x) = u_\star(x) + c$ for some constant c. But the boundary conditions force $c = 0$, and hence the energy functional will assume its minimum value if and only if $u = u_\star$.

Inhomogeneous Boundary Conditions

So far, we have restricted our attention to homogeneous boundary value problems. Inhomogeneous boundary conditions are a little trickier, since the spaces of allowable displacements and allowable strains are no longer vector spaces, and so the abstract theory, as developed in Chapter 7, is not directly applicable.

One way to circumvent this difficulty is to slightly modify the displacement function so as to satisfy homogeneous boundary conditions. Consider, for example, the inhomogeneous Dirichlet boundary value problem

$$K[u] = -\frac{d}{dx}\left(c(x)\,\frac{du}{dx}\right) = f(x), \qquad u(0) = \alpha, \qquad u(\ell) = \beta. \tag{11.96}$$

We shall choose a function $h(x)$ that satisfies the boundary conditions:

$$h(0) = \alpha, \qquad h(\ell) = \beta.$$

Note that we are *not* requiring h to satisfy the differential equation, and so one, but by no means the only, possible choice is the linear interpolating polynomial

$$h(x) = \alpha + \frac{\beta - \alpha}{\ell}\, x. \tag{11.97}$$

Since u and h have the same boundary values, their difference

$$\widetilde{u}(x) = u(x) - h(x) \tag{11.98}$$

satisfies the homogeneous Dirichlet boundary conditions

$$\widetilde{u}(0) = \widetilde{u}(\ell) = 0. \tag{11.99}$$

Moreover, by linearity, $\widetilde{u}$ satisfies the modified equation

$$K[\widetilde{u}] = K[u - h] = K[u] - K[h] = f - K[h] \equiv \widetilde{f},$$

or, explicitly,

$$-\frac{d}{dx}\left(c(x)\,\frac{d\widetilde{u}}{dx}\right) = \widetilde{f}(x), \qquad \text{where} \qquad \widetilde{f}(x) = f(x) + \frac{d}{dx}\left(c(x)\,\frac{dh}{dx}\right). \tag{11.100}$$

For the particular choice (11.97),

$$\widetilde{f}(x) = f(x) + \frac{\beta - \alpha}{\ell}\, c'(x).$$

Thus, we have managed to convert the original inhomogeneous problem for u into a homogeneous boundary value problem for $\widetilde{u}$. Once we have solved the latter, the solution to the original is simply reconstructed from the formula

$$u(x) = \widetilde{u}(x) + h(x). \tag{11.101}$$

We know that the homogeneous Dirichlet boundary value problem (11.99–100) is positive definite, and so we can characterize its solution by a minimum principle, namely as the minimizer of the quadratic energy functional

$$\mathcal{P}[\widetilde{u}] = \tfrac{1}{2}\| \widetilde{u}' \|^2 - \langle \widetilde{u}\,, \widetilde{f}\,\rangle = \int_0^\ell \big[\, \tfrac{1}{2}\, c(x)\, \widetilde{u}'(x)^2 - \widetilde{f}(x)\, \widetilde{u}(x)\,\big]\, dx. \tag{11.102}$$

Let us rewrite the minimization principle in terms of the original displacement function $u(x)$. Replacing $\widetilde{u}$ and $\widetilde{f}$ by their formulae (11.98, 100) yields

$$\begin{aligned} \mathcal{P}[\widetilde{u}] &= \tfrac{1}{2}\| u' - h' \|^2 - \langle u - h\,, f - K[h]\,\rangle \\ &= \big[\, \tfrac{1}{2}\| u' \|^2 - \langle u\,, f\,\rangle\,\big] - \big[\, \langle\!\langle u'\,, h'\,\rangle\!\rangle - \langle u\,, K[h]\,\rangle\,\big] + \big[\, \tfrac{1}{2}\| h' \|^2 + \langle h\,, f - K[h]\,\rangle\,\big] \\ &= \mathcal{P}[u] - \big[\, \langle\!\langle u'\,, h'\,\rangle\!\rangle - \langle u\,, K[h]\,\rangle\,\big] + C_0. \end{aligned} \tag{11.103}$$

In the middle expression, the last pair of terms depend only on the initial choice of $h(x)$, and not on $u(x)$; thus, once h has been selected, they can be regarded as a fixed constant, here denoted by C_0. The first pair of terms reproduces the quadratic energy functional (11.92) for the actual displacement $u(x)$. The middle terms can be explicitly evaluated:

$$\begin{aligned} \langle\!\langle u'\,, h'\,\rangle\!\rangle - \langle u\,, K[h]\,\rangle &= \int_0^\ell \big[\, c(x)\, h'(x)\, u'(x) + \big(\, c(x)\, h'(x)\,\big)'\, u(x)\,\big]\, dx \\ &= \int_0^\ell \frac{d}{dx}\big[\, c(x)\, h'(x)\, u(x)\,\big]\, dx = c(\ell)\, h'(\ell)\, u(\ell) - c(0)\, h'(0)\, u(0). \end{aligned} \tag{11.104}$$

In particular, if $u(x)$ satisfies the inhomogeneous Dirichlet boundary conditions $u(0) = \alpha$, $u(\ell) = \beta$, then

$$\langle\!\langle u'\,, h'\,\rangle\!\rangle - \langle u\,, K[h]\,\rangle = c(\ell)\, h'(\ell)\, \beta - c(0)\, h'(0)\, \alpha = C_1$$

also depends only on the interpolating function h and not on u. Therefore,

$$\mathcal{P}[\widetilde{u}] = \mathcal{P}[u] - C_1 + C_0$$

differ by a constant. We conclude that, if the function $\widetilde{u}$ minimizes $\mathcal{P}[\widetilde{u}]$, then $u = \widetilde{u} + h$ necessarily minimizes $\mathcal{P}[u]$. In this manner, we have characterized the solution to the inhomogeneous Dirichlet boundary value problem by the *same* minimization principle.

Theorem 11.10. *The solution $u_\star(x)$ to the Dirichlet boundary value problem*

$$-\frac{d}{dx}\left(c(x)\,\frac{du}{dx}\right) = f(x), \qquad u(0) = \alpha, \qquad u(\ell) = \beta,$$

is the unique C^2 *function that satisfies the indicated boundary conditions and minimizes the energy functional* $\mathcal{P}[u] = \displaystyle\int_0^\ell \big[\, \tfrac{1}{2}\, c(x)\, u'(x)^2 - f(x)\, u(x)\,\big]\, dx.$

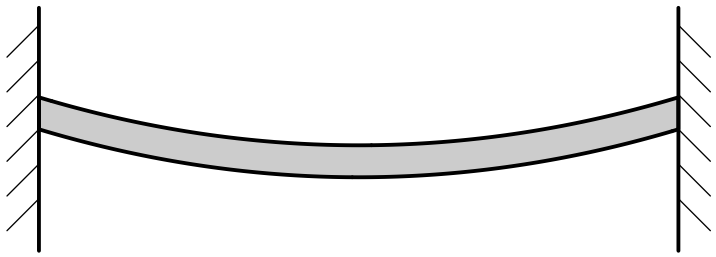


Figure 11.13. Bending of a Beam.

Warning: The inhomogeneous mixed boundary value problem is trickier, since the extra terms (11.104) *will* depend upon the value of $u(x)$. The details are worked out in Exercise ■.

11.4. Beams and Splines.

Unlike a bar, which can only stretch longitudinally, a *beam* is allowed to bend. To keep the geometry simple, we treat the case in which the beam is restricted to the $x\,y$ plane, as sketched in Figure 11.13. Let $0 \le x \le \ell$ represent the reference position along a horizontal beam of length ℓ. To further simplify the physics, we shall ignore stretching, and assume that the "atoms" in the beam can only move in the transverse direction, with $y = u(x)$ representing the vertical displacement of the "atom" that starts out at position x.

The *strain* in a beam depends on how much it is bent. Mathematically, bending is equal to the *curvature*[†] of the graph of the displacement function $u(x)$, and is computed by the usual calculus formula

$$\kappa = \frac{u''}{\big(\,1 + (u')^2\,\big)^{3/2}}\,. \tag{11.105}$$

Thus, for beams, the strain is a *nonlinear* function of displacement. Since we are still only willing to deal with linear systems, we shall suppress the nonlinearity by assuming that the beam is not bent too far; more specifically, we assume that the derivative $u'(x) \ll 1$ is small and so the tangent line is nearly horizontal. Under this assumption, the curvature function (11.105) is replaced by its linear approximation

$$\kappa \;\approx\; u''. \tag{11.106}$$

From now on, we will identify $v = D^2[u] = u''$ as the *strain* in a bending beam. The second derivative operator $L = D^2$ that maps displacement u to strain $v = L[u]$ thereby describes the beam's intrinsic (linearized) geometry.

The next step is to formulate a constitutive relation between stress and strain. Physically, the *stress* $w(x)$ represents the bending moment of the beam, defined as the product

[†] By definition, [**9**, **151**], the curvature of a curve at a point is equal to the reciprocal, $\kappa = 1/r$ of the radius of the osculating circle; see Exercise ■ for details.

of internal force and angular deflection. Our small bending assumption implies an elastic Hooke's law relation

$$w(x) = c(x)\, v(x) = c(x)\, \frac{d^2u}{dx^2}\ , \tag{11.107}$$

where the proportionality factor $c(x) > 0$ measures the *stiffness* of the beam at the point x. In particular, a uniform beam has constant stiffness, $c(x) \equiv c$.

Finally, the differential equation governing the equilibrium configuration of the beam will follow from a balance of the internal and external forces. To compute the internal force, we appeal to our general equilibrium framework, which tells us to apply the adjoint of the incidence operator $L = D^2$ to the strain, leading to the force balance law

$$L^*[\,v\,] = L^* \circ L[\,u\,] = f. \tag{11.108}$$

Let us compute the adjoint. We use the ordinary L^2 inner product on the space of displacements $u(x)$, and adopt a weighted inner product, based on the stiffness function $c(x)$, between strain functions:

$$\langle\, u\,, \widetilde{u}\,\rangle = \int_a^b u(x)\, \widetilde{u}(x)\, dx, \qquad\qquad \langle\!\langle\, v\,, \widetilde{v}\,\rangle\!\rangle = \int_a^b v(x)\, \widetilde{v}(x)\, c(x)\, dx. \tag{11.109}$$

According to the general adjoint equation (7.72), we need to equate

$$\int_0^\ell L[\,u\,]\, v\, c\, dx = \langle\!\langle\, L[\,u\,]\,, v\,\rangle\!\rangle = \langle\, u\,, L^*[\,v\,]\,\rangle = \int_0^\ell u\, L^*[\,v\,]\, dx. \tag{11.110}$$

As before, the computation relies on (in this case two) integrations by parts:

$$\begin{aligned} \langle\!\langle\, L[\,u\,]\,, v\,\rangle\!\rangle = \int_0^\ell \frac{d^2u}{dx^2}\ c\, v\, dx &= \left[\ \frac{du}{dx}\ c\, v\ \right]\Big|_{x=0}^{\ell} - \int_0^\ell \frac{du}{dx}\ \frac{d(c\, v)}{dx}\, dx \\ &= \left[\ \frac{du}{dx}\ c\, v - u\ \frac{d(c\, v)}{dx}\ \right]\Big|_{x=0}^{\ell} + \int_0^\ell u\ \frac{d^2(c\, v)}{dx^2}\, dx \end{aligned}.$$

Comparing with (11.110), we conclude that $L^*[\,v\,] = D^2(c\, v)$ provided the boundary terms vanish:

$$\begin{aligned} \left[\ \frac{du}{dx}\ c\, v - u\ \frac{d(c\, v)}{dx}\ \right]\Big|_{x=0}^{\ell} &= \left[\ \frac{du}{dx}\ w - u\ \frac{dw}{dx}\ \right]\Big|_{x=0}^{\ell} \\ &= \big[\, u'(\ell)\, w(\ell) - u(\ell)\, w'(\ell)\,\big] - \big[\, u'(0)\, w(0) - u(0)\, w'(0)\,\big] = 0. \end{aligned} \tag{11.111}$$

Thus, under suitable boundary conditions, the force balance equations are

$$L^*[\,v\,] = \frac{d^2(c\, v)}{dx^2} = f(x). \tag{11.112}$$

A justification of (11.112) based on physical principles can be found in [**164**]. Combining (11.107, 112), we conclude that the equilibrium configuration of the beam is characterized as a solution to the fourth order ordinary differential equation

$$\frac{d^2}{dx^2}\ \left(c(x)\, \frac{d^2u}{dx^2}\right) = f(x). \tag{11.113}$$

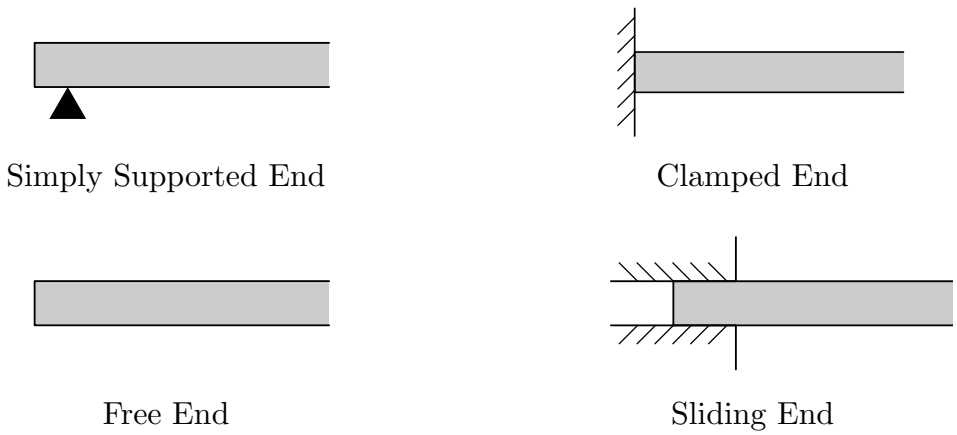


Figure 11.14. Boundary Conditions for a Beam.

As such, the general solution will depend upon 4 arbitrary constants, and so we need to impose a total of four boundary conditions — two at each end — in order to uniquely specify the equilibrium displacement. The (homogeneous) boundary conditions should be chosen so as to make the boundary terms in our integration by parts computation vanish, cf. (11.111). There are a variety of ways in which this can be arranged, and the most important possibilities are the following:

Self-Adjoint Boundary Conditions for a Beam

(a)	Simply supported end:	$u(0) = w(0) = 0,$
(b)	Fixed (clamped) end:	$u(0) = u'(0) = 0,$
(c)	Free end:	$w(0) = w'(0) = 0,$
(d)	Sliding end:	$u'(0) = w'(0) = 0.$

In these conditions, $w(x) = c(x)\, v(x) = c(x)\, u''(x)$ is the stress resulting from the displacement $u(x)$.

A second pair of boundary conditions must be imposed at the other end $x = \ell$. You can mix or match these conditions in any combination — for example, a pair of simply supported ends, or one free end and one fixed end, and so on. Inhomogeneous boundary conditions are also allowed and used to model applied displacements or applied forces at the ends. Yet another option is to consider a bendable circular ring, which is subject to *periodic boundary conditions*

$$u(0) = u(\ell), \qquad u'(0) = u'(\ell), \qquad w(0) = w(\ell), \qquad w'(0) = w'(\ell),$$

indicating that the ends of the beam have been welded together.

Let us concentrate our efforts on the uniform beam, of unit length $\ell = 1$, choosing units so that its stiffness $c(x) \equiv 1$. In the absence of external forcing, the differential

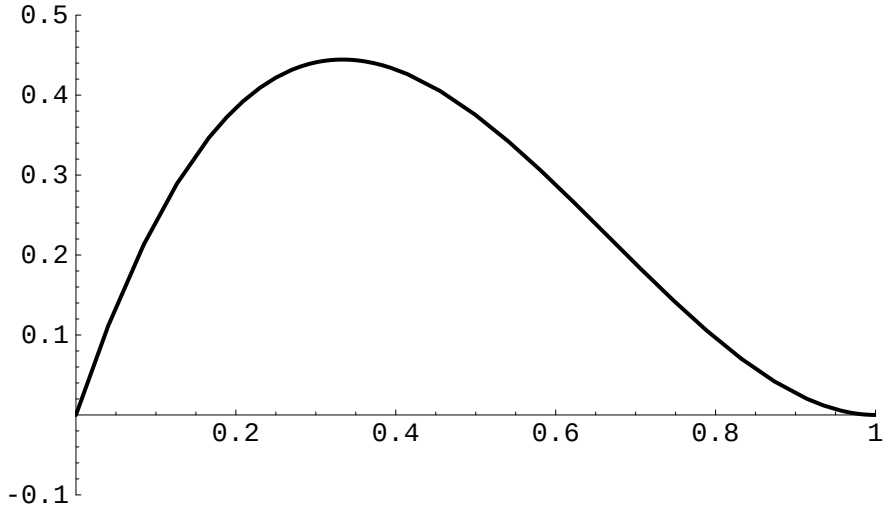


Figure 11.15. Hermite Cubic Spline.

equation (11.113) reduces to the elementary fourth order ordinary differential equation

$$\frac{d^4u}{dx^4} = 0. \tag{11.114}$$

The general solution is an arbitrary cubic polynomial,

$$u = a\,x^3 + b\,x^2 + c\,x + d. \tag{11.115}$$

Let us use this formula to solve a couple of representative boundary value problems.

First, suppose we clamp both ends of the beam, imposing the boundary conditions

$$u(0) = 0, \qquad u'(0) = \beta, \qquad u(1) = 0, \qquad u'(1) = 0, \tag{11.116}$$

so that the left end is tilted by a (small) angle $\tan^{-1}\beta$. We substitute the solution formula (11.115) into the boundary conditions (11.116) and solve for

$$a = \beta, \qquad b = -2\,\beta, \qquad c = \beta, \qquad d = 0.$$

The resulting solution

$$u(x) = \beta\,(x^3 - 2\,x^2 + x) = \beta\,x(1-x)^2 \tag{11.117}$$

is known as a *Hermite cubic spline*[†] and is graphed in Figure 11.15.

As a second example, suppose that we raise the left hand end of the beam without tilting, which corresponds to the boundary conditions

$$u(0) = \alpha, \qquad u'(0) = 0, \qquad u(1) = 0, \qquad u'(1) = 0. \tag{11.118}$$

[†] We first met Charles Hermite in Section 3.6, and the term "spline" will be explained shortly.

Substituting (11.115) and solving for a, b, c, d, we find that the solution is

$$u(x) = \alpha\,(1-x)^2\,(2\,x+1). \tag{11.119}$$

Observe that if we simultaneously raise and tilt the left end, so $u(0) = \alpha$, $u'(0) = \beta$, then we can simply use superposition to write the solution as the sum of (11.117) and (11.119):

$$u(x) = \alpha\,(1-x)^2\,(2\,x+1) + \beta\,x(1-x)^2.$$

To analyze a forced beam, we can adapt the Green's function approach. As we know, the Green's function will depend on the choice of (homogeneous) boundary conditions. Let us treat the case when the beam has two fixed ends, and so

$$u(0) = 0, \qquad u'(0) = 0, \qquad u(1) = 0, \qquad u'(1) = 0. \tag{11.120}$$

To construct the Green's function, we must solve the forced differential equation

$$\frac{d^4u}{dx^4} = \delta(x-y) \tag{11.121}$$

corresponding to a concentrated unit impulse applied at position y along the beam. Integrating (11.121) four times, using (11.45) with $n = 4$, we produce the general solution

$$u(x) = a\,x^3 + b\,x^2 + c\,x + d + \begin{cases} \frac{1}{6}\,(x-y)^3, & x > y, \\ 0, & x < y, \end{cases}$$

to the differential equation (11.121). The boundary conditions (11.120) require

$$\begin{aligned} u(0) &= d = 0, & u(1) &= a + b + \tfrac{1}{6}\,(1-y)^3 = 0, \\ u'(0) &= c = 0, & u'(1) &= 3\,a + 2\,b + \tfrac{1}{2}\,(1-y)^2 = 0, \end{aligned}$$

and hence

$$a = \tfrac{1}{3}\,(1-y)^3 - \tfrac{1}{2}\,(1-y)^2, \qquad b = -\tfrac{1}{2}\,(1-y)^3 + \tfrac{1}{2}\,(1-y)^2.$$

Therefore, the Green's function is

$$G(x,y) = \begin{cases} \frac{1}{6}\,x^2\,(1-y)^2\,(3\,y - x - 2\,x\,y), & x < y, \\ \frac{1}{6}\,y^2\,(1-x)^2\,(3\,x - y - 2\,x\,y), & x > y. \end{cases} \tag{11.122}$$

Observe that, as with the second order bar system, the Green's function is symmetric, $G(x,y) = G(y,x)$, which is a manifestation of the self-adjointness of the underlying boundary value problem, cf. (11.90). Symmetry implies that the deflection of the beam at position x due to a concentrated impulse force applied at position y is the same as the deflection at y due to an impulse force of the same magnitude applied at x.

As a function of x, the Green's function $G(x,y)$ satisfies the homogeneous differential equation (11.114) for all $x \neq y$. Its first and second derivatives $\partial G/\partial x, \partial^2 G/\partial x^2$ are continuous, while $\partial^3 G/\partial x^3$ has a unit jump discontinuity at $x = y$, which then produces the required delta function impulse in $\partial^4 G/\partial x^4$. The Green's function (11.122) is graphed

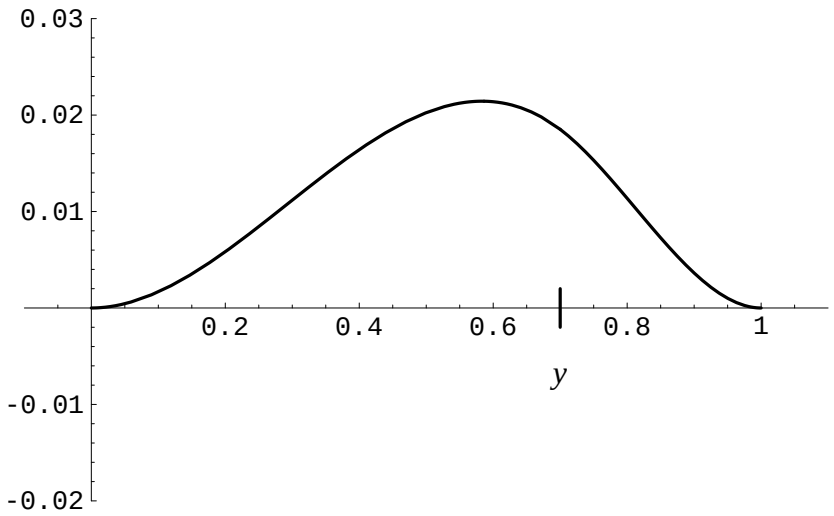


Figure 11.16. Green's Function for a Beam with Two Fixed Ends.

in Figure 11.16, and appears to be quite smooth. Evidently, the human eye cannot easily discern discontinuities in third order derivatives!

The solution to the forced boundary value problem

$$\frac{d^4u}{dx^4} = f(x), \qquad u(0) = u'(0) = u(1) = u'(1) = 0, \tag{11.123}$$

for a beam with fixed ends is then obtained by invoking the superposition principle. We view the forcing function as a linear superposition

$$f(x) = \int_0^\ell f(y)\,\delta(x-y)\,dx$$

of impulse delta forces. The solution is the self-same linear superposition of Green's function responses:

$$\begin{aligned} u(x) &= \int_0^1 G(x,y) f(y)\,dy \qquad\qquad (11.124)\\ &= \frac{1}{6}\int_0^x y^2\,(1-x)^2\,(3x-y-2xy)\,f(y)\,dy \;+\; \frac{1}{6}\int_x^1 x^2\,(1-y)^2\,(3y-x-2xy)\,f(y)\,dy. \end{aligned}$$

For example, under a constant unit downwards force $f(x) \equiv 1$, e.g., gravity, the deflection of the beam is given by

$$u(x) = \tfrac{1}{24}x^4 - \tfrac{1}{12}x^3 + \tfrac{1}{24}x^2 = \tfrac{1}{24}x^2(1-x)^2,$$

and graphed in Figure 11.17. Although we could, of course, obtain $u(x)$ by integrating the original differential equation (11.123) directly, writing the solution formula (11.124) as a single integral has evident advantages.

Since the beam operator $K = L^* \circ L$ assumes the standard self-adjoint, positive semi-definite form, the boundary value problem will be positive definite and hence stable if

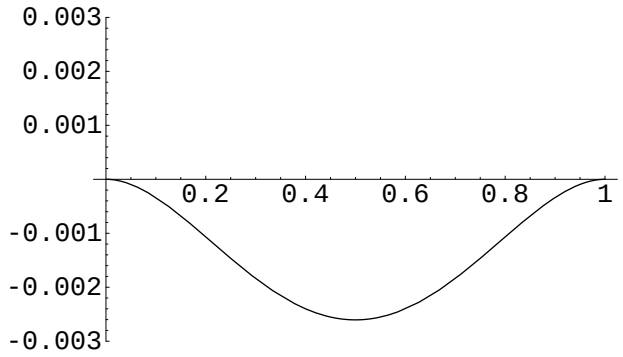


Figure 11.17. Deflection of a Uniform Beam under Gravity.

and only if $\ker L = \ker D^2 = \{0\}$ when restricted to the space of allowable displacement functions. Since the second derivative D^2 annihilates all linear polynomials

$$u(x) = \alpha + \beta\, x,$$

positive definiteness requires that no non-zero linear polynomials satisfy all four homogeneous boundary conditions. For example, any beam with one fixed end is stable since $u(x) \equiv 0$ is the only linear polynomial that satisfies $u(0) = u'(0) = 0$. On the other hand, a beam with two free ends is unstable since every linear polynomial displacement has zero stress $w(x) = u''(x) \equiv 0$, and so satisfies the boundary conditions $w(0) = w'(0) = w(\ell) = w'(\ell) = 0$. Similarly, a beam with a simply supported plus a free end is not positive definite since $u(x) = \beta\, x$ satisfies the four boundary conditions $u(0) = u'(0) = 0$, $w(\ell) = w'(\ell) = 0$. In the stable cases, the equilibrium solution can be characterized as the unique minimizer of the quadratic energy functional[†]

$$\mathcal{P}[u] = \tfrac{1}{2}\, \| L[u] \|^2 - \langle\, u\,, f\,\rangle = \int_a^b \big[\, \tfrac{1}{2}\, c(x)\, u''(x)^2 - f(x)\, u(x)\,\big]\, dx \tag{11.125}$$

among all C^4 functions satisfying the homogeneous boundary conditions. Inhomogeneous boundary conditions require some extra analysis, since the required integration by parts may introduce additional boundary contributions.

Splines

In pre–CAD (computer aided design) draftsmanship, a *spline* was a long, thin, flexible strip of wood that was used to draw a smooth curve through prescribed points. The points were marked by small pegs, and the spline rested on the pegs. The mathematical theory of splines was first developed in the 1940's by the Romanian mathematician Isaac Schoenberg as an attractive alternative to polynomial interpolation and approximation. Splines have since become ubiquitous in numerical analysis, in geometric modeling, in design and manufacturing, in computer graphics and animation, and in many other applications.

[†] Keep in mind that the norm on the strain functions $v = L[u] = u''$ is based on the weighted inner product $\langle\!\langle\, v\,, \tilde{v}\,\rangle\!\rangle$ in (11.109).

We suppose that the spline coincides with the graph of a function $y = u(x)$. The pegs are fixed at the prescribed data points $(x_0, y_0), \dots, (x_n, y_n)$, and this requires $u(x)$ to satisfy the interpolation conditions

$$u(x_j) = y_j, \qquad j = 0, \dots, n. \tag{11.126}$$

The *mesh points* $x_0 < x_1 < x_2 < \cdots < x_n$ are distinct and labeled in increasing order. The spline is modeled as an elastic beam, and so satisfies the homogeneous beam equation (11.114). Therefore,

$$u(x) = a_j + b_j\,(x - x_j) + c_j\,(x - x_j)^2 + d_j\,(x - x_j)^3, \qquad \begin{array}{c} x_j \le x \le x_{j+1}, \\ j = 0, \dots, n-1, \end{array} \tag{11.127}$$

is a piecewise cubic function — meaning that, between successive mesh points, it is a cubic polynomial, but not necessarily the same cubic on each subinterval. The fact that we write the formula (11.127) in terms of $x - x_j$ is merely for computational convenience.

Our problem is to determine the coefficients

$$a_j, \qquad b_j, \qquad c_j, \qquad d_j, \qquad\qquad j = 0, \dots, n-1.$$

Since there are n subintervals, there are a total of $4n$ coefficients, and so we require $4n$ equations to uniquely prescribe them. First, we need the spline to satisfy the interpolation conditions (11.126). Since it is defined by a different formula on each side of the mesh point, this results in a total of $2n$ conditions:

$$\begin{aligned} u(x_j^+) &= a_j = y_j, \\ u(x_{j+1}^-) &= a_j + b_j\,h_j + c_j\,h_j^2 + d_j\,h_j^3 = y_{j+1}, \end{aligned} \qquad j = 0, \dots, n-1, \tag{11.128}$$

where we abbreviate the length of the j^{th} subinterval by

$$h_j = x_{j+1} - x_j.$$

The next step is to require that the spline be as smooth as possible. The interpolation conditions (11.128) guarantee that $u(x)$ is continuous. The condition $u(x) \in \mathrm{C}^1$ be continuously differentiable requires that $u'(x)$ be continuous at the interior mesh points $x_1, \dots, x_{n-1}$, which imposes the $n-1$ additional conditions

$$b_j + 2\,c_j\,h_j + 3\,d_j\,h_j^2 = u'(x_{j+1}^-) = u'(x_{j+1}^+) = b_{j+1}, \qquad j = 0, \dots, n-2. \tag{11.129}$$

To make $u \in \mathrm{C}^2$, we impose $n-1$ further conditions

$$2\,c_j + 6\,d_j\,h_j = u''(x_{j+1}^-) = u''(x_{j+1}^+) = 2\,c_{j+1}, \qquad j = 0, \dots, n-2, \tag{11.130}$$

to ensure that u'' is continuous at the mesh points. We have now imposed a total of $4n-2$ conditions, namely (11.128–130), on the $4n$ coefficients. The two missing constraints will come from boundary conditions at the two endpoints, namely x_0 and x_n. There are three common types:

(*i*) *Natural boundary conditions*: $u''(x_0) = u''(x_n) = 0$, whereby

$$c_0 = 0, \qquad c_{n-1} + 3\,d_{n-1}\,h_{n-1} = 0. \tag{11.131}$$

Physically, this models a simply supported spline that rests freely on the first and last pegs.

(*ii*) *Clamped boundary conditions*: $u'(x_0) = \alpha$, $u'(x_n) = \beta$, where α, β, which could be 0, are fixed by the user. This requires

$$b_0 = \alpha, \qquad b_{n-1} + 2\,c_{n-1}\,h_{n-1} + 3\,d_{n-1}\,h_{n-1}^2 = \beta. \tag{11.132}$$

This corresponds to clamping the spline at prescribed angles at each end.

(*iii*) *Periodic boundary conditions*: $u'(x_0) = u'(x_n)$, $u''(x_0) = u''(x_n)$, so that

$$b_0 = b_{n-1} + 2\,c_{n-1}\,h_{n-1} + 3\,d_{n-1}\,h_{n-1}^2, \qquad c_0 = c_{n-1} + 3\,d_{n-1}\,h_{n-1}. \tag{11.133}$$

If we also require that the end interpolation values agree,

$$u(x_0) = y_0 = y_n = u(x_n), \tag{11.134}$$

then the resulting spline will be a periodic C^2 function, so $u(x+p) = u(x)$ with $p = x_n - x_0$ for all x. The periodic case is used to draw smooth closed curves; see below.

Theorem 11.11. *Suppose we are given mesh points $a = x_0 < x_1 < \cdots < x_n = b$, and corresponding data values $y_0, y_1, \ldots, y_n$, along with one of the three kinds of boundary conditions* (11.131), (11.132), *or* (11.133). *Then there exists a unique piecewise cubic spline function $u(x) \in C^2[a, b]$ that interpolates the data, $u(x_0) = y_0, \ldots, u(x_n) = y_n$, and satisfies the boundary conditions.*

Proof: We first discuss the natural case. The clamped case is left as an exercise for the reader, while the slightly harder periodic case will be treated at the end of the section. The first set of equations in (11.128) says that

$$a_j = y_j, \qquad j = 0, \ldots, n-1. \tag{11.135}$$

Next, (11.130–131) imply that

$$d_j = \frac{c_{j+1} - c_j}{3\,h_j}\,. \tag{11.136}$$

This equation also holds for $j = n-1$, provided that we make the convention that[†]

$$c_n = 0.$$

We now substitute (11.135–136) into the second set of equations in (11.128), and then solve the resulting equation for

$$b_j = \frac{y_{j+1} - y_j}{h_j} - \frac{(2\,c_j + c_{j+1})\,h_j}{3}\,. \tag{11.137}$$

[†] This is merely for convenience; there is no c_n used in the formula for the spline.

Substituting this result and (11.136) back into (11.129), and simplifying, we find

$$h_j\,c_j + 2\,(h_j + h_{j+1})\,c_{j+1} + h_{j+1}\,c_{j+2} = 3\left[\frac{y_{j+2} - y_{j+1}}{h_{j+1}} - \frac{y_{j+1} - y_j}{h_j}\right] = z_{j+1}, \quad (11.138)$$

where we introduce z_{j+1} as a shorthand for the quantity on the right hand side.

In the case of natural boundary conditions, we have

$$c_0 = 0, \qquad c_n = 0,$$

and so (11.138) constitutes a tridiagonal linear system

$$A\,\mathbf{c} = \mathbf{z}, \quad (11.139)$$

for the unknown coefficients $\mathbf{c} = \big(\,c_1, c_2, \ldots, c_{n-1}\,\big)^T$, with coefficient matrix

$$A = \begin{pmatrix} 2\,(h_0 + h_1) & h_1 & & & & & \\ h_1 & 2\,(h_1 + h_2) & h_2 & & & & \\ & h_2 & 2\,(h_2 + h_3) & h_3 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & h_{n-3} & 2\,(h_{n-3} + h_{n-2}) & h_{n-2} & \\ & & & & h_{n-2} & 2\,(h_{n-2} + h_{n-1}) \end{pmatrix} \quad (11.140)$$

and right hand side $\mathbf{z} = \big(\,z_1, z_2, \ldots, z_{n-1}\,\big)^T$. Once (11.140) has been solved, we will then use (11.135–137) to reconstruct the other spline coefficients a_j, b_j, d_j.

The key observation is that the coefficient matrix A is *strictly diagonally dominant*, cf. Definition 10.36, because all the $h_j > 0$, and so

$$2\,(h_{j-1} + h_j) > h_{j-1} + h_j.$$

Theorem 10.37 implies that A is nonsingular, and hence the tridiagonal linear system has a unique solution $\mathbf{c}$. This suffices to prove the theorem in the case of natural boundary conditions. *Q.E.D.*

To actually solve the linear system (11.139), we can apply our tridiagonal solution algorithm (1.63). Let us specialize to the most important case, when the mesh points are equally spaced in the interval $[a, b]$, so that

$$x_j = a + j\,h, \qquad \text{where} \qquad h = h_j = \frac{b - a}{n}\,, \qquad j = 0, \ldots, n-1.$$

In this case, the coefficient matrix $A = h\,B$ is equal to h times the tridiagonal matrix

$$B = \begin{pmatrix} 4 & 1 & & & & & \\ 1 & 4 & 1 & & & & \\ & 1 & 4 & 1 & & & \\ & & 1 & 4 & 1 & & \\ & & & 1 & 4 & 1 & \\ & & & & \ddots & \ddots & \ddots \end{pmatrix}$$

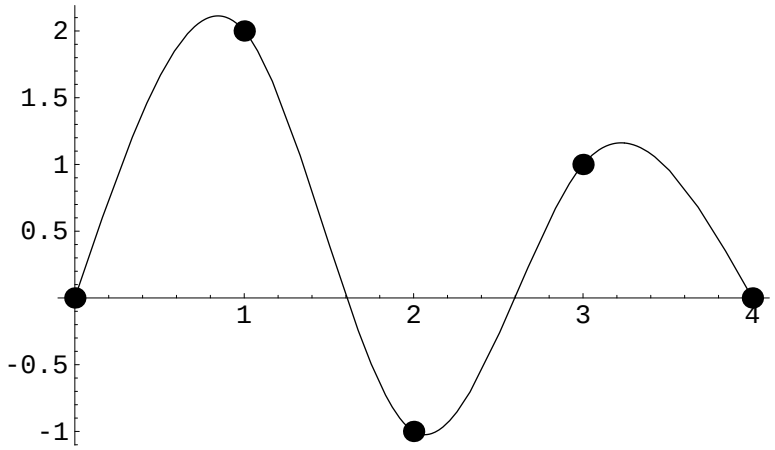


Figure 11.18. A Cubic Spline.

that first appeared in Example 1.37. Its LU factorization takes on an especially simple form, since most of the entries of L and U are essentially the same decimal numbers. This makes the implementation of the Forward and Back Substitution procedures almost trivial.

Figure 11.18 shows a particular example — a natural spline passing through the data points $(0,0)$, $(1,2)$, $(2,-1)$, $(3,1)$, $(4,0)$. As with the Green's function for the beam, the human eye is unable to discern the discontinuities in its third derivatives, and so the graph appears completely smooth, even though it is, in fact, only C^2.

In the periodic case, we set

$$a_{n+k} = a_n, \qquad b_{n+k} = b_n, \qquad c_{n+k} = c_n, \qquad d_{n+k} = d_n, \qquad z_{n+k} = z_n.$$

With this convention, the basic equations (11.135–138) are the same. In this case, the coefficient matrix for the linear system

$$A\,\mathbf{c} = \mathbf{z}, \qquad \text{with} \qquad \mathbf{c} = \big(\, c_0, c_1, \ldots, c_{n-1} \,\big)^T, \qquad \mathbf{z} = \big(\, z_0, z_1, \ldots, z_{n-1} \,\big)^T,$$

is of *circulant tridiagonal* form:

$$A = \begin{pmatrix} 2\,(h_{n-1}+h_0) & h_0 & & & & & h_{n-1} \\ h_0 & 2\,(h_0+h_1) & h_1 & & & & \\ & h_1 & 2\,(h_1+h_2) & h_2 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & h_{n-3} & 2\,(h_{n-3}+h_{n-2}) & h_{n-2} & \\ h_{n-1} & & & & h_{n-2} & 2\,(h_{n-2}+h_{n-1}) \end{pmatrix}. \tag{11.141}$$

Again A is strictly diagonally dominant, and so there is a unique solution $\mathbf{c}$, from which one reconstructs the spline, proving Theorem 11.11 in the periodic case. The LU factorization of tridiagonal circulant matrices was discussed in Exercise ■.

One immediate application of splines is curve fitting in computer aided design and graphics. The basic problem is to draw a smooth parametrized curve $\mathbf{u}(t) = \big(\, u(t), v(t) \,\big)^T$

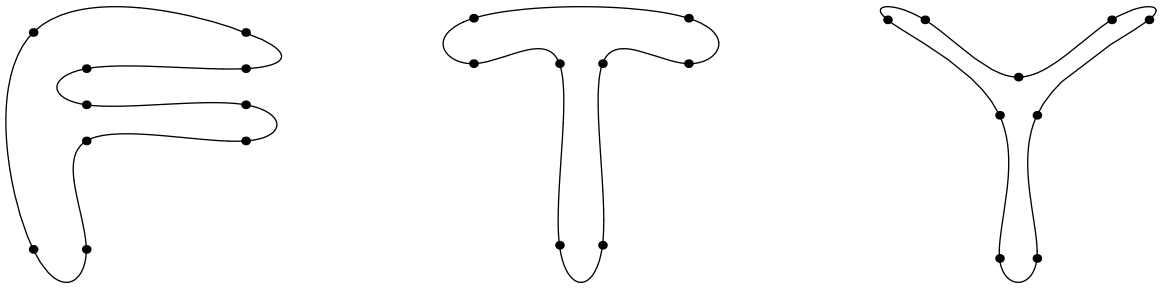


Figure 11.19. Three Sample Spline Letters.

that passes through a set of prescribed data points $\mathbf{x}_k = (\, x_k, y_k \,)^T$ in the plane. We have the freedom to choose the parameter value $t = t_k$ when the curve passes through the k^{th} point; the simplest and most common choice is to set $t_k = k$. We then construct the functions $x = u(t)$ and $y = v(t)$ as cubic splines interpolating the x and y coordinates of the data points, so $u(t_k) = x_k$, $v(t_k) = y_k$. For smooth closed curves, we require that both splines be periodic; for curves with ends, either natural or clamped boundary conditions are used.

Most computer graphics packages include one or more implementations of parametrized spline curves. The same idea also underlies modern font design for laser printing and typography (including the fonts used in this book). The great advantage of spline fonts over their bitmapped counterparts is that they can be readily scaled. Some sample letter shapes parametrized by periodic splines passing through the indicated data points are plotted in Figure 11.19. Better fits can be easily obtained by increasing the number of data points. Various extensions of the basic spline algorithms to space curves and surfaces are an essential component of modern computer graphics, design, and animation, [**57**, **146**].

11.5. Sturm–Liouville Boundary Value Problems.

The systems that govern the equilibrium configurations of bars are particular instances of a very general class of second order boundary value problems that was first systematically investigated by the nineteenth century French mathematicians Jacques Sturm and Joseph Liouville. Sturm–Liouville boundary value problems appear in a very wide range of applications, particularly in the analysis of partial differential equations by the method of separation of variables. A partial list of applications includes

(a) heat conduction in non-uniform bars;
(b) vibrations of non-uniform bars and strings;
(c) quantum mechanics — the one-dimensional Schrödinger equation;
(d) scattering theory — Hill's equation;
(e) oscillations of circular membranes (vibrations of drums) — Bessel's equation;
(f) oscillations of a sphere — Legendre's equation;
(g) thermodynamics of cylindrical and spherical bodies.

In this section, we will show how the class of Sturm–Liouville boundary value problems fits into our general equilibrium framework. However, the most interesting cases will be deferred until needed in our analysis of partial differential equations in Chapters 17 and 18.

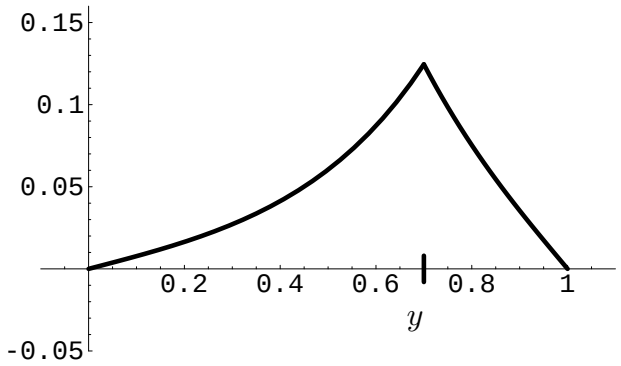


Figure 11.20. Green's Function for the Constant Coefficient Sturm–Liouville Problem.

The general *Sturm–Liouville boundary value problem* is based on a second order ordinary differential equation of the form

$$-\frac{d}{dx}\left(p(x)\,\frac{du}{dx}\right) + q(x)\,u = -p(x)\,\frac{d^2u}{dx^2} - p'(x)\,\frac{du}{dx} + q(x)\,u = f(x), \tag{11.142}$$

which is supplemented by Dirichlet, Neumann, mixed, or periodic boundary conditions. To be specific, let us concentrate on the case of homogeneous Dirichlet boundary conditions

$$u(a) = 0, \qquad u(b) = 0. \tag{11.143}$$

To avoid singular points of the differential equation (although we will later discover that most cases of interest in physics have one or more singular points) , we assume that $p(x) > 0$ for all $a \le x \le b$. To ensure positive definiteness of the Sturm–Liouville differential operator, we also assume $q(x) \ge 0$. These assumptions suffice to guarantee existence and uniqueness of the solution to the boundary value problem. A proof of the following theorem can be found in [**105**].

Theorem 11.12. *Let $p(x) > 0$ and $q(x) \ge 0$ for $a \le x \le b$. Then the Sturm–Liouville boundary value problem* (11.142–143) *admits a unique solution.*

Most Sturm–Liouville problems cannot be solved in terms of elementary functions. Indeed, most of the important special functions appearing in mathematical physics, including Bessel functions, Legendre functions, hypergeometric functions, and so on, first arise as solutions to particular Sturm–Liouville equations, [**129**].

Example 11.13. Consider the constant coefficient Sturm–Liouville boundary value problem

$$-u'' + \omega^2\,u = f(x), \qquad u(0) = u(1) = 0. \tag{11.144}$$

The functions $p(x) \equiv 1$ and $q(x) \equiv \omega^2 > 0$ are both constant. We will solve this problem by constructing the Green's function. Thus, we first consider the effect of a delta function inhomogeneity

$$-u'' + \omega^2\,u = \delta(x - y), \qquad u(0) = u(1) = 0. \tag{11.145}$$

Rather than try to integrate this differential equation directly, let us appeal to the defining properties of the Green's function. The general solution to the homogeneous equation is a linear combination of the two basic exponentials $e^{\omega x}$ and $e^{-\omega x}$, or better, the hyperbolic functions

$$\cosh \omega x = \frac{e^{\omega x} + e^{-\omega x}}{2}, \qquad \sinh \omega x = \frac{e^{\omega x} - e^{-\omega x}}{2}. \tag{11.146}$$

The solutions satisfying the first boundary condition are multiples of $\sinh \omega x$, while those satisfying the second boundary condition are multiples of $\sinh \omega (1 - x)$. Therefore, the solution to (11.145) has the form

$$G(x,y) = \begin{cases} a \sinh \omega x, & x < y, \\ b \sinh \omega (1 - x), & x > y. \end{cases} \tag{11.147}$$

Continuity of $G(x,y)$ at $x = y$ requires

$$a \sinh \omega y = b \sinh \omega (1 - y). \tag{11.148}$$

At $x = y$, the derivative $\partial G/\partial x$ must have a jump discontinuity of magnitude -1 in order that the second derivative term in (11.145) match the delta function. Since

$$\frac{\partial G}{\partial x}(x,y) = \begin{cases} a\,\omega \cosh \omega x, & x < y, \\ -b\,\omega \cosh \omega (1 - x), & x > y, \end{cases}$$

the jump condition requires

$$a\,\omega \cosh \omega y - 1 = -b\,\omega \cosh \omega (1 - y). \tag{11.149}$$

If we multiply (11.148) by $\omega \cosh \omega (1-y)$ and (11.149) by $\sinh \omega (1-y)$ and then add the results together, we find

$$\sinh \omega (1 - y) = a\,\omega \big[\sinh \omega y \, \cosh \omega (1 - y) + \cosh \omega y \, \sinh \omega (1 - y) \big] = a\,\omega \sinh \omega,$$

where we used the addition formula for the hyperbolic sine:

$$\sinh(\alpha + \beta) = \sinh \alpha \, \cosh \beta + \cosh \alpha \, \sinh \beta. \tag{11.150}$$

Therefore,

$$a = \frac{\sinh \omega (1 - y)}{\omega \sinh \omega}, \qquad b = \frac{\sinh \omega y}{\omega \sinh \omega},$$

and the Green's function is

$$G(x,y) = \begin{cases} \dfrac{\sinh \omega x \, \sinh \omega (1 - y)}{\omega \sinh \omega}, & x < y, \\ \dfrac{\sinh \omega (1 - x) \, \sinh \omega y}{\omega \sinh \omega}, & x > y. \end{cases} \tag{11.151}$$

Note that $G(x,y) = G(y,x)$ is symmetric, in accordance with the self-adjoint nature of the boundary value problem. A graph appears in Figure 11.20; note that the corner, indicating

a discontinuity in the first derivative, appears at the point $x = y$ where the impulse force is applied.

The general solution to the inhomogeneous boundary value problem (11.144) is given by the basic superposition formula (11.62), which becomes

$$\begin{aligned} u(x) &= \int_0^1 G(x,y) f(y)\, dx \\ &= \int_0^x \frac{\sinh \omega\,(1-x) \sinh \omega\, y}{\omega \sinh \omega}\, f(y)\, dy \;+\; \int_x^1 \frac{\sinh \omega\, x \, \sinh \omega\,(1-y)}{\omega \sinh \omega}\, f(y)\, dy. \end{aligned}$$

For example, under a constant unit force $f(x) \equiv 1$, the solution is

$$\begin{aligned} u(x) &= \int_0^x \frac{\sinh \omega\,(1-x) \sinh \omega\, y}{\omega \sinh \omega}\, dy \;+\; \int_x^1 \frac{\sinh \omega\, x \, \sinh \omega\,(1-y)}{\omega \sinh \omega}\, dy \\ &= \frac{\sinh \omega\,(1-x)\big(\cosh \omega\, x - 1\big)}{\omega^2 \sinh \omega} \;+\; \frac{\sinh \omega\, x \big(\cosh \omega\,(1-x) - 1\big)}{\omega^2 \sinh \omega} \\ &= \frac{1}{\omega^2} \;-\; \frac{\sinh \omega\, x + \sinh \omega\,(1-x)}{\omega^2 \sinh \omega}\,. \end{aligned} \tag{11.152}$$

For comparative purposes, the reader may wish to rederive this particular solution by a direct calculation, without appealing to the Green's function.

To place a Sturm–Liouville boundary value problem in our self-adjoint framework, we proceed as follows. (Exercise ■ serves to motivate the construction.) Consider the linear operator

$$L[u] = \begin{pmatrix} u' \\ u \end{pmatrix}$$

that maps $u(x)$ to the vector-valued function whose components are the function and its first derivative. For the homogeneous Dirichlet boundary conditions (11.143), the domain of L will be the vector space

$$U = \big\{\, u(x) \in \mathrm{C}^2[a,b] \;\big|\; u(a) = u(b) = 0 \,\big\}$$

consisting of all twice continuously differentiable functions that vanish at the endpoints. The target space of $L\colon U \to V$ consists of continuously differentiable vector-valued functions $\mathbf{v}(x) = (\, v_1(x), v_2(x)\,)^T$; we denote this vector space as $V = \mathrm{C}^1([a,b], \mathbb{R}^2)$.

To proceed, we must compute the adjoint of $L\colon U \to V$. To recover the Sturm–Liouville problem, we use the standard L^2 inner product (11.84) on U, but adopt a weighted inner product

$$\langle\!\langle\, \mathbf{v}\,, \mathbf{w}\,\rangle\!\rangle = \int_a^b \big[\, p(x)\, v_1(x)\, w_1(x) + q(x)\, v_2(x)\, w_2(x)\,\big]\, dx, \qquad \mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}, \tag{11.153}$$

on V. The positivity assumptions on the weight functions p, q ensure that this is a *bona fide* inner product. According to the defining equation (7.72), the adjoint $L^*\colon V \to U$ is required to satisfy

$$\langle\!\langle\, L[u]\,, \mathbf{v}\,\rangle\!\rangle = \langle\, u\,, L^*[\,\mathbf{v}\,]\,\rangle.$$

As usual, the adjoint computation relies on integration by parts. Here, we only need to manipulate the first summand:

$$\begin{aligned}\langle\!\langle\, L[u]\,, \mathbf{v}\,\rangle\!\rangle &= \int_a^b \big[\, p\,u'\,v_1 + q\,u\,v_2\,\big]\,dx \\ &= p(b)\,u(b)\,v_1(b) - p(a)\,u(a)\,v_1(a) + \int_a^b u\,\big[\,-(p\,v_1)' + q\,v_2\,\big]\,dx.\end{aligned}$$

The Dirichlet conditions (11.143) ensure that the boundary terms vanish, and therefore,

$$\langle\!\langle\, L[u]\,, \mathbf{v}\,\rangle\!\rangle = \int_a^b u\,\big[\,-(p\,v_1)' + q\,v_2\,\big]\,dx = \langle\, u\,, L^*[\,\mathbf{v}\,]\,\rangle.$$

We conclude that the adjoint operator is given by

$$L^*[\,\mathbf{v}\,] = -\,\frac{d(p\,v_1)}{dx} + q\,v_2.$$

The canonical self-adjoint combination

$$K[u] = L^* \circ L[u] = L^*\begin{pmatrix} u' \\ u \end{pmatrix} = -\,\frac{d}{dx}\begin{pmatrix} p\,\dfrac{du}{dx} \end{pmatrix} + q\,u \tag{11.154}$$

reproduces the Sturm–Liouville differential operator. Moreover, since $\ker L = \{0\}$ is trivial (why?), the boundary value problem is positive definite. Theorem 7.62 implies that the solution can be characterized as the unique minimizer of the quadratic functional

$$\mathcal{P}[u] = \tfrac{1}{2}\|\,L[u]\,\|^2 - \langle\, u\,, f\,\rangle = \int_a^b \big[\,\tfrac{1}{2}\,p(x)\,u'(x)^2 + \tfrac{1}{2}\,q(x)\,u(x)^2 - f(x)\,u(x)\,\big]\,dx \tag{11.155}$$

among all C^2 functions satisfying the prescribed boundary conditions. For example, the solution to the constant coefficient Sturm–Liouville problem (11.144) can be characterized as minimizing the quadratic functional

$$\mathcal{P}[u] = \int_0^1 \big[\,\tfrac{1}{2}\,u'^2 + \tfrac{1}{2}\,\omega^2\,u^2 - f\,u\,\big]\,dx$$

among all C^2 functions satisfying $u(0) = u(1) = 0$.

11.6. Finite Elements.

The characterization of the solution to a positive definite boundary value problem via a minimization principle inspires a very powerful and widely used numerical solution scheme, known as the *finite element method*. In this final section, we give a brief introduction to the finite element method in the context of one-dimensional boundary value problems involving ordinary differential equations. Extensions to boundary value problems in higher dimensions governed by partial differential equations will appear in Section 15.5.

The underlying idea is strikingly simple. We are trying to find the solution to a boundary value problem by minimizing a quadratic functional $\mathcal{P}[u]$ on an infinite-dimensional

vector space U. The solution $u_\star \in U$ to this minimization problem is found by solving a differential equation subject to specified boundary conditions. However, as we learned in Chapter 4, minimizing the functional on a *finite-dimensional subspace* $W \subset U$ is a problem in linear algebra, and, moreover, one that we already know how to solve! Of course, restricting the functional $\mathcal{P}[u]$ to the subspace W will not, barring luck, lead to the exact minimizer. Nevertheless, if we choose W to be a sufficiently "large" subspace, the resulting minimizer $w_\star \in W$ may very well provide a reasonable approximation to the actual solution $u_\star \in U$. A rigorous justification of this process, under appropriate hypotheses, requires a full analysis of the finite element method, and we refer the interested reader to [**157**, **180**]. Here we shall concentrate on trying to understand how to apply the method in practice.

To be a bit more explicit, consider the minimization principle

$$\mathcal{P}[u] = \tfrac{1}{2} \,\|\, L[u] \,\|^2 - \langle\, f \,,\, u \,\rangle \tag{11.156}$$

for the linear system

$$K[u] = f, \qquad \text{where} \qquad K = L^* \circ L,$$

representing our boundary value problem. The norm in (11.156) is typically based on some form of weighted inner product $\langle\!\langle\, v \,,\, \widetilde{v} \,\rangle\!\rangle$ on the space of strains $v = L[u] \in V$, while the inner product term $\langle\, f \,,\, u \,\rangle$ is typically (although not necessarily) unweighted on the space of displacements $u \in U$. The linear operator takes the self-adjoint form $K = L^* \circ L$, and must be positive definite — which requires $\ker L = \{0\}$. Without the positivity assumption, the boundary value problem has either no solutions, or infinitely many; in either event, the basic finite element method will not apply.

Rather than try to minimize $\mathcal{P}[u]$ on the entire function space U, we now seek to minimize it on a suitably chosen finite-dimensional subspace $W \subset U$. We begin by selecting a basis[†] $\varphi_1, \dots, \varphi_n$ of the subspace W. The general element of W is a (uniquely determined) linear combination

$$\varphi(x) = c_1\, \varphi_1(x) + \,\cdots\, + c_n\, \varphi_n(x) \tag{11.157}$$

of the basis functions. Our goal, then, is to determine the coefficients $c_1, \dots, c_n$ such that $\varphi(x)$ minimizes $\mathcal{P}[\varphi]$ among all such functions. Substituting (11.157) into (11.156) and expanding we find

$$\mathcal{P}[\varphi] = \frac{1}{2} \sum_{i,j\,=\,1}^{n} m_{ij}\, c_i\, c_j - \sum_{i\,=\,1}^{n} b_i\, c_i = \tfrac{1}{2}\, \mathbf{c}^T M\, \mathbf{c} - \mathbf{c}^T \mathbf{b}, \tag{11.158}$$

where

(a) $\mathbf{c} = (\, c_1, c_2, \dots, c_n \,)^T$ is the vector of unknown coefficients in (11.157),

(b) $M = (m_{ij})$ is the symmetric $n \times n$ matrix with entries

$$m_{ij} = \langle\!\langle\, L[\varphi_i] \,,\, L[\varphi_j] \,\rangle\!\rangle, \qquad i, j = 1, \dots, n, \tag{11.159}$$

[†] In this case, an orthonormal basis is not of any particular help.

(c) $\mathbf{b} = (\, b_1, b_2, \ldots, b_n \,)^T$ is the vector with entries

$$b_i = \langle\, f \,, \varphi_i \,\rangle, \qquad\qquad i = 1, \ldots, n. \tag{11.160}$$

Observe that, once we specify the basis functions φ_i, the coefficients m_{ij} and b_i are all known quantities. Therefore, we have reduced our original problem to a finite-dimensional problem of minimizing the quadratic function (11.158) over all possible vectors $\mathbf{c} \in \mathbb{R}^n$. The coefficient matrix M is, in fact, positive definite, since, by the preceding computation,

$$\mathbf{c}^T M \mathbf{c} = \sum_{i,j=1}^{n} m_{ij}\, c_i\, c_j = \| \, L[\, c_1\, \varphi_1(x) + \, \cdots \, + c_n\, \varphi_n \,] \, \|^2 = \| \, L[\varphi] \, \|^2 \, > \, 0 \tag{11.161}$$

as long as $L[\varphi] \neq 0$. Moreover, our positivity assumption implies that $L[\varphi] = 0$ if and only if $\varphi \equiv 0$, and hence (11.161) is indeed positive for all $\mathbf{c} \neq \mathbf{0}$. We can now invoke the original finite-dimensional minimization Theorem 4.1 to conclude that the unique minimizer to (11.158) is obtained by solving the associated linear system

$$M \mathbf{c} = \mathbf{b}. \tag{11.162}$$

Solving (11.162) relies on some form of Gaussian Elimination, or, alternatively, an iterative linear system solver, e.g., Gauss–Seidel or SOR.

This constitutes the basic abstract setting for the finite element method. The main issue, then, is how to effectively choose the finite-dimensional subspace W. Two candidates that might spring to mind are the space $\mathcal{P}^{(n)}$ of polynomials of degree $\leq n$, or the space $\mathcal{T}^{(n)}$ of trigonometric polynomials of degree $\leq n$, the focus of Chapter 12. However, for a variety of reasons, neither is well suited to the finite element method. One criterion is that the functions in W must satisfy the relevant boundary conditions — otherwise W would not be a subspace of U. More importantly, in order to obtain sufficient accuracy, the linear algebraic system (11.162) will typically be rather large, and so the coefficient matrix M should be as sparse as possible, i.e., have lots of zero entries. Otherwise, computing the solution will be too time-consuming to be of much practical value. Such considerations prove to be of absolutely crucial importance when applying the method to solve boundary value problems for partial differential equations in higher dimensions.

The really innovative contribution of the finite element method is to first (paradoxically) *enlarge* the space U of allowable functions upon which to minimize the quadratic functional $\mathcal{P}[u]$. The governing differential equation requires its solutions to have a certain degree of smoothness, whereas the associated minimization principle typically requires only half as many derivatives. Thus, for second order boundary value problems, including bars, (11.92), and general Sturm–Liouville problems, (11.155), $\mathcal{P}[u]$ only involves first order derivatives. It can be rigorously shown that the functional has the *same* minimizing solution, even if one allows (reasonable) functions that fail to have enough derivatives to satisfy the differential equation. Thus, one can try minimizing over subspaces containing fairly "rough" functions. Again, the justification of this method requires some deeper analysis, which lies beyond the scope of this introductory treatment.

For second order boundary value problems, a popular and effective choice of the finite-dimensional subspace is to use continuous, piecewise affine functions. Recall that a function

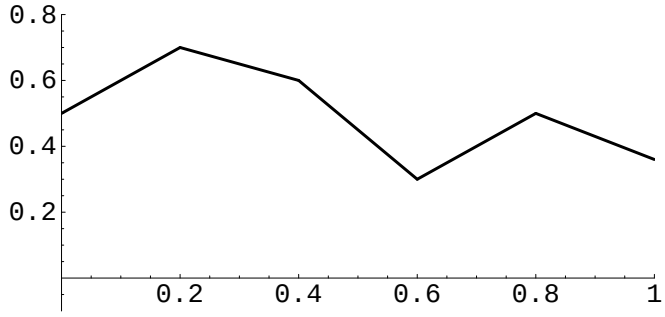


Figure 11.21. A Continuous Piecewise Affine Function.

is affine, $f(x) = a\,x + b$, if and only if its graph is a straight line. The function is *piecewise affine* if its graph consists of a finite number of straight line segments; a typical example is plotted in Figure 11.21. Continuity requires that the individual line segments be connected together end to end.

Given a boundary value problem on a bounded interval $[a, b]$, let us fix a finite collection of *mesh points*

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b.$$

The formulas simplify if one uses equally spaced mesh points, but this is not necessary for the method to apply. Let W denote the vector space consisting of all continuous, piecewise affine functions, with corners at the nodes, that satisfy the homogeneous boundary conditions. To be specific, let us treat the case of Dirichlet (fixed) boundary conditions

$$\varphi(a) = \varphi(b) = 0. \tag{11.163}$$

Thus, on each subinterval

$$\varphi(x) = c_j + b_j(x - x_j), \qquad \text{for} \qquad x_j \le x \le x_{j+1}, \qquad j = 0, \dots, n-1.$$

Continuity of $\varphi(x)$ requires

$$c_j = \varphi(x_j^+) = \varphi(x_j^-) = c_{j-1} + b_{j-1}\,h_{j-1}, \qquad j = 1, \dots, n-1, \tag{11.164}$$

where $h_{j-1} = x_j - x_{j-1}$ denotes the length of the j^{th} subinterval. The boundary conditions (11.163) require

$$\varphi(a) = c_0 = 0, \qquad \varphi(b) = c_{n-1} + h_{n-1}\,b_{n-1} = 0. \tag{11.165}$$

The function $\varphi(x)$ involves a total of $2n$ unspecified coefficients $c_0, \dots, c_{n-1}, b_0, \dots, b_{n-1}$. The continuity conditions (11.164) and the second boundary condition (11.165) uniquely determine the b_j. The first boundary condition specifies c_0, while the remaining $n-1$ coefficients $c_1 = \varphi(x_1), \dots, c_{n-1} = \varphi(x_{n-1})$ are arbitrary. We conclude that the finite element subspace W has dimension $n-1$, which is the number of interior mesh points.

Remark: Every function $\varphi(x)$ in our subspace has piecewise constant first derivative $w'(x)$. However, the jump discontinuities in $\varphi'(x)$ imply that its second derivative $\varphi''(x)$

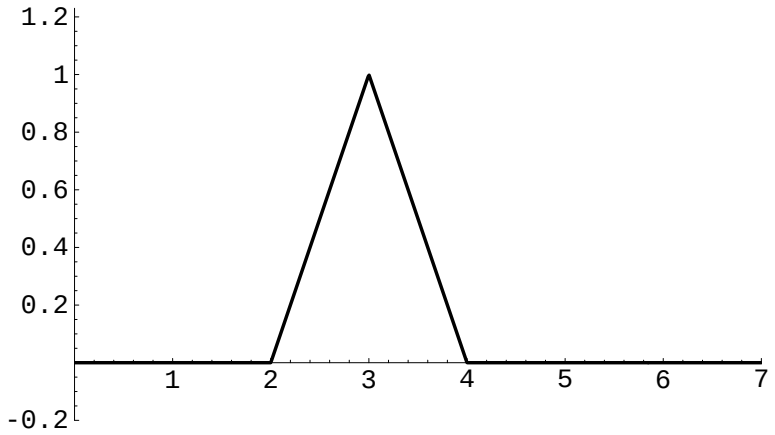


Figure 11.22. A Hat Function.

has a delta function impulse at each mesh point, and is therefore far from being a solution to the differential equation. Nevertheless, the finite element minimizer $\varphi_\star(x)$ will, in practice, provide a reasonable approximation to the actual solution $u_\star(x)$.

The most convenient basis for W consists of the *hat functions*, which are continuous, piecewise affine functions that interpolate the same basis data as the Lagrange polynomials (4.47), namely

$$\varphi_j(x_k) = \begin{cases} 1, & j = k, \\ 0, & j \neq k, \end{cases} \qquad \text{for} \qquad j = 1, \dots, n-1, \qquad k = 0, \dots, n.$$

The graph of a typical hat function appears in Figure 11.22. The explicit formula is easily established:

$$\varphi_j(x) = \begin{cases} \dfrac{x - x_{j-1}}{x_j - x_{j-1}}, & x_{j-1} \leq x \leq x_j, \\ \dfrac{x_{j+1} - x}{x_{j+1} - x_j}, & x_j \leq x \leq x_{j+1}, \\ 0, & x \leq x_{j-1} \quad \text{or} \quad x \geq x_{j+1}, \end{cases} \qquad j = 1, \dots, n-1. \tag{11.166}$$

An advantage of using these basis elements is that the resulting coefficient matrix (11.159) turns out to be tridiagonal. Therefore, the tridiagonal Gaussian Elimination algorithm in (1.63) will rapidly produce the solution to the linear system (11.162). Since the accuracy of the finite element solution increases with the number of mesh points, this solution scheme allows us to easily compute very accurate numerical approximations.

Example 11.14. Consider the equilibrium equations

$$K[u] = -\frac{d}{dx}\left(c(x)\,\frac{du}{dx}\right) = f(x), \qquad 0 < x < \ell,$$

for a non-uniform bar subject to homogeneous Dirichlet boundary conditions. In order to formulate a finite element approximation scheme, we begin with the minimization principle

(11.92) based on the quadratic functional

$$\mathcal{P}[u] = \tfrac{1}{2} \| u' \|^2 - \langle f, u \rangle = \int_0^\ell \big[\tfrac{1}{2} c(x) u'(x)^2 - f(x) u(x) \big] \, dx.$$

We divide the interval $[0, \ell]$ into n equal subintervals, each of length $h = \ell/n$. The resulting uniform mesh has

$$x_j = j h = \frac{j \ell}{n}, \qquad j = 0, \dots, n.$$

The corresponding finite element basis hat functions are explicitly given by

$$\varphi_j(x) = \begin{cases} (x - x_{j-1})/h, & x_{j-1} \le x \le x_j, \\ (x_{j+1} - x)/h, & x_j \le x \le x_{j+1}, \\ 0, & \text{otherwise}, \end{cases} \qquad j = 1, \dots, n-1. \tag{11.167}$$

The associated linear system (11.162) has coefficient matrix entries

$$m_{ij} = \langle\!\langle \varphi_i', \varphi_j' \rangle\!\rangle = \int_0^\ell \varphi_i'(x) \, \varphi_j'(x) \, c(x) \, dx, \qquad i, j = 1, \dots, n-1.$$

Since the function $\varphi_i(x)$ vanishes except on the interval $x_{i-1} < x < x_{i+1}$, while $\varphi_j(x)$ vanishes outside $x_{j-1} < x < x_{j+1}$, the integral will vanish unless $i = j$ or $i = j \pm 1$. Moreover,

$$\varphi_j'(x) = \begin{cases} 1/h, & x_{j-1} \le x \le x_j, \\ -1/h, & x_j \le x \le x_{j+1}, \\ 0, & \text{otherwise}, \end{cases} \qquad j = 1, \dots, n-1.$$

Therefore, the coefficient matrix has the tridiagonal form

$$M = \frac{1}{h^2} \begin{pmatrix} s_0 + s_1 & -s_1 & & & & \\ -s_1 & s_1 + s_2 & -s_2 & & & \\ & -s_2 & s_2 + s_3 & -s_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & -s_{n-3} & s_{n-3} + s_{n-2} & -s_{n-2} & \\ & & & -s_{n-2} & s_{n-2} + s_{n-1} \end{pmatrix}, \tag{11.168}$$

where

$$s_j = \int_{x_j}^{x_{j+1}} c(x) \, dx \tag{11.169}$$

is the total stiffness of the j^{th} subinterval. For example, in the homogeneous case $c(x) \equiv 1$, the coefficient matrix (11.168) reduces to the very special form

$$M = \frac{1}{h} \begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}. \tag{11.170}$$

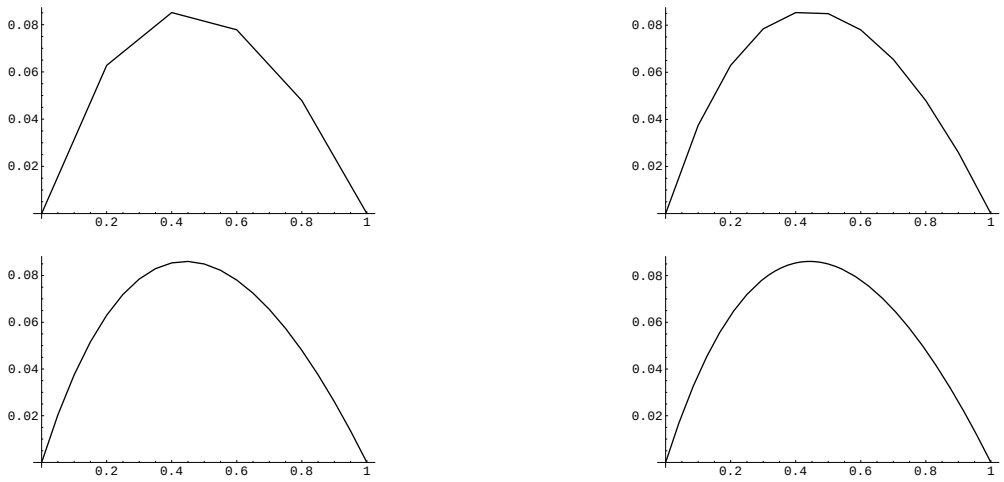


Figure 11.23. Finite Element Solution to (11.174).

The corresponding right hand side has entries

$$\begin{aligned} b_j = \langle\, f\,, \varphi_j\,\rangle &= \int_0^\ell f(x)\,\varphi_j(x)\,dx \\ &= \frac{1}{h}\left[\int_{x_{j-1}}^{x_j} (x - x_{j-1}) f(x)\,dx + \int_{x_j}^{x_{j+1}} (x_{j+1} - x) f(x)\,dx\right]. \end{aligned} \tag{11.171}$$

In this manner, we have assembled the basic ingredients for determining the finite element approximation to the solution.

In practice, we do not have to explicitly evaluate the integrals (11.169, 171), but may replace them by a suitably close numerical approximation. When $h \ll 1$ is small, then the integrals are taken over small intervals, and we can use the trapezoid rule†, [**30**, **151**], to approximate them:

$$s_j \;\approx\; \frac{h}{2}\big[\,c(x_j) + c(x_{j+1})\,\big], \qquad\qquad b_j \;\approx\; h\,f(x_j). \tag{11.172}$$

Remark: The j^{th} entry of the resulting finite element system $M\,\mathbf{c} = \mathbf{b}$ is, upon dividing by h, given by

$$-\,\frac{c_{j+1} - 2\,c_j + c_{j-1}}{h^2} \;=\; -\,\frac{u(x_{j+1}) - 2\,u(x_j) + u(x_{j-1})}{h^2} = -f(x_j). \tag{11.173}$$

The left hand side coincides with the standard finite difference approximation to minus the second derivative $-u''(x_j)$ at the mesh point x_j. (Details concerning finite differences can be found in Section 14.6.) As a result, for this particular differential equation, the finite element and finite difference numerical solution methods happen to coincide.

† One might be tempted use more accurate numerical integration procedures, but the improvement in accuracy of the final answer is not very significant, particularly if the step size h is small.

Example 11.15. Consider the boundary value problem

$$-\frac{d}{dx}\,(x+1)\,\frac{du}{dx} = 1, \qquad u(0) = 0, \qquad u(1) = 0. \tag{11.174}$$

The explicit solution is easily found by direct integration:

$$u(x) = -x + \frac{\log(x+1)}{\log 2}\,. \tag{11.175}$$

It minimizes the associated quadratic functional

$$\mathcal{P}[\,u\,] = \int_0^{\ell} \left[\, \tfrac{1}{2}\,(x+1)\,u'(x)^2 - u(x)\,\right] dx \tag{11.176}$$

over all possible functions $u \in \mathrm{C}^1$ that satisfy the given boundary conditions. The finite element system (11.162) has coefficient matrix given by (11.168) and right hand side (11.171), where

$$s_j = \int_{x_j}^{x_{j+1}} (1+x)\,dx \;= h\,(1+x_j) + \tfrac{1}{2}\,h^2 = h + h^2\,\left(j + \frac{1}{2}\right), \qquad b_j = \int_{x_j}^{x_{j+1}} 1\,dx \;= h.$$

The resulting solution is plotted in Figure 11.23. The first three graphs contain, respectively, 5, 10, 20 points in the mesh, so that $h = .2, .1, .05$, while the last plots the exact solution (11.175). Even when computed on rather coarse meshes, the finite element approximation is quite respectable.

Example 11.16. Consider the Sturm–Liouville boundary value problem

$$-u'' + (x+1)u = x\,e^x, \qquad u(0) = 0, \qquad u(1) = 0. \tag{11.177}$$

The solution minimizes the quadratic functional (11.155), which in this particular case is

$$\mathcal{P}[\,u\,] = \int_0^1 \left[\, \tfrac{1}{2}\,u'(x)^2 + \tfrac{1}{2}\,(x+1)\,u(x)^2 - e^x\,u(x)\,\right] dx, \tag{11.178}$$

over all functions $u(x)$ that satisfy the boundary conditions. We lay out a uniform mesh of step size $h = 1/n$. The corresponding finite element basis hat functions as in (11.167). The matrix entries are given by[†]

$$m_{ij} = \int_0^1 \left[\, \varphi_i'(x)\,\varphi_j'(x) + (x+1)\,\varphi_i(x)\,\varphi_j(x)\,\right] dx \;\approx \begin{cases} \dfrac{2}{h} + \dfrac{2\,h}{3}\,(x_i+1), & i = j, \\ -\dfrac{1}{h} + \dfrac{h}{6}\,(x_i+1), & |\,i-j\,| = 1, \\ 0, & \text{otherwise}, \end{cases}$$

[†] The integration is made easier by noting that the integrand is zero except on a small subinterval. Since the function $x+1$ (but not φ_i or φ_j) does not vary significantly on this subinterval, it can be approximated by its value $1+x_i$ at a mesh point. A similar simplification is used in the ensuing integral for b_i.

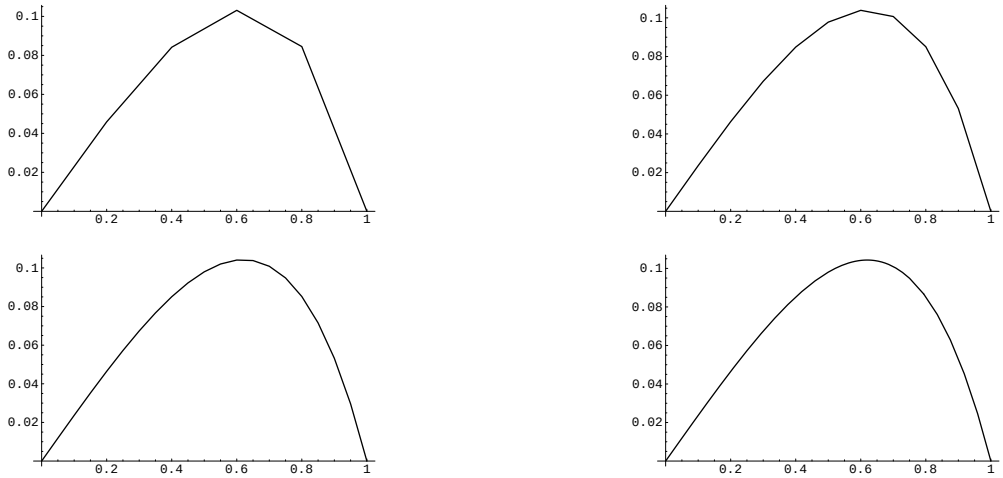


Figure 11.24. Finite Element Solution to (11.177).

while

$$b_i = \langle\, x\, e^x\,, \varphi_i\,\rangle = \int_0^1 x\, e^x\, \varphi_i(x)\, dx \;\approx\; x_i\, e^{x_i}\, h.$$

The resulting solution is plotted in Figure 11.24. As in the previous figure, the first three graphs contain, respectively, 5, 10, 20 points in the mesh, while the last plots the exact solution, which can be expressed in terms of Airy functions, cf. [**129**].

So far, we have only treated homogeneous boundary conditions. An inhomogeneous boundary value problem does not immediately fit into our framework since the set of functions satisfying the boundary conditions does *not* form a vector space. As discussed at the end of Section 11.3, one way to get around this problem is to replace $u(x)$ by $\widetilde{u}(x) = u(x) - h(x)$, where $h(x)$ is any convenient function that satisfies the boundary conditions. For example, for the inhomogeneous Dirichlet conditions

$$u(a) = \alpha, \qquad\qquad u(b) = \beta,$$

we can subtract off the affine function

$$h(x) = \frac{(\beta - \alpha)x + \alpha\, b - \beta\, a}{b - a}.$$

Another option is to choose an appropriate combination of elements at the endpoints:

$$h(x) = \alpha\, \varphi_0(x) + \beta\, \varphi_n(x).$$

Linearity implies that the difference $\widetilde{u}(x) = u(x) - h(x)$ satisfies the amended differential equation

$$K[\,\widetilde{u}\,] = \widetilde{f}, \qquad\qquad \text{where} \qquad\qquad \widetilde{f} = f - K[h],$$

now supplemented by homogeneous boundary conditions. The modified boundary value problem can then be solved by the standard finite element method. Further details are left as a project for the motivated student.

Finally, one can employ other functions beyond the piecewise affine hat functions (11.166) to span finite element subspace. Another popular choice, which is essential for higher order boundary value problems such as beams, is to use splines. Thus, once we have chosen our mesh points, we can let $\varphi_j(x)$ be the basis B–splines discussed in Exercise ■. Since $\varphi_j(x) = 0$ for $x \leq x_{j-2}$ or $x \geq x_{j+2}$, the resulting coefficient matrix (11.159) is *pentadiagonal*, which means $m_{ij} = 0$ whenever $|\, i - j \,| > 2$. Pentadiagonal matrices are not quite as pleasant as their tridiagonal cousins, but are still rather sparse. Positive definiteness of M implies that an iterative solution technique, e.g., SOR, can effectively and rapidly solve the linear system, and thereby produce the finite element spline approximation to the boundary value problem.

Weak Solutions

There is an alternative way of introducing the finite element solution method, which also applies when there is no convenient minimization principle available, based on an important analytical extension of the usual notion of what constitutes a solution to a differential equation. One reformulates the differential equation as an integral equation. The resulting "weak solutions", which include non-classical solutions with singularities and discontinuities, are particularly appropriate in the study of discontinuous and non-smooth physical phenomena, such as shock waves, cracks and dislocations in elastic media, singularities in liquid crystals, and so on; see [**171**] and Section 22.1 for details. The weak solution approach has the advantage that it applies even to equations that do not possess an associated minimization principle. However, the convergence of the induced finite element scheme is harder to justify, and, indeed, not always valid.

The starting point is a trivial observation: the only element of an inner product space which is orthogonal to every other element is zero. More precisely:

Lemma 11.17. *If V is an inner product space, then $\langle\, \mathbf{w}\,, \mathbf{v}\,\rangle = 0$ for all $\mathbf{v} \in V$ if and only if $\mathbf{w} = \mathbf{0}$.*

Proof: Choose $\mathbf{v} = \mathbf{w}$. The orthogonality condition implies $0 = \langle\, \mathbf{w}\,, \mathbf{w}\,\rangle = \|\, \mathbf{w}\,\|^2$, and so $\mathbf{w} = \mathbf{0}$. *Q.E.D.*

Note that the result is equally valid in both finite- and infinite-dimensional vector spaces. Suppose we are trying to solve a linear† system

$$K[\,\mathbf{u}\,] = \mathbf{f}, \tag{11.179}$$

where $K\colon U \to V$ is a linear operator between inner product spaces. Using the lemma, this can be reformulated as requiring

$$\langle\!\langle\, K[\,\mathbf{u}\,]\,, \mathbf{v}\,\rangle\!\rangle = \langle\!\langle\, \mathbf{f}\,, \mathbf{v}\,\rangle\!\rangle \qquad \text{for all} \qquad \mathbf{v} \in V.$$

According to the definition (7.72), one can replace K by its adjoint $K^*\colon W \to V$, and require

$$\langle\, \mathbf{u}\,, K^*[\,\mathbf{v}\,]\,\rangle = \langle\!\langle\, \mathbf{f}\,, \mathbf{v}\,\rangle\!\rangle \qquad \text{for all} \qquad \mathbf{v} \in V. \tag{11.180}$$

† The method also straightforwardly extends to nonlinear systems.

The latter is called the *weak formulation* of our original equation. The general philosophy is that one can check whether $\mathbf{u}$ is a weak solution to the system by evaluating it on various *test elements* $\mathbf{v}$ using the weak form (11.180) of the system.

In the finite-dimensional situation, when K is merely multiplication by some matrix, the weak formulation is an unnecessary complication, and not of use. However, in the infinite-dimensional situation, when K is a differential operator, then the original boundary value problem $K[u] = f$ requires that u be sufficiently differentiable, whereas the weak version

$$\langle\, u\,, K^*[\,\varphi\,]\,\rangle = \langle\!\langle\, f\,, \varphi\,\rangle\!\rangle \qquad \text{for all} \qquad \varphi$$

requires only that the *test function* $\varphi(x)$ be smooth. As a result, weak solutions are not restricted to be smooth functions possessing the required number of derivatives.

Example 11.18. Consider the homogeneous Dirichlet boundary value problem

$$K[u] = -\frac{d}{dx}\left(c(x)\,\frac{du}{dx}\right) = f(x), \qquad 0 < x < \ell, \qquad u(0) = u(\ell) = 0,$$

for a nonuniform bar. Its weak version is obtained by integration by parts. We initially restrict to test functions which vanish at the boundary $\varphi(0) = \varphi(\ell) = 0$. This requirement will eliminate any boundary terms in the integration by parts computation

$$\begin{aligned}\langle\, K[u]\,, \varphi\,\rangle = \int_0^\ell \left[-\frac{d}{dx}\left(c(x)\,\frac{du}{dx}\right)\,\varphi(x)\right] dx &= -\int_0^\ell c(x)\,\frac{du}{dx}\,\frac{d\varphi}{dx}\,dx \\ &= \int_0^\ell f(x)\,\varphi(x)\,dx = \langle\, f\,, \varphi\,\rangle.\end{aligned} \tag{11.181}$$

This "semi-weak" formulation is known in mechanics as the *principle of virtual work*, [**152**]. For example, the Green's function of the boundary value problem does not qualify as a classical solution since it is not twice continuously differentiable, but can be formulated as a weak solution satisfying the virtual work equation with right hand side defined by the delta forcing function.

A second integration by parts produces the weak form (11.180) of the differential equation:

$$\langle\, u\,, K[\,\varphi\,]\,\rangle = -\int_0^\ell u(x)\,\frac{d}{dx}\left(c(x)\,\frac{d\varphi}{dx}\right) dx = \int_0^\ell f(x)\,\varphi(x)\,dx = \langle\, f\,, \varphi\,\rangle. \tag{11.182}$$

Now, even discontinuous functions $u(x)$ are allowed as weak solutions. The goal is to find $u(x)$ such that this condition holds for all smooth test functions $\varphi(x)$. For example, any function $u(x)$ which satisfies the differential equation except at points of discontinuity qualifies as a weak solution.

In a *finite element* or *Galerkin approximation* to the weak solution, one restricts attention to a finite-dimensional subspace W spanned by functions $\varphi_1, \dots, \varphi_{n-1}$, and requires that the approximate solution

$$\varphi(x) = c_1\,\varphi_1(x) + \,\cdots\, + c_{n-1}\,\varphi_{n-1}(x) \tag{11.183}$$

satisfy the orthogonality condition (11.180) only for elements $\varphi \in W$ of the subspace. As usual, this only needs to be checked on the basis elements. Substituting (11.183) into the semi-weak form of the system, (11.181), produces a linear system of equations of the form

$$\langle\, w\,, K[\varphi_i]\,\rangle = \sum_{i=1}^{n} m_{ij}\, c_j = b_i = \langle\, f\,, \varphi_i\,\rangle, \qquad\qquad i = 1, \dots, n. \tag{11.184}$$

The reader will recognize this as exactly the same finite element linear system (11.162) derived through the minimization approach. Therefore, for a self-adjoint boundary value problem, the weak formulation and the minimization principle, when restricted to the finite-dimensional subspace W, lead to exactly the same equations for the finite element approximation to the solution.

In non-self-adjoint scenarios, the weak formulation is still applicable even though there is no underlying minimization principle. On the other hand, there is no guarantee that either the original boundary value problem or its finite element approximation have a solution. Indeed, it is entirely possible that the boundary value problem has a solution, but the finite element matrix system does not. Even more worrying are cases in which the finite element system has a solution, but there is, in fact, no actual solution to the boundary value problem! In such cases, one is usually tipped off by the non-convergence of the approximations as the mesh size goes to zero. Nevertheless, in many situations, the weak solution approach leads to a perfectly acceptable numerical approximation to the true solution to the system. Further analytical details and applications of weak solutions can be found in [**69**, **171**].

Chapter 12

Fourier Series

Just before 1800, the French mathematician/physicist/engineer Jean Baptiste Joseph Fourier made an astonishing discovery. As a result of his investigations into the partial differential equations modeling vibration and heat propagation in bodies, Fourier was led to claim that "every" function could be represented by an infinite series of elementary trigonometric functions — sines and cosines. As an example, consider the sound produced by a musical instrument, e.g., piano, violin, trumpet, oboe, or drum. Decomposing the signal into its trigonometric constituents reveals the fundamental frequencies (tones, overtones, etc.) that are combined to produce its distinctive timbre. The Fourier decomposition lies at the heart of modern electronic music; a synthesizer combines pure sine and cosine tones to reproduce the diverse sounds of instruments, both natural and artificial, according to Fourier's general prescription.

Fourier's claim was so remarkable and unexpected that most of the leading mathematicians of the time did not believe him. Nevertheless, it was not long before scientists came to appreciate the power and far-ranging applicability of Fourier's method, thereby opening up vast new realms of physics, engineering, and elsewhere, to mathematical analysis. Indeed, Fourier's discovery easily ranks in the "top ten" mathematical advances of all time, a list that would include Newton's invention of the calculus, and Gauss and Riemann's establishment of differential geometry that, 70 years later, became the foundation of Einstein's general relativity. Fourier analysis is an essential component of much of modern applied (and pure) mathematics. It forms an exceptionally powerful analytical tool for solving a broad range of partial differential equations. Applications in pure mathematics, physics and engineering are almost too numerous to catalogue — typing in "Fourier" in the subject index of a modern science library will dramatically demonstrate just how ubiquitous these methods are. Fourier analysis lies at the heart of signal processing, including audio, speech, images, videos, seismic data, radio transmissions, and so on. Many modern technological advances, including television, music CD's and DVD's, video movies, computer graphics, image processing, and fingerprint analysis and storage, are, in one way or another, founded upon the many ramifications of Fourier's discovery. In your career as a mathematician, scientist or engineer, you will find that Fourier theory, like calculus and linear algebra, is one of the most basic and essential tools in your mathematical arsenal. Mastery of the subject is unavoidable.

Furthermore, a remarkably large fraction of modern pure mathematics is the result of subsequent attempts to place Fourier series on a firm mathematical foundation. Thus, all of the student's "favorite" analytical tools, including the definition of a function, the ε–δ definition of limit and continuity, convergence properties in function space, including uni-

form convergence, weak convergence, etc., the modern theory of integration and measure, generalized functions such as the delta function, and many others, all owe a profound debt to the prolonged struggle to establish a rigorous framework for Fourier analysis. Even more remarkably, modern set theory, and, as a result, mathematical logic and foundations, can be traced directly back to Cantor's attempts to understand the sets upon which Fourier series converge!

As we will appreciate, Fourier series are, in fact, a very natural outgrowth of the basic linear algebra constructions that we have already developed for analyzing discrete dynamical processes. The Fourier representation of a function is a continuous counterpart of the eigenvector expansions used to solve linear systems of ordinary differential equations. The fundamental partial differential equations governing heat propagation and vibrations in continuous media can be viewed as the function space counterparts of such discrete systems. In the continuous realm, solutions are expressed as linear combinations of simple "separable" solutions constructed from the eigenvalues and eigenfunctions of an associated self-adjoint boundary value problem. The efficacy of Fourier analysis rests on the orthogonality properties of the trigonometric functions, which is a direct consequence of their status as eigenfunctions. So, Fourier series can be rightly viewed as a function space version of the finite-dimensional spectral theory of symmetric matrices and orthogonal eigenvector bases. The main complication is that we must now deal with infinite series rather than finite sums, and so convergence issues that do not appear in the finite-dimensional situation become of paramount importance.

Once we have established the proper theoretical background, the trigonometric Fourier series will no longer be a special, isolated phenomenon, but, rather, in its natural context as the *simplest* representative of a broad class of orthogonal eigenfunction expansions based on self-adjoint boundary value problems. Modern and classical extensions of the Fourier method, including Fourier integrals, discrete Fourier series, wavelets, Bessel functions, spherical harmonics, as well as the entire apparatus of modern quantum mechanics, all rest on the same basic theoretical foundation, and so gaining familiarity with the general theory and abstract eigenfunction framework will be essential. Many of the most important cases used in modern physical and engineering applications will appear in the ensuing chapters.

We begin our development of Fourier methods with a section that will explain why Fourier series naturally appear when we move from discrete systems of ordinary differential equations to the partial differential equations that govern the dynamics of continuous media. The reader uninterested in motivations can safely omit this section as the same material reappears in Chapter 14 when we completely analyze the dynamical partial differential equations that lead to Fourier methods. Beginning in Section 12.2, we shall review, omitting proofs, the most basic computational techniques in Fourier series, for both ordinary and generalized functions. In the final section, we include an abbreviated introduction to the analytical background required to develop a rigorous foundation for Fourier series methods.

12.1. Dynamical Equations of Continuous Media.

The purpose of this section is to discover why Fourier series arise naturally when we move from discrete systems of ordinary differential equations to the partial differential

equations that govern the dynamics of continuous mechanical systems. In our reconstrucdtion of Fourier's thought processes, let us start by reviewing what we have learned.

In Chapter 6, we characterized the equilibrium equations of discrete mechanical and electrical systems as a linear algebraic system

$$K\,\mathbf{u} = \mathbf{f} \tag{12.1}$$

with symmetric, positive (semi-)definite coefficient matrix K. There are two principal types of dynamical systems associated with such equilibrium equations. Free vibrations are governed by Newton's Law, which leads to a second order system of ordinary differential equations (9.59), of the form

$$\frac{d^2\mathbf{u}}{dt^2} = -K\,\mathbf{u}. \tag{12.2}$$

On the other hand, the gradient flow equations (9.21), namely

$$\frac{d\mathbf{u}}{dt} = -K\,\mathbf{u}, \tag{12.3}$$

are designed to decrease the quadratic energy function $q(\mathbf{u}) = \frac{1}{2}\,\mathbf{u}^T K\,\mathbf{u}$ as rapidly as possible. In each case, the solution to the system was made by imposing a particular *ansatz* or inspired guess[†] for the basic solutions. In the case of a gradient flow, the solutions are of exponential form $e^{-\lambda t}\,\mathbf{v}$, while for vibrations they are of trigonometric form $\cos(\omega t)\,\mathbf{v}$ or $\sin(\omega t)\,\mathbf{v}$ with $\omega^2 = \lambda$. In either case, substituting the relevant solution ansatz reduces the dynamical system to the algebraic eigenvalue problem

$$K\,\mathbf{v} = \lambda\,\mathbf{v} \tag{12.4}$$

for the matrix K. Each eigenvalue and eigenvector creates a particular solution or natural mode, and the general solution to the dynamical system can be expressed as a linear superposition of these fundamental modes. The remarkable fact is that the *same* mathematical framework, suitably reinterpreted, carries over directly to the continuous realm!

In Chapter 11, we developed the equilibrium equations governing one-dimensional continuous media — bars, beams, etc. The solution is now a function $u(x)$ representing, say, displacement of the bar, while the positive (semi-)definite matrix is replaced by a certin positive (semi-)definite linear operator $K[u]$. Formally, then, under an external forcing function $f(x)$, the equilibrium system can be written the abstract form

$$K[u] = f. \tag{12.5}$$

By analogy with (12.1, 3), the corresponding gradient flow system will thus be of the form[‡]

$$\frac{\partial u}{\partial t} = -K[u]. \tag{12.6}$$

[†] See the footnote in Example 7.32 for an explanation of this term.

[‡] Since $u(t, x)$ now depends upon time as well as position, we switch from ordinary to partial derivative notation.

Such partial differential equations model diffusion processes in which a quadratic energy functional is decreasing as rapidly as possible. A good physical example is the flow of heat in a body; the heat disperses throughout the body so as to decrease the thermal energy as quickly as it can, tending (in the absence of external heat sources) to thermal equilibrium. Other physical processes modeled by (12.6) include diffusion of chemicals (solvents, pollutants, etc.), and of populations (animals, bacteria, people, etc.).

The simplest and most instructive example is a uniform periodic (or circular) bar of length 2π. As we saw in Chapter 11, the equilibrium equation for the temperature $u(x)$ takes the form

$$K[u] = -u'' = f, \qquad u(-\pi) = u(\pi), \qquad u'(-\pi) = u'(\pi), \tag{12.7}$$

associated with the positive semi-definite differential operator

$$K = D^* \circ D = (-D)\, D = -D^2 \tag{12.8}$$

acting on the space of 2π periodic functions. The corresponding gradient flow (12.6) is the partial differential equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}\,, \qquad u(t,-\pi) = u(t,\pi), \qquad \frac{\partial u}{\partial x}\,(t,-\pi) = \frac{\partial u}{\partial x}\,(t,\pi), \tag{12.9}$$

known as the *heat equation* since it models (among other diffusion processes) heat flow. The solution $u(t,x)$ represents the temperature at position x and time t, and turns out to be uniquely prescribed by the initial distribution:

$$u(0,x) = f(x), \qquad -\pi \le x \le \pi. \tag{12.10}$$

Heat naturally flows from hot to cold, and so the fact that it can be described by a gradient flow should not be surprising; a derivation of (12.9) from physical principles will appear in Chapter 14. Solving the periodic heat equation was the seminal problem that led Fourier to develop the profound theory that now bears his name.

As in the discrete version, the elemental solutions to a diffusion equation (12.6) are found by introducing an exponential ansatz:

$$u(t,x) = e^{-\lambda t}\, v(x), \tag{12.11}$$

in which we replace the eigenvector $\mathbf{v}$ by a function $v(x)$. These are often referred to as *separable solutions* to indicate that they are the product of a function of t alone times a function of x alone. We substitute the solution formula (12.11) into the dynamical equations (12.6). We compute

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial t}\left[\, e^{-\lambda t}\, v(x)\,\right] = -\lambda\, e^{-\lambda t}\, v(x), \quad \text{while} \quad -K[u] = -K\left[\, e^{-\lambda t}\, v(x)\,\right] = -e^{-\lambda t}\, K[v],$$

since the exponential factor is a function of t, while K only involves differentiation with respect to x. Equating these two expressions and canceling the common exponential factor, we conclude that $v(x)$ must solve a boundary value problem of the form

$$K[v] = \lambda\, v. \tag{12.12}$$

We interpret λ as the *eigenvalue* and $v(x)$ as the corresponding *eigenfunction* for the operator K subject to the relevant boundary conditions. Each eigenvalue and eigenfunction pair will produce a solution (12.11) to the partial differential equation, and the general solution can be built up through linear superposition.

For example, substitution of the exponential ansatz (12.11) into the periodic heat equation (12.9) leads to the eigenvalue problem

$$v'' + \lambda v = 0, \qquad v(-\pi) = v(\pi), \qquad v'(-\pi) = v'(\pi). \tag{12.13}$$

for the *eigenfunction* $v(x)$. Now, it is not hard to show that if $\lambda < 0$ or λ is complex, then the only periodic solution to (12.13) is the trivial solution $v(x) \equiv 0$. Thus, all eigenvalues must be real and non-negative: $\lambda \geq 0$. This is not an accident — as we will discuss in detail in Section 14.7, it is a direct consequence of the positive semi-definiteness of the underlying differential operator (12.8). When $\lambda = 0$, periodicity singles out the nonzero constants $v(x) \equiv c \neq 0$ as the associated eigenfunctions. If $\lambda = \omega^2 > 0$, then the general solution to the differential equation (12.13) is a linear combination

$$v(x) = a \cos \omega x + b \sin \omega x$$

of the basis solutions. A nonzero function of this form will satisfy the 2π periodic boundary conditions if and only if $\omega = k$ is an integer. Therefore, the eigenvalues

$$\lambda = k^2, \qquad 0 \leq k \in \mathbb{N},$$

are the squares of positive integers. Each positive eigenvalue $\lambda = k^2 > 0$ admits two linearly independent eigenfunctions, namely $\sin k x$ and $\cos k x$, while the zero eigenvalue $\lambda = 0$ has only one, the constant function 1. We conclude that the basic trigonometric functions

$$1, \qquad \cos x, \qquad \sin x, \qquad \cos 2x, \qquad \sin 2x, \qquad \cos 3x, \qquad \dots \tag{12.14}$$

form a complete system of independent eigenfunctions for the periodic boundary value problem (12.13).

By construction, each eigenfunction gives rise to a particular solution (12.11) to the periodic heat equation (12.9). We have therefore discovered an infinite collection of independent solutions:

$$u_k(x) = e^{-k^2 t} \cos k x, \qquad \widetilde{u}_k(x) = e^{-k^2 t} \sin k x, \qquad k = 0, 1, 2, 3, \dots .$$

Linear superposition tells us that finite linear combinations of solutions are also solutions. However, these will *not* suffice to describe the general solution to the problem, and so we are led to propose an infinite series[†]

$$u(t, x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} \Big[a_k e^{-k^2 t} \cos k x + b_k e^{-k^2 t} \sin k x \Big] \tag{12.15}$$

[†] For technical reasons, one takes the basic null eigenfunction to be $\frac{1}{2}$ instead of 1. The explanation will be revealed in the following section.

to represent the general solution to the periodic heat equation. As in the discrete version, the coefficients a_k, b_k are found by substituting the solution formula into the initial condition (12.10), whence

$$u(0,x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} \big[\, a_k \cos k\,x + b_k \sin k\,x \,\big] = f(x). \tag{12.16}$$

The result is the *Fourier series* representation of the initial temperature distribution. Once we have prescribed the Fourier coefficients a_k, b_k, (12.15) provides an explicit formula for the solution to the periodic initial-boundary value problem for the heat equation.

However, since we are dealing with infinite series, the preceding is purely a formal construction, and requires some serious mathematical analysis to place it on a firm footing. The key questions are

- First, when does such an infinite trigonometric series converge?
- Second, what kinds of functions $f(x)$ can be represented by a convergent Fourier series?
- Third, if we have such an f, how do we determine its Fourier coefficients a_k, b_k?
- And lastly, since we are trying to solve differential equations, can we safely differentiate a Fourier series?

These are the fundamental questions of Fourier analysis, and must be properly dealt with before we can make any serious progress towards solving the heat equation.

A similar analysis applies to a second order dynamical system of the Newtonian form

$$\frac{\partial^2 u}{\partial t^2} = -\,K[\,u\,]. \tag{12.17}$$

Such differential equations are used to describe the free vibrations of continuous mechanical systems, such as bars, strings, and, in higher dimensions, membranes, solid bodies, fluids, etc. For example, the vibration system (12.17) corresponding to the differential operator (12.8) is the *wave equation*

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}\,. \tag{12.18}$$

The wave equation models stretching vibrations of a bar, sound vibrations in a column of air, e.g., inside a wind instrument, transverse vibrations of a string, e.g., a violin string, surfaces waves on a fluid, electromagnetic waves, and a wide variety of other vibrational and wave phenomena.

As always, we need to impose suitable boundary conditions in order to proceed. Consider, for example, the wave equation with homogeneous Dirichlet boundary conditions

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}\,, \qquad\qquad u(t,0) = 0, \qquad\qquad u(t,\ell) = 0, \tag{12.19}$$

that models, for instance, the vibrations of a uniform violin string whose ends are tied down. Adapting our discrete trigonometric ansatz, we are naturally led to look for a separable solution of the form

$$u(t,x) = \cos(\omega\,t)\,v(x) \tag{12.20}$$

in which ω represents the vibrational frequency. Substituting into the wave equation and the associated boundary conditions, we deduce that $v(x)$ must be a solution to the eigenvalue problem

$$\frac{d^2v}{dx^2} + \omega^2\, v = 0, \qquad\qquad v(0) = 0 = v(\ell), \tag{12.21}$$

in which $\omega^2 = \lambda$ plays the role of the eigenvalue. We require a *nonzero* solution to this linear boundary value problem, and this requires ω^2 to be strictly positive. As above, this can be checked by directly solving the boundary value problem, but is, in fact, a consequence of positive definiteness; see Section 14.7 for details. Assuming $\omega^2 > 0$, the general solution to the differential equation is a trigonometric function

$$v(x) = a\cos\omega\, x + b\sin\omega\, x.$$

The boundary condition at $x = 0$ requires $a = 0$, and so

$$v(x) = b\sin\omega\, x.$$

The second boundary condition requires

$$v(\ell) = b\sin\omega\,\ell = 0.$$

Assuming $b \neq 0$, as otherwise the solution is trivial, $\omega\,\ell$ must be an integer multiple of π. Thus, the natural frequencies of vibration are

$$\omega_k = \frac{k\,\pi}{\ell}\,, \qquad\qquad k = 1, 2, 3, \ldots\,.$$

The corresponding eigenfunctions are

$$v_k(x) = \sin\frac{k\,\pi\, x}{\ell}\,, \qquad\qquad k = 1, 2, 3, \ldots\,. \tag{12.22}$$

Thus, we find the following natural modes of vibration of the wave equation:

$$u_k(t,x) = \cos\frac{k\,\pi\, t}{\ell}\,\sin\frac{k\,\pi\, x}{\ell}\,, \qquad\qquad \widetilde{u}_k(t,x) = \sin\frac{k\,\pi\, t}{\ell}\,\sin\frac{k\,\pi\, x}{\ell}\,.$$

Each solution represents a spatially periodic standing wave form. We expect to write the general solution to the boundary value problem as an infinite series

$$u(t,x) = \sum_{k=1}^{\infty}\left(b_k\,\cos\frac{k\,\pi\, t}{\ell}\,\sin\frac{k\,\pi\, x}{\ell} + d_k\,\sin\frac{k\,\pi\, t}{\ell}\sin\frac{k\,\pi\, x}{\ell}\right) \tag{12.23}$$

in the natural modes. Interestingly, in this case at each fixed t, there are no cosine terms, and so we have a more specialized type of Fourier series. The same convergence issues for such *Fourier sine series* arise. It turns out that the general theory of Fourier series will also cover Fourier sine series.

We have now completed our brief introduction to the dynamical equations of continuous media and the Fourier series method of solution. The student should now be sufficiently motivated, and it is time to delve into the underlying Fourier theory. In Chapter 14 we will return to the applications to the one-dimensional heat and wave equations.

12.2. Fourier Series.

While the need to solve physically interesting partial differential equations served as our (and Fourier's) initial motivation, the remarkable range of applications qualifies Fourier's discovery as one of the most important in all of mathematics. We therefore take some time to properly develop the basic theory of Fourier series and, in the following chapter, a number of important extensions. Then, properly equipped, we will be in a position to return to the source — solving partial differential equations.

The starting point is the need to represent a given function $f(x)$, defined for $-\pi \leq x \leq \pi$, as a convergent series in the elementary trigonometric functions:

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} \left[a_k \cos k\,x + b_k \sin k\,x \right]. \tag{12.24}$$

The first order of business is to determine the formulae for the Fourier coefficients a_k, b_k. The key is orthogonality. We already observed, in Example 5.12, that the trigonometric functions are orthogonal with respect to the rescaled L^2 inner product

$$\langle\, f\,, g\,\rangle = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x)\, g(x)\, dx \tag{12.25}$$

on the interval[†] $[-\pi, \pi]$. The explicit orthogonality relations are

$$\begin{aligned} \langle \cos k\,x\,, \cos l\,x\,\rangle = \langle \sin k\,x\,, \sin l\,x\,\rangle = 0, \qquad & \text{for} \quad k \neq l, \\ \langle \cos k\,x\,, \sin l\,x\,\rangle = 0, \qquad & \text{for all} \quad k, l, \\ \| \,1\, \| = \sqrt{2}\,, \qquad \| \cos k\,x\, \| = \| \sin k\,x\, \| = 1, \qquad & \text{for} \quad k \neq 0, \end{aligned} \tag{12.26}$$

where k and l indicate non-negative integers.

Remark: If we were to replace the constant function 1 by $\frac{1}{\sqrt{2}}$, then the resulting functions would form an orthonormal system. However, this extra $\sqrt{2}$ turns out to be utterly annoying, and is best omitted from the outset.

Remark: Orthogonality of the trigonometric functions is not an accident, but follows from their status as eigenfunctions for the self-adjoint boundary value problem (12.13). The general result, to be presented in Section 14.7, is the function space analog of the orthogonality of eigenvectors of symmetric matrices, cf. Theorem 8.20.

If we ignore convergence issues for the moment, then the orthogonality relations (12.26) serve to prescribe the Fourier coefficients: Taking the inner product of both sides

[†] We have chosen the interval $[-\pi, \pi]$ for convenience. A common alternative is the interval $[0, 2\pi]$. In fact, since the trigonometric functions are 2π periodic, any interval of length 2π will serve equally well. Adapting Fourier series to intervals of other lengths will be discussed in Section 12.4.

with $\cos l\,x$ for $l > 0$, and invoking the underlying linearity[‡] of the inner product, yields

$$\begin{aligned}\langle\, f\,, \cos l\,x\,\rangle &= \frac{a_0}{2}\,\langle\, 1\,, \cos l\,x\,\rangle \;+\; \sum_{k=1}^{\infty}\,\big[\, a_k\,\langle\, \cos k\,x\,, \cos l\,x\,\rangle + b_k\,\langle\, \sin k\,x\,, \cos l\,x\,\rangle\,\big]\\ &= a_l\,\langle\, \cos l\,x\,, \cos l\,x\,\rangle = a_l,\end{aligned}$$

since, by the orthogonality relations (12.26), all terms but the l^{th} vanish. This serves to prescribe the Fourier coefficient a_l. A similar manipulation with $\sin l\,x$ fixes $b_l = \langle\, f\,, \sin l\,x\,\rangle$, while taking the inner product with the constant function 1 gives

$$\langle\, f\,, 1\,\rangle = \frac{a_0}{2}\,\langle\, 1\,, 1\,\rangle \;+\; \sum_{k=1}^{\infty}\,\big[\, a_k\,\langle\, \cos k\,x\,, 1\,\rangle + b_k\,\langle\, \sin k\,x\,, 1\,\rangle\,\big] = \frac{a_0}{2}\,\|\,1\,\|^2 = a_0,$$

which agrees with the preceding formula for a_l when $l = 0$, and explains why we include the extra factor of $\frac{1}{2}$ in the constant term. Thus, *if the Fourier series converges to the function $f(x)$, then its coefficients are prescribed by taking inner products with the basic trigonometric functions*. The alert reader may recognize the preceding argument — it is the function space version of our derivation or the fundametnal orthonormal and orthogonal basis formulae (5.4, 7), which are valid in any inner product space. The key difference here is that we are dealing with infinite series instead of finite sums, and copnvergence issues must be properly addressed. However, we defer these more delicate considerations until after we have gained some basic familiarity with how Fourier series work in practice.

Let us summarize where we are with the following fundamental definition.

Definition 12.1. The *Fourier series* of a function $f(x)$ defined on $-\pi \le x \le \pi$ is the infinite trigonometric series

$$f(x) \quad\sim\quad \frac{a_0}{2} \;+\; \sum_{k=1}^{\infty}\,\big[\, a_k\,\cos k\,x + b_k\,\sin k\,x\,\big]\,, \tag{12.27}$$

whose coefficients are given by the inner product formulae

$$\begin{aligned} a_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos k\,x\,dx, \qquad & k &= 0, 1, 2, 3, \ldots,\\ b_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin k\,x\,dx, \qquad & k &= 1, 2, 3, \ldots. \end{aligned} \tag{12.28}$$

Note that the function $f(x)$ cannot be completely arbitrary, since, at the vary least, the integrals in the coefficient formulae must be well defined and finite. Even if the the coefficients (12.28) are finite, there is no guarantee that the resulting Fourier series converges, and, even if it converges, no guarantee that it converges to the original function $f(x)$. For these reasons, we use the $\sim$ symbol instead of an equals sign when writing down a Fourier series. Before tackling these key issues, let us look at an elementary example.

[‡] More rigorously, linearity only applies to finite linear combinations, not infinite series. Here, thought, we are just trying to establish and motivate the basic formulae, and can safely defer such technical complications until the final section.

Example 12.2. Consider the function $f(x) = x$. We may compute its Fourier coefficients directly, employing integration by parts to evaluate the integrals:

$$a_0 = \frac{1}{\pi}\int_{-\pi}^{\pi} x\,dx = 0, \qquad a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} x\cos k\,x\,dx = \frac{1}{\pi}\left[\frac{x\,\sin k\,x}{k} + \frac{\cos k\,x}{k^2}\right]\Bigg|_{x=-\pi}^{\pi} = 0,$$

$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} x\sin k\,x\,dx = \frac{1}{\pi}\left[-\frac{x\,\cos k\,x}{k} + \frac{\sin k\,x}{k^2}\right]\Bigg|_{x=-\pi}^{\pi} = \frac{2}{k}\,(-1)^{k+1}. \tag{12.29}$$

Therefore, the Fourier cosine coefficients of the function x all vanish, $a_k = 0$, and its Fourier series is

$$x \quad\sim\quad 2\left(\sin x \;-\; \frac{\sin 2\,x}{2} \;+\; \frac{\sin 3\,x}{3} \;-\; \frac{\sin 4\,x}{4} \;+\; \cdots\right). \tag{12.30}$$

Convergence of this series is *not* an elementary matter. Standard tests, including the ratio and root tests, fail to apply. Even if we know that the series converges (which it does — for all x), it is certainly not obvious what function it converges to. Indeed, it *cannot* converge to the function $f(x) = x$ for all values of x. If we substitute $x = \pi$, then every term in the series is zero, and so the Fourier series converges to 0 — which is not the same as $f(\pi) = \pi$.

The n^{th} *partial sum* of a Fourier series is the trigonometric polynomial[†]

$$s_n(x) \;=\; \frac{a_0}{2} + \sum_{k=1}^{n}\,\left[\,a_k\,\cos k\,x + b_k\,\sin k\,x\,\right]. \tag{12.31}$$

By definition, the Fourier series *converges* at a point x if and only if the partial sums have a limit:

$$\lim_{n\to\infty} s_n(x) = \widetilde{f}(x), \tag{12.32}$$

which may or may not equal the value of the original function $f(x)$. Thus, a key requirement is to formulate conditions on the function $f(x)$ that guarantee that the Fourier series converges, and, even more importantly, the limiting sum reproduces the original function: $\widetilde{f}(x) = f(x)$. This will all be done in detail below.

Remark: The passage from trigonometric polynomials to Fourier series is similar to the passage from polynomials to power series. A power series

$$f(x) \quad\sim\quad c_0 + c_1\,x + \;\cdots\; + c_n\,x^n + \;\cdots\; = \sum_{k=0}^{\infty} c_k\,x^k$$

can be viewed as an infinite linear combination of the basic monomials $1, x, x^2, x^3, \ldots$. According to Taylor's formula, (C.3), the coefficients $c_k = \dfrac{f^{(k)}(0)}{k!}$ are given in terms of

[†] The reason for the term "trigonometric polynomial" was discussed at length in Example 2.17(c).

the derivatives of the function at the origin. The partial sums

$$s_n(x) = c_0 + c_1\, x + \cdots + c_n\, x^n = \sum_{k=0}^{n} c_k\, x^k$$

of a power series are ordinary polynomials, and the same convergence issues arise.

Although superficially similar, in actuality the two theories are profoundly different. Indeed, while the theory of power series was well established in the early days of the calculus, there remain, to this day, unresolved foundational issues in Fourier theory. A power series either converges everywhere, or on an interval centered at 0, or nowhere except at 0. (See Section 16.2 for additional details.) On the other hand, a Fourier series can converge on quite bizarre sets. In fact, the detailed analysis of the convergence properties of Fourier series led the nineteenth century German mathematician Georg Cantor to formulate modern set theory, and, thus, played a seminal role in the establishment of the foundations of modern mathematics. Secondly, when a power series converges, it converges to an analytic function, which is infinitely differentiable, and whose derivatives are represented by the power series obtained by termwise differentiation. Fourier series may converge, not only to periodic continuous functions, but also to a wide variety of discontinuous functions and, even, when suitably interpreted, to generalized functions like the delta function! Therefore, the termwise differentiation of a Fourier series is a nontrivial issue.

Once one comprehends how different the two subjects are, one begins to understand why Fourier's astonishing claims were initially widely disbelieved. Before the advent of Fourier, mathematicians only accepted analytic functions as the genuine article. The fact that Fourier series can converge to nonanalytic, even discontinuous functions was extremely disconcerting, and resulted in a complete re-evaluation of function theory, culminating in the modern definition of function that you now learn in first year calculus. Only through the combined efforts of many of the leading mathematicians of the nineteenth century was a rigorous theory of Fourier series firmly established; see Section 12.5 for the main details and the advanced text [**193**] for a comprehensive treatment.

Periodic Extensions

The trigonometric constituents (12.14) of a Fourier series are all periodic functions of period 2π. Therefore, if the series converges, the limiting function $\widetilde{f}(x)$ must also be periodic of period 2π:

$$\widetilde{f}(x + 2\pi) = \widetilde{f}(x) \qquad \text{for all} \qquad x \in \mathbb{R}.$$

A Fourier series can only converge to a 2π periodic function. So it was unreasonable to expect the Fourier series (12.30) to converge to the non-periodic to $f(x) = x$ everywhere. Rather, it should converge to its periodic extension, as we now define.

Lemma 12.3. *If $f(x)$ is any function defined for $-\pi < x \le \pi$, then there is a unique 2π periodic function $\widetilde{f}$, known as the 2π* periodic extension *of f, that satisfies $\widetilde{f}(x) = f(x)$ for all $-\pi < x \le \pi$.*

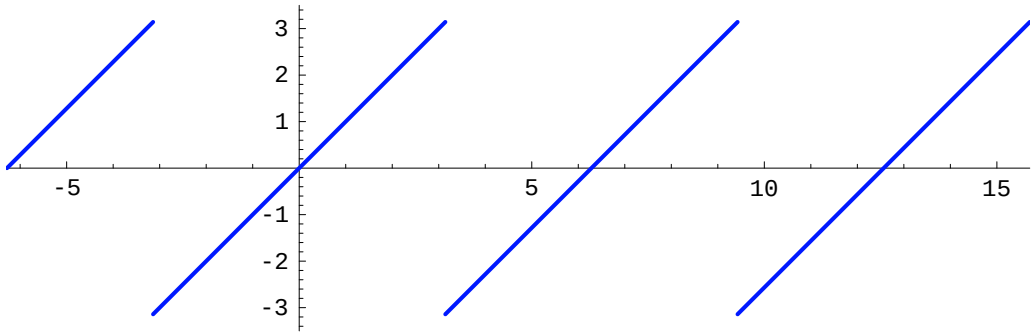


Figure 12.1. Periodic extension of x.

Proof: Pictorially, the graph of the periodic extension of a function $f(x)$ is obtained by repeatedly copying that part of the graph of f between $-\pi$ and π to adjacent intervals of length 2π; Figure 12.1 shows a simple example. More formally, given $x \in \mathbb{R}$, there is a unique integer m so that $(2m-1)\pi < x \leq (2m+1)\pi$. Periodicity of $\widetilde{f}$ leads us to define

$$\widetilde{f}(x) = \widetilde{f}(x - 2m\pi) = f(x - 2m\pi), \tag{12.33}$$

noting that if $-\pi < x \leq \pi$, then $m = 0$ and hence $\widetilde{f}(x) = f(x)$ for such x. The proof that the resulting function $\widetilde{f}$ is 2π periodic is left as Exercise ∎. *Q.E.D.*

Remark: The construction of the periodic extension of Lemma 12.3 uses the value $f(\pi)$ at the right endpoint and requires $\widetilde{f}(-\pi) = \widetilde{f}(\pi) = f(\pi)$. One could, alternatively, require $\widetilde{f}(\pi) = \widetilde{f}(-\pi) = f(-\pi)$, which, if $f(-\pi) \neq f(\pi)$, leads to a slightly different 2π periodic extension of the function. There is no *a priori* reason to prefer one over the other. In fact, for Fourier theory, as we shall discover, one should use neither, but rather an "average" of the two. Thus, the preferred Fourier periodic extension $\widetilde{f}(x)$ will satisfy

$$\widetilde{f}(\pi) = \widetilde{f}(-\pi) = \tfrac{1}{2}\big[\, f(\pi) + f(-\pi) \,\big], \tag{12.34}$$

which then fixes its values at the odd multiples of π.

Example 12.4. The 2π periodic extension $\widetilde{f}(x)$ of $f(x) = x$ is the "sawtooth" function graphed in Figure 12.1. It agrees with x between $-\pi$ and π. Since $f(\pi) = \pi$, $f(-\pi) = -\pi$, the Fourier extension (12.34) sets $\widetilde{f}(k\pi) = 0$ for any odd integer k. Explicitly,

$$\widetilde{f}(x) = \begin{cases} x - 2m\pi, & (2m-1)\pi < x < (2m+1)\pi, \\ 0, & x = (2m-1)\pi, \end{cases} \qquad \text{where } m \text{ is any integer.}$$

With this convention, it can be proved that the Fourier series (12.30) converges everywhere to the 2π periodic extension $\widetilde{f}(x)$. In particular,

$$2\sum_{k=1}^{\infty} (-1)^{k+1}\, \frac{\sin k x}{k} = \begin{cases} x, & -\pi < x < \pi, \\ 0, & x = \pm\pi. \end{cases} \tag{12.35}$$

Even this very simple example has remarkable and nontrivial consequences. For instance, if we substitute $x = \frac{1}{2}\pi$ in (12.30) and divide by 2, we obtain *Gregory's series*

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \cdots . \tag{12.36}$$

While this striking formula predates Fourier theory — it was, in fact, first discovered by Leibniz — a direct proof is not easy.

Remark: While numerologically fascinating, Gregory's series is of scant practical use for actually computing π since its rate of convergence is painfully slow. The reader may wish to try adding up terms to see how far out one needs to go to accurately compute even the first two decimal digits of π. Round-off errors will eventually interfere with any attempt to compute the complete summation to any reasonable degree of accuracy.

Piecewise Continuous Functions

As we shall see, all continuously differentiable, 2π periodic functions can be represented as convergent Fourier series. More generally, we can allow the function to have some simple discontinuities. Although not the most general class of functions that possess convergent Fourier series, such "piecewise continuous" functions will suffice for all the applications we consider in this text.

Definition 12.5. A function $f(x)$ is said to be *piecewise continuous* on an interval $[a, b]$ if it is defined and continuous except possibly at a finite number of points $a \leq x_1 < x_2 < \ldots < x_n \leq b$. At each point of discontinuity, the left and right hand limits[†]

$$f(x_k^-) = \lim_{x \to x_k^-} f(x), \qquad f(x_k^+) = \lim_{x \to x_k^+} f(x),$$

exist. Note that we do not require that $f(x)$ be defined at x_k. Even if $f(x_k)$ is defined, it does not necessarily equal either the left or the right hand limit.

A function $f(x)$ defined for all $x \in \mathbb{R}$ is piecewise continuous provided it is piecewise continuous on every bounded interval. In particular, a 2π periodic function $\widetilde{f}(x)$ is piecewise continuous if and only if it is piecewise continuous on the interval $[-\pi, \pi]$.

A representative graph of a piecewise continuous function appears in Figure 12.2. The points x_k are known as *jump discontinuities* of $f(x)$ and the difference

$$\beta_k = f(x_k^+) - f(x_k^-) = \lim_{x \to x_k^+} f(x) - \lim_{x \to x_k^-} f(x) \tag{12.37}$$

between the left and right hand limits is the *magnitude* of the jump, cf. (11.48). If $\beta_k = 0$, and so the right and left hand limits agree, then the discontinuity is *removable* since redefining $f(x_k) = f(x_k^+) = f(x_k^-)$ makes f continuous at x_k. We will assume, without significant loss of generality, that our functions have no removable discontinuities.

[†] At the endpoints a, b we only require one of the limits, namely $f(a^+)$ and $f(b^-)$, to exist.

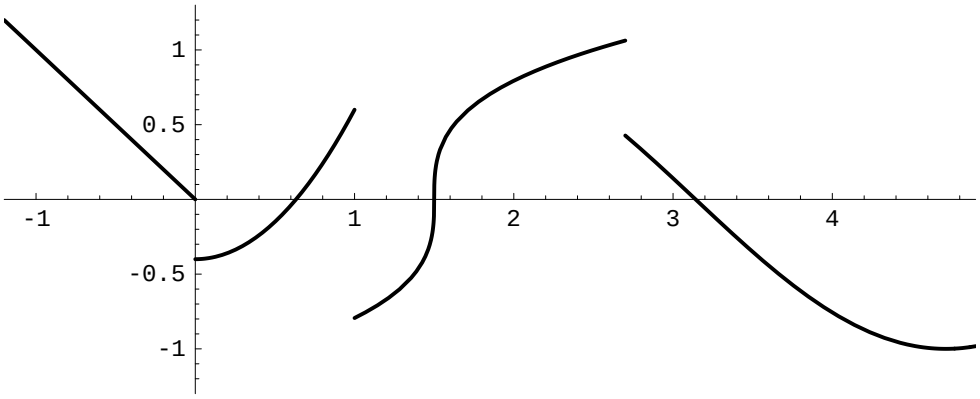


Figure 12.2. Piecewise Continuous Function.

The simplest example of a piecewise continuous function is the step function

$$\sigma(x) = \begin{cases} 1, & x > 0, \\ 0, & x < 0. \end{cases} \tag{12.38}$$

It has a single jump discontinuity at $x = 0$ of magnitude 1, and is continuous — indeed, constant — everywhere else. If we translate and scale the step function, we obtain a function

$$h(x) = \beta\, \sigma(x - y) = \begin{cases} \beta, & x > y, \\ 0, & x < y, \end{cases} \tag{12.39}$$

with a single jump discontinuity of magnitude β at the point $x = y$.

If $f(x)$ is any piecewise continuous function, then its Fourier coefficients are well-defined — the integrals (12.28) exist and are finite. Continuity, however, is not enough to ensure convergence of the resulting Fourier series.

Definition 12.6. A function $f(x)$ is called *piecewise* C^1 on an interval $[a, b]$ if it is defined, continuous and continuously differentiable except possibly at a finite number of points $a \le x_1 < x_2 < \ldots < x_n \le b$. At each exceptional point, the left and right hand limits[†] exist:

$$f(x_k^-) = \lim_{x \to x_k^-} f(x), \qquad f(x_k^+) = \lim_{x \to x_k^+} f(x),$$
$$f'(x_k^-) = \lim_{x \to x_k^-} f'(x), \qquad f'(x_k^+) = \lim_{x \to x_k^+} f'(x).$$

See Figure 12.3 for a representative graph. For a piecewise continuous C^1 function, an exceptional point x_k is either

- a *jump discontinuity* of f, but where the left and right hand derivatives exist, or
- a *corner*, meaning a point where f is continuous, so $f(x_k^-) = f(x_k^+)$, but has different left and right hand derivatives: $f'(x_k^-) \neq f'(x_k^+)$.

[†] As before, at the endpoints we only require the appropriate one-sided limits, namely $f(a^+)$, $f'(a^+)$ and $f(b^-), f'(b^-)$, to exist.

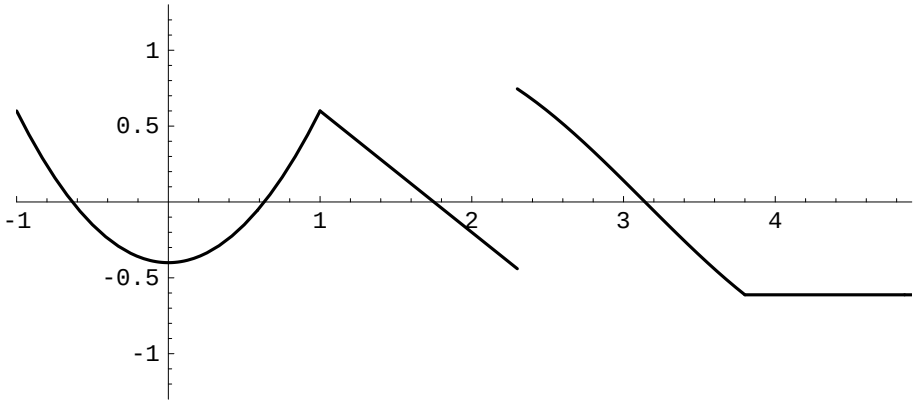


Figure 12.3. Piecewise C^1 Function.

Thus, at each point, including jump discontinuities, the graph of $f(x)$ has well-defined right and left tangent lines. For example, the function $f(x) = |x|$ is piecewise C^1 since it is continuous everywhere and has a corner at $x = 0$, with $f'(0^+) = +1$, $f'(0^-) = -1$.

There is an analogous definition of a piecewise C^n function. One requires that the function has n continuous derivatives, except at a finite number of points. Moreover, at every point, the function has well-defined right and left hand limits of all its derivatives up to order n.

The Convergence Theorem

We are now able to state the fundamental convergence theorem for Fourier series.

Theorem 12.7. *If $\widetilde{f}(x)$ is any 2π periodic, piecewise C^1 function, then, for any $x \in \mathbb{R}$, its Fourier series converges to*

$$\begin{cases} \widetilde{f}(x), & \text{if } \widetilde{f} \text{ is continuous at } x, \\ \frac{1}{2}\big[\,\widetilde{f}(x^+) + \widetilde{f}(x^-)\,\big], & \text{if } x \text{ is a jump discontinuity.} \end{cases}$$

Thus, the Fourier series converges, as expected, to $f(x)$ at all points of continuity; at discontinuities, the Fourier series can't decide whether to converge to the right or left hand limit, and so ends up "splitting the difference" by converging to their average; see Figure 12.4. If we redefine $\widetilde{f}(x)$ at its jump discontinuities to have the average limiting value, so

$$\widetilde{f}(x) = \tfrac{1}{2}\,\big[\,\widetilde{f}(x^+) + \widetilde{f}(x^-)\,\big], \tag{12.40}$$

— an equation that automatically holds at all points of continuity — then Theorem 12.7 would say that the Fourier series converges to $\widetilde{f}(x)$ everywhere. We will discuss the ideas underlying the proof of the Convergence Theorem 12.7 at the end of Section 12.5.

Example 12.8. Let $\sigma(x)$ denote the step function (12.38). Its Fourier coefficients

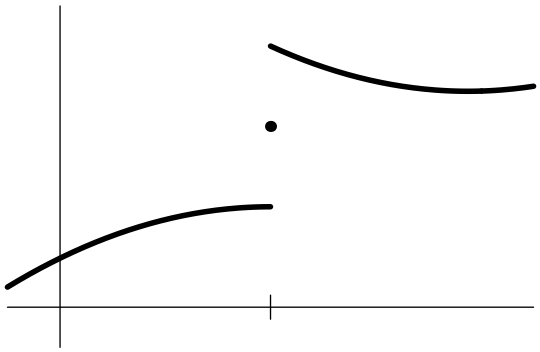


Figure 12.4. Splitting the Difference.

are easily computed:

$$a_0 = \frac{1}{\pi}\int_{-\pi}^{\pi} \sigma(x)\,dx = \frac{1}{\pi}\int_{0}^{\pi} dx = 1,$$
$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} \sigma(x)\cos k\,x\,dx = \frac{1}{\pi}\int_{0}^{\pi} \cos k\,x\,dx = 0,$$
$$b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} \sigma(x)\sin k\,x\,dx = \frac{1}{\pi}\int_{0}^{\pi} \sin k\,x\,dx = \begin{cases} \dfrac{2}{k\,\pi}, & k = 2l+1 \text{ odd}, \\ 0, & k = 2l \text{ even}. \end{cases}$$

Therefore, the Fourier series for the step function is

$$\sigma(x) \quad \sim \quad \frac{1}{2} \;+\; \frac{2}{\pi}\left(\sin x \;+\; \frac{\sin 3\,x}{3} \;+\; \frac{\sin 5\,x}{5} \;+\; \frac{\sin 7\,x}{7} \;+\; \cdots\right). \tag{12.41}$$

According to Theorem 12.7, the Fourier series will converge to the $2\,\pi$ periodic extension of the step function:

$$\widetilde{\sigma}(x) = \begin{cases} 0, & (2m-1)\,\pi < x < 2m\,\pi, \\ 1, & 2m\,\pi < x < (2m+1)\,\pi, \\ \frac{1}{2}, & x = m\,\pi, \end{cases} \qquad \text{where } m \text{ is any integer},$$

which is plotted in Figure 12.5. Observe that, in accordance with Theorem 12.7, $\widetilde{\sigma}(x)$ takes the midpoint value $\frac{1}{2}$ at the jump discontinuities $0, \pm\pi, \pm 2\,\pi, \ldots$.

It is instructive to investigate the convergence of this particular Fourier series in some detail. Figure 12.6 displays a graph of the first few partial sums, taking, respectively, $n = 3, 5$, and 10 terms. The reader will notice that away from the discontinuities, the series does appear to be converging, albeit slowly. However, near the jumps there is a consistent overshoot of about 9%. The region where the overshoot occurs becomes narrower and narrower as the number of terms increases, but the magnitude of the overshoot persists no matter how many terms are summed up. This was first noted by the American physicist Josiah Gibbs, and is now known as the *Gibbs phenomenon* in his honor. The Gibbs overshoot is a manifestation of the subtle non-uniform convergence of the Fourier series.

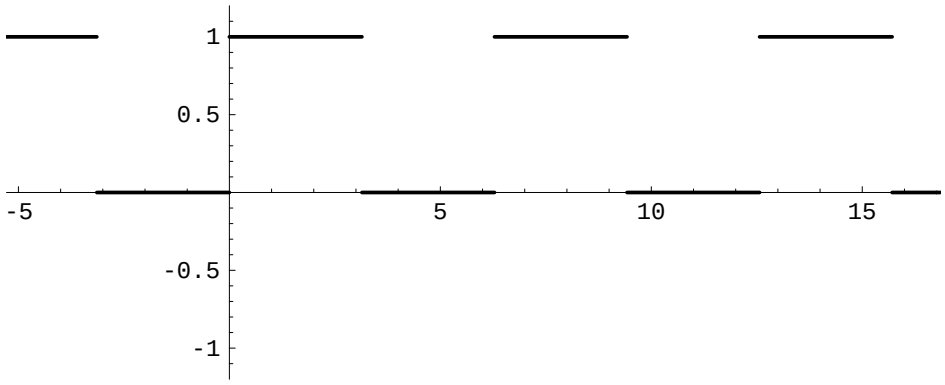


Figure 12.5. Periodic Step Function.

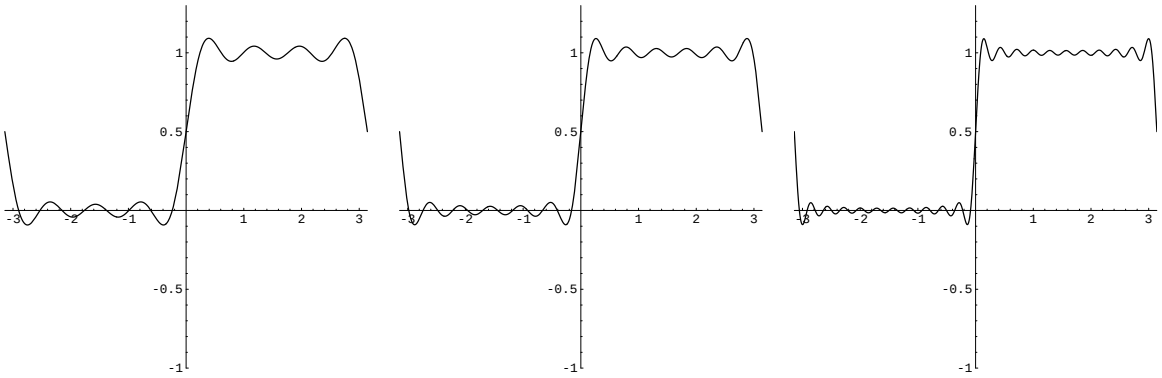


Figure 12.6. Gibbs Phenomenon.

Even and Odd Functions

We already noted that the Fourier cosine coefficients of the function $f(x) = x$ are all 0. This is not an accident, but rather a direct consequence of the fact that x is an odd function. Recall first the basic definition:

Definition 12.9. A function is called *even* if $f(-x) = f(x)$. A function is *odd* if $f(-x) = -f(x)$.

For example, the functions 1, $\cos k\,x$, and x^2 are all even, whereas x, $\sin k\,x$, and $\operatorname{sign} x$ are odd. We require two elementary lemmas, whose proofs are left to the reader.

Lemma 12.10. *The sum, $f(x) + g(x)$, of two even functions is even; the sum of two odd functions is odd. The product $f(x)\,g(x)$ of two even functions, or of two odd functions, is an even function. The product of an even and an odd function is odd.*

Remark: *Every* function can be represented as the sum of an even and an odd function; see Exercise ■.

Lemma 12.11. *If $f(x)$ is odd and integrable on the symmetric interval $[-a, a]$, then $\displaystyle\int_{-a}^{a} f(x)\,dx = 0$. If $f(x)$ is even and integrable, then $\displaystyle\int_{-a}^{a} f(x)\,dx = 2\int_{0}^{a} f(x)\,dx$.*

The next result is an immediate consequence of applying Lemmas 12.10 and 12.11 to the Fourier integrals (12.28).

Proposition 12.12. *If $f(x)$ is even, then its Fourier sine coefficients all vanish, $b_k = 0$, and so f can be represented by a* Fourier cosine series

$$f(x) \quad \sim \quad \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos k\,x\,, \tag{12.42}$$

where

$$a_k = \frac{2}{\pi}\int_0^{\pi} f(x)\cos k\,x\,dx, \qquad k = 0, 1, 2, 3, \ldots\,. \tag{12.43}$$

If $f(x)$ is odd, then its Fourier cosine coefficients vanish, $a_k = 0$, and so f can be represented by a Fourier sine series

$$f(x) \quad \sim \quad \sum_{k=1}^{\infty} b_k \sin k\,x\,, \tag{12.44}$$

where

$$b_k = \frac{2}{\pi}\int_0^{\pi} f(x)\sin k\,x\,dx, \qquad k = 1, 2, 3, \ldots\,. \tag{12.45}$$

Conversely, a convergent Fourier cosine (*respectively, sine*) *series always represents an even* (*respectively, odd*) *function.*

Example 12.13. The absolute value $f(x) = |\,x\,|$ is an even function, and hence has a Fourier cosine series. The coefficients are

$$a_0 = \frac{2}{\pi}\int_0^{\pi} x\,dx = \pi, \tag{12.46}$$

$$a_k = \frac{2}{\pi}\int_0^{\pi} x\cos k\,x\,dx = \frac{2}{\pi}\left[\,\frac{x\,\sin k\,x}{k} + \frac{\cos k\,x}{k^2}\,\right]_{x=0}^{\pi} = \begin{cases} 0, & 0 \neq k \text{ even}, \\ -\dfrac{4}{k^2\,\pi}\,, & k \text{ odd}. \end{cases}$$

Therefore

$$|\,x\,| \quad \sim \quad \frac{\pi}{2} - \frac{4}{\pi}\left(\cos x + \frac{\cos 3\,x}{9} + \frac{\cos 5\,x}{25} + \frac{\cos 7\,x}{49} + \cdots\right). \tag{12.47}$$

According to Theorem 12.7, this Fourier cosine series converges to the $2\,\pi$ periodic extension of $|\,x\,|$, the "sawtooth function" graphed in Figure 12.7.

In particular, if we substitute $x = 0$, we obtain another interesting series

$$\frac{\pi^2}{8} = 1 + \frac{1}{9} + \frac{1}{25} + \frac{1}{49} + \cdots = \sum_{n=0}^{\infty} \frac{1}{(2\,n+1)^2}\,. \tag{12.48}$$

It converges faster than Gregory's series (12.36), and, while far from optimal in this regards, can be used to compute reasonable approximations to π. One can further manipulate this

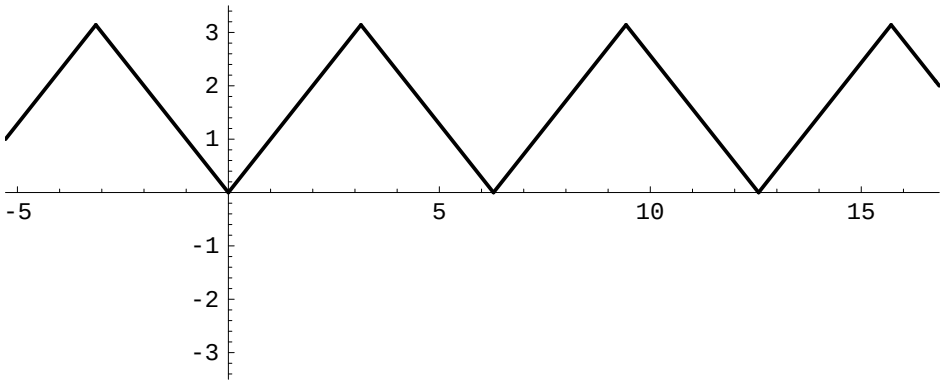


Figure 12.7. Periodic extension of $|\,x\,|$.

result to compute the sum of the series

$$S = \sum_{n=1}^{\infty} \frac{1}{n^2} = 1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \frac{1}{25} + \frac{1}{36} + \frac{1}{49} + \cdots .$$

We note that

$$\frac{S}{4} = \sum_{n=1}^{\infty} \frac{1}{4n^2} = \sum_{n=1}^{\infty} \frac{1}{(2n)^2} = \frac{1}{4} + \frac{1}{16} + \frac{1}{36} + \frac{1}{64} + \cdots .$$

Therefore, by (12.48),

$$\frac{3}{4}\,S = S - \frac{S}{4} = 1 + \frac{1}{9} + \frac{1}{25} + \frac{1}{49} + \cdots = \frac{\pi^2}{8}\,,$$

from which we conclude that

$$S = \sum_{n=1}^{\infty} \frac{1}{n^2} = 1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \frac{1}{25} + \cdots = \frac{\pi^2}{6}\,. \tag{12.49}$$

Remark: The most famous function in number theory — and the source of the most outstanding problem in mathematics, the Riemann hypothesis — is the *Riemann zeta function*

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}\,. \tag{12.50}$$

Formula (12.49) shows that $\zeta(2) = \frac{1}{6}\,\pi^2$. In fact, the value of the zeta function at any *even* positive integer $s = 2n$ can be written as a rational polynomial in π.

If $f(x)$ is any function defined on $[\,0, \pi\,]$, then its *Fourier cosine series* is defined by the formulas (12.42–43); the resulting series represents its even, 2π periodic extension. For example, the cosine series of $f(x) = x$ is given in (12.47); indeed the even, 2π periodic extension of x coincides with the 2π periodic extension of $|\,x\,|$. Similarly, the formulas (12.44, 45) define its *Fourier sine series*, representing its odd, 2π periodic extension. ∎

Complex Fourier Series

An alternative, and often more convenient, approach to Fourier series is to use complex exponentials instead of sines and cosines. Indeed, Euler's formula

$$e^{\mathrm{i}\,k\,x} = \cos k\,x + \mathrm{i}\,\sin k\,x, \qquad e^{-\mathrm{i}\,k\,x} = \cos k\,x - \mathrm{i}\,\sin k\,x, \tag{12.51}$$

shows how to write the trigonometric functions

$$\cos k\,x = \frac{e^{\mathrm{i}\,k\,x} + e^{-\mathrm{i}\,k\,x}}{2}\,, \qquad \sin k\,x = \frac{e^{\mathrm{i}\,k\,x} - e^{-\mathrm{i}\,k\,x}}{2\,\mathrm{i}}\,, \tag{12.52}$$

in terms of complex exponentials. Orthonormality with respect to the rescaled L^2 Hermitian inner product

$$\langle\, f\,, g\,\rangle = \frac{1}{2\,\pi}\,\int_{-\pi}^{\pi} f(x)\,\overline{g(x)}\,dx\,, \tag{12.53}$$

was proved by direct computation in Example 3.45:

$$\begin{aligned} \langle\, e^{\mathrm{i}\,k\,x}\,, e^{\mathrm{i}\,l\,x}\,\rangle &= \frac{1}{2\,\pi}\,\int_{-\pi}^{\pi} e^{\mathrm{i}\,(k-l)x}\,dx = \begin{cases} 1, & k = l, \\ 0, & k \neq l, \end{cases} \\ \|\, e^{\mathrm{i}\,k\,x}\,\|^2 &= \frac{1}{2\,\pi}\,\int_{-\pi}^{\pi} |\, e^{\mathrm{i}\,k\,x}\,|^2\,dx = 1. \end{aligned} \tag{12.54}$$

Again, orthogonality follows from their status as (complex) eigenfunctions for the periodic boundary value problem (12.13).

The *complex Fourier series* for a (piecewise continuous) real or complex function f is

$$f(x) \;\sim\; \sum_{k=-\infty}^{\infty} c_k\,e^{\mathrm{i}\,k\,x} = \;\cdots + c_{-2}\,e^{-2\,\mathrm{i}\,x} + c_{-1}\,e^{-\mathrm{i}\,x} + c_0 + c_1\,e^{\mathrm{i}\,x} + c_2\,e^{2\,\mathrm{i}\,x} + \;\cdots\;. \tag{12.55}$$

The orthonormality formulae (12.53) imply that the *complex Fourier coefficients* are obtained by taking the inner products

$$c_k = \langle\, f\,, e^{\mathrm{i}\,k\,x}\,\rangle = \frac{1}{2\,\pi}\,\int_{-\pi}^{\pi} f(x)\,e^{-\mathrm{i}\,k\,x}\,dx \tag{12.56}$$

with the associated complex exponential. Pay attention to the minus sign in the integrated exponential — the result of taking the complex conjugate of the second argument in the inner product (12.53). It should be emphasized that the real (12.27) and complex (12.55) Fourier formulae are just two different ways of writing the *same* series! Indeed, if we apply Euler's formula (12.51) to (12.56) and compare with the real Fourier formulae (12.28), we find that the real and complex Fourier coefficients are related by

$$\begin{aligned} a_k &= c_k + c_{-k}, & c_k &= \tfrac{1}{2}(a_k - \mathrm{i}\,b_k), \\ b_k &= \mathrm{i}\,(c_k - c_{-k}), & c_{-k} &= \tfrac{1}{2}(a_k + \mathrm{i}\,b_k), \end{aligned} \qquad k = 0, 1, 2, \ldots\;. \tag{12.57}$$

Remark: We already see one advantage of the complex version. The constant function $1 = e^{0\,\mathrm{i}\,x}$ no longer plays an anomalous role — the annoying factor of $\frac{1}{2}$ in the real Fourier series (12.27) has mysteriously disappeared!

Example 12.14. For the step function $\sigma(x)$ considered in Example 12.8, the complex Fourier coefficients are

$$c_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} \sigma(x)\, e^{-\mathrm{i}\,k\,x}\, dx = \frac{1}{2\pi}\int_{0}^{\pi} e^{-\mathrm{i}\,k\,x}\, dx = \begin{cases} \frac{1}{2}, & k = 0, \\ 0, & 0 \neq k \text{ even}, \\ \dfrac{1}{\mathrm{i}\,k\,\pi}, & k \text{ odd}. \end{cases}$$

Therefore, the step function has the complex Fourier series

$$\sigma(x) \quad \sim \quad \frac{1}{2} \;-\; \frac{\mathrm{i}}{\pi} \sum_{l=-\infty}^{\infty} \frac{e^{(2l+1)\,\mathrm{i}\,x}}{2l+1}\,.$$

You should convince yourself that this is *exactly the same series* as the real Fourier series (12.41). We are merely rewriting it using complex exponentials instead of real sines and cosines.

Example 12.15. Let us find the Fourier series for the exponential function $e^{a\,x}$. It is much easier to evaluate the integrals for the complex Fourier coefficients, and so

$$\begin{aligned} c_k &= \langle\, e^{a\,x}\,, e^{\mathrm{i}\,k\,x}\,\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{(a-\mathrm{i}\,k)\,x}\, dx = \frac{e^{(a-\mathrm{i}\,k)\,x}}{2\pi\,(a-\mathrm{i}\,k)}\,\Bigg|_{x=-\pi}^{\pi} \\ &= \frac{e^{(a-\mathrm{i}\,k)\,\pi} - e^{-(a-\mathrm{i}\,k)\,\pi}}{2\pi\,(a-\mathrm{i}\,k)} = (-1)^k\,\frac{e^{a\,\pi} - e^{-a\,\pi}}{2\pi\,(a-\mathrm{i}\,k)} = \frac{(-1)^k(a+\mathrm{i}\,k)\,\sinh a\,\pi}{\pi\,(a^2+k^2)}\,. \end{aligned}$$

Therefore, the desired Fourier series is

$$e^{a\,x} \quad \sim \quad \frac{\sinh a\,\pi}{\pi} \sum_{k=-\infty}^{\infty} \frac{(-1)^k(a+\mathrm{i}\,k)}{a^2+k^2}\, e^{\mathrm{i}\,k\,x}. \tag{12.58}$$

As an exercise, the reader should try writing this as a real Fourier series, either by breaking up the complex series into its real and imaginary parts, or by direct evaluation of the real coefficients via their integral formulae (12.28). According to Theorem 12.7 (which is equally valid for complex Fourier series) the Fourier series converges to the 2π periodic extension of the exponential function, graphed in Figure 12.8.

The Delta Function

Fourier series can even be used to represent more general objects than mere functions. The most important example is the delta function $\delta(x)$. Using its characterizing properties

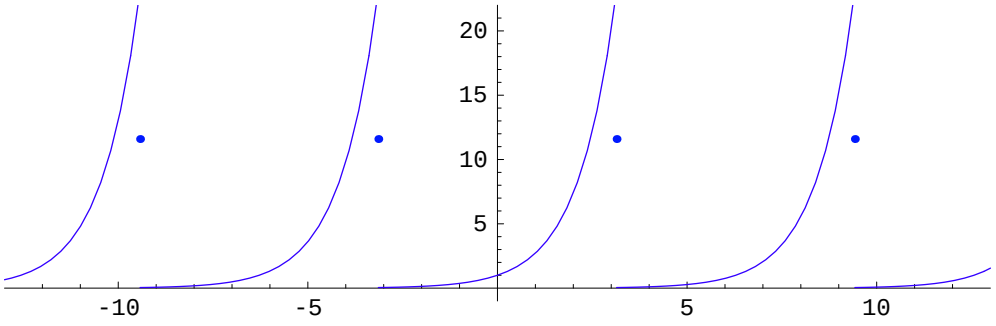


Figure 12.8. Periodic Extension of e^x.

(11.36), the real Fourier coefficients are computed as

$$\begin{aligned} a_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} \delta(x)\cos k\,x\,dx = \frac{1}{\pi}\cos k\,0 = \frac{1}{\pi}\,, \\ b_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} \delta(x)\sin k\,x\,dx = \frac{1}{\pi}\sin k\,0 = 0. \end{aligned} \tag{12.59}$$

Therefore,

$$\delta(x) \quad \sim \quad \frac{1}{2\,\pi} \;+\; \frac{1}{\pi}\,\big(\cos x + \cos 2\,x + \cos 3\,x + \;\cdots\;\big). \tag{12.60}$$

Since $\delta(x)$ is an even function, it should come as no surprise that it has a cosine series.

To understand in what sense this series converges to the delta function, it will help to rewrite it in complex form

$$\delta(x) \quad \sim \quad \frac{1}{2\,\pi}\sum_{k=-\infty}^{\infty} e^{\,\mathrm{i}\,k\,x} = \frac{1}{2\,\pi}\,\big(\;\cdots + e^{-2\,\mathrm{i}\,x} + e^{-\,\mathrm{i}\,x} + 1 + e^{\,\mathrm{i}\,x} + e^{2\,\mathrm{i}\,x} + \;\cdots\;\big). \tag{12.61}$$

where the complex Fourier coefficients are computed[†] as

$$c_k = \frac{1}{2\,\pi}\int_{-\pi}^{\pi} \delta(x)\,e^{-\,\mathrm{i}\,k\,x}\,dx = \frac{1}{2\,\pi}\,.$$

The n^{th} partial sum

$$s_n(x) = \frac{1}{2\,\pi}\sum_{k=-n}^{n} e^{\,\mathrm{i}\,k\,x} = \frac{1}{2\,\pi}\,\big(e^{-\,\mathrm{i}\,n\,x} + \;\cdots\; + e^{-\,\mathrm{i}\,x} + 1 + e^{\,\mathrm{i}\,x} + \;\cdots\; + e^{\,\mathrm{i}\,n\,x}\big)$$

can, in fact, be explicitly evaluated. Recall the formula for the sum of a geometric series

$$\sum_{k=0}^{m} a\,r^k = a + a\,r + a\,r^2 + \;\cdots\; + a\,r^m = a\,\left(\frac{r^{m+1}-1}{r-1}\right). \tag{12.62}$$

[†] Or, we could use (12.57).

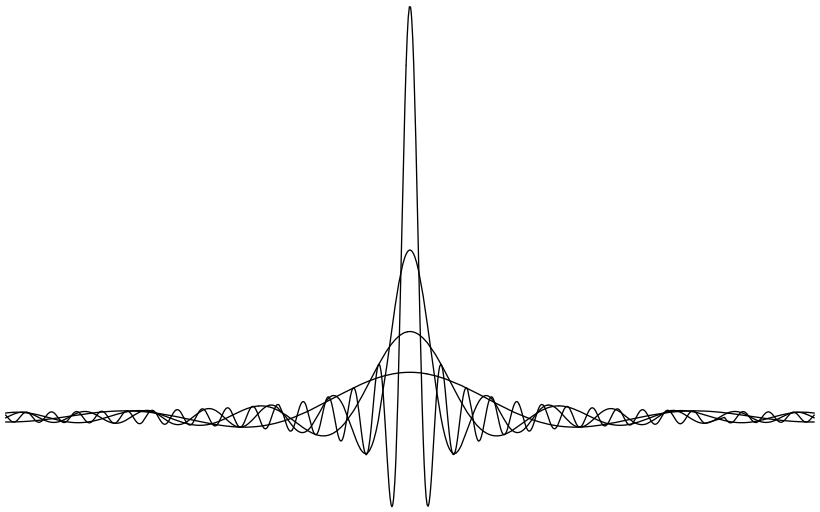


Figure 12.9. Partial Fourier Sums Approximating the Delta Function.

The partial sum $s_n(x)$ has this form, with $m+1 = 2n+1$ summands, initial term $a = e^{-\mathrm{i}\,nx}$, and ratio $r = e^{\mathrm{i}\,x}$. Therefore,

$$\begin{aligned} s_n(x) = \frac{1}{2\pi} \sum_{k=-n}^{n} e^{\mathrm{i}\,k\,x} &= \frac{1}{2\pi}\, e^{-\mathrm{i}\,nx} \left(\frac{e^{\mathrm{i}\,(2n+1)x} - 1}{e^{\mathrm{i}\,x} - 1} \right) = \frac{1}{2\pi}\, \frac{e^{\mathrm{i}\,(n+1)x} - e^{-\mathrm{i}\,nx}}{e^{\mathrm{i}\,x} - 1} \\ &= \frac{1}{2\pi}\, \frac{e^{\mathrm{i}\left(n+\frac{1}{2}\right)x} - e^{-\mathrm{i}\left(n+\frac{1}{2}\right)x}}{e^{\mathrm{i}\,x/2} - e^{-\mathrm{i}\,x/2}} = \frac{1}{2\pi}\, \frac{\sin\left(n+\frac{1}{2}\right)x}{\sin\frac{1}{2}\,x}\,. \end{aligned} \tag{12.63}$$

In this computation, to pass from the first to the second line, we multiplied numerator and denominator by $e^{-\mathrm{i}\,x/2}$, after which we used the formula (3.86) for the sine function in terms of complex exponentials. Incidentally, (12.63) is equivalent to the intriguing trigonometric summation formula

$$s_n(x) = \frac{1}{2\pi} + \frac{1}{\pi}\left(\cos x + \cos 2\,x + \cos 3\,x + \,\cdots\, + \cos n\,x\right) = \frac{1}{2\pi}\, \frac{\sin\left(n+\frac{1}{2}\right)x}{\sin\frac{1}{2}\,x}\,. \tag{12.64}$$

Graphs of the partial sums $s_n(x)$ for several values of n are displayed in Figure 12.9. Note that the spike, at $x = 0$, progressively becomes taller and thinner, converging to an infinitely tall, infinitely thin delta spike. Indeed, by l'Hôpital's Rule,

$$\lim_{x\to 0}\, \frac{1}{2\pi}\, \frac{\sin\left(n+\frac{1}{2}\right)x}{\sin\frac{1}{2}\,x} = \lim_{x\to 0}\, \frac{1}{2\pi}\, \frac{\left(n+\frac{1}{2}\right)\cos\left(n+\frac{1}{2}\right)x}{\frac{1}{2}\cos\frac{1}{2}\,x} = \frac{n+\frac{1}{2}}{\pi} \;\longrightarrow\; \infty \quad \text{as} \quad n \to \infty.$$

(An elementary proof of this formula is to note that, at $x = 0$, every term in the original sum (12.64) is equal to 1.) Furthermore, the integrals remain fixed,

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} s_n(x)\,dx = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\sin\left(n+\frac{1}{2}\right)x}{\sin\frac{1}{2}\,x}\,dx = \frac{1}{2\pi} \int_{-\pi}^{\pi} \sum_{k=-n}^{n} e^{\mathrm{i}\,k\,x}\,dx = 1, \tag{12.65}$$

as required for convergence to the delta function. However, away from the spike, the partial sums do *not* go to zero! Rather, they oscillate more and more rapidly, maintaining an overall amplitude of $\frac{1}{2\pi}\csc\frac{1}{2}x = 1/\big(2\pi\sin\frac{1}{2}x\big)$. As n gets large, the amplitude function appears as an envelope of the increasingly rapid oscillations. Roughly speaking, the fact that $s_n(x) \to \delta(x)$ as $n \to \infty$ means that the "infinitely fast" oscillations somehow cancel each other out, and the net effect is zero away from the spike at $x = 0$. Thus, the convergence of the Fourier sums to $\delta(x)$ is much more subtle than in the original limiting definition (11.31). The technical term is *w*eak convergence, which plays an very important role in advanced mathematical analysis, [**153**]; see Exercise ∎ below for additional details.

Remark: Although we stated that the Fourier series (12.60, 61) represent the delta function, this is not entirely correct. Remember that a Fourier series converges to the 2π periodic extension of the original function. Therefore, (12.61) actually represents the periodic extension of the delta function:

$$\widetilde{\delta}(x) = \cdots + \delta(x+4\pi) + \delta(x+2\pi) + \delta(x) + \delta(x-2\pi) + \delta(x-4\pi) + \delta(x-6\pi) + \cdots, \quad (12.66)$$

consisting of a periodic array of delta spikes concentrated at all integer multiples of 2π.

12.3. Differentiation and Integration.

If a series of functions converges "nicely" then one expects to be able to integrate and differentiate it term by term; the resulting series should converge to the integral and derivative of the original sum. Integration and differentiation of power series is always valid within the range of convergence, and is used extensively in the construction of series solutions of differential equations, series for integrals of non-elementary functions, and so on. The interested reader can consult Appendix C for further details.

As we now appreciate, the convergence of Fourier series is a much more delicate matter, and so one must be considerably more careful with their differentiation and integration. Nevertheless, in favorable situations, both operations lead to valid results, and are quite useful for constructing Fourier series of more complicated functions. It is a remarkable, profound fact that Fourier analysis is completely compatible with the calculus of generalized functions that we developed in Chapter 11. For instance, differentiating the Fourier series for a piecewise C^1 function leads to the Fourier series for the differentiated function that has delta functions of the appropriate magnitude appearing at each jump discontinuity. This fact reassures us that the rather mysterious construction of delta functions and their generalizations is indeed the right way to extend calculus to functions which do not possess derivatives in the ordinary sense.

Integration of Fourier Series

Integration is a smoothing operation — the integrated function is always nicer than the original. Therefore, we should anticipate being able to integrate Fourier series without difficulty. There is, however, one complication: the integral of a periodic function is not necessarily periodic. The simplest example is the constant function 1, which is certainly periodic, but its integral, namely x, is not. On the other hand, integrals of all the other periodic sine and cosine functions appearing in the Fourier series are periodic. Thus, only

the constant term might cause us difficulty when we try to integrate a Fourier series (12.27). According to (2.4), the constant term

$$\frac{a_0}{2} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)\, dx \tag{12.67}$$

is the *mean* or *average* of the function $f(x)$ on the interval $[-\pi, \pi]$. A function has no constant term in its Fourier series if and only if it has mean zero. It is easily shown, cf. Exercise ■, that the mean zero functions are precisely the ones that remain periodic upon integration.

Lemma 12.16. *If $f(x)$ is 2π periodic, then its integral $g(x) = \int_0^x f(y)\, dy$ is 2π periodic if and only if $\int_{-\pi}^{\pi} f(x)\, dx = 0$, so that f has mean zero on the interval $[-\pi, \pi]$.*

In particular, Lemma 12.11 implies that all odd functions automatically have mean zero, and hence periodic integrals.

Since

$$\int \cos k\,x\, dx = \frac{\sin k\,x}{k}\,, \qquad \int \sin k\,x\, dx = -\frac{\cos k\,x}{k}\,, \tag{12.68}$$

termwise integration of a Fourier series without constant term is straightforward. The resulting Fourier series is given precisely as follows.

Theorem 12.17. *If f is piecewise continuous, 2π periodic, and has mean zero, then its Fourier series*

$$f(x) \quad \sim \quad \sum_{k=1}^{\infty} \left[\, a_k \cos k\,x + b_k \sin k\,x \,\right],$$

can be integrated term by term, to produce the Fourier series

$$g(x) = \int_0^x f(y)\, dy \quad \sim \quad m \,+\, \sum_{k=1}^{\infty} \left[-\frac{b_k}{k} \cos k\,x + \frac{a_k}{k} \sin k\,x \right], \tag{12.69}$$

for its periodic integral. The constant term

$$m = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(x)\, dx$$

is the mean of the integrated function.

In many situations, the integration formula (12.69) provides a very convenient alternative to the direct derivation of the Fourier coefficients.

Example 12.18. The function $f(x) = x$ is odd, and so has mean zero: $\int_{-\pi}^{\pi} x\, dx = 0$. Let us integrate its Fourier series

$$x \;\sim\; 2 \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} \sin k\,x \tag{12.70}$$

that we found in Example 12.2. The result is the Fourier series

$$\begin{aligned} \frac{1}{2}\,x^2 \;&\sim\; \frac{\pi^2}{6} \;-\; 2\,\sum_{k=1}^{\infty}\,\frac{(-1)^{k-1}}{k^2}\,\cos k\,x \\ &=\; \frac{\pi^2}{6} \;-\; 2\left(\cos x \;-\; \frac{\cos 2\,x}{4} \;+\; \frac{\cos 3\,x}{9} \;-\; \frac{\cos 4\,x}{16} \;+\; \cdots\;\right), \end{aligned} \tag{12.71}$$

whose the constant term is the mean of the left hand side:

$$\frac{1}{2\,\pi}\int_{-\pi}^{\pi}\frac{x^2}{2}\,dx \;=\; \frac{\pi^2}{6}\,.$$

Let us revisit the derivation of the integrated Fourier series from a slightly different standpoint. If we were to integrate each trigonometric summand in a Fourier series (12.27) from 0 to x, we would obtain

$$\int_0^x \cos k\,y\,dy \;=\; \frac{\sin k\,x}{k}\,, \qquad \text{whereas} \qquad \int_0^x \sin k\,y\,dy \;=\; \frac{1}{k} - \frac{\cos k\,x}{k}\,.$$

The extra $1/k$ terms arising from the definite sine integrals do not appear explicitly in our previous form for the integrated Fourier series, (12.69), and so must be hidden in the constant term m. We deduce that the mean value of the integrated function can be computed using the Fourier sine coefficients of f via the formula

$$\frac{1}{2\,\pi}\int_{-\pi}^{\pi} g(x)\,dx = m = \sum_{k=1}^{\infty}\,\frac{b_k}{k}. \tag{12.72}$$

For example, the result of integrating both sides of the Fourier series (12.70) for $f(x) = x$ from 0 to x is

$$\frac{x^2}{2} \;\sim\; 2\,\sum_{k=1}^{\infty}\,\frac{(-1)^{k-1}}{k^2}\,(1-\cos k\,x).$$

The constant terms sum up to yield the mean value of the integrated function:

$$2\left(1 - \frac{1}{4} + \frac{1}{9} - \frac{1}{16} + \,\cdots\,\right) \;=\; 2\,\sum_{k=1}^{\infty}\,\frac{(-1)^{k-1}}{k^2} \;=\; \frac{1}{2\,\pi}\int_{-\pi}^{\pi}\frac{x^2}{2}\,dx \;=\; \frac{\pi^2}{6}\,, \tag{12.73}$$

which reproduces a formula established in Exercise ■.

More generally, if $f(x)$ does not have mean zero, its Fourier series has a nonzero constant term,

$$f(x) \;\sim\; \frac{a_0}{2} + \sum_{k=1}^{\infty}\,\left[\,a_k\,\cos k\,x + b_k\,\sin k\,x\,\right].$$

In this case, the result of integration will be

$$g(x) = \int_0^x f(y)\,dy \;\sim\; \frac{a_0}{2}\,x + m + \sum_{k=1}^{\infty}\,\left[\,-\,\frac{b_k}{k}\,\cos k\,x + \frac{a_k}{k}\,\sin k\,x\,\right], \tag{12.74}$$

where m is given in (12.72). The right hand side is not, strictly speaking, a Fourier series. There are two ways to interpret this formula within the Fourier framework. Either we can write (12.74) as the Fourier series for the difference

$$g(x) - \frac{a_0}{2}\, x \;\sim\; m + \sum_{k=1}^{\infty} \left[-\frac{b_k}{k}\, \cos k\, x + \frac{a_k}{k}\, \sin k\, x \right], \tag{12.75}$$

which is a 2π periodic function, cf. Exercise ■. Alternatively, one can replace x by its Fourier series (12.30), and the result will be the Fourier series for the 2π periodic extension of the integral $g(x) = \displaystyle\int_0^x f(y)\, dy$.

Differentiation of Fourier Series

Differentiation has the opposite effect to integration. Differentiation makes a function worse. Therefore, to justify taking the derivative of a Fourier series, we need to know that the differentiated function remains reasonably nice. Since we need the derivative $f'(x)$ to be piecewise C^1 for the convergence Theorem 12.7 to be applicable, we must require that $f(x)$ itself be continuous and piecewise C^2.

Theorem 12.19. *If f is 2π periodic, continuous, and piecewise C^2, then its Fourier series can be differentiated term by term, to produce the Fourier series for its derivative*

$$f'(x) \quad\sim\quad \sum_{k=1}^{\infty} \big[\, k\, b_k \cos k\, x - k\, a_k \sin k\, x \,\big]. \tag{12.76}$$

Example 12.20. The derivative (11.52) of the absolute value function $f(x) = |\, x\,|$ is the sign function

$$\frac{d}{dx}\, |\, x\,| = \operatorname{sign} x = \begin{cases} +1, & x > 0 \\ -1, & x < 0. \end{cases}$$

Therefore, if we differentiate its Fourier series (12.47), we obtain the Fourier series

$$\operatorname{sign} x \;\sim\; \frac{4}{\pi}\left(\sin x \,+\, \frac{\sin 3\, x}{3} \,+\, \frac{\sin 5\, x}{5} \,+\, \frac{\sin 7\, x}{7} \,+\, \cdots \right). \tag{12.77}$$

Note that $\operatorname{sign} x = \sigma(x) - \sigma(-x)$ is the difference of two step functions. Indeed, subtracting the step function Fourier series (12.41) at x from the same series at $-x$ reproduces (12.77).

Example 12.21. If we differentiate the Fourier series

$$x \;\sim\; 2 \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} \sin k\, x \;=\; 2\left(\sin x \,-\, \frac{\sin 2\, x}{2} \,+\, \frac{\sin 3\, x}{3} \,-\, \frac{\sin 4\, x}{4} \,+\, \cdots \right),$$

we obtain an apparent contradiction:

$$1 \;\sim\; 2 \sum_{k=1}^{\infty} (-1)^{k+1} \cos k\, x = 2\cos x - 2\cos 2\, x + 2\cos 3\, x - 2\cos 4\, x + \,\cdots\,. \tag{12.78}$$

But the Fourier series for 1 just consists of a single constant term! (Why?)

The resolution of this paradox is not difficult. The Fourier series (12.30) does *not* converge to x, but rather to its periodic extension $\widetilde{f}(x)$, which has a jump discontinuity of magnitude 2π at odd multiples of π; see Figure 12.1. Thus, Theorem 12.19 is *not* directly applicable. Nevertheless, we can assign a consistent interpretation to the differentiated series. The derivative $\widetilde{f}\,'(x)$ of the periodic extension is *not* equal to the constant function 1, but, rather, has an additional delta function concentrated at each jump discontinuity:

$$\widetilde{f}\,'(x) = 1 - 2\pi \sum_{j=-\infty}^{\infty} \delta\big(x - (2j+1)\pi\big) = 1 - 2\pi\,\widetilde{\delta}(x-\pi),$$

where $\widetilde{\delta}$ denotes the 2π periodic extension of the delta function, cf. (12.66). The differentiated Fourier series (12.78) does, in fact, converge to this modified distributional derivative! Indeed, differentiation and integration of Fourier series is entirely compatible with the calculus of generalized functions, as will be borne out in yet another example.

Example 12.22. Let us differentiate the Fourier series (12.41) for the step function and see if we end up with the Fourier series (12.60) for the delta function. We find (12.41)

$$\frac{d}{dx}\,\sigma(x) \quad \sim \quad \frac{2}{\pi}\big(\cos x + \cos 3x + \cos 5x + \cos 7x + \,\cdots\,\big), \tag{12.79}$$

which does *not* agree with (12.60) — half the terms are missing! The explanation is similar to the preceding example: the 2π periodic extension of the step function has two jump discontinuities, of magnitudes $+1$ at even multiples of π and -1 at odd multiples. Therefore, its derivative is the difference of the 2π periodic extension of the delta function at 0, with Fourier series (12.60) minus the 2π periodic extension of the delta function at π, with Fourier series

$$\delta(x-\pi) \quad \sim \quad \frac{1}{2\pi} \,+\, \frac{1}{\pi}\big(-\cos x + \cos 2x - \cos 3x + \,\cdots\,\big)$$

derived in Exercise ■. The difference of these two delta function series produces (12.79).

12.4. Change of Scale.

So far, we have only dealt with Fourier series on the standard interval of length 2π. (We chose $[-\pi, \pi]$ for convenience, but all of the results and formulas are easily adapted to any other interval of the same length, e.g., $[0, 2\pi]$.) Since physical objects like bars and strings do not all come in this particular length, we need to understand how to adapt the formulas to more general intervals. The basic idea is to rescale the variable so as to stretch or contract the standard interval†.

† The same device was already used, in Section 5.4, to adapt the orthogonal Legendre polynomials to other intervals.

Any symmetric interval $[-\ell\,,\ell\,]$ of length $2\,\ell$ can be rescaled to the standard interval $[-\pi,\pi]$ by using the linear change of variables

$$x = \frac{\ell}{\pi}\,y, \qquad \text{so that} \qquad -\,\pi \le y \le \pi \qquad \text{whenever} \qquad -\,\ell \le x \le \ell. \tag{12.80}$$

Given a function $f(x)$ defined on $[-\ell\,,\ell\,]$, the *rescaled function* $F(y) = f\left(\dfrac{\ell}{\pi}\,y\right)$ lives on $[-\pi,\pi]$. Let

$$F(y) \quad \sim \quad \frac{a_0}{2} \;+\; \sum_{k=1}^{\infty} \left[\, a_k \cos k\,y + b_k \sin k\,y \,\right],$$

be the standard Fourier series for $F(y)$, so that

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} F(y)\cos k\,y\,dy, \qquad\qquad b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} F(y)\sin k\,y\,dy. \tag{12.81}$$

Then, reverting to the unscaled variable x, we deduce that

$$f(x) \quad \sim \quad \frac{a_0}{2} \;+\; \sum_{k=1}^{\infty} \left[\, a_k \cos\frac{k\,\pi\,x}{\ell} + b_k \sin\frac{k\,\pi\,x}{\ell} \,\right]. \tag{12.82}$$

The Fourier coefficients a_k, b_k can be computed directly from $f(x)$. Indeed, replacing the integration variable in (12.81) by $y = \pi\,x/\ell$, and noting that $dy = (\,\pi/\ell\,)\,dx$, we deduce the adapted formulae

$$a_k = \frac{1}{\ell}\int_{-\ell}^{\ell} f(x)\,\cos\frac{k\,\pi\,x}{\ell}\;dx, \qquad\qquad b_k = \frac{1}{\ell}\int_{-\ell}^{\ell} f(x)\,\sin\frac{k\,\pi\,x}{\ell}\;dx, \tag{12.83}$$

for the Fourier coefficients of $f(x)$ on the interval $[-\ell\,,\ell\,]$.

All of the convergence results, integration and differentiation formulae, etc., that are valid for the interval $[-\pi,\pi]$ carry over, essentially unchanged, to Fourier series on nonstandard intervals. In particular, adapting our basic convergence Theorem 12.7, we conclude that if $f(x)$ is piecewise C^1, then its rescaled Fourier series (12.82) converges to its $2\,\ell$ periodic extension $\widetilde{f}(x)$, subject to the proviso that $\widetilde{f}(x)$ takes on the midpoint values at all jump discontinuities.

Example 12.23. Let us compute the Fourier series for the function $f(x) = x$ on the interval $-1 \le x \le 1$. Since f is odd, only the sine coefficients will be nonzero. We have

$$b_k = \int_{-1}^{1} x\sin k\,\pi\,x\,dx = \left[\, -\,\frac{x\cos k\,\pi\,x}{k\,\pi} + \frac{\sin k\,\pi\,x}{(k\,\pi)^2} \,\right]_{x=-1}^{1} = \frac{2(-1)^{k+1}}{k\,\pi}\,.$$

The resulting Fourier series is

$$x \quad \sim \quad \frac{2}{\pi}\left(\sin\pi\,x \;-\; \frac{\sin 2\,\pi\,x}{2} \;+\; \frac{\sin 3\pi\,x}{3} \;-\; \cdots\right).$$

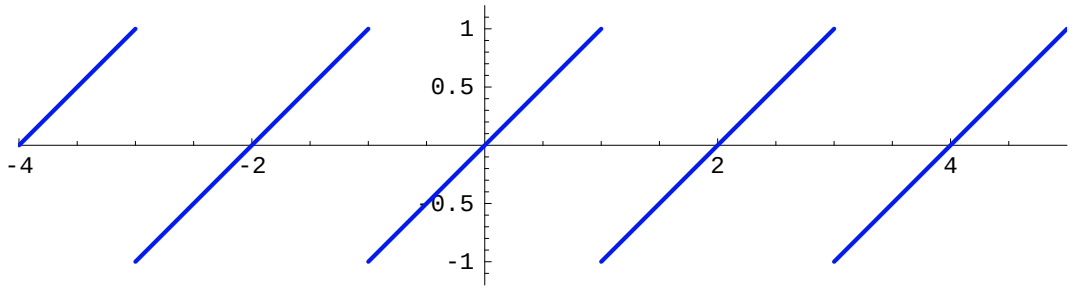


Figure 12.10. 2 Periodic Extension of x.

The series converges to the 2 periodic extension of the function x, namely

$$\widetilde{f}(x) = \begin{cases} x - 2m, & 2m-1 < x < 2m+1, \\ 0, & x = m, \end{cases} \qquad \text{where } m \text{ is an arbitrary integer,}$$

plotted in Figure 12.10.

We can similarly reformulate complex Fourier series on the nonstandard interval $[-\ell\,, \ell\,]$. Using (12.80) to rescale the variables in (12.55), we find

$$f(x) \quad \sim \quad \sum_{k=-\infty}^{\infty} c_k\, e^{\,\mathrm{i}\,k\pi\,x/\ell}, \qquad \text{where} \qquad c_k = \frac{1}{2\,\ell} \int_{-\ell}^{\ell} f(x)\, e^{-\,\mathrm{i}\,k\pi\,x/\ell}\, dx. \tag{12.84}$$

Again, this is merely an alternative way of writing the real Fourier series (12.82).

When dealing with a more general interval $[a, b]$, there are two options. The first is to take a function $f(x)$ defined for $a \le x \le b$ and periodically extend it to a function $\widetilde{f}(x)$ that agrees with $f(x)$ on $[a, b]$ and has period $b - a$. One can then compute the Fourier series (12.82) for its periodic extension $\widetilde{f}(x)$ on the symmetric interval $[\,\frac{1}{2}(a-b), \frac{1}{2}(b-a)\,]$ of width $2\,\ell = b - a$; the resulting Fourier series will (under the appropriate hypotheses) converge to $\widetilde{f}(x)$ and hence agree with $f(x)$ on the original interval. An alternative approach is to translate the interval by an amount $\frac{1}{2}(a+b)$ so as to make it symmetric; this is accomplished by the change of variables $\widehat{x} = x - \frac{1}{2}(a+b)$. an additional rescaling will convert the interval into $[-\pi, \pi]$. The two methods are essentially equivalent, and full details are left to the reader.

12.5. Convergence of the Fourier Series.

The purpose of this final section is to establish some basic convergence results for Fourier series. This is not a purely theoretical exercise, since convergence considerations impinge directly upon a variety of applications of Fourier series. One particularly important consequence is the connection between smoothness of a function and the decay rate of its high order Fourier coefficients — a result that is exploited in signal and image denoising and in the analytical properties of solutions to partial differential equations.

Be forewarned: the material in this section is more mathematical than we are used to, and the more applied reader may consider omitting it on a first reading. However, a full

understanding of the scope of Fourier analysis as well as its limitations does requires some familiarity with the underlying theory. Moreover, the required techniques and proofs serve as an excellent introduction to some of the most important tools of modern mathematical analysis. Any effort expended to assimilate this material will be more than amply rewarded in your later career.

Unlike power series, which converge to analytic functions on the interval of convergence, and diverge elsewhere (the only tricky point being whether or not the series converges at the endpoints), the convergence of a Fourier series is a much more subtle matter, and still not understood in complete generality. A large part of the difficulty stems from the intricacies of convergence in infinite-dimensional function spaces. Let us therefore begin with a brief discussion of the fundamental issues.

Convergence in Vector Spaces

We assume that you are familiar with the usual calculus definition of the limit of a sequence of real numbers: $\lim_{n \to \infty} a_n = a^\star$. In any finite-dimensional vector space, e.g., $\mathbb{R}^m$, there is essentially only one way for a sequence of vectors $\mathbf{v}^{(0)}, \mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots \in \mathbb{R}^m$ to converge, which is guaranteed by any one of the following equivalent criteria:

(*a*) The vectors converge: $\mathbf{v}^{(n)} \longrightarrow \mathbf{v}^\star \in \mathbb{R}^m$ as $n \to \infty$.

(*b*) The individual components of $\mathbf{v}^{(n)} = (v_1^{(n)}, \ldots, v_m^{(n)})$ converge, so $\lim_{n \to \infty} v_i^{(n)} = v_i^\star$ for all $i = 1, \ldots, m$.

(*c*) The difference in norms goes to zero: $\| \mathbf{v}^{(n)} - \mathbf{v}^\star \| \longrightarrow 0$ as $n \to \infty$.

The last requirement, known as *convergence in norm*, does not, in fact, depend on which norm is chosen. Indeed, Theorem 3.17 implies that, on a finite-dimensional vector space, all norms are essentially equivalent, and if one norm goes to zero, so does any other norm.

The analogous convergence criteria are certainly *not the same* in infinite-dimensional vector spaces. There are, in fact, a bewildering variety of convergence mechanisms in function space, that include pointwise convergence, uniform convergence, convergence in norm, weak convergence, and many others. All play a significant role in advanced mathematical analysis, and hence all are deserving of study. Here, though, we shall be content to learn just the most basic aspects of convergence of the Fourier series, leaving further details to more advanced texts, e.g., [**62**, **153**, **193**].

The most basic convergence mechanism for a sequence of functions $v_n(x)$ is called *pointwise convergence*, which requires that

$$\lim_{n \to \infty} v_n(x) = v_\star(x) \qquad \text{for all} \quad x. \tag{12.85}$$

In other words, the functions' values at each individual point converge in the usual sense. Pointwise convergence is the function space version of the convergence of the components of a vector. Indeed, pointwise convergence immediately implies component-wise convergence of the sample vectors $\mathbf{v}^{(n)} = \big(v_n(x_1), \ldots, v_n(x_m) \big)^T \in \mathbb{R}^m$ for any choice of sample points $x_1, \ldots, x_m$.

On the other hand, *convergence in norm* of the function sequence requires

$$\lim_{n \to \infty} \| v_n - v_\star \| \; = \; 0,$$

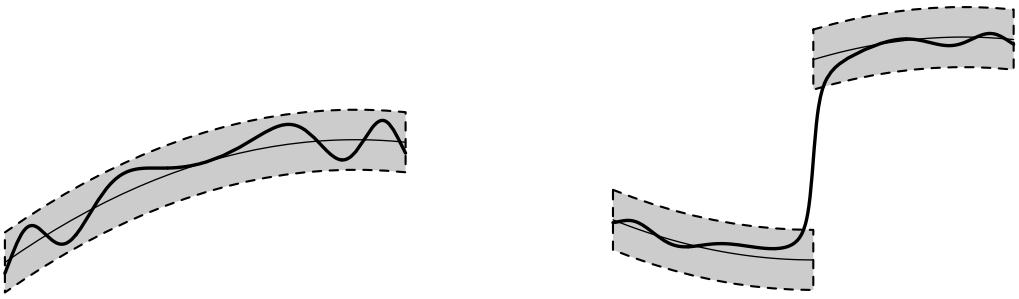


Figure 12.11. Uniform and Non-Uniform Convergence of Functions.

where $\| \cdot \|$ is a prescribed norm on the function space. As we have learned, not all norms on an infinite-dimensional function space are equivalent: a function might be small in one norm, but large in another. As a result, convergence in norm *will* depend upon the choice of norm. Moreover, convergence in norm does not necessarily imply pointwise convergence or vice versa. A variety of examples can be found in the exercises.

Uniform Convergence

Proving uniform convergence of a Fourier series is reasonably straightforward, and so we will begin there. You no doubt first saw the concept of a uniformly convergent sequence of functions in your calculus course, although chances are it didn't leave much of an impression. In Fourier analysis, uniform convergence begins to play an increasingly important role, and is worth studying in earnest. For the record, let us restate the basic definition.

Definition 12.24. A sequence of functions $v_n(x)$ is said to converge *uniformly* to a function $v_\star(x)$ on a subset $I \subset \mathbb{R}$ if, for every $\varepsilon > 0$, there exists an integer $N = N(\varepsilon)$ such that

$$| v_n(x) - v_\star(x) | < \varepsilon \qquad \text{for all } \ x \in I \ \text{ and all } \ n \geq N. \tag{12.86}$$

The key point — and the reason for the term "uniform convergence" — is that the integer N depends only upon ε and not on the point $x \in I$. Roughly speaking, the sequence converges uniformly if and only if for any small ε, the graphs of the functions eventually lie inside a band of width 2ε centered around the graph of the limiting function; see Figure 12.11. Functions may converge pointwise, but non-uniformly: the Gibbs phenomenon is the prototypical example of a nonuniformly convergent sequence: For a given $\varepsilon > 0$, the closer x is to the discontinuity, the larger n must be chosen so that the inequality in (12.86) holds, and hence there is *no* consistent choice of N that makes (12.86) valid for all x and all $n \geq N$. A detailed discussion of these issues, including the proofs of the basic theorems, can be found in any basic real analysis text, e.g., [**9**, **152**, **153**].

A key consequence of uniform convergence is that it preserves continuity.

Theorem 12.25. *If $v_n(x) \to v_\star(x)$ converges uniformly, and each $v_n(x)$ is continuous, then $v_\star(x)$ is also a continuous function.*

The proof is by contradiction. Intuitively, if $v_\star(x)$ were to have a discontinuity, then, as sketched in Figure 12.11, a sufficiently small band around its graph would not connect together, and this prevents the graph of any continuous function, such as $v_n(x)$, from remaining entirely within the band. Rigorous details can be found in [**9**].

Warning: A sequence of continuous functions can converge *non-uniformly* to a continuous function. An example is the sequence $v_n(x) = \dfrac{2nx}{1+n^2x^2}$, which converges pointwise to $v_\star(x) \equiv 0$ (why?) but not uniformly since $\max |\, v_n(x) \,| = v_n\left(\frac{1}{n}\right) = 1$, which implies that (12.86) cannot hold when $\varepsilon < 1$.

The convergence (pointwise, uniform, in norm, etc.) of a series $\displaystyle\sum_{k=1}^{\infty} u_k(x)$ is, by definition, governed by the convergence of its sequence of *partial sums*

$$v_n(x) = \sum_{k=1}^{n} u_k(x). \tag{12.87}$$

The most useful test for uniform convergence of series of functions is known as the *Weierstrass M–test*, in honor of the nineteenth century German mathematician Karl Weierstrass, known as the "father of modern analysis".

Theorem 12.26. *Let $I \subset \mathbb{R}$. Suppose the functions $u_k(x)$ are bounded by*

$$|\, u_k(x) \,| \le m_k \qquad \text{for all} \qquad x \in I, \tag{12.88}$$

where the $m_k \ge 0$ are fixed positive constants. If the series

$$\sum_{k=1}^{\infty} m_k < \infty \tag{12.89}$$

converges, then the series

$$\sum_{k=1}^{\infty} u_k(x) = f(x) \tag{12.90}$$

converges uniformly and absolutely[†] *to a function $f(x)$ for all $x \in I$. In particular, if the summands $u_k(x)$ in Theorem 12.26 are continuous, so is the sum $f(x)$.*

With some care, we are allowed to manipulate uniformly convergent series just like finite sums. Thus, if (12.90) is a uniformly convergent series, so is the term-wise product

$$\sum_{k=1}^{\infty} g(x)\, u_k(x) = g(x)\, f(x) \tag{12.91}$$

[†] Recall that a series $\displaystyle\sum_{n=1}^{\infty} a_n = a^\star$ is said to converge *absolutely* if and only if $\displaystyle\sum_{n=1}^{\infty} |\, a_n \,|$ converges, [**9**].

with any bounded function: $|\, g(x)\,| \le C$ for $x \in I$. We can integrate a uniformly convergent series term by term[‡], and the resulting integrated series

$$\int_a^x \left(\sum_{k=1}^{\infty} u_k(y) \right) dy = \sum_{k=1}^{\infty} \int_a^x u_k(y)\, dy = \int_a^x f(y)\, dy \tag{12.92}$$

is uniformly convergent. Differentiation is also allowed — but only when the differentiated series converges uniformly.

Proposition 12.27. *If* $\displaystyle\sum_{k=1}^{\infty} u_k'(x) = g(x)$ *is a uniformly convergent series, then* $\displaystyle\sum_{k=1}^{\infty} u_k(x) = f(x)$ *is also uniformly convergent, and, moreover,* $f'(x) = g(x)$.

We are particularly interested in applying these results to Fourier series, which, for convenience, we take in complex form

$$f(x) \;\sim \sum_{k=-\infty}^{\infty} c_k\, e^{\,\mathrm{i}\,k\,x}. \tag{12.93}$$

Since x is real, $\left|\, e^{\,\mathrm{i}\,k\,x} \,\right| \le 1$, and hence the individual summands are bounded by

$$\left|\, c_k\, e^{\,\mathrm{i}\,k\,x} \,\right| \le |\, c_k\,| \qquad \text{for all } x.$$

Applying the Weierstrass M–test, we immediately deduce the basic result on uniform convergence of Fourier series.

Theorem 12.28. *If the Fourier coefficients* c_k *satisfy*

$$\sum_{k=-\infty}^{\infty} |\, c_k\,| < \infty, \tag{12.94}$$

then the Fourier series (12.93) *converges uniformly to a continuous function* $\widetilde{f}(x)$ *having the same Fourier coefficients:* $c_k = \langle\, f\,, e^{\,\mathrm{i}\,k\,x}\,\rangle = \langle\, \widetilde{f}\,, e^{\,\mathrm{i}\,k\,x}\,\rangle$.

Proof: Uniform convergence and continuity of the limiting function follow from Theorem 12.26. To show that the c_k actually are the Fourier coefficients of the sum, we multiply the Fourier series by $e^{-\,\mathrm{i}\,k\,x}$ and integrate term by term from $-\pi$ to π. As in (12.91, 92), both operations are valid thanks to the uniform convergence of the series. *Q.E.D.*

The one thing that the theorem does not guarantee is that the original function $f(x)$ used to compute the Fourier coefficients c_k is the *same* as the function $\widetilde{f}(x)$ obtained by summing the resulting Fourier series! Indeed, this may very well not be the case. As we know, the function that the series converges to is necessarily 2π periodic. Thus, at the very least, $\widetilde{f}(x)$ will be the 2π periodic extension of $f(x)$. But even this may not suffice.

[‡] Assuming that the individual functions are all integrable.

Two functions $f(x)$ and $\widehat{f}(x)$ that have the same values except for a finite set of points $x_1, \dots, x_m$ have the same Fourier coefficients. (Why?) More generally, two functions which agree everywhere outside a set of "measure zero" will have the same Fourier coefficients. In this way, a convergent Fourier series singles out a distinguished representative from a collection of essentially equivalent 2π periodic functions.

Remark: The term "measure" refers to a rigorous generalization of the notion of the length of an interval to more general subsets $S \subset \mathbb{R}$. In particular, S has *measure zero* if it can be covered by a collection of intervals of arbitrarily small total length. For example, any collection of finitely many points, or even countably many points, e.g., the rational numbers, has measure zero. The proper development of the notion of measure, and the consequential Lebesgue theory of integration, is properly studied in a course in real analysis, [**152**, **153**].

As a consequence of Theorem 12.28, Fourier series cannot converge uniformly when discontinuities are present. Non-uniform convergence is typically manifested by some form of Gibbs pehnomenon at the discontinuities. However, it can be proved, [**32**, **62**, **193**], that even when the function fails to be everywhere continuous, its Fourier series is uniformly convergent on any closed subset of continuity.

Theorem 12.29. *Let $f(x)$ be 2π periodic and piecewise C^1. If f is continuous for $a < x < b$, then its Fourier series converges uniformly to $f(x)$ on any closed subinterval $a + \delta \leq x \leq b - \delta$, with $\delta > 0$.*

For example, the Fourier series (12.41) for the step function does converge uniformly if we stay away from the discontinuities; for instance, by restriction to a subinterval of the form $[\,\delta, \pi - \delta\,]$ or $[\,-\pi + \delta, -\delta\,]$ for any $0 < \delta < \frac{1}{2}\pi$. This reconfirms our observation that the nonuniform Gibbs behavior becomes progressively more and more localized at the discontinuities.

Smoothness and Decay

The uniform convergence criterion (12.94) requires, at the very least, that the Fourier coefficients decay to zero: $c_k \to 0$ as $k \to \pm\infty$. In fact, the Fourier coefficients cannot tend to zero too slowly. For example, the individual summands of the infinite series

$$\sum_{k=-\infty}^{\infty} \frac{1}{|\,k\,|^{\alpha}} \tag{12.95}$$

go to 0 as $k \to \infty$ whenever $\alpha > 0$, but the series only converges when $\alpha > 1$. (This follows from the standard integral convergence test for series, [**9**, **153**].) Thus, if we can bound the Fourier coefficients by

$$|\,c_k\,| \;\leq\; \frac{M}{|\,k\,|^{\alpha}} \qquad \text{for all} \qquad |\,k\,| \gg 0, \tag{12.96}$$

for some power $\alpha > 1$ and some positive constant $M > 0$, then the Weierstrass M test will guarantee that the Fourier series converges uniformly to a continuous function.

An important consequence of the differentiation formulae (12.76) for Fourier series is the fact that the faster the Fourier coefficients of a function tend to zero as $k \to \infty$, the smoother the function is. Thus, one can detect the degree of smoothness of a function by seeing how rapidly its Fourier coefficients decay to zero. More rigorously:

Theorem 12.30. *If the Fourier coefficients satisfy*

$$\sum_{k=-\infty}^{\infty} k^n \, |\, c_k \,| \;<\; \infty, \tag{12.97}$$

then the Fourier series (12.55) *converges to an n times continuously differentiable 2π periodic function $f(x) \in \mathrm{C}^n$. Moreover, for any $m \leq n$, the m times differentiated Fourier series converges uniformly to the corresponding derivative $f^{(m)}(x)$.*

Proof: This is an immediate consequence of Proposition 12.27 combined with Theorem 12.28. Application of the Weierstrass M test to the differentiated Fourier series based on our hypothesis (12.97) serves to complete the proof. *Q.E.D.*

Corollary 12.31. *If the Fourier coefficients satisfy* (12.96) *for some $\alpha > n+1$, then the function $f(x)$ is n times continuously differentiable.*

Thus, stated roughly, the smaller its high frequency Fourier coefficients, the smoother the function. If the Fourier coefficients go to zero faster than any power of k, e.g., exponentially fast, then the function is infinitely differentiable. Analyticity is a little more delicate, and we refer the reader to [**62**, **193**] for details.

Example 12.32. The 2π periodic extension of the function $|\, x \,|$ is continuous with piecewise continuous first derivative. Its Fourier coefficients (12.46) satisfy the estimate (12.96) for $\alpha = 2$, which is not quite fast enough to ensure a continuous second derivative. On the other hand, the Fourier coefficients (12.29) of the step function $\sigma(x)$ only tend to zero as $1/k$, so $\alpha = 1$, reflecting the fact that its periodic extension is only piecewise continuous. Finally, the Fourier coefficients (12.59) for the delta function do not tend to zero at all, indicative of the fact that it is not an ordinary function, and its Fourier series does not converge in the standard sense.

Hilbert Space

In order to make further progress, we must take a little detour. The proper setting for the rigorous theory of Fourier series turns out to be the most important function space in modern physics and modern analysis, known as *Hilbert space* in honor of the great German mathematician David Hilbert. The precise definition of this infinite-dimensional inner product space is rather technical, but a rough version goes as follows:

Definition 12.33. A complex-valued function $f(x)$ is called *square-integrable* on the interval $[\, -\pi, \pi \,]$ if it satisfies

$$\| \, f \, \|^2 = \frac{1}{2\,\pi} \int_{-\pi}^{\pi} |\, f(x) \,|^2 \, dx < \infty. \tag{12.98}$$

The *Hilbert space* $\mathrm{L}^2 = \mathrm{L}^2[-\pi, \pi]$ is the vector space consisting of all complex-valued *square-integrable* functions.

Note that (12.98) is the L^2 norm based on the standard Hermitian inner product

$$\langle\, f\,, g\,\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\,\overline{g(x)}\,dx. \tag{12.99}$$

The triangle inequality

$$\|\, c\,f + d\,g\,\| \;\le\; |\,c\,|\,\|\,f\,\| + |\,d\,|\,\|\,g\,\|,$$

implies that the Hilbert space is, as claimed, a complex vector space, i.e., if $f, g \in \mathrm{L}^2$, so $\|\,f\,\|, \|\,g\,\| < \infty$, then any linear combination $c\,f + d\,g \in \mathrm{L}^2$ since $\|\,c\,f + d\,g\,\| < \infty$. The Cauchy–Schwarz inequality

$$|\,\langle\, f\,, g\,\rangle\,| \le \|\,f\,\|\,\|\,g\,\|,$$

implies that the inner product of two square-integrable functions is well-defined and finite. In particular, the Fourier coefficients of a function $f(x)$ are defined as inner products

$$c_k = \langle\, f\,, e^{\,\mathrm{i}\,k\,x}\,\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\,e^{-\,\mathrm{i}\,k\,x}\,dx$$

of f with the complex exponentials (which are continuous and so in L^2), and hence are well-defined for any $f \in \mathrm{L}^2$.

There are some interesting analytical subtleties that arise when one tries to prescribe precisely which functions are to be admitted to Hilbert space. Every piecewise continuous function belongs to L^2. But some functions with singularities are also members. For example, the power function $|\,x\,|^{-\alpha}$ belongs to L^2 for any $\alpha < \frac{1}{2}$, but not if $\alpha \ge \frac{1}{2}$.

Analysis requires limiting procedures, and Hilbert space must be "complete" in the sense that appropriately convergent[†] sequences of functions have a limit. The completeness requirement is not elementary, and relies on the development of the more sophisticated Lebesgue theory of integration, which was formalized in the early part of the twentieth century by the French mathematician Henri Lebesgue. Any function which is square-integrable in the Lebesgue sense is admitted into L^2. This includes such non-piecewise continuous functions as $\sin\dfrac{1}{x}$ and $x^{-1/3}$, as well as the strange function

$$r(x) = \begin{cases} 1 & \text{if } x \text{ is a rational number,} \\ 0 & \text{if } x \text{ is irrational.} \end{cases} \tag{12.100}$$

One soon discovers that square-integrable functions can be quite bizarre.

[†] The precise technical requirement is that every *Cauchy sequence* of functions $v_k(x) \in \mathrm{L}^2$ converges to a function $v_\star(x) \in \mathrm{L}^2$; see Exercise ■ for details.

A second complication is that (12.98) does not, strictly speaking, define a norm once we allow discontinuous functions into the fold. For example, the piecewise continuous function

$$f_0(x) = \begin{cases} 1, & x = 0, \\ 0, & x \neq 0, \end{cases} \tag{12.101}$$

has norm zero, $\| f_0 \| = 0$, even though it is not zero everywhere. Indeed, any function which is zero except on a set of measure zero also has norm zero, including the function (12.100). Therefore, in order to make (12.98) into a legitimate norm on Hilbert space, we must agree to identify any two functions which have the same values except on a set of measure zero. For instance, the zero function 0 and the preceding examples $f_0(x)$ and $r(x)$ are all viewed as defining the *same* element of Hilbert space. Thus, although we treat them as if they were ordinary functions, each element of Hilbert space is not, in fact, a function, but, rather, an equivalence class of functions all differing on a set of measure zero. All this might strike the applied reader as becoming much too abstract and arcane. In practice, you will not lose much by assuming that the "functions" in L^2 are always piecewise continuous and square-integrable. Nevertheless, the full analytical power of Hilbert space theory is only unleashed by including completely general functions in L^2.

After its invention by pure mathematicians around the turn of the twentieth century, physicists in the 1920's suddenly realized that Hilbert space was the correct setting to establish the modern theory of quantum mechanics. A quantum mechanical *wave function* is a element[†] $\varphi \in \mathrm{L}^2$ that has unit norm: $\| \varphi \| = 1$. Thus, the set of wave functions is merely the unit sphere in Hilbert space. Quantum mechanics endows each physical wave function with a probabilistic interpretation. Suppose the wave function represents a single subatomic particle — photon, electron, etc. The modulus $|\, \varphi(x) \,|$ of the wave function quantifies the probability of finding the particle at the position x. More correctly, the probability that the particle resides in a prescribed interval $[\, a, b \,]$ is equal to $\sqrt{\dfrac{1}{2\pi} \displaystyle\int_a^b |\, \varphi(x) \,|^2 \, dx}$. In particular, the wave function has unit norm

$$\| \varphi \| = \sqrt{\frac{1}{2\pi} \int_{-\pi}^{\pi} |\, \varphi(x) \,|^2 \, dx} = 1$$

because the particle must certainly, i.e., with probability 1, be *somewhere*!

Convergence in Norm

We are now in a position to discuss convergence in norm of the Fourier series. We begin with the basic definition, which makes sense on any normed vector space.

Definition 12.34. Let V be a normed vector space. A sequence $\mathbf{v}^{(n)}$ is said to *converge in norm* to $\mathbf{v}^\star \in V$ if $\| \mathbf{v}^{(n)} - \mathbf{v}^\star \| \to 0$ as $n \to \infty$.

[†] Here we are acting as if the physical space were represented by the one-dimensional interval $[\, -\pi, \pi \,]$. The more apt case of three-dimensional physical space is developed analogously, replacing the single integral by a triple integral over all of $\mathbb{R}^3$.

As we noted earlier, on finite-dimensional vector spaces, convergence in norm is equivalent to ordinary convergence. On the other hand, on infinite-dimensional function spaces, convergence in norm is very different from pointwise convergence. For instance, it is possible, cf. Exercise ■, to construct a sequence of functions that converges in norm to 0, but does not converge pointwise *anywhere*!

We are particularly interested in the convergence in norm of the Fourier series of a square integrable function $f(x) \in \mathrm{L}^2$. Let

$$s_n(x) = \sum_{k=-n}^{n} c_k \, e^{\mathrm{i}\, k\, x} \tag{12.102}$$

be the n^{th} partial sum of its Fourier series (12.55). The partial sum (12.102) belongs to the subspace $\mathcal{T}^{(n)} \subset \mathrm{L}^2$ of all trigonometric polynomials of degree at most n, spanned by $e^{-\mathrm{i}\, n x}, \dots, e^{\mathrm{i}\, n x}$. It is, in fact, distinguished as the function in $\mathcal{T}^{(n)}$ that lies the closest to f, where the distance between functions is measured by the L^2 norm of their difference: $\| f - g \|$. This important characterization of the Fourier partial sums is, in fact, an immediate consequence of the orthonormality of the trigonometric basis.

Theorem 12.35. *The n^{th} order Fourier partial sum $s_n \in \mathcal{T}^{(n)}$ is the best least squares approximation to $f \in \mathrm{L}^2$, meaning that it minimizes the distance, as measured by the L^2 norm of the difference*

$$\| f - p_n \|^2 = \frac{1}{2\,\pi} \int_{-\pi}^{\pi} | \, f(x) - p_n(x) \, |^2 \, dx, \tag{12.103}$$

among all possible degree n trigonometric polynomials

$$p_n(x) = \sum_{k=-n}^{n} d_k \, e^{\mathrm{i}\, k\, x} \in \mathcal{T}^{(n)}. \tag{12.104}$$

Proof: The proof is, in fact, an exact replica of that of the finite-dimensional Theorems 5.37 and 5.39. Note first that, owing to the orthonormality of the basis exponentials, (12.54), we can compute the norm of a trigonometric polynomial (12.104) by summing the squared moduli of its Fourier coefficients:

$$\| \, p_n \, \|^2 = \langle \, p_n \, , p_n \, \rangle = \sum_{k,l=-n}^{n} d_k \, \overline{d_l} \, \langle \, e^{\mathrm{i}\, k\, x} \, , e^{\mathrm{i}\, l\, x} \, \rangle = \sum_{k=-n}^{n} | \, d_k \, |^2,$$

reproducing our standard formula (5.5) for the norm with respect to an orthonormal basis in this situation. Therefore, employing the identity in Exercise 3.6.42(*a*),

$$\| f - p_n \|^2 = \| f \|^2 - 2 \, \mathrm{Re} \, \langle \, f \, , p_n \, \rangle + \| \, p_n \, \|^2 = \| f \|^2 - 2 \, \mathrm{Re} \sum_{k=-n}^{n} \overline{d_k} \, \langle \, f \, , e^{\mathrm{i}\, k\, x} \, \rangle + \| \, p_n \, \|^2$$

$$= \| f \|^2 - 2 \sum_{k=-n}^{n} \mathrm{Re} \, \big(\, c_k \, \overline{d_k} \, \big) + \sum_{k=-n}^{n} | \, d_k \, |^2 = \| f \|^2 - \sum_{k=-n}^{n} | \, c_k \, |^2 + \sum_{k=-n}^{n} | \, d_k - c_k \, |^2;$$

the last equality results from adding and subtracting the squared norm

$$\| s_n \|^2 = \sum_{k=-n}^{n} | c_k |^2 \tag{12.105}$$

of the Fourier partial sum. We conclude that

$$\| f - p_n \|^2 = \| f \|^2 - \| s_n \|^2 + \sum_{k=-n}^{n} | d_k - c_k |^2. \tag{12.106}$$

The first and second terms on the right hand side of (12.106) are uniquely determined by $f(x)$ and hence cannot be altered by the choice of trigonometric polynomial $p_n(x)$, which only affects the final summation. Since the latter is a sum of nonnegative quantities, it is minimized by setting all the summands to zero, i.e., setting $d_k = c_k$. We conclude that $\| f - p_n \|$ is minimized if and only if $d_k = c_k$ are the Fourier coefficients, and hence the least squares minimizer is the Fourier partial sum: $p_n(x) = s_n(x)$. *Q.E.D.*

Setting $p_n = s_n$, so $d_k = c_k$, in (12.106), we conclude that the least squares error for the Fourier partial sum is

$$0 \le \| f - s_n \|^2 = \| f \|^2 - \| s_n \|^2 = \| f \|^2 - \sum_{k=-n}^{n} | c_k |^2.$$

Therefore, the Fourier coefficients of the function f must satisfy the basic inequality

$$\sum_{k=-n}^{n} | c_k |^2 \le \| f \|^2.$$

Consider what happens in the limit as $n \to \infty$. Since we are summing a sequence of non-negative numbers with uniformly bounded partial sums, the limiting summation must exist, and be subject to the same bound. We have thus proved *Bessel's inequality*:

$$\sum_{k=-\infty}^{\infty} | c_k |^2 \le \| f \|^2, \tag{12.107}$$

which is an important waystation on the road to the general theory. Now, as noted earlier, if a series is to converge, the individual summands must go to zero: $| c_k |^2 \to 0$. Therefore, Bessel's inequality immediately implies the following simplified form of the *Riemann–Lebesgue Lemma*.

Lemma 12.36. *If $f \in \mathrm{L}^2$ is square integrable, then its Fourier coefficients satisfy*

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)\, e^{-\mathrm{i}\, k\, x}\, dx \quad \longrightarrow \quad 0 \qquad \text{as} \qquad | k | \to \infty, \tag{12.108}$$

which is equivalent to the decay of the real Fourier coefficients

$$\left.\begin{aligned} a_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos k\,x\,dx \\ b_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin k\,x\,dx \end{aligned}\right\} \quad \longrightarrow \quad 0 \qquad \text{as} \qquad k \to \infty. \tag{12.109}$$

Remark: As before, the convergence of the sum (12.107) requires that the coefficients c_k cannot tend to zero too slowly. For instance, assuming the power bound (12.96), namely

$$|\,c_k\,| \le M\,|\,k\,|^{-\alpha}, \text{ then requiring } \alpha > \tfrac{1}{2} \text{ is enough to ensure that } \sum_{k=-\infty}^{\infty} |\,c_k\,|^2 < \infty.$$

Thus, as we should expect, convergence in norm imposes less restrictive requirements on the decay of the Fourier coefficients than uniform convergence — which needed $\alpha > 1$. Indeed, a Fourier series may very well converge in norm to a discontinuous function, which is not possible under uniform convergence. In fact, there even exist bizarre continuous functions whose Fourier series do not converge uniformly, even failing to converge at all at some points. A deep result says that the Fourier series of a continuous function converges except possibly on a set of measure zero, [**193**]. Again, the subtle details of the convergence of Fourier series are rather delicate, and lack of space and analytical savvy prevents us from delving any further into these topics.

Completeness

As we know, specification of a basis enables you to prescribe all elements of a finite-dimensional vector space as linear combinations of the basis elements. The number of basis elements dictates the dimension. In an infinite-dimensional vector space, there are, by definition, infinitely many linearly independent elements, and no finite collection can serve to describe the entire space. The question then arises to what extent an infinite collection of linearly independent elements can be considered as a basis for the vector space. Mere counting will no longer suffice, since omitting one, or two, or any finite number — or even certain infinite subcollections — from a purported basis will still leave infinitely many linearly independent elements; but, clearly, the reduced collection should, in some sense, no longer serve as a complete basis. The curse of infinity strikes again! For example, while the complete trigonometric collection $1, \cos x, \sin x, \cos 2\,x, \sin 2\,x, \dots$ will represent any $2\,\pi$ periodic L^2 function as a Fourier series, the subcollection $\cos x, \sin x, \cos 2\,x, \sin 2\,x, \dots$ can only represent functions with mean zero, while the subcollection $\sin x, \sin 2\,x, \dots$ only represents odd functions. All three consist of infinitely many linearly independent functions, but only the first could possibly be deemed a basis of L^2. In general, just because we have found a infinite collection of independent elements in an infinite-dimensional vector space, how do we know that we have enough, and are not missing one or two or 10,000 or even infinitely many additional independent elements?

The concept of "completeness" serves to properly formalize the notion of a "basis" of an infinite-dimensional vector space. We shall discuss completeness in a general, abstract setting, but the key example is, of course, the Hilbert space L^2 and the system of

trigonometric (or complex exponential) functions forming a Fourier series. Other important examples arising in later applications include wavelets, Bessel functions, Legendre polynomials, spherical harmonics, and other systems of eigenfunctions of self-adjoint boundary value problems.

For simplicity, we only define completeness in the case of orthonormal systems. (Similar arguments will clearly apply to orthogonal systems, but normality helps to streamline the presentation.) Let V be an infinite-dimensional complex[†] inner product space. Suppose that $u_1, u_2, u_3, \ldots \in V$ form an orthonormal collection of elements of V, so

$$\langle\, u_i \,,\, u_j \,\rangle = \begin{cases} 1 & i = j, \\ 0, & i \neq j. \end{cases} \tag{12.110}$$

A straightforward argument, cf. Proposition 5.4, proves that the u_i are linearly independent. Given $f \in V$, we form its *generalized Fourier series*

$$f \quad \sim \quad \sum_{k=1}^{\infty} c_k\, u_k, \qquad \text{where} \qquad c_k = \langle\, f \,,\, u_k \,\rangle, \tag{12.111}$$

which is our usual orthonormal basis coefficient formula (5.4), and is obtained by formally taking the inner product of the series with u_k and invoking the orthonormality conditions (12.110).

Definition 12.37. An orthonormal system $u_1, u_2, u_3, \ldots \in V$ is called *complete* if the generalized Fourier series (12.111) of any $f \in V$ converges in norm to f:

$$\| f - s_n \| \;\longrightarrow\; 0, \qquad \text{as } n \to \infty, \qquad \text{where} \qquad s_n = \sum_{k=1}^{n} c_k\, u_k \tag{12.112}$$

is the n^{th} partial sum of the generalized Fourier series (12.111).

Thus, completeness requires that every element can be arbitrarily closely approximated (in norm) by a suitable linear combination of the basis elements. A complete orthonormal system should be viewed as the infinite-dimensional version of an orthonormal basis of a finite-dimensional vector space.

The key result for classical Fourier series is that the complex exponentials, or, equivalently, the trigonometric functions, form a complete system. An indication of its proof will appear below.

Theorem 12.38. *The complex exponentials $e^{\,\mathrm{i}\,k\,x}$, $k = 0, \pm 1, \pm 2, \ldots$, form a complete orthonormal system in $\mathrm{L}^2 = \mathrm{L}^2[-\pi, \pi]$. In other words, if $s_n(x)$ denotes the n^{th} partial sum of the Fourier series of the square-integrable function $f(x) \in \mathrm{L}^2$, then $\lim_{n \to \infty} \| f - s_n \| = 0$.*

[†] The results are equally valid in real inner product spaces, with slightly simpler proofs.

In order to understand completeness, let us describe some equivalent characterizations. The *Plancherel formula* is the infinite-dimensional counterpart of our formula (5.5) for the norm of a vector in terms of its coordinates with respect to an orthonormal basis.

Theorem 12.39. *The orthonormal system $u_1, u_2, u_3, \ldots \in V$ is complete if and only if the Plancherel formula*

$$\| f \|^2 = \sum_{k=1}^{\infty} | c_k |^2 = \sum_{k=1}^{\infty} \langle f , u_k \rangle^2, \tag{12.113}$$

holds for every $f \in V$.

Proof: We begin by computing[†] the Hermitian norm

$$\| f - s_n \|^2 = \| f \|^2 - 2 \operatorname{Re} \langle f , s_n \rangle + \| s_n \|^2.$$

Substituting the formula (12.112) for the partial sums, we find, by orthonormality,

$$\| s_n \|^2 = \sum_{k=1}^{n} | c_k |^2, \qquad \text{while} \qquad \langle f , s_n \rangle = \sum_{k=1}^{n} \overline{c_k} \, \langle f , u_k \rangle = \sum_{k=1}^{n} | c_k |^2.$$

Therefore,

$$0 \leq \| f - s_n \|^2 = \| f \|^2 - \sum_{k=1}^{n} | c_k |^2. \tag{12.114}$$

The fact that the left hand side of (12.114) is non-negative for all n implies the abstract form of *Bessel inequality*

$$\sum_{k=1}^{\infty} | c_k |^2 \ \leq \ \| f \|^2, \tag{12.115}$$

which is valid for *any* orthonormal system of elements in an inner product space. The trigonometric Bessel inequality (12.107) is a particular case of this general result. As we noted above, Bessel's inequality implies that the generalized Fourier coefficients $c_k \to 0$ must tend to zero reasonably rapidly in order that the sum of their squares converges.

Plancherel's Theorem 12.39, thus, states that the system of functions is complete if and only if the Bessel inequality is, in fact, an equality! Indeed, letting $n \to \infty$ in (12.114), we have

$$\lim_{n \to \infty} \| f - s_n \|^2 \ = \ \| f \|^2 - \sum_{k=1}^{\infty} | c_k |^2.$$

Therefore, the completeness condition (12.112) holds if and only if the right hand side vanishes, which is the Plancherel identity (12.113). *Q.E.D.*

[†] We are, in essence, repeating the proofs of Theorem 12.35 and the subsequent trigonometric Bessel inequality (12.107) in a more abstract setting.

Corollary 12.40. *Let $f, g \in V$. Then their Fourier coefficients $c_k = \langle f, \varphi_k \rangle, d_k = \langle g, \varphi_k \rangle$ satisfy* Parseval's identity

$$\langle f, g \rangle = \sum_{k=1}^{\infty} c_k \, \overline{d_k}\,. \tag{12.116}$$

Proof: Using the identity in Exercise 3.6.42(*b*),

$$\langle f, g \rangle = \tfrac{1}{4} \left(\| f + g \|^2 - \| f - g \|^2 + \mathrm{i} \, \| f + \mathrm{i}\, g \|^2 - \mathrm{i} \, \| f - \mathrm{i}\, g \|^2 \right).$$

Parseval's identity results from applying the Plancherel formula (12.113) to each term on the right hand side:

$$\langle f, g \rangle = \frac{1}{4} \sum_{k=-\infty}^{\infty} \left(|\, c_k + d_k \,|^2 - |\, c_k - d_k \,|^2 + \mathrm{i} \, |\, c_k + \mathrm{i}\, d_k \,|^2 - \mathrm{i} \, |\, c_k - \mathrm{i}\, d_k \,|^2 \right) = \sum_{k=-\infty}^{\infty} c_k \, \overline{d_k},$$

by a straightforward algebraic manipulation. *Q.E.D.*

In particular, in the case of the complex exponential basis of $\mathrm{L}^2[-\pi, \pi]$, the Plancherel and Parseval formulae tell us that

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} |\, f(x) \,|^2 \, dx = \sum_{k=-\infty}^{\infty} |\, c_k \,|^2, \qquad \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) \, \overline{g(x)} \, dx = \sum_{k=-\infty}^{\infty} c_k \, \overline{d_k}\,, \tag{12.117}$$

in which $c_k = \langle f, e^{\mathrm{i}\, k\, k} \rangle$, $d_k = \langle g, e^{\mathrm{i}\, k\, k} \rangle$ are the ordinary Fourier coefficients of the complex-valued functions $f(x)$ and $g(x)$. Note that the Plancherel formula is a special case of the Parseval identity, obtained by setting $f = g$. In Exercise ■, you are asked to rewrite these formulas in terms of the real Fourier coefficients.

Completeness also tells us that a function is uniquely determined by its Fourier coefficients.

Proposition 12.41. *If the orthonormal system $u_1, u_2, \ldots \in V$ is complete, then the only element $f \in V$ with all zero Fourier coefficients, $0 = c_1 = c_2 = \cdots$, is the zero element: $f = 0$. More generally, two elements $f, g \in V$ have the same Fourier coefficients if and only if they are the same: $f = g$.*

Proof: The proof is an immediate consequence of the Plancherel formula. Indeed, if $c_k = 0$, then (12.113) implies that $\| f \| = 0$. The second statement follows by applying the first to their difference $f - g$. *Q.E.D.*

Another way of stating this result is that the only function which is orthogonal to every element of a complete orthonormal system is the zero function[†]. Interpreted in yet another way, a complete orthonormal system is maximal in the sense that no further orthonormal elements can be appended to it.

[†] Or, to be more technically accurate, any function which is zero outside a set of measure zero.

Let us now discuss the completeness of the Fourier trigonometric/complex exponential functions. We shall prove the completeness criterion only for continuous functions, leaving the harder general proof to the references, [**62**, **193**]. According to Theorem 12.28, if $f(x)$ is continuous, 2π periodic, and piecewise C^1, its Fourier series converges uniformly to $f(x)$, so

$$f(x) = \sum_{k=-\infty}^{\infty} c_k\, e^{\mathrm{i}\,k\,x} \qquad \text{for all} \qquad -\pi \le x \le \pi.$$

The same holds for its complex conjugate $\overline{f(x)}$. Therefore,

$$|\, f(x)\,|^2 = f(x)\,\overline{f(x)} = f(x) \sum_{k=-\infty}^{\infty} \overline{c}_k\, e^{-\mathrm{i}\,k\,x} = \sum_{k=-\infty}^{\infty} \overline{c}_k\, f(x)\, e^{-\mathrm{i}\,k\,x},$$

which also converges uniformly by (12.91). Equation (12.92) permits us to integrate both sides from $-\pi$ to π, yielding

$$\| f \|^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |\, f(x)\,|^2\, dx = \sum_{k=-\infty}^{\infty} \frac{1}{2\pi} \int_{-\pi}^{\pi} \overline{c}_k\, f(x)\, e^{-\mathrm{i}\,k\,x}\, dx = \sum_{k=-\infty}^{\infty} c_k\, \overline{c}_k = \sum_{k=-\infty}^{\infty} |\, c_k\,|^2.$$

Therefore, Plancherel's identity (12.113) holds for any continuous function. With some additional technical work, this result is used to establish the validity of Plancherel's formula for all $f \in L^2$, the key step being to suitably approximate f by continuous functions. With this in hand, completeness is an immediate consequence of Theorem 12.39. *Q.E.D.*

Pointwise Convergence

Let us finally turn to the proof of the Pointwise Convergence Theorem 12.7. The goal is to prove that, under the appropriate hypotheses on $f(x)$, namely 2π periodic and piecewise C^1, the limit of the partial Fourier sums is

$$\lim_{n \to \infty} s_n(x) = \tfrac{1}{2}\,\big[\, f(x^+) + f(x^-)\,\big]. \tag{12.118}$$

We begin by substituting the formulae (12.56) for the complex Fourier coefficients into the formula (12.102) for the n^{th} partial sum:

$$\begin{aligned} s_n(x) = \sum_{k=-n}^{n} c_k\, e^{\mathrm{i}\,k\,x} &= \sum_{k=-n}^{n} \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} f(y)\, e^{-\mathrm{i}\,k\,y}\, dy \right) e^{\mathrm{i}\,k\,x} \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(y) \sum_{k=-n}^{n} e^{\mathrm{i}\,k(x-y)}\, dy. \end{aligned}$$

We can then use the geometric summation formula (12.63) to evaluate the result:

$$\begin{aligned} s_n(x) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f(y)\, \frac{\sin\left(n + \frac{1}{2}\right)(x-y)}{\sin \frac{1}{2}(x-y)}\, dy \\ &= \frac{1}{2\pi} \int_{x-\pi}^{x+\pi} f(x+y)\, \frac{\sin\left(n + \frac{1}{2}\right) y}{\sin \frac{1}{2} y}\, dy = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x+y)\, \frac{\sin\left(n + \frac{1}{2}\right) y}{\sin \frac{1}{2} y}\, dy. \end{aligned}$$

The second equality is the result of changing the integration variable from y to $x + y$; the final equality follows since the integrand is 2π periodic, and so its integrals over *any* interval of length 2π all have the same value; see Exercise ■.

Thus, to prove (12.118), it suffices to show that

$$\begin{aligned} \lim_{n \to \infty} \frac{1}{\pi} \int_0^{\pi} f(x+y) \, \frac{\sin\left(n + \frac{1}{2}\right) y}{\sin \frac{1}{2} y} \, dy &= f(x^+), \\ \lim_{n \to \infty} \frac{1}{\pi} \int_{-\pi}^{0} f(x+y) \, \frac{\sin\left(n + \frac{1}{2}\right) y}{\sin \frac{1}{2} y} \, dy &= f(x^-). \end{aligned} \tag{12.119}$$

The proofs of the two formulae are identical, and so we concentrate on the first. Since the integrand is even,

$$\frac{1}{\pi} \int_0^{\pi} \frac{\sin\left(n + \frac{1}{2}\right) y}{\sin \frac{1}{2} y} \, dy = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\sin\left(n + \frac{1}{2}\right) y}{\sin \frac{1}{2} y} \, dy = 1,$$

by equation (12.65). Multiplying this formula by $f(x^+)$ and substracting off the right hand side leads to

$$\lim_{n \to \infty} \frac{1}{\pi} \int_0^{\pi} \frac{f(x+y) - f(x^+)}{\sin \frac{1}{2} y} \sin\left(n + \tfrac{1}{2}\right) y \; dy = 0, \tag{12.120}$$

which we now proceed to prove. We claim that, for each fixed value of x, the function

$$g(y) = \frac{f(x+y) - f(x^+)}{\sin \frac{1}{2} y}$$

is piecewise continuous for all $0 \leq y \leq \pi$. Owing to our hypotheses on $f(x)$, the only problematic point is when $y = 0$, but then, by l'Hôpital's rule (for one-sided limits),

$$\lim_{y \to 0^+} g(y) = \lim_{y \to 0^+} \frac{f(x+y) - f(x^+)}{\sin \frac{1}{2} y} = \lim_{y \to 0^+} \frac{f'(x+y)}{\frac{1}{2} \cos \frac{1}{2} y} = 2\, f'(x^+).$$

Consequently, (12.120) will be established if we can show that

$$\lim_{n \to \infty} \frac{1}{\pi} \int_0^{\pi} g(y) \, \sin\left(n + \tfrac{1}{2}\right) y \; dy = 0 \tag{12.121}$$

whenever g is piecewise continuous. Were it not for the extra $\frac{1}{2}$, this would immediately follow from the simplified Riemann–Lebesgue Lemma 12.36. More honestly, we can invoke the addition formula for $\sin\left(n + \frac{1}{2}\right) y$ to write

$$\frac{1}{\pi} \int_0^{\pi} g(y) \sin\left(n + \tfrac{1}{2}\right) y \, dy = \frac{1}{\pi} \int_0^{\pi} \left(g(y) \sin \tfrac{1}{2}\, y \right) \cos n\, y \, dy + \frac{1}{\pi} \int_0^{\pi} \left(g(y) \cos \tfrac{1}{2}\, y \right) \sin n\, y \, dy$$

The first integral is the n^{th} Fourier cosine coefficient for the piecewise continuous function $g(y) \sin \frac{1}{2}\, y$, while the second integral is the n^{th} Fourier sine coefficient for the piecewise continuous function $g(y) \cos \frac{1}{2}\, y$. Lemma 12.36 implies that both of these converge to zero as $n \to \infty$, and hence (12.121) holds. This completes the proof, establishing pointwise convergence of the Fourier series. *Q.E.D.*

Remark: An alternative approach to the last part of the proof is to use the general *Riemann–Lebesgue Lemma*, whose proof can be found in [**62**, **193**].

Lemma 12.42. *Suppose $g(x)$ is piecewise continuous on $[a, b]$. Then*

$$0 = \lim_{\omega \to \infty} \int_a^b g(x)\, e^{\,\mathrm{i}\,\omega\, x}\, dx = \int_a^b g(x) \cos \omega\, x\, dx + \mathrm{i} \int_a^b g(x) \sin \omega\, x\, dx. \tag{12.122}$$

Intuitively, as the frequency ω gets larger and larger, the increasingly rapid oscillations in $e^{\,\mathrm{i}\,\omega\, x}$ tend to cancel each other out. In mathematical language, the Riemann–Lebesgue formula (12.122) says that, as $\omega \to \infty$, the integrand $g(x)\, e^{\,\mathrm{i}\,\omega\, x}$ converges weakly to 0; we saw the same phenomenon in the weak convergence of the Fourier series of the delta function (12.61).

Chapter 13

Fourier Analysis

In addition to their inestimable importance in mathematics and its applications, Fourier series also serve as the entry point into the wonderful world of Fourier analysis and its wide-ranging extensions and generalizations. An entire industry is devoted to further developing the theory and enlarging the scope of applications of Fourier–inspired methods. New directions in Fourier analysis continue to be discovered and exploited in a broad range of physical, mathematical, engineering, chemical, biological, financial, and other systems. In this chapter, we will concentrate on four of the most important variants: discrete Fourier sums leading to the Fast Fourier Transform (FFT); the modern theory of wavelets; the Fourier transform; and, finally, its cousin, the Laplace transform. In addition, more general types of eigenfunction expansions associated with partial differential equations in higher dimensions will appear in the following chapters.

Modern digital media, such as CD's, DVD's and MP3's, are based on discrete data, not continuous functions. One typically samples an analog signal at equally spaced time intervals, and then works exclusively with the resulting discrete (digital) data. The associated discrete Fourier representation re-expresses the data in terms of sampled complex exponentials; it can, in fact, be handled by finite-dimensional vector space methods, and so, technically, belongs back in the linear algebra portion of this text. However, the insight gained from the classical continuous Fourier theory proves to be essential in understanding and analyzing its discrete digital counterpart. An important application of discrete Fourier sums is in signal and image processing. Basic data compression and noise removal algorithms are applied to the sample's discrete Fourier coefficients, acting on the observation that noise tends to accumulate in the high frequency Fourier modes, while most important features are concentrated at low frequencies. The first Section 13.1 develops the basic Fourier theory in this discrete setting, culminating in the Fast Fourier Transform (FFT), which produces an efficient numerical algorithm for passing between a signal and its discrete Fourier coefficients.

One of the inherent limitations of classical Fourier methods, both continuous and discrete, is that they are not well adapted to localized data. (In physics, this lack of localization is the basis of the Heisenberg Uncertainty Principle.) As a result, Fourier-based signal processing algorithms tend to be inaccurate and/or inefficient when confronting highly localized signals or images. In the second section, we introduce the modern theory of wavelets, which is a recent extension of Fourier analysis that more naturally incorporates multiple scales and localization. Wavelets are playing an increasingly dominant role in many modern applications; for instance, the new JPEG digital image compression format is based on wavelets, as are the computerized FBI fingerprint data used in law enforcement

in the United States.

Spectral analysis of non-periodic functions defined on the entire real line requires replacing the Fourier series by a limiting Fourier integral. The resulting Fourier transform plays an essential role in functional analysis, in the analysis of ordinary and partial differential equations, and in quantum mechanics, data analysis, signal processing, and many other applied fields. In Section 13.3, we introduce the most important applied features of the Fourier transform.

The closely related Laplace transform is a basic tool in engineering applications. To mathematicians, the Fourier transform is the more fundamental of the two, while the Laplace transform is viewed as a certain real specialization. Both transforms change differentiation into multiplication, thereby converting linear differential equations into algebraic equations. The Fourier transform is primarily used for solving boundary value problems on the real line, while initial value problems, particularly those involving discontinuous forcing terms, are effectively handled by the Laplace transform.

13.1. Discrete Fourier Analysis and the Fast Fourier Transform.

In modern digital media — audio, still images or video — continuous signals are sampled at discrete time intervals before being processed. Fourier analysis decomposes the sampled signal into its fundamental periodic constituents — sines and cosines, or, more conveniently, complex exponentials. The crucial fact, upon which all of modern signal processing is based, is that the sampled complex exponentials form an orthogonal basis. The section introduces the Discrete Fourier Transform, and concludes with an introduction to the Fast Fourier Transform, an efficient algorithm for computing the discrete Fourier representation and reconstructing the signal from its Fourier coefficients.

We will concentrate on the one-dimensional version here. Let $f(x)$ be a function representing the signal, defined on an interval $a \leq x \leq b$. Our computer can only store its measured values at a finite number of *sample points* $a \leq x_0 < x_1 < \cdots < x_n \leq b$. In the simplest and, by far, the most common case, the sample points are equally spaced, and so

$$x_j = a + j\,h, \qquad j = 0, \ldots, n, \qquad \text{where} \qquad h = \frac{b-a}{n}$$

indicates the sample rate. In signal processing applications, x represents time instead of space, and the x_j are the times at which we sample the signal $f(x)$. Sample rates can be very high, e.g., every 10–20 milliseconds in current speech recognition systems.

For simplicity, we adopt the "standard" interval of $0 \leq x \leq 2\pi$, and the n equally spaced sample points[†]

$$x_0 = 0, \qquad x_1 = \frac{2\pi}{n}, \qquad x_2 = \frac{4\pi}{n}, \quad \ldots \quad x_j = \frac{2j\pi}{n}, \quad \ldots \quad x_{n-1} = \frac{2(n-1)\pi}{n}. \tag{13.1}$$

(Signals defined on other intervals can be handled by simply rescaling the interval to have length 2π.) Sampling a (complex-valued) signal or function $f(x)$ produces the *sample*

[†] We will find it convenient to omit the final sample point $x_n = 2\pi$ from consideration.

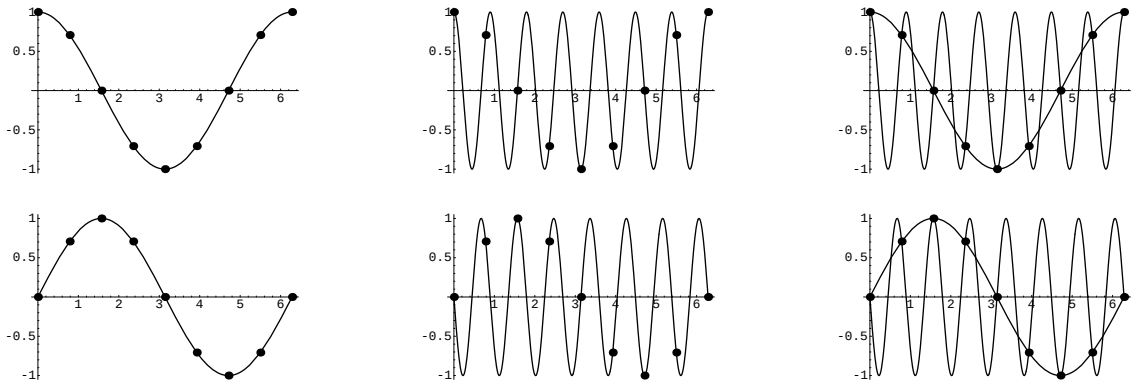


Figure 13.1. Sampling $e^{-\mathrm{i}\,x}$ and $e^{7\,\mathrm{i}\,x}$ on $n = 8$ sample points.

vector

$$\mathbf{f} = \big(\, f_0, f_1, \dots, f_{n-1} \,\big)^T = \big(\, f(x_0), f(x_1), \dots, f(x_{n-1}) \,\big)^T,$$

where

$$f_j = f(x_j) = f\left(\frac{2\,j\,\pi}{n}\right). \tag{13.2}$$

Sampling cannot distinguish between functions that have the same values at all of the sample points — from the sampler's point of view they are identical. For example, the periodic complex exponential function

$$f(x) = e^{\mathrm{i}\,n\,x} = \cos n\,x + \mathrm{i}\,\sin n\,x$$

has sampled values

$$f_j = f\left(\frac{2\,j\,\pi}{n}\right) = \exp\left(\mathrm{i}\,n\,\frac{2\,j\,\pi}{n}\right) = e^{2\,j\,\pi\,\mathrm{i}} = 1 \qquad \text{for all} \qquad j = 0, \dots, n-1,$$

and hence is indistinguishable from the constant function $c(x) \equiv 1$ — both lead to the *same* sample vector $(\,1, 1, \dots, 1\,)^T$. This has the important implication that sampling at n equally spaced sample points *cannot* detect periodic signals of frequency n. More generally, the two complex exponential signals

$$e^{\mathrm{i}\,(k+n)x} \qquad \text{and} \qquad e^{\mathrm{i}\,k\,x}$$

are also indistinguishable when sampled. This has the important consequence that we need only use the first n periodic complex exponential functions

$$f_0(x) = 1, \qquad f_1(x) = e^{\mathrm{i}\,x}, \qquad f_2(x) = e^{2\,\mathrm{i}\,x}, \qquad \dots \qquad f_{n-1}(x) = e^{(n-1)\,\mathrm{i}\,x}, \tag{13.3}$$

in order to represent any $2\,\pi$ periodic sampled signal. In particular, exponentials $e^{-\mathrm{i}\,k\,x}$ of "negative" frequency can all be converted into positive versions, namely $e^{\mathrm{i}\,(n-k)x}$, by the same sampling argument. For example,

$$e^{-\mathrm{i}\,x} = \cos x - \mathrm{i}\,\sin x \qquad \text{and} \qquad e^{(n-1)\,\mathrm{i}\,x} = \cos(n-1)\,x + \mathrm{i}\,\sin(n-1)\,x$$

have identical values on the sample points (13.1). However, off of the sample points, they are quite different; the former is slowly varying, while the latter represents a high frequency oscillation. In Figure 13.1, we compare $e^{-\mathrm{i}\,x}$ and $e^{7\,\mathrm{i}\,x}$ when there are $n = 8$ sample values, indicated by the dots on the graphs. The top row compares the real parts, $\cos x$ and $\cos 7\,x$, while the bottom row compares the imaginary parts, $\sin x$ and $-\sin 7\,x$. Note that both functions have the same pattern of sample values, even though their overall behavior is strikingly different.

This effect is commonly referred to as *aliasing*[†]. If you view a moving particle under a stroboscopic light that flashes only eight times, you would be unable to determine which of the two graphs the particle was following. Aliasing is the cause of a well-known artifact in movies: spoked wheels can appear to be rotating backwards when our brain interprets the discretization of the high frequency forward motion imposed by the frames of the film as an equivalently discretized low frequency motion in reverse. Aliasing also has important implications for the design of music CD's. We must sample an audio signal at a sufficiently high rate that all audible frequencies can be adequately represented. In fact, human appreciation of music also relies on inaudible high frequency tones, and so a much higher sample rate is actually used in commercial CD design. But the sample rate that was selected remains controversial; hi fi aficionados complain that it was not set high enough to fully reproduce the musical quality of an analog LP record!

The *discrete Fourier representation* decomposes a sampled function $f(x)$ into a linear combination of complex exponentials. Since we cannot distinguish sampled exponentials of frequency higher than n, we only need consider a finite linear combination

$$f(x) \;\sim\; p(x) \;=\; c_0 + c_1\, e^{\,\mathrm{i}\,x} + c_2\, e^{2\,\mathrm{i}\,x} + \;\cdots\; + c_{n-1}\, e^{(n-1)\,\mathrm{i}\,x} \;=\; \sum_{k=0}^{n-1} c_k\, e^{\,\mathrm{i}\,k\,x} \qquad (13.4)$$

of the first n exponentials (13.3). The symbol $\sim$ in (13.4) means that the function $f(x)$ and the sum $p(x)$ agree on the sample points:

$$f(x_j) = p(x_j), \qquad\qquad j = 0, \ldots, n-1. \qquad (13.5)$$

Therefore, $p(x)$ can be viewed as a (complex-valued) *interpolating trigonometric polynomial* of degree $\leq n-1$ for the sample data $f_j = f(x_j)$.

Remark: If $f(x)$ is real, then $p(x)$ is also real on the sample points, but may very well be complex-valued in between. To avoid this unsatisfying state of affairs, we will usually discard its imaginary component, and regard the real part of $p(x)$ as "the" interpolating trigonometric polynomial. On the other hand, sticking with a purely real construction unnecessarily complicates the analysis, and so we will retain the complex exponential form (13.4) of the discrete Fourier sum.

[†] In computer graphics, the term "aliasing" is used in a much broader sense that covers a variety of artifacts introduced by discretization — particularly, the jagged appearance of lines and smooth curves on a digital monitor.

Since we are working in the finite-dimensional vector space $\mathbb{C}^n$ throughout, we may reformulate the discrete Fourier series in vectorial form. Sampling the basic exponentials (13.3) produces the complex vectors

$$\begin{aligned}\boldsymbol{\omega}_k &= \left(e^{\mathrm{i}\,k\,x_0}, e^{\mathrm{i}\,k\,x_1}, e^{\mathrm{i}\,k\,x_2}, \dots, e^{\mathrm{i}\,k\,x_n} \right)^T \\ &= \left(1, e^{2\,k\,\pi\,\mathrm{i}/n}, e^{4\,k\,\pi\,\mathrm{i}/n}, \dots, e^{2\,(n-1)\,k\,\pi\,\mathrm{i}/n} \right)^T, \qquad k = 0, \dots, n-1. \end{aligned} \tag{13.6}$$

The interpolation conditions (13.5) can be recast in the equivalent vector form

$$\mathbf{f} = c_0\,\boldsymbol{\omega}_0 + c_1\,\boldsymbol{\omega}_1 + \,\cdots\, + c_{n-1}\,\boldsymbol{\omega}_{n-1}. \tag{13.7}$$

In other words, to compute the discrete Fourier coefficients $c_0, \dots, c_{n-1}$ of f, all we need to do is rewrite its sample vector $\mathbf{f}$ as a linear combination of the sampled exponential vectors $\boldsymbol{\omega}_0, \dots, \boldsymbol{\omega}_{n-1}$.

Now, as with continuous Fourier series, the absolutely crucial property is the orthonormality of the basis elements $\boldsymbol{\omega}_0, \dots, \boldsymbol{\omega}_{n-1}$. Were it not for the power of orthogonality, Fourier analysis might have remained a mere mathematical curiosity, rather than today's indispensable tool.

Proposition 13.1. *The sampled exponential vectors $\boldsymbol{\omega}_0, \dots, \boldsymbol{\omega}_{n-1}$ form an orthonormal basis of $\mathbb{C}^n$ with respect to the inner product*

$$\langle\, \mathbf{f}\,, \mathbf{g}\,\rangle \;=\; \frac{1}{n}\sum_{j=0}^{n-1} f_j\,\overline{g_j} \;=\; \frac{1}{n}\sum_{j=0}^{n-1} f(x_j)\,\overline{g(x_j)}\,, \qquad \mathbf{f}, \mathbf{g} \in \mathbb{C}^n. \tag{13.8}$$

Remark: The inner product (13.8) is a rescaled version of the standard Hermitian dot product (3.90) between complex vectors. We can interpret the inner product between the sample vectors $\mathbf{f}, \mathbf{g}$ as the *average* of the sampled values of the product signal $f(x)\,\overline{g(x)}$.

Remark: As usual, orthogonality is no accident. Just as the complex exponentials are eigenfunctions for a self-adjoint boundary value problem, so their discrete sampled counterparts are eigenvectors for a self-adjoint matrix eigenvalue problem; details can be found in Exercise 8.4.12. Here, though, to keep the discussion on track, we shall outline a direct proof.

Proof: The crux of the matter relies on properties of the remarkable complex numbers

$$\zeta_n = e^{2\,\pi\,\mathrm{i}/n} = \cos\frac{2\,\pi}{n} + \mathrm{i}\,\sin\frac{2\,\pi}{n}\,, \qquad \text{where} \qquad n = 1, 2, 3, \dots\,. \tag{13.9}$$

Particular cases include

$$\zeta_2 = -1, \qquad \zeta_3 = -\tfrac{\sqrt{3}}{2} + \tfrac{1}{2}\,\mathrm{i}\,, \qquad \zeta_4 = \mathrm{i}\,, \qquad \text{and} \qquad \zeta_8 = \tfrac{\sqrt{2}}{2} + \tfrac{\sqrt{2}}{2}\,\mathrm{i}\,. \tag{13.10}$$

The n^{th} power of ζ_n is

$$\zeta_n^n = \left(e^{2\,\pi\,\mathrm{i}/n}\right)^n = e^{2\,\pi\,\mathrm{i}} = 1,$$

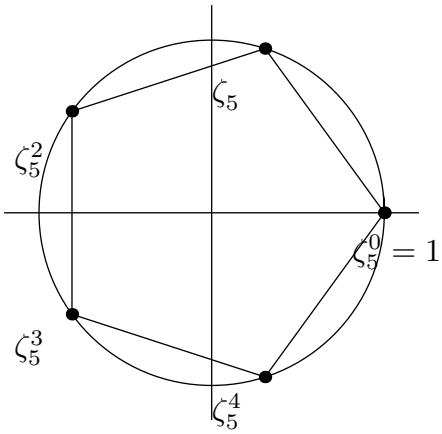


Figure 13.2. The Fifth Roots of Unity.

and hence ζ_n is one of the complex n^{th} *roots of unity*: $\zeta_n = \sqrt[n]{1}$. There are, in fact, n different complex n^{th} roots of 1, including 1 itself, namely the powers of ζ_n:

$$\zeta_n^k = e^{2\,k\,\pi\,\mathrm{i}\,/n} = \cos\frac{2\,k\,\pi}{n} + \mathrm{i}\,\sin\frac{2\,k\,\pi}{n}\,, \qquad\qquad k = 0, \dots, n-1. \tag{13.11}$$

Since it generates all the others, ζ_n is known as the *primitive* n^{th} *root of unity*. Geometrically, the n^{th} roots (13.11) lie on the vertices of a regular unit n–gon in the complex plane; see Figure 13.2. The primitive root ζ_n is the first vertex we encounter as we go around the n–gon in a counterclockwise direction, starting at 1. Continuing around, the other roots appear in their natural order $\zeta_n^2, \zeta_n^3, \dots, \zeta_n^{n-1}$, and finishing back at $\zeta_n^n = 1$. The complex conjugate of ζ_n is the "last" n^{th} root

$$e^{-2\,\pi\,\mathrm{i}\,/n} = \overline{\zeta}_n = \frac{1}{\zeta_n} = \zeta_n^{n-1} = e^{2(n-1)\pi\,\mathrm{i}\,/n}. \tag{13.12}$$

The complex numbers (13.11) are a complete set of roots of the polynomial $z^n - 1$, which can therefore be factored:

$$z^n - 1 = (z-1)(z-\zeta_n)(z-\zeta_n^2)\ \cdots\ (z-\zeta_n^{n-1}).$$

On the other hand, elementary algebra provides us with the real factorization

$$z^n - 1 = (z-1)(1 + z + z^2 + \ \cdots\ + z^{n-1}).$$

Comparing the two, we conclude that

$$1 + z + z^2 + \ \cdots\ + z^{n-1} = (z-\zeta_n)(z-\zeta_n^2)\cdots(z-\zeta_n^{n-1}).$$

Substituting $z = \zeta_n^k$ into both sides of this identity, we deduce the useful formula

$$1 + \zeta_n^k + \zeta_n^{2\,k} + \ \cdots\ + \zeta_n^{(n-1)\,k} = \begin{cases} n, & k = 0, \\ 0, & 0 < k < n. \end{cases} \tag{13.13}$$

Since $\zeta_n^{n+k} = \zeta_n^k$, this formula can easily be extended to general integers k; the sum is equal to n if n evenly divides k and is 0 otherwise.

Now, let us apply what we've learned to prove Proposition 13.1. First, in view of (13.11), the sampled exponential vectors (13.6) can all be written in terms of the n^{th} roots of unity:

$$\boldsymbol{\omega}_k = \big(\, 1, \zeta_n^k, \zeta_n^{2\,k}, \zeta_n^{3\,k}, \ \dots\ , \zeta_n^{(n-1)\,k} \,\big)^T, \qquad\qquad k = 0, \dots, n-1. \tag{13.14}$$

Therefore, applying (13.12, 13), we conclude that

$$\langle\, \boldsymbol{\omega}_k \,, \boldsymbol{\omega}_l \,\rangle = \frac{1}{n} \sum_{j=0}^{n-1} \zeta_n^{j\,k}\, \overline{\zeta_n^{j\,l}} = \frac{1}{n} \sum_{j=0}^{n-1} \zeta_n^{j(k-l)} = \begin{cases} 1, & k = l, \\ 0, & k \neq l, \end{cases} \qquad 0 \leq k, l < n,$$

which establishes orthonormality of the sampled exponential vectors. *Q.E.D.*

Orthonormality of the basis vectors implies that we can immediately compute the Fourier coefficients in the discrete Fourier sum (13.4) by taking inner products:

$$c_k = \langle\, \mathbf{f} \,, \boldsymbol{\omega}_k \,\rangle = \frac{1}{n} \sum_{j=0}^{n-1} f_j\, \overline{e^{\,\mathrm{i}\,k\,x_j}} = \frac{1}{n} \sum_{j=0}^{n-1} f_j\, e^{-\,\mathrm{i}\,k\,x_j} = \frac{1}{n} \sum_{j=0}^{n-1} \zeta_n^{-\,j\,k}\, f_j. \tag{13.15}$$

In other words, the discrete Fourier coefficient c_k is obtained by averaging the sampled values of the product function $f(x)\, e^{-\,\mathrm{i}\,k\,x}$. The passage from a signal to its Fourier coefficients is known as the *Discrete Fourier Transform* or DFT for short. The reverse procedure of reconstructing a signal from its discrete Fourier coefficients via the sum (13.4) (or (13.7)) is known as the *Inverse Discrete Fourier Transform* or IDFT. The Discrete Fourier Transform and its inverse define mutually inverse linear transformations on the space $\mathbb{C}^n$, whose matrix representations can be found in Exercise 5.7.9.

Example 13.2. If $n = 4$, then $\zeta_4 = \mathrm{i}$. The corresponding sampled exponential vectors

$$\boldsymbol{\omega}_0 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \boldsymbol{\omega}_1 = \begin{pmatrix} 1 \\ \mathrm{i} \\ -1 \\ -\,\mathrm{i} \end{pmatrix}, \qquad \boldsymbol{\omega}_2 = \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}, \qquad \boldsymbol{\omega}_3 = \begin{pmatrix} 1 \\ -\,\mathrm{i} \\ -1 \\ \mathrm{i} \end{pmatrix},$$

form an orthonormal basis of $\mathbb{C}^4$ with respect to the averaged Hermitian dot product

$$\langle\, \mathbf{v} \,, \mathbf{w} \,\rangle = \tfrac{1}{4} \left(v_0\, \overline{w_0} + v_1\, \overline{w_1} + v_2\, \overline{w_2} + v_3\, \overline{w_3} \right), \qquad \text{where} \qquad \mathbf{v} = \begin{pmatrix} v_0 \\ v_1 \\ v_2 \\ v_3 \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \end{pmatrix}.$$

Given the sampled function values

$$f_0 = f(0), \qquad f_1 = f\big(\tfrac{1}{2}\,\pi\big), \qquad f_2 = f(\pi), \qquad f_3 = f\big(\tfrac{3}{2}\,\pi\big),$$

we construct the discrete Fourier representation

$$\mathbf{f} = c_0\, \boldsymbol{\omega}_0 + c_1\, \boldsymbol{\omega}_1 + c_2\, \boldsymbol{\omega}_2 + c_3\, \boldsymbol{\omega}_3, \tag{13.16}$$

where

$$\begin{aligned} c_0 &= \langle\, \mathbf{f} \,, \boldsymbol{\omega}_0 \,\rangle = \tfrac{1}{4}(f_0 + f_1 + f_2 + f_3), & c_1 &= \langle\, \mathbf{f} \,, \boldsymbol{\omega}_1 \,\rangle = \tfrac{1}{4}(f_0 - \mathrm{i}\, f_1 - f_2 + \mathrm{i}\, f_3), \\ c_2 &= \langle\, \mathbf{f} \,, \boldsymbol{\omega}_2 \,\rangle = \tfrac{1}{4}(f_0 - f_1 + f_2 - f_3), & c_3 &= \langle\, \mathbf{f} \,, \boldsymbol{\omega}_3 \,\rangle = \tfrac{1}{4}(f_0 + \mathrm{i}\, f_1 - f_2 - \mathrm{i}\, f_3). \end{aligned}$$

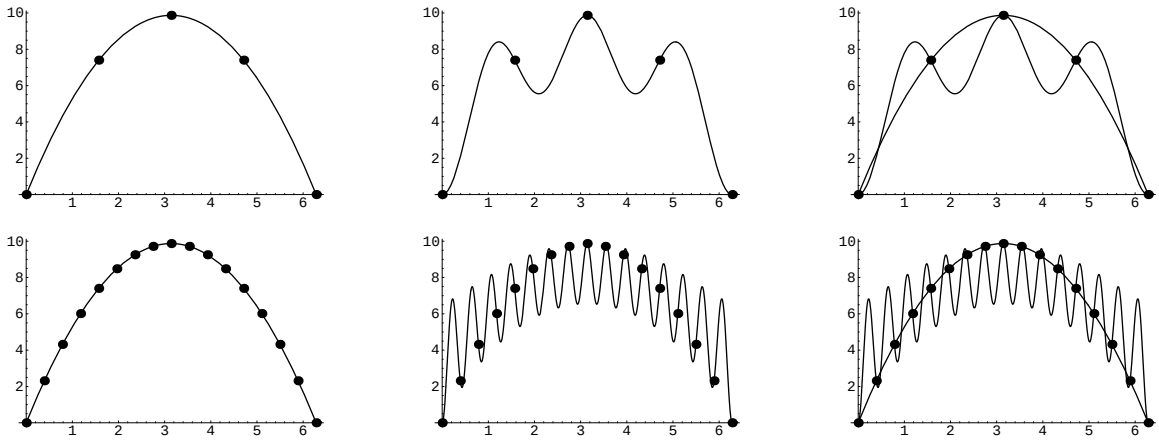


Figure 13.3. The Discrete Fourier Representation of $x^2 - 2\pi x$.

We interpret this decomposition as the complex exponential interpolant

$$f(x) \;\sim\; p(x) = c_0 + c_1 e^{\,\mathrm{i}\,x} + c_2 e^{2\,\mathrm{i}\,x} + c_3 e^{3\,\mathrm{i}\,x}$$

that agrees with $f(x)$ on the sample points.

For instance, if

$$f(x) = 2\pi x - x^2,$$

then

$$f_0 = 0., \qquad f_1 = 7.4022, \qquad f_2 = 9.8696, \qquad f_3 = 7.4022,$$

and hence

$$c_0 = 6.1685, \qquad c_1 = -2.4674, \qquad c_2 = -1.2337, \qquad c_3 = -2.4674.$$

Therefore, the interpolating trigonometric polynomial is given by the real part of

$$p(x) = 6.1685 - 2.4674\, e^{\,\mathrm{i}\,x} - 1.2337\, e^{2\,\mathrm{i}\,x} - 2.4674\, e^{3\,\mathrm{i}\,x}, \tag{13.17}$$

namely,

$$\mathrm{Re}\, p(x) = 6.1685 - 2.4674\, \cos x - 1.2337\, \cos 2x - 2.4674\, \cos 3x. \tag{13.18}$$

In Figure 13.3 we compare the function, with the interpolation points indicated, and discrete Fourier representations (13.18) for both $n = 4$ and $n = 16$ points. The resulting graphs point out a significant difficulty with the Discrete Fourier Transform as developed so far. While the trigonometric polynomials do indeed correctly match the sampled function values, their pronounced oscillatory behavior makes them completely unsuitable for interpolation away from the sample points.

However, this difficulty can be rectified by being a little more clever. The problem is that we have not been paying sufficient attention to the frequencies that are represented in the Fourier sum. Indeed, the graphs in Figure 13.3 might remind you of our earlier observation that, due to aliasing, low and high frequency exponentials can have the same sample data, but differ wildly in between the sample points. While the first half of the

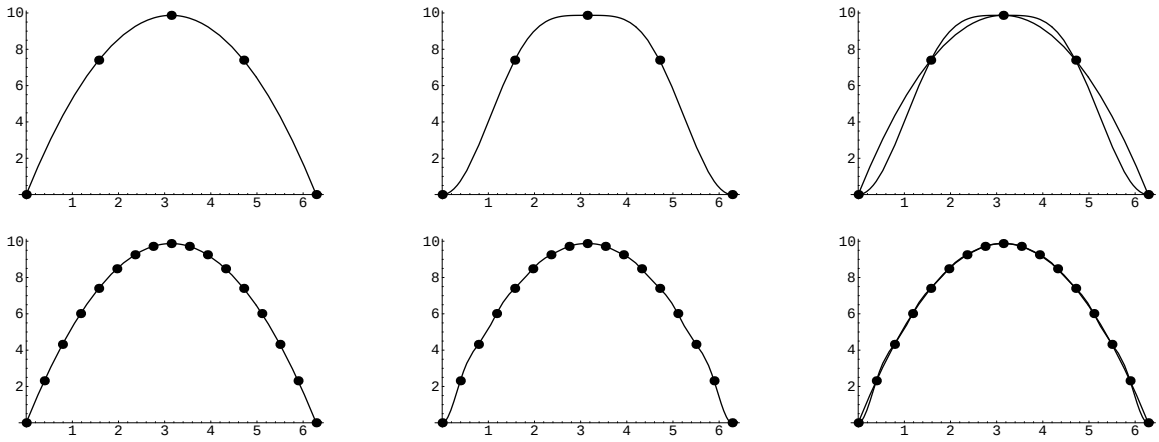


Figure 13.4. The Low Frequency Discrete Fourier Representation of $x^2 - 2\pi x$.

summands in (13.4) represent relatively low frequencies, the second half do not, and can be replaced by equivalent lower frequency, and hence less oscillatory exponentials. Namely, if $0 < k \leq \frac{1}{2}n$, then $e^{-\mathrm{i}kx}$ and $e^{\mathrm{i}(n-k)x}$ have the same sample values, but the former is of lower frequency than the latter. Thus, for interpolatory purposes, we should replace the second half of the summands in the Fourier sum (13.4) by their low frequency alternatives. If $n = 2m + 1$ is odd, then we take

$$\widehat{p}(x) \;=\; c_{-m}\, e^{-\mathrm{i}mx} + \cdots + c_{-1}\, e^{-\mathrm{i}x} + c_0 + c_1\, e^{\mathrm{i}x} + \cdots + c_m\, e^{\mathrm{i}mx} \;=\; \sum_{k=-m}^{m} c_k\, e^{\mathrm{i}kx} \tag{13.19}$$

as the equivalent low frequency interpolant. If $n = 2m$ is even — which is the most common case occurring in applications — then

$$\widehat{p}(x) \;=\; c_{-m}\, e^{-\mathrm{i}mx} + \cdots + c_{-1}\, e^{-\mathrm{i}x} + c_0 + c_1\, e^{\mathrm{i}x} + \cdots + c_{m-1}\, e^{\mathrm{i}(m-1)x} \;=\; \sum_{k=-m}^{m-1} c_k\, e^{\mathrm{i}kx} \tag{13.20}$$

will be our choice. (It is a matter of personal taste whether to use $e^{-\mathrm{i}mx}$ or $e^{\mathrm{i}mx}$ to represent the highest frequency term.) In both cases, the Fourier coefficients with negative indices are the same as their high frequency alternatives:

$$c_{-k} = c_{n-k} = \langle\, \mathbf{f}\,, \boldsymbol{\omega}_{n-k}\,\rangle = \langle\, \mathbf{f}\,, \boldsymbol{\omega}_{-k}\,\rangle, \tag{13.21}$$

where $\boldsymbol{\omega}_{-k} = \boldsymbol{\omega}_{n-k}$ is the sample vector for $e^{-\mathrm{i}kx} \sim e^{\mathrm{i}(n-k)x}$.

Returning to the previous example, for interpolating purposes, we should replace (13.17) by the equivalent low frequency interpolant

$$\widehat{p}(x) = -1.2337\, e^{-2\mathrm{i}x} - 2.4674\, e^{-\mathrm{i}x} + 6.1685 - 2.4674\, e^{\mathrm{i}x}, \tag{13.22}$$

with real part

$$\operatorname{Re} \widehat{p}(x) = 6.1685 - 4.9348 \cos x - 1.2337 \cos 2x.$$

Graphs of the $n = 4$ and 16 low frequency trigonometric interpolants can be seen in Figure 13.4. Thus, by utilizing only the lowest frequency exponentials, we successfully

suppress the aliasing artifacts, resulting in a quite reasonable trigonometric interpolant to the given function.

Remark: The low frequency version also serves to unravel the reality of the Fourier representation of a real function $f(x)$. Since $\boldsymbol{\omega}_{-k} = \overline{\boldsymbol{\omega}_k}$, formula (13.21) implies that $c_{-k} = \overline{c_k}$, and so the common frequency terms

$$c_{-k}\, e^{-\mathrm{i}\, k x} + c_k\, e^{\mathrm{i}\, k x} = a_k \cos k x + b_k \sin k x$$

add up to a real trigonometric function. Therefore, the odd n interpolant (13.19) is a real trigonometric polynomial, whereas in the even version (13.20) only the highest frequency term $c_{-m}\, e^{-\mathrm{i}\, m x}$ produces a complex term — which is, in fact, 0 on the sample points.

Compression and Noise Removal

In a typical experimental signal, noise primarily affects the high frequency modes, while the authentic features tend to appear in the low frequencies. Think of the hiss and static you hear on an AM radio station or a low quality audio recording. Thus, a very simple, but effective, method for denoising a corrupted signal is to decompose it into its Fourier modes, as in (13.4), and then discard the high frequency constituents. A similar idea underlies the Dolby® recording system used on most movie soundtracks: during the recording process, the high frequency modes are artificially boosted, so that scaling them back when showing the movie in the theater has the effect of eliminating much of the extraneous noise. The one design issue is the specification of a cut-off between low and high frequency, that is, between signal and noise. This choice will depend upon the properties of the measured signal, and is left to the discretion of the signal processor.

A correct implementation of the denoising procedure is facilitated by using the unaliased forms (13.19, 20) of the trigonometric interpolant, in which the low frequency summands only appear when $|\, k\,|$ is small. In this version, to eliminate high frequency components, we replace the full summation by

$$q_l(x) \;=\; \sum_{k=-l}^{l} c_k\, e^{\mathrm{i}\, k x}, \tag{13.23}$$

where $l < \frac{1}{2}\,(n+1)$ specifies the selected cut-off frequency between signal and noise. The $2l+1 \ll n$ low frequency Fourier modes retained in (13.23) will, in favorable situations, capture the essential features of the original signal while simultaneously eliminating the high frequency noise.

In Figure 13.5 we display a sample signal followed by the same signal corrupted by adding in random noise. We use $n = 2^8 = 256$ sample points in the discrete Fourier representation, and to remove the noise, we retain only the $2l+1 = 11$ lowest frequency modes. In other words, instead of all $n = 512$ Fourier coefficients $c_{-256}, \dots, c_{-1}, c_0, c_1, \dots, c_{255}$, we only compute the 11 lowest order ones $c_{-5}, \dots, c_5$. Summing up just those 11 exponentials produces the denoised signal $q(x) = c_{-5}\, e^{-5\,\mathrm{i}\, x} + \cdots + c_5\, e^{5\,\mathrm{i}\, x}$. To compare, we plot both the original signal and the denoised version on the same graph. In this case, the maximal deviation is less than .15 over the entire interval $[\,0, 2\,\pi\,]$.

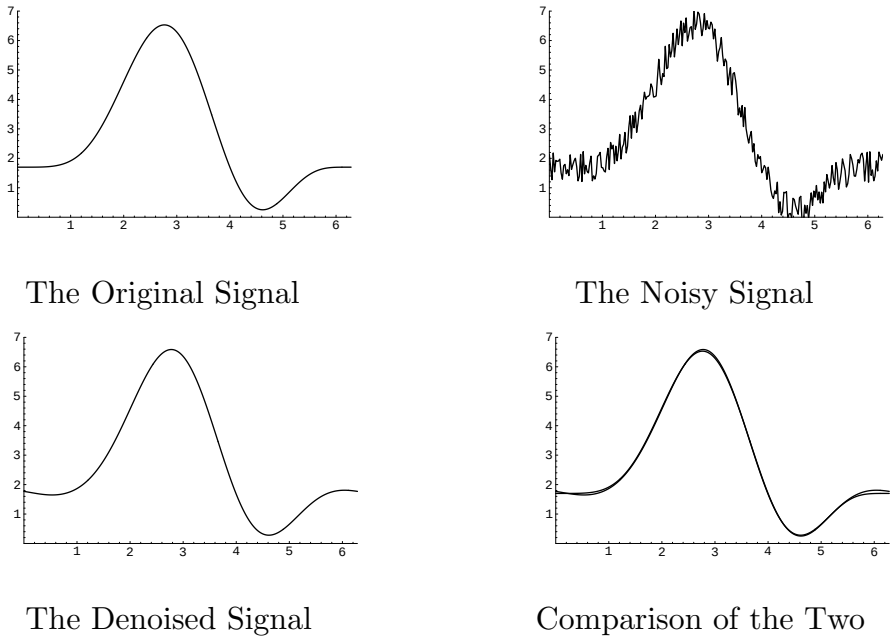


Figure 13.5. Denoising a Signal.

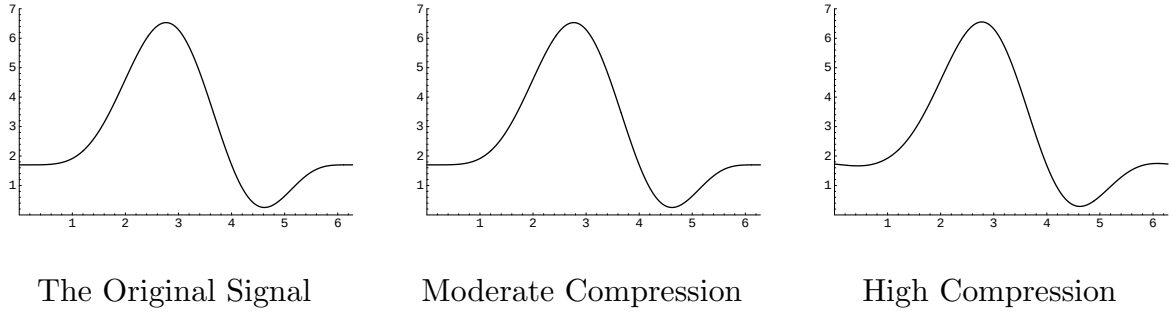


Figure 13.6. Compressing a Signal.

The same idea underlies many data compression algorithms for audio recordings, digital images and, particularly, video. The goal is efficient storage and/or transmission of the signal. As before, we expect all the important features to be contained in the low frequency constituents, and so discarding the high frequency terms will, in favorable situations, not lead to any noticeable degradation of the signal or image. Thus, to compress a signal (and, simultaneously, remove high frequency noise), we retain only its low frequency discrete Fourier coefficients. The signal is reconstructed by summing the associated truncated discrete Fourier series (13.23). A mathematical justification of Fourier-based compression algorithms relies on the fact that the Fourier coefficients of smooth functions tend rapidly to zero — the smoother the function, the faster the decay rate. Thus, the small high frequency Fourier coefficients will be of negligible importance.

In Figure 13.6, the same signal is compressed by retaining, respectively, $2l + 1 = 21$ and $2l + 1 = 7$ Fourier coefficients only instead of all $n = 512$ that would be required for complete accuracy. For the case of moderate compression, the maximal deviation between the signal and the compressed version is less than 1.5×10^{-4} over the entire interval,

while even the highly compressed version deviates at most .05 from the original signal. Of course, the lack of any fine scale features in this particular signal means that a very high compression can be achieved — the more complicated or detailed the original signal, the more Fourier modes need to be retained for accurate reproduction.

The Fast Fourier Transform

While one may admire an algorithm for its intrinsic beauty, in the real world, the bottom line is always efficiency of implementation: the less total computation, the faster the processing, and hence the more extensive the range of applications. Orthogonality is the first and most important feature of many practical linear algebra algorithms, and is *the* critical feature of Fourier analysis. Still, even the power of orthogonality reaches its limits when it comes to dealing with truly large scale problems such as three-dimensional medical imaging or video processing. In the early 1960's, James Cooley and John Tukey, [**43**], discovered† a much more efficient approach to the Discrete Fourier Transform, exploiting the rather special structure of the sampled exponential vectors. The resulting algorithm is known as the *Fast Fourier Transform*, often abbreviated FFT, and its discovery launched the modern revolution in digital signal and data processing, [**29**, **30**].

In general, computing all the discrete Fourier coefficients (13.15) of an n times sampled signal requires a total of n^2 complex multiplications and $n^2 - n$ complex additions. Note also that each complex addition

$$z + w = (x + \mathrm{i}\, y) + (u + \mathrm{i}\, v) = (x + u) + \mathrm{i}\,(y + v) \tag{13.24}$$

generally requires two real additions, while each complex multiplication

$$z\, w = (x + \mathrm{i}\, y)\,(u + \mathrm{i}\, v) = (x\, u - y\, v) + \mathrm{i}\,(x\, v + y\, u) \tag{13.25}$$

requires 4 real multiplications and 2 real additions, or, by employing the alternative formula

$$x\, v + y\, u = (x + y)\,(u + v) - x\, u - y\, v \tag{13.26}$$

for the imaginary part, 3 real multiplications and 5 real additions. (The choice of formula (13.25) or (13.26) will depend upon the processor's relative speeds of multiplication and addition.) Similarly, given the Fourier coefficients $c_0, \dots, c_{n-1}$, reconstruction of the sampled signal via (13.4) requires $n^2 - n$ complex multiplications and $n^2 - n$ complex additions. As a result, both computations become quite labor intensive for large n. Extending these ideas to multi-dimensional data only exacerbates the problem.

In order to explain the method without undue complication, we return to the original, aliased form of the discrete Fourier representation (13.4). (Once one understands how the FFT works, one can easily adapt the algorithm to the low frequency version (13.20).) The seminal observation is that if the number of sample points

$$n = 2\, m$$

† In fact, the key ideas can be found in Gauss' hand computations in the early 1800's, but his insight was not fully appreciated until modern computers arrived on the scene.

is even, then the primitive m^{th} root of unity $\zeta_m = \sqrt[m]{1}$ equals the square of the primitive n^{th} root:

$$\zeta_m = \zeta_n^2.$$

We use this fact to split the summation (13.15) for the order n discrete Fourier coefficients into two parts, collecting together the even and the odd powers of ζ_n^k:

$$\begin{aligned} c_k &= \frac{1}{n}\left(f_0 + f_1\,\zeta_n^{-k} + f_2\,\zeta_n^{-2\,k} + \,\cdots\, + f_{n-1}\,\zeta_n^{-(n-1)\,k}\right) \\ &= \frac{1}{n}\left(f_0 + f_2\,\zeta_n^{-2\,k} + f_4\,\zeta_n^{-4\,k} + \,\cdots\, + f_{2m-2}\,\zeta_n^{-(2\,m-2)\,k}\right) + \\ &\qquad + \zeta_n^{-k}\,\frac{1}{n}\left(f_1 + f_3\,\zeta_n^{-2\,k} + f_5\,\zeta_n^{-4\,k} + \,\cdots\, + f_{2m-1}\,\zeta_n^{-(2\,m-2)\,k}\right) \\ &= \frac{1}{2}\left\{ \frac{1}{m}\left(f_0 + f_2\,\zeta_m^{-k} + f_4\,\zeta_m^{-2\,k} + \,\cdots\, + f_{2m-2}\,\zeta_m^{-(m-1)\,k}\right)\right\} + \\ &\qquad + \frac{\zeta_n^{-k}}{2}\left\{ \frac{1}{m}\left(f_1 + f_3\,\zeta_m^{-k} + f_5\,\zeta_m^{-2\,k} + \,\cdots\, + f_{2m-1}\,\zeta_m^{-(m-1)\,k}\right)\right\}. \end{aligned} \tag{13.27}$$

Now, observe that the expressions in braces are the order m Fourier coefficients for the sample data

$$\begin{aligned} \mathbf{f}^e &= \big(f_0, f_2, f_4, \ldots, f_{2m-2}\big)^T = \big(f(x_0), f(x_2), f(x_4), \ldots, f(x_{2m-2})\big)^T, \\ \mathbf{f}^o &= \big(f_1, f_3, f_5, \ldots, f_{2m-1}\big)^T = \big(f(x_1), f(x_3), f(x_5), \ldots, f(x_{2m-1})\big)^T. \end{aligned} \tag{13.28}$$

Note that $\mathbf{f}^e$ is obtained by sampling $f(x)$ on the *even* sample points x_{2j}, while $\mathbf{f}^o$ is obtained by sampling the same function $f(x)$, but now at the *odd* sample points x_{2j+1}. In other words, we are splitting the original sampled signal into two "half-sampled" signals obtained by sampling on every other point. The even and odd Fourier coefficients are

$$\begin{aligned} c_k^e &= \frac{1}{m}\left(f_0 + f_2\,\zeta_m^{-k} + f_4\,\zeta_m^{-2\,k} + \,\cdots\, + f_{2m-2}\,\zeta_m^{-(m-1)\,k}\right), \\ c_k^o &= \frac{1}{m}\left(f_1 + f_3\,\zeta_m^{-k} + f_5\,\zeta_m^{-2\,k} + \,\cdots\, + f_{2m-1}\,\zeta_m^{-(m-1)\,k}\right), \end{aligned} \qquad k = 0, \ldots, m-1. \tag{13.29}$$

Since they contain just m data values, both the even and odd samples require only m distinct Fourier coefficients, and we adopt the identification

$$c_{k+m}^e = c_k^e, \qquad c_{k+m}^o = c_k^o, \qquad k = 0, \ldots, m-1. \tag{13.30}$$

Therefore, the order $n = 2\,m$ discrete Fourier coefficients (13.27) can be constructed from a pair of order m discrete Fourier coefficients via

$$c_k = \tfrac{1}{2}\big(c_k^e + \zeta_n^{-k}\,c_k^o\big), \qquad k = 0, \ldots, n-1. \tag{13.31}$$

Now if $m = 2\,l$ is also even, then we can play the same game on the order m Fourier coefficients (13.29), reconstructing each of them from a pair of order l discrete Fourier coefficients — obtained by sampling the signal at every fourth point. If $n = 2^r$ is a power of 2, then this game can be played all the way back to the start, beginning with the trivial order 1 discrete Fourier representation, which just samples the function at a single point.

The result is the desired algorithm. After some rearrangement of the basic steps, we arrive at the Fast Fourier Transform, which we now present in its final form.

We begin with a sampled signal on $n = 2^r$ sample points. To efficiently program the Fast Fourier Transform, it helps to write out each index $0 \le j < 2^r$ in its binary (as opposed to decimal) representation

$$j = j_{r-1}\, j_{r-2}\, \cdots\, j_2\, j_1\, j_0, \qquad \text{where} \qquad j_\nu = 0 \text{ or } 1; \tag{13.32}$$

the notation is shorthand for its r digit binary expansion

$$j = j_0 + 2\, j_1 + 4\, j_2 + 8\, j_3 + \, \cdots \, + 2^{r-1}\, j_{r-1}.$$

We then define the *bit reversal* map

$$\rho(j_{r-1}\, j_{r-2}\, \cdots\, j_2\, j_1\, j_0) = j_0\, j_1\, j_2\, \cdots\, j_{r-2}\, j_{r-1}. \tag{13.33}$$

For instance, if $r = 5$, and $j = 13$, with 5 digit binary representation 01101, then $\rho(j) = 22$ has the reversed binary representation 10110. Note especially that the bit reversal map $\rho = \rho_r$ depends upon the original choice of $r = \log_2 n$.

Secondly, for each $0 \le k < r$, define the maps

$$\begin{aligned} \alpha_k(j) &= j_{r-1}\, \cdots\, j_{k+1}\, 0\, j_{k-1}\, \cdots\, j_0, \\ \beta_k(j) &= j_{r-1}\, \cdots\, j_{k+1}\, 1\, j_{k-1}\, \cdots\, j_0 = \alpha_k(j) + 2^k, \end{aligned} \qquad \text{for} \qquad j = j_{r-1}\, j_{r-2}\, \cdots\, j_1\, j_0. \tag{13.34}$$

In other words, $\alpha_k(j)$ sets the k^{th} binary digit of j to 0, while $\beta_k(j)$ sets it to 1. In the preceding example, $\alpha_2(13) = 9$, with binary form 01001, while $\beta_2(13) = 13$ with binary form 01101. The bit operations (13.33, 34) are especially easy to implement on modern binary computers.

Given a sampled signal $f_0, \dots, f_{n-1}$, its discrete Fourier coefficients $c_0, \dots, c_{n-1}$ are computed by the following iterative algorithm:

$$c_j^{(0)} = f_{\rho(j)}, \qquad c_j^{(k+1)} = \tfrac{1}{2}\big(c_{\alpha_k(j)}^{(k)} + \zeta_{2^{k+1}}^{-j}\, c_{\beta_k(j)}^{(k)} \big), \qquad \begin{matrix} j = 0, \dots, n-1, \\ k = 0, \dots, r-1, \end{matrix} \tag{13.35}$$

in which $\zeta_{2^{k+1}}$ is the primitive 2^{k+1} root of unity. The final output of the iterative procedure, namely

$$c_j = c_j^{(r)}, \qquad j = 0, \dots, n-1, \tag{13.36}$$

are the discrete Fourier coefficients of our signal. The preprocessing step of the algorithm, where we define $c_j^{(0)}$, produces a more convenient rearrangement of the sample values. The subsequent steps successively combine the Fourier coefficients of the appropriate even and odd sampled subsignals together, reproducing (13.27) in a different notation. The following example should help make the overall process clearer.

Example 13.3. Consider the case $r = 3$, and so our signal has $n = 2^3 = 8$ sampled values $f_0, f_1, \dots, f_7$. We begin the process by rearranging the sample values

$$c_0^{(0)} = f_0, \;\; c_1^{(0)} = f_4, \;\; c_2^{(0)} = f_2, \;\; c_3^{(0)} = f_6, \;\; c_4^{(0)} = f_1, \;\; c_5^{(0)} = f_5, \;\; c_6^{(0)} = f_3, \;\; c_7^{(0)} = f_7,$$

in the order specified by the bit reversal map ρ. For instance $\rho(3) = 6$, or, in binary notation, $\rho(011) = 110$.

The first stage of the iteration is based on $\zeta_2 = -1$. Equation (13.35) gives

$$c_0^{(1)} = \tfrac{1}{2}(c_0^{(0)} + c_1^{(0)}), \quad c_1^{(1)} = \tfrac{1}{2}(c_0^{(0)} - c_1^{(0)}), \quad c_2^{(1)} = \tfrac{1}{2}(c_2^{(0)} + c_3^{(0)}), \quad c_3^{(1)} = \tfrac{1}{2}(c_2^{(0)} - c_3^{(0)}),$$
$$c_4^{(1)} = \tfrac{1}{2}(c_4^{(0)} + c_5^{(0)}), \quad c_5^{(1)} = \tfrac{1}{2}(c_4^{(0)} - c_5^{(0)}), \quad c_6^{(1)} = \tfrac{1}{2}(c_6^{(0)} + c_7^{(0)}), \quad c_7^{(1)} = \tfrac{1}{2}(c_6^{(0)} - c_7^{(0)}),$$

where we combine successive pairs of the rearranged sample values. The second stage of the iteration has $k = 1$ with $\zeta_4 = \mathrm{i}$. We find

$$c_0^{(2)} = \tfrac{1}{2}(c_0^{(1)} + c_2^{(1)}), \quad c_1^{(2)} = \tfrac{1}{2}(c_1^{(1)} - \mathrm{i}\, c_3^{(1)}), \quad c_2^{(2)} = \tfrac{1}{2}(c_0^{(1)} - c_2^{(1)}), \quad c_3^{(2)} = \tfrac{1}{2}(c_1^{(1)} + \mathrm{i}\, c_3^{(1)}),$$
$$c_4^{(2)} = \tfrac{1}{2}(c_4^{(1)} + c_6^{(1)}), \quad c_5^{(2)} = \tfrac{1}{2}(c_5^{(1)} - \mathrm{i}\, c_7^{(1)}), \quad c_6^{(2)} = \tfrac{1}{2}(c_4^{(1)} - c_6^{(1)}), \quad c_7^{(2)} = \tfrac{1}{2}(c_5^{(1)} + \mathrm{i}\, c_7^{(1)}).$$

Note that the indices of the combined pairs of coefficients differ by 2. In the last step, where $k = 2$ and $\zeta_8 = \frac{\sqrt{2}}{2}(1 + \mathrm{i})$, we combine coefficients whose indices differ by $4 = 2^2$; the final output

$$\begin{aligned}
c_0 &= c_0^{(3)} = \tfrac{1}{2}(c_0^{(2)} + c_4^{(2)}), & c_4 &= c_4^{(3)} = \tfrac{1}{2}(c_0^{(2)} - c_4^{(2)}),\\
c_1 &= c_1^{(3)} = \tfrac{1}{2}\big(c_1^{(2)} + \tfrac{\sqrt{2}}{2}(1 - \mathrm{i})\, c_5^{(2)}\big), & c_5 &= c_5^{(3)} = \tfrac{1}{2}\big(c_1^{(2)} - \tfrac{\sqrt{2}}{2}(1 - \mathrm{i})\, c_5^{(2)}\big),\\
c_2 &= c_2^{(3)} = \tfrac{1}{2}\big(c_2^{(2)} - \mathrm{i}\, c_6^{(2)}\big), & c_6 &= c_6^{(3)} = \tfrac{1}{2}\big(c_2^{(2)} + \mathrm{i}\, c_6^{(2)}\big),\\
c_3 &= c_3^{(3)} = \tfrac{1}{2}\big(c_3^{(2)} - \tfrac{\sqrt{2}}{2}(1 + \mathrm{i})\, c_7^{(2)}\big), & c_7 &= c_7^{(3)} = \tfrac{1}{2}\big(c_3^{(2)} + \tfrac{\sqrt{2}}{2}(1 + \mathrm{i})\, c_7^{(2)}\big),
\end{aligned}$$

is the complete set of discrete Fourier coefficients.

Let us count the number of arithmetic operations required in the Fast Fourier Transform algorithm. At each stage in the computation, we must perform $n = 2^r$ complex additions/subtractions and the same number of complex multiplications. (Actually, the number of multiplications is slightly smaller since multiplications by ± 1 and $\pm \mathrm{i}$ are extremely simple. However, this does not significantly alter the final operations count.) There are $r = \log_2 n$ stages, and so we require a total of $r\, n = n \log_2 n$ complex additions/subtractions and the same number of multiplications. Now, when n is large, $n \log_2 n$ is *significantly* smaller than n^2, which is the number of operations required for the direct algorithm. For instance, if $n = 2^{10} = 1,024$, then $n^2 = 1,048,576$, while $n \log_2 n = 10,240$ — a net savings of 99%. As a result, many large scale computations that would be intractable using the direct approach are immediately brought into the realm of feasibility. This is the reason why all modern implementations of the Discrete Fourier Transform are based on the FFT algorithm and its variants.

The reconstruction of the signal from the discrete Fourier coefficients $c_0, \dots, c_{n-1}$ is speeded up in exactly the same manner. The only differences are that we replace $\zeta_n^{-1} = \overline{\zeta_n}$ by ζ_n, and drop the factors of $\frac{1}{2}$ since there is no need to divide by n in the final result (13.4). Therefore, we apply the slightly modified iterative procedure

$$f_j^{(0)} = c_{\rho(j)}, \qquad f_j^{(k+1)} = f_{\alpha_k(j)}^{(k)} + \zeta_{2^{k+1}}^{j}\, f_{\beta_k(j)}^{(k)}, \qquad \begin{matrix} j = 0, \dots, n-1, \\ k = 0, \dots, r-1, \end{matrix} \tag{13.37}$$

and finish with

$$f(x_j) = f_j = f_j^{(r)}, \qquad j = 0, \dots, n-1. \tag{13.38}$$

Example 13.4. The reconstruction formulae in the case of $n = 8 = 2^3$ Fourier coefficients $c_0, \dots, c_7$, which were computed in Example 13.3, can be implemented as follows. First, we rearrange the Fourier coefficients in bit reversed order:

$$f_0^{(0)} = c_0, \;\; f_1^{(0)} = c_4, \;\; f_2^{(0)} = c_2, \;\; f_3^{(0)} = c_6, \;\; f_4^{(0)} = c_1, \;\; f_5^{(0)} = c_5, \;\; f_6^{(0)} = c_3, \;\; f_7^{(0)} = c_7,$$

Then we begin combining them in successive pairs:

$$\begin{aligned} f_0^{(1)} &= f_0^{(0)} + f_1^{(0)}, \quad f_1^{(1)} = f_0^{(0)} - f_1^{(0)}, \quad f_2^{(1)} = f_2^{(0)} + f_3^{(0)}, \quad f_3^{(1)} = f_2^{(0)} - f_3^{(0)}, \\ f_4^{(1)} &= f_4^{(0)} + f_5^{(0)}, \quad f_5^{(1)} = f_4^{(0)} - f_5^{(0)}, \quad f_6^{(1)} = f_6^{(0)} + f_7^{(0)}, \quad f_7^{(1)} = f_6^{(0)} - f_7^{(0)}. \end{aligned}$$

Next,

$$\begin{aligned} f_0^{(2)} &= f_0^{(1)} + f_2^{(1)}, \quad f_1^{(2)} = f_1^{(1)} + \mathrm{i}\, f_3^{(1)}, \quad f_2^{(2)} = f_0^{(1)} - f_2^{(1)}, \quad f_3^{(2)} = f_1^{(1)} - \mathrm{i}\, f_3^{(1)}, \\ f_4^{(2)} &= f_4^{(1)} + f_6^{(1)}, \quad f_5^{(2)} = f_5^{(1)} + \mathrm{i}\, f_7^{(1)}, \quad f_6^{(2)} = f_4^{(1)} - f_6^{(1)}, \quad f_7^{(2)} = f_5^{(1)} - \mathrm{i}\, f_7^{(1)}. \end{aligned}$$

Finally, the sampled signal values are

$$\begin{aligned} f(x_0) &= f_0^{(3)} = f_0^{(2)} + f_4^{(2)}, & f(x_4) &= f_4^{(3)} = f_0^{(2)} - f_4^{(2)}, \\ f(x_1) &= f_1^{(3)} = f_1^{(2)} + \tfrac{\sqrt{2}}{2}\,(1 + \mathrm{i}\,)\, f_5^{(2)}, & f(x_5) &= f_5^{(3)} = f_1^{(2)} - \tfrac{\sqrt{2}}{2}\,(1 + \mathrm{i}\,)\, f_5^{(2)}, \\ f(x_2) &= f_2^{(3)} = f_2^{(2)} + \mathrm{i}\, f_6^{(2)}, & f(x_6) &= f_6^{(3)} = f_2^{(2)} - \mathrm{i}\, f_6^{(2)}, \\ f(x_3) &= f_3^{(3)} = f_3^{(2)} - \tfrac{\sqrt{2}}{2}\,(1 - \mathrm{i}\,)\, f_7^{(2)}, & f(x_7) &= f_7^{(3)} = f_3^{(2)} + \tfrac{\sqrt{2}}{2}\,(1 - \mathrm{i}\,)\, f_7^{(2)}. \end{aligned}$$

13.2. Wavelets.

Trigonometric Fourier series, both continuous and discrete, are amazingly powerful, but they do suffer from one potentially serious defect. The basis functions $e^{\,\mathrm{i}\,k\,x} = \cos k\,x + \mathrm{i}\,\sin k\,x$ are spread out over the entire interval $[-\pi, \pi]$, and so are not well-suited to processing localized signals — meaning data that are concentrated in a relatively small regions. Indeed, the most concentrated data of all — a single delta function — has every Fourier component of equal magnitude in its Fourier series (12.61) and its high degree of localization is completely obscured. Ideally, one would like to construct a system of functions that is orthogonal, and so has all the advantages of the Fourier trigonometric functions, but, in addition, adapts to localized structures in signals. This dream was the inspiration for the development of the modern theory of wavelets.

The Haar Wavelets

Although the modern era of wavelets started in the mid 1980's, the simplest example of a wavelet basis was discovered by the Hungarian mathematician Alfréd Haar in 1910, [**86**]. We consider the space of functions (signals) defined the interval $[0, 1]$, equipped with the standard L^2 inner product

$$\langle\, f\,, g\,\rangle = \int_0^1 f(x)\, g(x)\, dx. \tag{13.39}$$

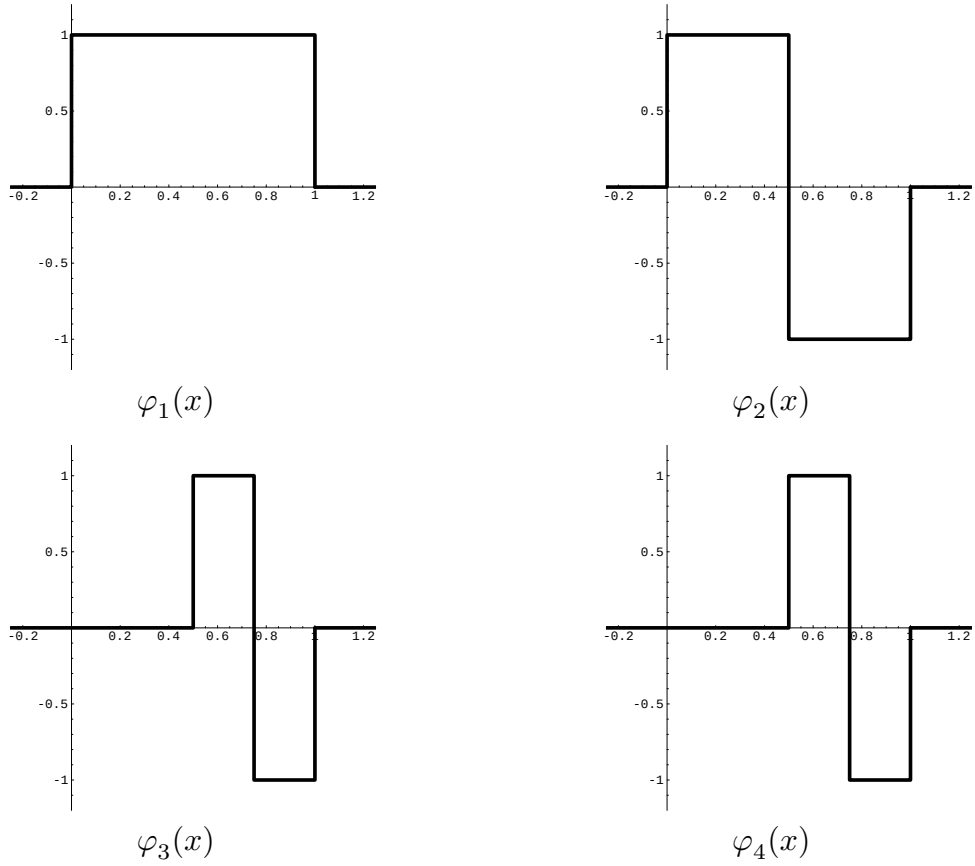


Figure 13.7. The First Four Haar Wavelets.

This choice is merely for convenience, being slightly better suited to our construction than $[-\pi, \pi]$ or $[0, 2\pi]$. Moreover, the usual scaling arguments can be used to adapt the wavelet formulas to any other interval.

The *Haar wavelets* are certain piecewise constant functions. The first four are graphed in Figure 13.7 The first is the *box function*

$$\varphi_1(x) = \varphi(x) = \begin{cases} 1, & 0 < x \leq 1, \\ 0, & \text{otherwise}, \end{cases} \tag{13.40}$$

known as the *scaling function*, for reasons that shall appear shortly. Although we are only interested in the value of $\varphi(x)$ on the interval $[0, 1]$, it will be convenient to extend it, and all the other wavelets, to be zero outside the basic interval. Its values at the points of discontinuity, i.e., $0, 1$, is not critical, but, unlike the Fourier series midpoint value, it will be more convenient to consistently choose the left hand limiting value. The second Haar function

$$\varphi_2(x) = w(x) = \begin{cases} 1, & 0 < x \leq \frac{1}{2}, \\ -1, & \frac{1}{2} < x \leq 1, \\ 0, & \text{otherwise}, \end{cases} \tag{13.41}$$

is known as the *mother wavelet*. The third and fourth Haar functions are compressed

versions of the mother wavelet:

$$\varphi_3(x) = w(2x) = \begin{cases} 1, & 0 < x \le \frac{1}{4}, \\ -1, & \frac{1}{4} < x \le \frac{1}{2}, \\ 0, & \text{otherwise}, \end{cases} \qquad \varphi_4(x) = w(2x-1) = \begin{cases} 1, & \frac{1}{2} < x \le \frac{3}{4}, \\ -1, & \frac{3}{4} < x \le 1, \\ 0, & \text{otherwise}, \end{cases}$$

called *daughter wavelets*. One can easily check, by direct evaluation of the integrals, that the four Haar wavelet functions are orthogonal with respect to the L^2 inner product (13.39).

The scaling transformation $x \mapsto 2x$ serves to compress the wavelet function, while the translation $2x \mapsto 2x - 1$ moves the compressed version to the right by a half a unit. Furthermore, we can represent the mother wavelet by compressing and translating the scaling function,:

$$w(x) = \varphi(2x) - \varphi(2x-1). \tag{13.42}$$

It is these two operations of scaling and compression — coupled with the all-important orthogonality — that underlies the power of wavelets.

The Haar wavelets have an evident discretization. If we decompose the interval $(0, 1]$ into the four subintervals

$$\left(0\,,\tfrac{1}{4}\right], \qquad \left(\tfrac{1}{4}\,,\tfrac{1}{2}\right], \qquad \left(\tfrac{1}{2}\,,\tfrac{3}{4}\right], \qquad \left(\tfrac{3}{4}\,,1\right], \tag{13.43}$$

on which the four wavelet functions are constant, then we can represent each of them by a vector in $\mathbb{R}^4$ whose entries are the values of each wavelet function sampled at the left endpoint of each subinterval. In this manner, we obtain the wavelet sample vectors

$$\mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \mathbf{v}_2 = \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}, \qquad \mathbf{v}_3 = \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix}, \qquad \mathbf{v}_4 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \end{pmatrix}, \tag{13.44}$$

that form the orthogonal wavelet basis of $\mathbb{R}^4$ we encountered in Examples 2.35 and 5.10. Orthogonality of the vectors (13.44) with respect to the standard Euclidean dot product is equivalent to orthogonality of the Haar wavelet functions with respect to the inner product (13.39). Indeed, if

$$f(x) \;\sim\; \mathbf{f} = (f_1, f_2, f_3, f_4) \qquad \text{and} \qquad g(x) \;\sim\; \mathbf{g} = (g_1, g_2, g_3, g_4)$$

are *piecewise constant* real functions that achieve the indicated values on the four subintervals (13.43), then their L^2 inner product

$$\langle\, f\,, g\,\rangle = \int_0^1 f(x)\, g(x)\, dx = \tfrac{1}{4}\left(\, f_1\, g_1 + f_2\, g_2 + f_3\, g_3 + f_4\, g_4\,\right) = \tfrac{1}{4}\, \mathbf{f} \cdot \mathbf{g},$$

is equal to the averaged dot product of their sample values — the real form of the inner product (13.8) that was used in the discrete Fourier transform.

Since the vectors (13.44) form an orthogonal basis of $\mathbb{R}^4$, we can uniquely decompose any such piecewise constant function as a linear combination of wavelets

$$f(x) = c_1\, \varphi_1(x) + c_2\, \varphi_2(x) + c_3\, \varphi_3(x) + c_4\, \varphi_4(x),$$

or, equivalently, in terms of the sample vectors,

$$\mathbf{f} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 + c_4\mathbf{v}_4.$$

The required coefficients

$$c_k = \frac{\langle\, f\,,\varphi_k\,\rangle}{\|\,\varphi_k\,\|^2} = \frac{\mathbf{f}\cdot\mathbf{v}_k}{\|\,\mathbf{v}_k\,\|^2}$$

are fixed by our usual orthogonality formula (5.7). Explicitly,

$$\begin{aligned} c_1 &= \tfrac{1}{4}\,(f_1 + f_2 + f_3 + f_4), & c_3 &= \tfrac{1}{2}\,(f_1 - f_2),\\ c_2 &= \tfrac{1}{4}\,(f_1 + f_2 - f_3 - f_4), & c_4 &= \tfrac{1}{2}\,(f_3 - f_4). \end{aligned}$$

Before proceeding to the more general case, let us introduce an important analytical definition that quantifies precisely how localized a function is.

Definition 13.5. The *support* of a function $f(x)$, written $\operatorname{supp} f$, is the closure of the set where $f(x) \neq 0$.

Thus, a point will belong to the support of $f(x)$, provided f is not zero there, or at least is not zero at nearby points. More precisely:

Lemma 13.6. *If $f(a) \neq 0$, then $a \in \operatorname{supp} f$. More generally, a point $a \in \operatorname{supp} f$ if and only if there exist a convergent sequence $x_n \to a$ such that $f(x_n) \neq 0$. Conversely, $a \notin \operatorname{supp} f$ if and only if $f(x) \equiv 0$ on an interval $a - \delta < x < a + \delta$ for some $\delta > 0$.*

Intuitively, the smaller the support of a function, the more localized it is. For example, the support of the Haar mother wavelet (13.41) is $\operatorname{supp} w = [0, 1]$ — the point $x = 0$ is included, even though $w(0) = 0$, because $w(x) \neq 0$ at nearby points. The two daughter wavelets have smaller support:

$$\operatorname{supp}\varphi_3 = \left[0\,,\tfrac{1}{2}\right], \qquad \operatorname{supp}\varphi_4 = \left[\tfrac{1}{2}\,,1\right],$$

and so are twice as localized. An extreme case is the delta function, whose support is a single point. In contrast, the support of the Fourier trigonometric basis functions is all of $\mathbb{R}$, since they only vanish at isolated points.

The effect of scalings and translations on the support of a function is easily discerned.

Lemma 13.7. *If* $\operatorname{supp} f = [a, b]$, *and*

$$g(x) = f(r\,x - \delta), \qquad \textit{then} \qquad \operatorname{supp} g = \left[\,\frac{a+\delta}{r}\,,\frac{b+\delta}{r}\,\right].$$

In other words, scaling x by a factor r compresses the support of the function by a factor $1/r$, while translating x translates the support of the function.

The key requirement for a wavelet basis is that it contains functions with arbitrarily small support. To this end, the full Haar wavelet basis is obtained from the mother wavelet by iterating the scaling and translation processes. We begin with the scaling function

$$\varphi(x), \tag{13.45}$$

from which we construct the mother wavelet via (13.42). For any "generation" $j \geq 0$, we form the wavelet offspring by first compressing the mother wavelet so that its support fits into an interval of length 2^{-j},

$$w_{j,0}(x) = w(2^j x), \qquad \text{so that} \qquad \operatorname{supp} w_{j,0} = [0, 2^{-j}\,], \tag{13.46}$$

and then translating $w_{j,0}$ so as to fill up the entire interval $[0,1]$ by 2^j subintervals, each of length 2^{-j}, defining

$$w_{j,k}(x) = w_{j,0}(x-k) = w(2^j x - k), \qquad \text{where} \qquad k = 0, 1, \dots 2^j - 1. \tag{13.47}$$

Lemma 13.7 implies that $\operatorname{supp} w_{j,k} = [\, 2^{-j}\, k, 2^{-j}\,(k+1)\,]$, and so the combined supports of all the j^{th} generation of wavelets is the entire interval: $\bigcup_{k=0}^{2^j-1} \operatorname{supp} w_{j,k} = [0,1]$. The primal generation, $j = 0$, just consists of the mother wavelet

$$w_{0,0}(x) = w(x).$$

The first generation, $j = 1$, consists of the two daughter wavelets already introduced as φ_3 and φ_4, namely

$$w_{1,0}(x) = w(2x), \qquad\qquad w_{1,1}(x) = w(2x-1).$$

The second generation, $j = 2$, appends four additional granddaughter wavelets to our basis:

$$w_{2,0}(x) = w(4x), \quad w_{2,1}(x) = w(4x-1), \quad w_{2,2}(x) = w(4x-2), \quad w_{2,3}(x) = w(4x-3).$$

The 8 Haar wavelets $\varphi, w_{0,0}, w_{1,0}, w_{1,1}, w_{2,0}, w_{2,1}, w_{2,2}, w_{2,3}$ are constant on the 8 subintervals of length $\frac{1}{8}$, taking the successive sample values indicated by the columns of the matrix

$$W_8 = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & -1 & 0 & 0 & 0 \\ 1 & 1 & -1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & -1 & 0 & 0 & -1 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 & -1 & 0 \\ 1 & -1 & 0 & -1 & 0 & 0 & 0 & 1 \\ 1 & -1 & 0 & -1 & 0 & 0 & 0 & -1 \end{pmatrix}. \tag{13.48}$$

Orthogonality of the wavelets is manifested in the orthogonality of the columns of W_8. (Unfortunately, the usual terminological constraints of Definition 5.18 prevent us from calling W_8 an orthogonal matrix because its columns are not orthonormal!)

The n^{th} stage consists of 2^{n+1} different wavelet functions comprising the scaling functions and all the generations up to the n^{th} : $w_0(x) = \varphi(x)$ and $w_{j,k}(x)$ for $0 \leq j \leq n$ and $0 \leq k < 2^j$. They are all constant on each subinterval of length 2^{-n-1}.

Theorem 13.8. *The wavelet functions $\varphi(x)$, $w_{j,k}(x)$ form an orthogonal system with respect to the inner product* (13.39).

Proof: First, note that each wavelet $w_{j,k}(x)$ is equal to $+1$ on an interval of length 2^{-j-1} and to -1 on an adjacent interval of the same length. Therefore,

$$\langle\, w_{j,k}\,,\varphi\,\rangle = \int_0^1 w_{j,k}(x)\,dx = 0, \tag{13.49}$$

since the $+1$ and -1 contributions cancel each other. If two different wavelets $w_{j,k}$ and $w_{l,m}$ with, say $j \le l$, have supports which are either disjoint, or just overlap at a single point, then their product $w_{j,k}(x)\,w_{l,m}(x) \equiv 0$, and so their inner product is clearly zero:

$$\langle\, w_{j,k}\,, w_{l,m}\,\rangle = \int_0^1 w_{j,k}(x)\,w_{l,m}(x)\,dx = 0.$$

Otherwise, except in the case when the two wavelets are identical, the support of $w_{l,m}$ is entirely contained in an interval where $w_{j,k}$ is constant and so $w_{j,k}(x)\,w_{l,m}(x) = \pm\, w_{l,m}(x)$. Therefore, by (13.49),

$$\langle\, w_{j,k}\,, w_{l,m}\,\rangle = \int_0^1 w_{j,k}(x)\,w_{l,m}(x)\,dx = \pm \int_0^1 w_{l,m}(x)\,dx = 0.$$

Finally, we compute

$$\|\,\varphi\,\|^2 = \int_0^1 dx = 1, \qquad\qquad \|\,w_{j,k}\,\|^2 = \int_0^1 w_{j,k}(x)^2\,dx = 2^{-j}. \tag{13.50}$$

The second formula follows from the fact that $|\,w_{j,k}(x)\,| = 1$ on an interval of length 2^{-j} and is 0 elsewhere. *Q.E.D.*

In direct analogy with the trigonometric Fourier series, the *wavelet series* of a signal $f(x)$ is given by

$$f(x) \;\sim\; c_0\,\varphi(x) \;+\; \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} c_{j,k}\,w_{j,k}(x). \tag{13.51}$$

Orthogonality implies that the wavelet coefficients $c_0, c_{j,k}$ can be immediately computed using the standard inner product formula coupled with (13.50):

$$\begin{aligned} c_0 &= \frac{\langle\, f\,,\varphi\,\rangle}{\|\,\varphi\,\|^2} = \int_0^1 f(x)\,dx, \\ c_{j,k} &= \frac{\langle\, f\,, w_{j,k}\,\rangle}{\|\,w_{j,k}\,\|^2} = 2^j \int_{2^{-j}k}^{2^{-j}k+2^{-j-1}} f(x)\,dx \;-\; 2^j \int_{2^{-j}k+2^{-j-1}}^{2^{-j}(k+1)} f(x)\,dx. \end{aligned} \tag{13.52}$$

The convergence properties of the wavelet series (13.51) are similar to those of Fourier series; details can be found [**50**].

Example 13.9. In Figure 13.8, we plot the Haar expansions of the signal in the first plot. The next plots show the partial sums over $j = 0, \dots, r$ with $r = 2, 3, 4, 5, 6$. We have used a discontinuous signal to demonstrate that there is no nonuniform Gibbs phenomenon in a Haar wavelet expansion. Indeed, since the wavelets are themselves discontinuous, they do not have any difficult uniformly converging to a discontinuous function. On the other hand, it takes quite a few wavelets to begin to accurately reproduce the signal. In the last plot, we combine a total of $2^6 = 64$ Haar wavelets, which is considerably more than would be required in a comparably accurate Fourier expansion (excluding points very close to the discontinuity).

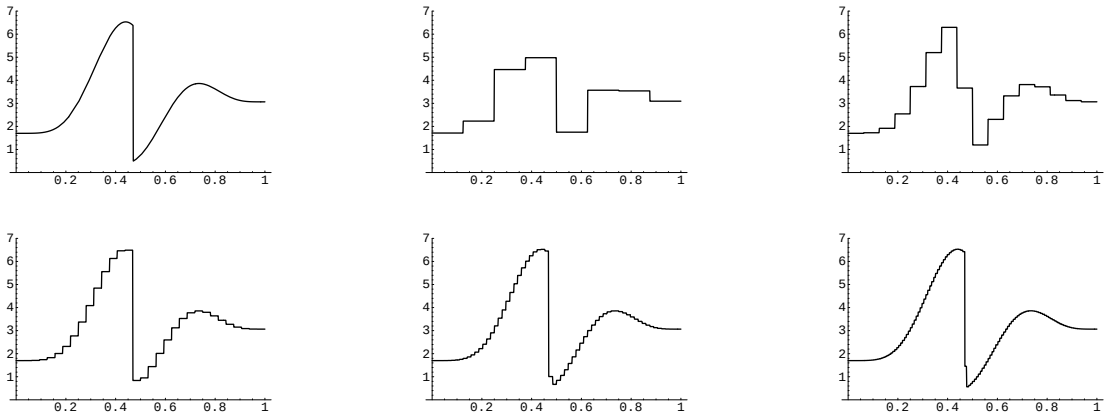


Figure 13.8. Haar Wavelet Expansion.

Remark: To the novice, there may appear to be many more wavelets than trigonometric functions. But this is just another illusion of the magic show of infinite dimensional space. The point is that they both form a countably infinite set of functions, and so could, if necessary, but less conveniently, be numbered in order $1, 2, 3, \dots$. On the other hand, accurate reproduction of functions usually does require many more Haar wavelets. This handicap makes the Haar system of less use in practical situations, and served to motivate the search for a more sophisticated choice of wavelet basis.

Just as the discrete Fourier representation arises from a sampled version of the full Fourier series, so there is a discrete wavelet transformation for suitably sampled signals. To take full advantage of the wavelet basis, we sample the signal $f(x)$ at $n = 2^r$ equally spaced sample points $x_k = k/2^r$, for $k = 0, \dots, n-1$, on the interval $[0, 1]$. As before, we can identify the sampled signal with the vector

$$\mathbf{f} = \big(f(x_0), f(x_1), \dots, f(x_{n-1}) \big)^T = \big(f_0, f_1, \dots, f_{n-1} \big)^T \in \mathbb{R}^n. \tag{13.53}$$

Since we are only sampling on intervals of length 2^{-r}, the discrete wavelet transform of our sampled signal will only use the first $n = 2^r$ wavelets $\varphi(x)$ and $w_{j,k}(x)$ for $j = 0, \dots, r-1$ and $k = 0, \dots, 2^j - 1$. Let $\mathbf{w}_0 \sim \varphi(x)$ and $\mathbf{w}_{j,k} \sim w_{j,k}(x)$ denote the corresponding sampled wavelet vectors; all of their entries are either $+1$, -1, or 0. (When $r = 3$, so $n = 8$, these are the columns of the wavelet matrix (13.48).) Moreover, orthogonality of the wavelets immediately implies that the wavelet vectors form an orthogonal basis of $\mathbb{R}^n$ with $n = 2^r$. We can decompose our sample vector (13.53) as a linear combination of the sampled wavelets,

$$\mathbf{f} = \widehat{c}_0 \, \mathbf{w}_0 \; + \sum_{j=0}^{r-1} \sum_{k=0}^{2^j - 1} \widehat{c}_{j,k} \, \mathbf{w}_{j,k}, \tag{13.54}$$

where, by our usual orthogonality formulae,

$$\begin{aligned} \widehat{c}_0 &= \frac{\langle\, f\,,\mathbf{w}_0\,\rangle}{\|\,\mathbf{w}_0\,\|^2} = \frac{1}{2^r}\,\sum_{i\,=\,0}^{2^r-1}\,f_i, \\ \widehat{c}_{j,k} &= \frac{\langle\, f\,,\mathbf{w}_{j,k}\,\rangle}{\|\,\mathbf{w}_{j,k}\,\|^2} = 2^{j-r}\,\left(\sum_{i\,=\,k}^{k+2^{r-j-1}-1}\,f_i\;-\;\sum_{i\,=\,k+2^{r-j-1}}^{k+2^{r-j}-1}\,f_i\right). \end{aligned} \tag{13.55}$$

These are the basic formulae connecting the functions $f(x)$, or, rather, its sample vector $\mathbf{f}$, and its *discrete wavelet transform* consisting of the 2^r coefficients $\widehat{c}_0, \widehat{c}_{j,k}$. The reconstructed function

$$\widetilde{f}(x) = \widehat{c}_0\,\varphi(x) + \sum_{j\,=\,0}^{r-1}\,\sum_{k\,=\,0}^{2^j-1}\,\widehat{c}_{j,k}\,w_{j,k}(x) \tag{13.56}$$

is constant on each subinterval of length 2^{-r}, and has the same value

$$\widetilde{f}(x) = \widetilde{f}(x_i) = f(x_i) = f_i, \qquad\qquad x_i = 2^{-r}\,i \le x < x_{i+1} = 2^{-r}\,(i+1),$$

as our signal at the left hand endpoint of the interval. In other words, we are interpolating the sample points by a piecewise *constant* (and thus discontinuous) function.

Modern Wavelets

The main defect of the Haar wavelets is that they do not provide a very efficient means of representing even very simple functions — it takes quite a large number of wavelets to reproduce signals with any degree of precision. The reason for this is that the Haar wavelets are piecewise constant, and so even an affine function $y = \alpha\,x + \beta$ requires many sample values, and hence a relatively extensive collection of Haar wavelets, to be accurately reproduced. In particular, compression and denoising algorithms based on Haar wavelets are either insufficiently precise or hopelessly inefficient, and hence of minor practical value.

For a long time it was thought that it was impossible to simultaneously achieve the requirements of localization, orthogonality and accurate reproduction of simple functions. The breakthrough came in 1988, when, in her Ph.D. thesis, the Dutch mathematician Ingrid Daubechies produced the first examples of wavelet bases that realized all three basic criteria. Since then, wavelets have developed into a sophisticated and burgeoning industry with major impact on modern technology. Significant applications include compression, storage and recognition of fingerprints in the FBI's data base, and the JPEG2000 image format, which, unlike earlier Fourier-based JPEG standards, incorporates wavelet technology in its image compression and reconstruction algorithms. In this section, we will present a brief outline of the basic ideas underlying Daubechies' remarkable construction.

The recipe for any wavelet system involves two basic ingredients — a scaling function and a mother wavelet. The latter can be constructed from the scaling function by a prescription similar to that in (13.42), and therefore we first concentrate on the properties of the scaling function. The key requirement is that the scaling function must solve a

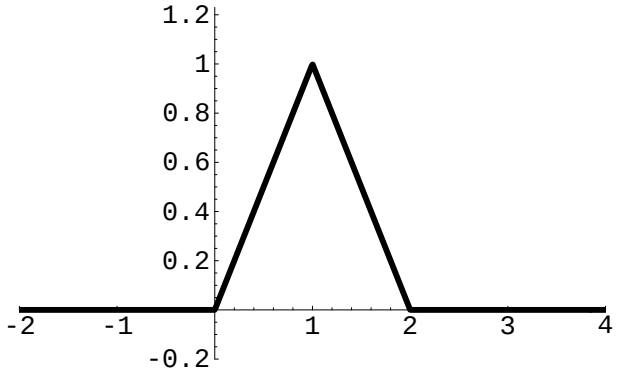


Figure 13.9. The Hat Function.

dilation equation of the form

$$\varphi(x) = \sum_{k=0}^{p} c_k \, \varphi(2\,x - k) = c_0 \, \varphi(2\,x) + c_1 \, \varphi(2\,x - 1) + \, \cdots \, + c_p \, \varphi(2\,x - p) \tag{13.57}$$

for some collection of constants $c_0, \ldots, c_p$. The dilation equation relates the function $\varphi(x)$ to a finite linear combination of its compressed translates. The coefficients $c_0, \ldots, c_p$ are not arbitrary, since the properties of orthogonality and localization will impose certain rather stringent requirements.

Example 13.10. The Haar or box scaling function (13.40) satisfies the dilation equation (13.57) with $c_0 = c_1 = 1$, namely

$$\varphi(x) = \varphi(2\,x) + \varphi(2\,x - 1). \tag{13.58}$$

We recommend that you convince yourself of the validity of this identity before continuing.

Example 13.11. Another example of a scaling function is the *hat function*

$$\varphi(x) = \begin{cases} x, & 0 \le x \le 1, \\ 2 - x, & 1 \le x \le 2, \\ 0, & \text{otherwise}, \end{cases} \tag{13.59}$$

graphed in Figure 13.9, whose variants play a starring role in the finite element method, cf. (11.167). The hat function satisfies the dilation equation

$$\varphi(x) = \tfrac{1}{2} \, \varphi(2\,x) + \varphi(2\,x - 1) + \tfrac{1}{2} \, \varphi(2\,x - 2), \tag{13.60}$$

which is (13.57) with $c_0 = \frac{1}{2}, c_1 = 1, c_2 = \frac{1}{2}$. Again, the reader should be able to check this identity by hand.

The dilation equation (13.57) is a kind of *functional equation*, and, as such, is not so easy to solve. Indeed, the mathematics of functional equations remains much less well developed than that of differential equations or integral equations. Even to prove that (nonzero) solutions exist is a nontrivial analytical problem. Since we already know two

explicit examples, let us defer the discussion of solution techniques until we understand how the dilation equation can be used to construct a wavelet basis.

Given a solution to the dilation equation, we define the *mother wavelet* to be

$$\begin{aligned} w(x) &= \sum_{k=0}^{p} (-1)^k c_{p-k}\, \varphi(2\,x - k) \\ &= c_p\, \varphi(2\,x) - c_{p-1}\, \varphi(2\,x - 1) + c_{p-2}\, \varphi(2\,x - 2) + \,\cdots\, \pm\, c_0\, \varphi(2\,x - p), \end{aligned} \tag{13.61}$$

This formula directly generalizes the Haar wavelet relation (13.42), in light of its dilation equation (13.58). The daughter wavelets are then all found, as in the Haar basis, by iteratively compressing and translating the mother wavelet:

$$w_{j,k}(x) = w(2^j\, x - k). \tag{13.62}$$

In the general framework, we do not necessarily restrict our attention to the interval $[\,0,1\,]$ and so j and k can, in principle, be arbitrary integers.

Let us investigate what sort of conditions should be imposed on the dilation coefficients $c_0, \dots, c_p$ in order that we obtain a viable wavelet basis by this construction. First, localization of the wavelets requires that the scaling function has bounded support, and so $\varphi(x) \equiv 0$ when x lies outside some bounded interval $[\,a,b\,]$. If we integrate both sides of (13.57), we find

$$\int_a^b \varphi(x)\, dx = \int_{-\infty}^{\infty} \varphi(x)\, dx = \sum_{k=0}^{p} c_k \int_{-\infty}^{\infty} \varphi(2\,x - k)\, dx. \tag{13.63}$$

Now using the change of variables $y = 2\,x - k$, with $dx = \frac{1}{2}\, dy$, we find

$$\int_{-\infty}^{\infty} \varphi(2\,x - k)\, dx = \frac{1}{2} \int_{-\infty}^{\infty} \varphi(y)\, dy = \frac{1}{2} \int_a^b \varphi(x)\, dx, \tag{13.64}$$

where we revert to x as our (dummy) integration variable. We substitute this result back into (13.63). Assuming that $\displaystyle\int_a^b \varphi(x)\, dx \neq 0$, we discover that the dilation coefficients must satisfy

$$c_0 + \,\cdots\, + c_p = 2. \tag{13.65}$$

Example 13.12. Once we impose the constraint (13.65), the very simplest version of the dilation equation is

$$\varphi(x) = 2\, \varphi(2\,x) \tag{13.66}$$

where $c_0 = 2$ is the only (nonzero) coefficient. Up to constant multiple, the only "solutions" of the functional equation (13.66) with bounded support are scalar multiples of the delta function $\delta(x)$, which follows from the identity in Exercise 11.2.5. Other solutions, such as $\varphi(x) = 1/x$, are not localized, and thus not useful for constructing a wavelet basis.

The second condition we require is orthogonality of the wavelets. For simplicity, we only consider the standard L^2 inner product[†]

$$\langle\, f\,, g\,\rangle = \int_{-\infty}^{\infty} f(x)\, g(x)\, dx.$$

It turns out that the orthogonality of the complete wavelet system is guaranteed once we know that the scaling function $\varphi(x)$ is orthogonal to all its integer translates:

$$\langle\, \varphi(x)\,, \varphi(x-m)\,\rangle = \int_{-\infty}^{\infty} \varphi(x)\, \varphi(x-m)\, dx = 0 \qquad \text{for all} \qquad m \neq 0. \tag{13.67}$$

We first note the formula

$$\begin{aligned} \langle\, \varphi(2\,x-k)\,, \varphi(2\,x-l)\,\rangle &= \int_{-\infty}^{\infty} \varphi(2\,x-k)\, \varphi(2\,x-l)\, dx \\ &= \frac{1}{2}\int_{-\infty}^{\infty} \varphi(x)\, \varphi(x+k-l)\, dx = \frac{1}{2}\,\langle\, \varphi(x)\,, \varphi(x+k-l)\,\rangle \end{aligned} \tag{13.68}$$

follows from the same change of variables $y = 2\,x - k$ used in (13.64). Therefore, since φ satisfies the dilation equation (13.57),

$$\langle\, \varphi(x)\,, \varphi(x-m)\,\rangle = \left\langle \sum_{j=0}^{p} c_j\, \varphi(2\,x-j)\,,\, \sum_{k=0}^{p} c_k\, \varphi(2\,x-2\,m-k) \right\rangle \tag{13.69}$$

$$= \sum_{j,k=0}^{p} c_j\, c_k\, \langle\, \varphi(2\,x-j)\,, \varphi(2\,x-2\,m-k)\,\rangle = \frac{1}{2}\sum_{j,k=0}^{p} c_j\, c_k\, \langle\, \varphi(x)\,, \varphi(x+j-2\,m-k)\,\rangle.$$

If we require orthogonality (13.67) of all the integer translates of φ, then the left hand side of this identity will be 0 unless $m = 0$, while only the summands with $j = 2\,m + k$ will be nonzero on the right. Therefore, orthogonality requires that

$$\sum_{0\,\le\,k\,\le\,p-2m} c_{2m+k}\, c_k = \begin{cases} 2, & m = 0, \\ 0, & m \neq 0. \end{cases} \tag{13.70}$$

The algebraic equations (13.65, 70) for the dilation coefficients are the key requirements for the construction of an orthogonal wavelet basis.

For example, if we have just two nonzero coefficients c_0, c_1, then (13.65, 70) reduce to

$$c_0 + c_1 = 2, \qquad\qquad c_0^2 + c_1^2 = 2,$$

and so $c_0 = c_1 = 1$ is the only solution, resulting in the Haar dilation equation (13.58). If we have three coefficients c_0, c_1, c_2, then (13.65), (13.70) require

$$c_0 + c_1 + c_2 = 2, \qquad\qquad c_0^2 + c_1^2 + c_2^2 = 2, \qquad\qquad c_0\, c_2 = 0.$$

[†] In all instances, the functions have bounded support, and so the inner product integral can be reduced to an integral over a finite interval where both f and g are nonzero.

Thus either $c_2 = 0$, $c_0 = c_1 = 1$, and we are back to the Haar case, or $c_0 = 0$, $c_1 = c_2 = 1$, and the resulting dilation equation is a simple reformulation of the Haar case; see Exercise ■. In particular, the hat function (13.59) does *not* give rise to orthogonal wavelets.

The remarkable fact, discovered by Daubechies, is that there *is* a nontrivial solution for four (and, indeed, any even number) of nonzero coefficients c_0, c_1, c_2, c_3. The basic equations (13.65), (13.70) require

$$c_0 + c_1 + c_2 + c_3 = 2, \qquad c_0^2 + c_1^2 + c_2^2 + c_3^2 = 2, \qquad c_0\, c_2 + c_1\, c_3 = 0. \tag{13.71}$$

The particular values

$$c_0 = \tfrac{1+\sqrt{3}}{4}, \qquad c_1 = \tfrac{3+\sqrt{3}}{4}, \qquad c_2 = \tfrac{3-\sqrt{3}}{4}, \qquad c_3 = \tfrac{1-\sqrt{3}}{4}, \tag{13.72}$$

solve (13.71). These coefficients correspond to the *Daubechies dilation equation*

$$\varphi(x) = \tfrac{1+\sqrt{3}}{4}\,\varphi(2\,x) + \tfrac{3+\sqrt{3}}{4}\,\varphi(2\,x-1) + \tfrac{3-\sqrt{3}}{4}\,\varphi(2\,x-2) + \tfrac{1-\sqrt{3}}{4}\,\varphi(2\,x-3). \tag{13.73}$$

Any nonzero solution of bounded support to this remarkable functional equation will give rise to a scaling function $\varphi(x)$, a mother wavelet

$$w(x) = \tfrac{1-\sqrt{3}}{4}\,\varphi(2\,x) - \tfrac{3-\sqrt{3}}{4}\,\varphi(2\,x-1) + \tfrac{3+\sqrt{3}}{4}\,\varphi(2\,x-2) - \tfrac{1+\sqrt{3}}{4}\,\varphi(2\,x-3), \tag{13.74}$$

and then, by compression and translation (13.62), the complete system of orthogonal wavelets $w_{j,k}(x)$.

Before explaining how to solve the Daubechies dilation equation, let us complete the proof of orthogonality. It is easy to see that, by translation invariance, since $\varphi(x)$ and $\varphi(x-m)$ are orthogonal for any $m \neq 0$, so are $\varphi(x-k)$ and $\varphi(x-l)$ for any $k \neq l$. Next we prove orthogonality of $\varphi(x-m)$ and $w(x)$:

$$\begin{aligned}
\langle\, w(x)\,, \varphi(x-m)\,\rangle &= \left\langle\, \sum_{j=0}^{p} (-1)^{j+1}\, c_j\, \varphi(2x-1+j)\,,\, \sum_{k=0}^{p} c_k\, \varphi(2x-2m-k)\,\right\rangle \\
&= \sum_{j,k=0}^{p} (-1)^{j+1}\, c_j\, c_k\, \langle\, \varphi(2x-1+j)\,, \varphi(2x-2m-k)\,\rangle \\
&= \frac{1}{2} \sum_{j,k=0}^{p} (-1)^{j+1}\, c_j\, c_k\, \langle\, \varphi(x)\,, \varphi(x-1+j-2m-k)\,\rangle,
\end{aligned}$$

using (13.68). By orthogonality (13.67) of the translates of φ, the only summands that are nonzero are when $j = 2m + k + 1$; the resulting coefficient of $\| \varphi(x) \|^2$ is

$$\sum_k (-1)^k\, c_{1-2m-k}\, c_k = 0,$$

where the sum is over all $0 \leq k \leq p$ such that $0 \leq 1 - 2m - k \leq p$. Each term in the sum appears twice, with opposite signs, and hence the result is always zero — no matter what the coefficients $c_0, \dots, c_p$ are! The proof of orthogonality of the translates $w(x-m)$ of the mother wavelet, along with all her wavelet descendants $w(2^j\,x - k)$, relies on a similar argument, and the details are left as an exercise for the reader.

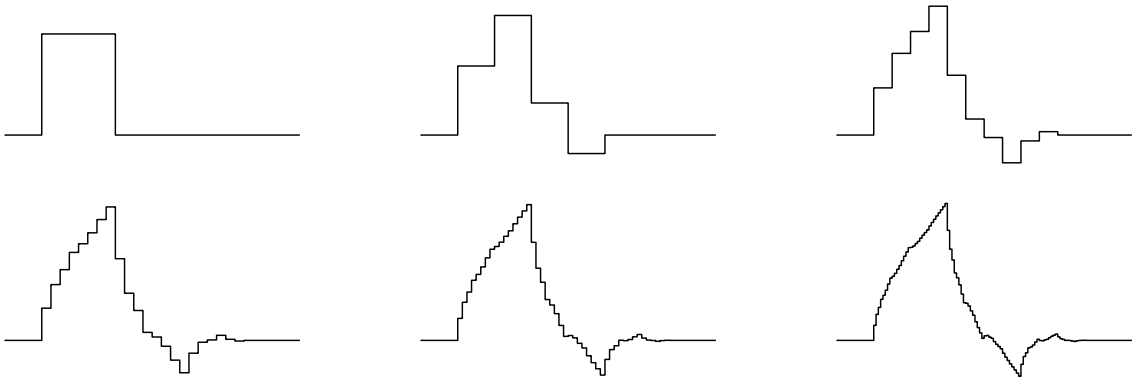


Figure 13.10. Approximating the Daubechies Wavelet.

Solving the Dilation Equation

Let us next discuss how to solve the dilation equation (13.57). The solution we are after does not have an elementary formula, and we require a slightly sophisticated approach to recover it. The key observation is that (13.57) has the form of a fixed point equation

$$\varphi = F[\varphi],$$

not in ordinary Euclidean space, but in an infinite-dimensional function space. With luck, the fixed point (or, more correctly, fixed function) will be stable, and so starting with a suitable initial guess $\varphi_0(x)$, the successive iterates

$$\varphi_{n+1} = F[\varphi_n]$$

will converge to the desired solution: $\varphi_n(x) \longrightarrow \varphi(x)$. In detail, the iterative version of the dilation equation (13.57) reads

$$\varphi_{n+1}(x) = \sum_{k=0}^{p} c_k \, \varphi_n(2x - k), \qquad n = 0, 1, 2, \dots . \tag{13.75}$$

Before attempting to prove convergence of this iterative procedure to the Daubechies scaling function, let us experimentally investigate what happens.

A reasonable choice for the initial guess might be the Haar scaling or box function

$$\varphi_0(x) = \begin{cases} 1, & 0 < t \leq 1. \\ 0, & \text{otherwise.} \end{cases}$$

In Figure 13.10 we graph the next 5 iterates $\varphi_1(x), \dots, \varphi_5(x)$. There clearly appears to be converging to some function $\varphi(x)$, although the final result does look a little bizarre. Bolstered by this preliminary experimental evidence, we can now try to prove convergence of the iterative scheme. This turns out to be true; a fully rigorous proof relies on the Fourier transform, [**50**], but is a little too advanced for this text and will be omitted.

Theorem 13.13. *The functions converge $\varphi_n(x)$ defined by the iterative functional equation* (13.75) *converge uniformly to a continuous function $\varphi(x)$, called the* Daubechies scaling function.

Once we have established convergence, we are now able to verify that the scaling function and consequential system of wavelets form an orthogonal system of functions.

Proposition 13.14. *All integer translates $\varphi(x-k)$, for $k \in \mathbb{Z}$ of the Daubechies scaling function, and all wavelets $w_{j,k}(x) = w(2^j x - k)$, $j \geq 0$, are mutually orthogonal functions with respect to the* L^2 *inner product. Moreover, $\| \varphi \|^2 = 1$, while $\| w_{j,k} \|^2 = 2^{-j}$.*

Proof: As noted earlier, the orthogonality of the entire wavelet system will follow once we know the orthogonality (13.67) of the scaling function and its integer translates. We use induction to prove that this holds for all the iterates $\varphi_n(x)$, and so, in view of uniform convergence, the limiting scaling function also satisfies this property. Details are relegated to Exercise ■. *Q.E.D.*

In practical computations, the limiting procedure for constructing the scaling function is not so convenient, and an alternative means of computing its values is employed. The starting point is to determine its values at integer points. First, the initial box function has values $\varphi_0(m) = 0$ for all integers $m \in \mathbb{Z}$ except $\varphi_0(1) = 1$. The iterative functional equation (13.75) will then produce the values of the iterates $\varphi_n(m)$ at integer points $m \in \mathbb{Z}$. A simple induction will convince you that $\varphi_n(m) = 0$ except for $m = 1$ and $m = 2$, and, therefore, by (13.75),

$$\varphi_{n+1}(1) = \tfrac{3+\sqrt{3}}{4}\,\varphi_n(1) + \tfrac{1+\sqrt{3}}{4}\,\varphi_n(2), \qquad \varphi_{n+1}(2) = \tfrac{1-\sqrt{3}}{4}\,\varphi_n(1) + \tfrac{3-\sqrt{3}}{4}\,\varphi_n(2),$$

since all other terms are 0. This has the form of a linear iterative system

$$\mathbf{v}^{(n+1)} = A\,\mathbf{v}^{(n)} \tag{13.76}$$

with coefficient matrix

$$A = \begin{pmatrix} \frac{3+\sqrt{3}}{4} & \frac{1+\sqrt{3}}{4} \\ \frac{1-\sqrt{3}}{4} & \frac{3-\sqrt{3}}{4} \end{pmatrix} \qquad \text{and where} \qquad \mathbf{v}^{(n)} = \begin{pmatrix} \varphi_n(1) \\ \varphi_n(2) \end{pmatrix}.$$

Referring back to Chapter 10, the solution to such an iterative system is specified by the eigenvalues and eigenvectors of the coefficient matrix, which are

$$\lambda_1 = 1, \qquad \mathbf{v}_1 = \begin{pmatrix} \frac{1+\sqrt{3}}{4} \\ \frac{1-\sqrt{3}}{4} \end{pmatrix}, \qquad \lambda_2 = \tfrac{1}{2}, \qquad \mathbf{v}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

We write the initial condition as a linear combination of the eigenvectors

$$\mathbf{v}^{(0)} = \begin{pmatrix} \varphi_0(1) \\ \varphi_0(2) \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} = 2\,\mathbf{v}_1 - \frac{1-\sqrt{3}}{2}\,\mathbf{v}_2.$$

The solution is

$$\mathbf{v}^{(n)} = A^n\,\mathbf{v}^{(0)} = 2\,A^n\mathbf{v}_1 - \frac{1-\sqrt{3}}{2}\,A^n\mathbf{v}_2 = 2\,\mathbf{v}_1 - \frac{1}{2^n}\,\frac{1-\sqrt{3}}{2}\,\mathbf{v}_2.$$

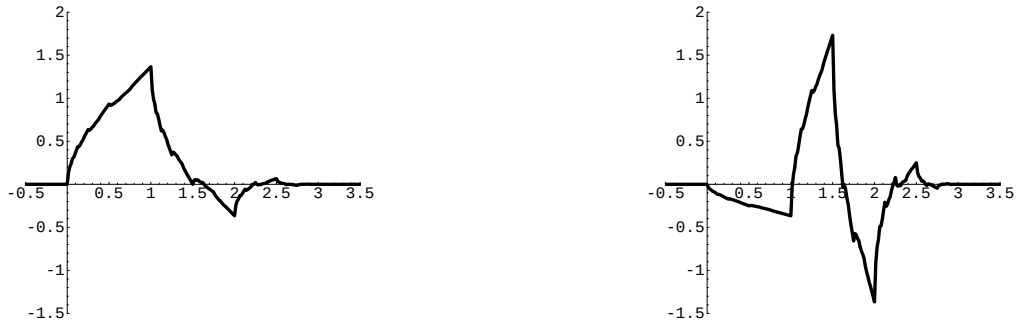


Figure 13.11. The Daubechies Scaling Function and Mother Wavelet.

The limiting vector

$$\begin{pmatrix} \varphi(1) \\ \varphi(2) \end{pmatrix} = \lim_{n \to \infty} \mathbf{v}^{(n)} = 2\,\mathbf{v}_1 = \begin{pmatrix} \frac{1+\sqrt{3}}{2} \\ \frac{1-\sqrt{3}}{2} \end{pmatrix}$$

gives the desired values of the scaling function:

$$\varphi(1) = \frac{1-\sqrt{3}}{2} = 1.366025\dots\,, \qquad \varphi(2) = \frac{-1+\sqrt{3}}{2} = -\,.366025\dots\,, \qquad \varphi(m) = 0, \quad \text{for all} \quad m \neq 1, 2. \tag{13.77}$$

With this in hand, the Daubechies dilation equation (13.73) then prescribes the function values $\varphi\big(\frac{1}{2}\,m\big)$ at all half integers, because when $x = \frac{1}{2}\,m$ then $2\,x - k = m - k$ is an integer. Once we know its values at the half integers, we can re-use equation (13.73) to give its values at quarter integers $\frac{1}{4}\,m$. Continuing onwards, we determine the values of $\varphi(x)$ at all *dyadic points*, meaning rational numbers of the form $x = m/2^j$ for $m, j \in \mathbb{Z}$. Continuity will then prescribe its value at any other $x \in \mathbb{R}$ since x can be written as the limit of dyadic numbers x_n — namely those obtained by truncating its binary (base 2) expansion at the n^{th} digit beyond the decimal (or, rather "binary") point. But, in practice, this latter step is unnecessary, since all computers are ultimately based on the binary number system, and so only dyadic numbers actually reside in a computer's memory. Thus, there is no real need to determine the value of φ at non-dyadic points.

The preceding scheme was used to produce the graphs of the Daubechies scaling function in Figure 13.11. It is continuous, but non-differentiable function — and its graph has a very jagged, fractal-like appearance when viewed at close range. The Daubechies scaling function is, in fact, a close relative of the famous example of a continuous, nowhere differentiable function originally due to Weierstrass, [**125**, **152**], whose construction also relies on a similar scaling argument.

With the values of the Daubechies scaling function on a sufficiently dense set of dyadic points in hand, the consequential values of the mother wavelet are given by formula (13.74). Note that $\operatorname{supp}\varphi = \operatorname{supp} w = [0, 3]$. The daughter wavelets are then found by the usual compression and translation procedure (13.62).

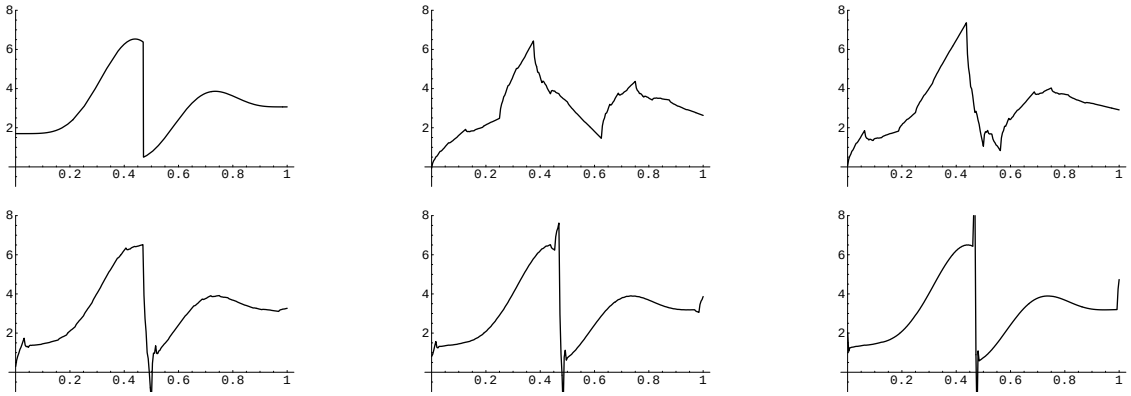


Figure 13.12. Daubechies Wavelet Expansion.

The Daubechies wavelet expansion of a function whose support is contained in[†] $[0,1]$ is then given by

$$f(x) \;\sim\; c_0\,\varphi(x) + \sum_{j=0}^{\infty} \sum_{k=-2}^{2^j-1} c_{j,k}\, w_{j,k}(x). \tag{13.78}$$

The inner summation begins at $k = -2$ so as to include *all* the wavelet offspring $w_{j,k}$ whose support has a nontrivial intersection with the interval $[0,1]$. The wavelet coefficients $c_0, c_{j,k}$ are computed by the usual orthogonality formula

$$\begin{aligned} c_0 &= \langle\, f\,, \varphi\,\rangle = \int_0^3 f(x)\,\varphi(x)\,dx, \\ c_{j,k} &= \langle\, f\,, w_{j,k}\,\rangle = 2^j \int_{2^{-j}\,k}^{2^{-j}\,(k+3)} f(x)\,w_{j,k}(x)\,dx = \int_0^3 f\big(\,2^{-j}\,(x+k)\,\big)\,w(x)\,dx, \end{aligned} \tag{13.79}$$

where we agree that $f(x) = 0$ whenever $x < 0$ or $x > 1$. In practice, one employs a numerical integration procedure, e.g., the trapezoid rule, [**9**], based on dyadic nodes to speedily evaluate the integrals (13.79). A proof of completeness of the resulting wavelet basis functions can be found in [**50**]. Compression and denoising algorithms based on retaining only low frequency modes proceed as before, and are left as exercises for the reader to implement.

Example 13.15. In Figure 13.12, we plot the Daubechies wavelet expansions of the same signal for Example 13.9. The first plot is the original signal, and the following show the partial sums of (13.78) over $j = 0, \dots, r$ with $r = 2, 3, 4, 5, 6$. Unlike the Haar expansion, the Daubechies wavelets do exhibit the nonuniform Gibbs phenomenon at the interior discontinuity as well as the endpoints, since the function is set to 0 outside the interval $[0,1]$. Indeed, the Daubechies wavelets are continuous, and so cannot converge uniformly to a discontinuous function.

[†] For functions with larger support, one should include additional terms in the expansion corresponding to further translates of the wavelets so as to cover the entire support of the function. Alternatively, one can translate and rescale x to fit the function's support inside $[0,1]$.

13.3. The Fourier Transform.

Fourier series and their ilk are designed to solve boundary value problems on bounded intervals. The extension of Fourier methods to unbounded intervals — the entire real line — leads naturally to the Fourier transform, which is a powerful mathematical tool for the analysis of non-periodic functions. The Fourier transform is of fundamental importance in a broad range of applications, including both ordinary and partial differential equations, quantum mechanics, signal processing, control theory, and probability, to name but a few.

We begin by motivating the Fourier transform as a limiting case of Fourier series. Although the rigorous details are rather exacting, the underlying idea is not so difficult. Let $f(x)$ be a reasonably nice function defined for all $-\infty < x < \infty$. The goal is to construct a Fourier expansion for $f(x)$ in terms of basic trigonometric functions. One evident approach is to construct its Fourier series on progressively larger and larger intervals, and then take the limit as the intervals' length becomes infinite. The limiting process converts the Fourier sums into integrals, and the resulting representation of a function is renamed the Fourier transform. Since we are dealing with an infinite interval, there are no longer any periodicity requirements on the function $f(x)$. Moreover, the frequencies represented in the Fourier transform are no longer constrained by the length of the interval, and so we are effectively decomposing a quite general, non-periodic function into a continuous superposition of trigonometric functions of all possible frequencies.

Let us present the details of this construction in a more concrete form. The computations will be significantly simpler if we work with the complex version of the Fourier series from the outset. Our starting point is the rescaled Fourier series (12.84) on a symmetric interval $[\,-\ell\,,\ell\,]$ of length $2\,\ell$, which we rewrite in the adapted form

$$f(x) \quad \sim \quad \sum_{\nu=-\infty}^{\infty} \sqrt{\frac{\pi}{2}}\; \frac{\widehat{f}_{\ell}(k_\nu)}{\ell}\; e^{\,\mathrm{i}\,k_\nu x}. \tag{13.80}$$

The sum is over the discrete collection of frequencies

$$k_\nu = \frac{\pi\,\nu}{\ell}, \qquad\qquad \nu = 0, \pm 1, \pm 2, \dots\,, \tag{13.81}$$

corresponding to those trigonometric functions that have period $2\,\ell$. For reasons that will soon become apparent, the Fourier coefficients of f are now denoted as

$$c_\nu = \langle\, f\,, e^{\,\mathrm{i}\,k_\nu\,x}\,\rangle = \frac{1}{2\,\ell}\int_{-\ell}^{\ell} f(x)\, e^{-\,\mathrm{i}\,k_\nu\,x}\,dx = \sqrt{\frac{\pi}{2}}\;\frac{\widehat{f}(k_\nu)}{\ell},$$

so that

$$\widehat{f}(k_\nu) = \frac{1}{\sqrt{2\,\pi}}\int_{-\ell}^{\ell} f(x)\, e^{-\,\mathrm{i}\,k_\nu\,x}\,dx. \tag{13.82}$$

This reformulation of the basic Fourier series formula allows us to smoothly pass to the limit when the intervals' length $\ell \to \infty$.

On an interval of length $2\,\ell$, the frequencies (13.81) required to represent a function in Fourier series form are equally distributed, with interfrequency spacing

$$\Delta k = k_{\nu+1} - k_\nu = \frac{\pi}{\ell}\,.$$

As $\ell \to \infty$, the spacing $\Delta k \to 0$, and so the relevant frequencies become more and more densely packed in the space of all possible frequencies: $-\infty < k < \infty$. In the limit, we anticipate that *all* possible frequencies will be represented. Indeed, letting $k_\nu = k$ be arbitrary in (13.82), and sending $\ell \to \infty$, results in the infinite integral

$$\widehat{f}(k) = \frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{\infty} f(x)\, e^{-\,\mathrm{i}\,k\,x}\, dx \tag{13.83}$$

known as the *Fourier transform* of the function $f(x)$. If $f(x)$ is a reasonably nice function, e.g., piecewise continuous and decayin to 0 reasonably quickly as $|\,x\,| \to \infty$, its Fourier transform $\widehat{f}(k)$ is defined for all possible frequencies $-\infty < k < \infty$. This formula will sometimes conveniently be abbreviated as

$$\widehat{f}(k) = \mathcal{F}[\,f(x)\,], \tag{13.84}$$

where $\mathcal{F}$ is the *Fourier transform operator*. The extra $\sqrt{2\,\pi}$ factor in front of the integral, which is sometimes omitted, is for later convenience.

To reconstruct the function from its Fourier transform, we employ a similar limiting procedure on the Fourier series (13.80), which we first rewrite in a more suggestive form:

$$f(x) \;\sim\; \frac{1}{\sqrt{2\,\pi}} \sum_{\nu\,=\,-\infty}^{\infty} \widehat{f}_{\,\ell}(k_\nu)\, e^{\,\mathrm{i}\,k_\nu\,x}\, \Delta k. \tag{13.85}$$

For each fixed value of x, the right hand side has the form of a Riemann sum, [**9**, **153**], over the entire frequency space $-\infty < k < \infty$, for the function $g_\ell(k) = \widehat{f}_{\,\ell}(k)\, e^{\,\mathrm{i}\,k\,x}$. Under reasonable hypotheses, as $\ell \to \infty$, the functions $\widehat{f}_{\,\ell}(k) \to \widehat{f}(k)$ converge to the Fourier transform; moreover, the interfrequency spacing $\Delta k \to 0$ and so one expects the Riemann sums to converge to the integral

$$f(x) \;\sim\; \frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{\infty} \widehat{f}(k)\, e^{\,\mathrm{i}\,k\,x}\, dk, \tag{13.86}$$

known as the *inverse Fourier transform*, that serves to recover the original signal from its Fourier transform. In abbreviated form

$$f(x) = \mathcal{F}^{-1}[\,\widehat{f}(k)\,] \tag{13.87}$$

is the inverse of the Fourier transform operator (13.84). In this manner, the Fourier series (13.85) becomes a Fourier integral that reconstructs the function $f(x)$ as a (continuous) superposition of complex exponentials $e^{\,\mathrm{i}\,k\,x}$ of *all* possible frequencies. The contribution of each such exponential is the value of the Fourier transform $\widehat{f}(k)$ at the exponential's frequency.

It is also worth pointing out that both the Fourier transform (13.84) and its inverse (13.87) define linear maps on function space. This means that the Fourier transform of the sum of two functions is the sum of their individual transforms, while multiplying a function by a constant multiplies its Fourier transform bby the same factor:

$$\begin{aligned} &\mathcal{F}[\,f(x) + g(x)\,] = \mathcal{F}[\,f(x)\,] + \mathcal{F}[\,g(x)\,] = \widehat{f}(k) + \widehat{g}(k), \\ &\mathcal{F}[\,c\,f(x)\,] = c\,\mathcal{F}[\,f(x)\,] = c\,\widehat{f}(k). \end{aligned} \tag{13.88}$$

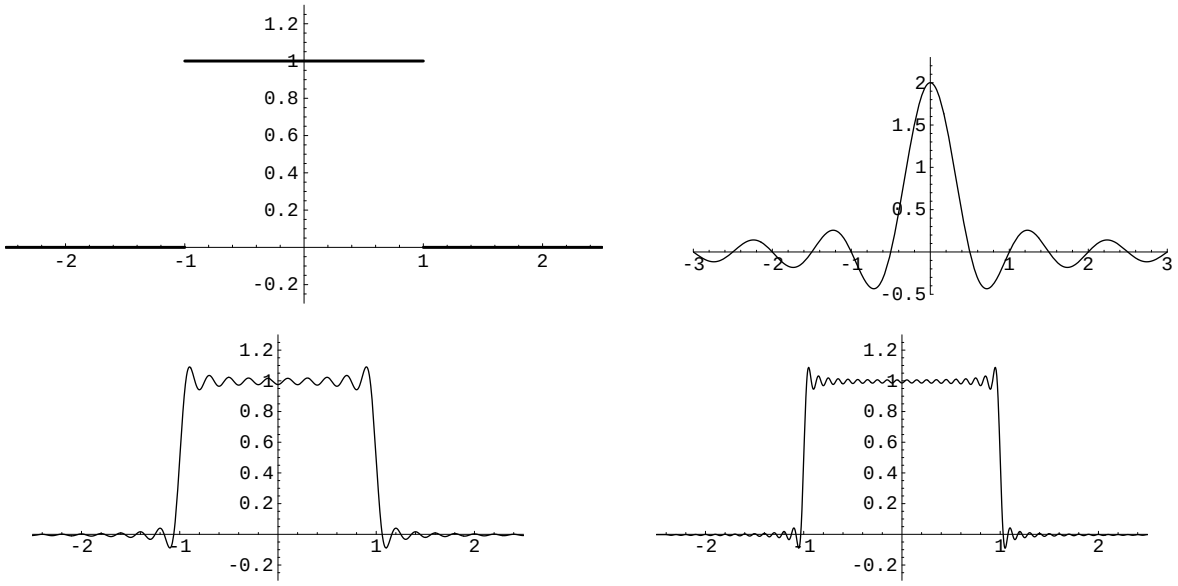


Figure 13.13. Fourier Transform of a Rectangular Pulse.

A similar statement hold for the inverse Fourier transform $\mathcal{F}^{-1}$.

Recapitulating, by letting the length of the interval go to ∞, the discrete Fourier series has become a continuous Fourier integral, while the Fourier coefficients, which were defined only at a discrete collection of possible frequencies, have become an entire function $\widehat{f}(k)$ defined on all of frequency space $k \in \mathbb{R}$. The reconstruction of $f(x)$ from its Fourier transform $\widehat{f}(k)$ via (13.86) can be rigorously justified under suitable hypotheses. For example, if $f(x)$ is piecewise C^1 on all of $\mathbb{R}$ and $f(x) \to 0$ decays reasonably rapidly as $|\, x \,| \to \infty$ ensuring that its Fourier integral (13.83) converges, then it can be proved that the inverse Fourier integral (13.86) will converge to $f(x)$ at all points of continuity, and to the midpoint $\frac{1}{2}(f(x^-) + f(x^+))$ at jump discontinuities — just like a Fourier series. In particular, its Fourier transform $\widehat{f}(k) \to 0$ must also decay as $|\, k \,| \to 0$, implying that (as with Fourier series) the very high frequency modes make negligible contributions to the reconstruction of the signal. A more precise, general result will be formulated in Theorem 13.28 below.

Example 13.16. The Fourier transform of a rectangular pulse or box function

$$f(x) = \sigma(x+a) - \sigma(x-a) = \begin{cases} 1, & -a < x < a, \\ 0, & |\, x \,| > a. \end{cases} \tag{13.89}$$

of width $2a$ is easily computed:

$$\widehat{f}(k) = \frac{1}{\sqrt{2\,\pi}} \int_{-a}^{a} e^{-\,\mathrm{i}\,k\,x}\, dx = \frac{e^{\,\mathrm{i}\,k a} - e^{-\,\mathrm{i}\,k a}}{\sqrt{2\,\pi}\;\mathrm{i}\,k} = \sqrt{\frac{2}{\pi}}\;\frac{\sin a\,k}{k}\,. \tag{13.90}$$

On the other hand, the reconstruction of the pulse via the inverse transform (13.86) tells

us that

$$\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{e^{\,\mathrm{i}\,k\,x}\,\sin a\,k}{k}\,dk = f(x) = \begin{cases} 1, & -a < x < a, \\ \frac{1}{2}, & x = \pm a, \\ 0, & |\,x\,| > a. \end{cases} \tag{13.91}$$

Note the convergence to the middle of the jump discontinuities at $x = \pm a$. Splitting this complex integral into its real and imaginary parts, we deduce a pair of striking trigonometric integral identities

$$\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{\cos k\,x\,\sin a\,k}{k}\,dk = \begin{cases} 1, & -a < x < a, \\ \frac{1}{2}, & x = \pm a, \\ 0, & |\,x\,| > a, \end{cases} \qquad \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{\sin k\,x\,\sin a\,k}{k}\,dk = 0. \tag{13.92}$$

Just as many Fourier series yield nontrivial summation formulae, the reconstruction of a function from its Fourier transform often leads to nontrivial integral identities. One *cannot* compute the integral (13.91) by the Fundamental Theorem of Calculus, since there is no elementary function[†] whose derivative equals the integrand. Moreover, it is not even clear that the integral converges; indeed, the amplitude of the oscillatory integrand decays like $1/|\,k\,|$, but the latter function does not have a convergent integral, and so the usual comparison test for infinite integrals, [**9**, **45**, **165**], fails to apply. Thus, the convergence of the integral is marginal at best: the trigonometric oscillations somehow overcome the slow rate of decay of $1/k$ and thereby induce the (conditional) convergence of the integral! In Figure 13.13 we display the box function with $a = 1$, its Fourier transform, along with a reconstruction obtained by numerically integrating (13.92). Since we are dealing with an infinite integral, we must break off the numerical integrator by restricting it to a finite interval. The first graph is obtained by integrating from $-5 \le k \le 5$ while the second is from $-10 \le k \le 10$. The non-uniform convergence of the integral leads to the appearance of a Gibbs phenomenon at the two discontinuities, similar to that of a Fourier series.

Example 13.17. Consider an exponentially decaying right-handed pulse[‡]

$$f_r(x) = \begin{cases} e^{-a\,x}, & x > 0, \\ 0, & x < 0, \end{cases} \tag{13.93}$$

where $a > 0$. We compute its Fourier transform directly from the definition:

$$\widehat{f}(k) = \frac{1}{\sqrt{2\,\pi}}\int_0^{\infty} e^{-a\,x}\,e^{-\,\mathrm{i}\,k\,x}\,dx = -\,\frac{1}{\sqrt{2\,\pi}}\,\frac{e^{-(a+\,\mathrm{i}\,k)x}}{a+\,\mathrm{i}\,k}\,\Big|_{x=0}^{\infty} = \frac{1}{\sqrt{2\,\pi}\,(a+\,\mathrm{i}\,k)}\,.$$

[†] One can use Euler's formula (3.84) to reduce the integrand to one of the form $e^{\alpha\,k}/k$, but it can be proved, [**int**], that there is no formula for $\int (e^{\alpha\,k}/k)\,dk$ in terms of elementary functions.

[‡] Note that we can't Fourier transform the entire exponential function $e^{-a\,x}$ because it does not go to zero at both $\pm\infty$, which is required for the integral (13.83) to converge.

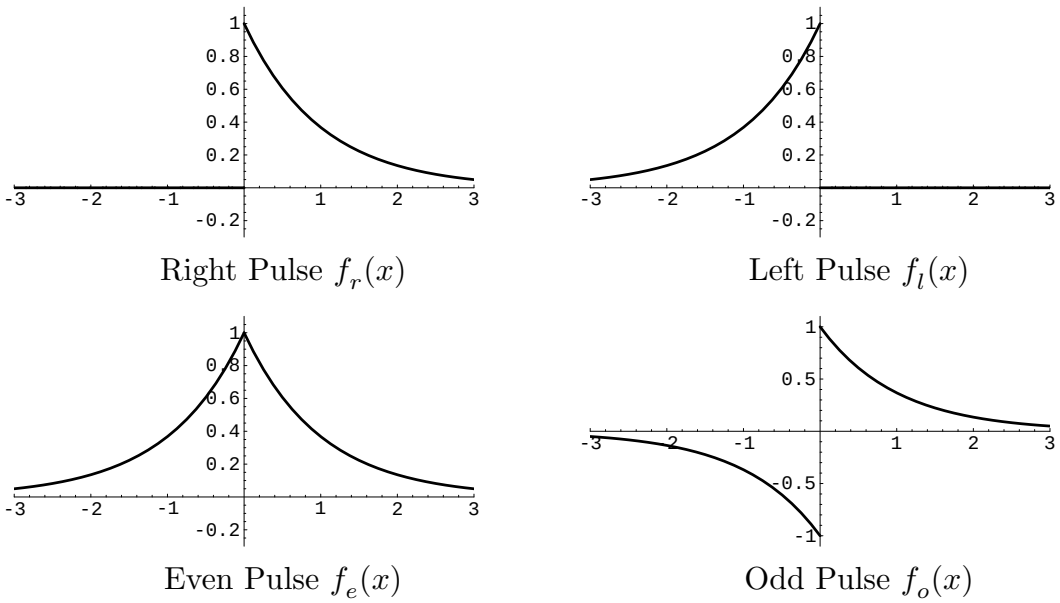


Figure 13.14. Exponential Pulses.

As in the preceding example, the inverse Fourier transform for this function produces a nontrivial complex integral identity:

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{e^{\mathrm{i}\,k\,x}}{a+\mathrm{i}\,k}\,dk=\begin{cases}e^{-a\,x}, & x>0,\\ \frac{1}{2}, & x=0,\\ 0, & x<0.\end{cases} \tag{13.94}$$

Similarly, a pulse that decays to the left,

$$f_l(x)=\begin{cases}e^{a\,x}, & x<0,\\ 0, & x>0,\end{cases} \tag{13.95}$$

where $a>0$ is still positive, has Fourier transform

$$\widehat{f}(k)=\frac{1}{\sqrt{2\,\pi}\,(a-\mathrm{i}\,k)}\,. \tag{13.96}$$

This also follows from the general fact that the Fourier transform of $f(-x)$ is $\widehat{f}(-k)$; see Exercise ∎. The even exponentially decaying pulse

$$f_e(x)=e^{-a\,|\,x\,|} \tag{13.97}$$

is merely the sum of left and right pulses: $f_e=f_r+f_l$. Thus, by linearity,

$$\widehat{f}_e(k)=\frac{1}{\sqrt{2\,\pi}\,(a+\mathrm{i}\,k)}\;+\;\frac{1}{\sqrt{2\,\pi}\,(a-\mathrm{i}\,k)}=\sqrt{\frac{2}{\pi}}\,\frac{a}{k^2+a^2}\,, \tag{13.98}$$

The result is real and even because $f_e(x)$ is a real even function; see Exercise ∎. The inverse Fourier transform (13.86) produces another nontrivial integral identity:

$$e^{-a\,|\,x\,|}=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{a\,e^{\mathrm{i}\,k\,x}}{k^2+a^2}\,dk=\frac{a}{\pi}\int_{-\infty}^{\infty}\frac{\cos k\,x}{k^2+a^2}\;dk. \tag{13.99}$$

(The imaginary part of the integral vanishes because the integrand is odd.) On the other hand, the odd exponentially decaying pulse,

$$f_o(x) = (\operatorname{sign} x)\, e^{-a\,|\,x\,|} = \begin{cases} e^{-a\,x}, & x > 0, \\ -\,e^{a\,x}, & x < 0, \end{cases} \tag{13.100}$$

is the difference of the right and left pulses, $f_o = f_r - f_l$, and has purely imaginary and odd Fourier transform

$$\widehat{f}_o(k) = \frac{1}{\sqrt{2\,\pi}\,(a + \mathrm{i}\,k)} \;-\; \frac{1}{\sqrt{2\,\pi}\,(a - \mathrm{i}\,k)} = -\,\mathrm{i}\,\sqrt{\frac{2}{\pi}}\,\frac{k}{k^2 + a^2}\,. \tag{13.101}$$

The inverse transform is

$$(\operatorname{sign} x)\, e^{-a\,|\,x\,|} = -\,\frac{\mathrm{i}}{\pi}\,\int_{-\infty}^{\infty} \frac{k\,e^{\,\mathrm{i}\,k\,x}}{k^2 + a^2}\,dk = \frac{1}{\pi}\,\int_{-\infty}^{\infty} \frac{k \sin k\,x}{k^2 + a^2}\;dk. \tag{13.102}$$

As a final example, consider the rational function

$$f(x) = \frac{1}{x^2 + c^2}\,, \qquad \text{where} \qquad c > 0. \tag{13.103}$$

Its Fourier transform requires integrating

$$\widehat{f}(k) = \frac{1}{\sqrt{2\,\pi}}\,\int_{-\infty}^{\infty} \frac{e^{-\,\mathrm{i}\,k\,x}}{x^2 + a^2}\;dx. \tag{13.104}$$

The indefinite integral (anti-derivative) does not appear in basic integration tables, and, in fact, cannot be done in terms of elementary functions. However, we have just managed to evaluate this particular integral! Look at (13.99). If we change x to k and k to $-x$, then we exactly recover the integral (13.104) up to a factor of $a\sqrt{2/\pi}$. Therefore, in accordance with Theorem 13.18, we conclude that the Fourier transform of (13.103) is

$$\widehat{f}(k) = \sqrt{\frac{\pi}{2}}\;\frac{e^{-a\,|\,k\,|}}{a}\,. \tag{13.105}$$

This last example is indicative of an important general fact. The reader has no doubt already noted the remarkable similarity between the Fourier transform (13.83) and its inverse (13.86). Indeed, the only difference is that the former has a minus sign in the exponential. This implies the following *symmetry principle* relating the direct and inverse Fourier transforms.

Theorem 13.18. *If the Fourier transform of the function $f(x)$ is $\widehat{f}(k)$, then the Fourier transform of $\widehat{f}(x)$ is $f(-k)$.*

Symmetry allows us to reduce the tabulation of Fourier transforms by half. For instance, referring back to Example 13.16, we deduce that the Fourier transform of the function

$$f(x) = \sqrt{\frac{2}{\pi}}\;\frac{\sin a\,x}{x}$$

is

$$\widehat{f}(k) = \sigma(-k+a) - \sigma(-k-a) = \sigma(k+a) - \sigma(k-a) = \begin{cases} 1, & -a < k < a, \\ \frac{1}{2}, & k = \pm a, \\ 0, & |\, k \,| > a. \end{cases} \tag{13.106}$$

Note that, by linearity, we can divide both $f(x)$ and $\widehat{f}(k)$ by $\sqrt{2/\pi}$ to deduce the Fourier transform of $\dfrac{\sin a\, x}{x}$.

Warning: Some authors leave out the $\sqrt{2\,\pi}$ factor in the definition (13.83) of the Fourier transform $\widehat{f}(k)$. This alternative convention does have a slight advantage of eliminating many $\sqrt{2\,\pi}$ factors in the Fourier transforms of elementary functions. However, this requires an extra such factor in the reconstruction formula (13.86), which is achieved by replacing $\sqrt{2\,\pi}$ by $2\,\pi$. A significant disadvantage is that the resulting formulae for the Fourier transform and its inverse are no as symmetric, and so the symmetry principle of Theorem 13.18 requires some modification. On the other hand, convolution (to be discussed below) is a little easier without the extra factor. When consulting any particular reference book, the reader *always* needs to check which convention is being used.

All of the functions in Example 13.17 required $a > 0$ for the Fourier integrals to converge. The functions that emerge in the limit as a goes to 0 are of fundamental importance. Let us start with the odd exponential pulse (13.100). When $a \to 0$, the function $f_o(x)$ converges to the *sign function*

$$f(x) = \operatorname{sign} x = \sigma(x) - \sigma(-x) = \begin{cases} +1, & x > 0, \\ -1, & x < 0. \end{cases} \tag{13.107}$$

Taking the limit of the Fourier transform (13.101) leads to

$$\widehat{f}(k) = -\,\mathrm{i}\,\sqrt{\frac{2}{\pi}}\,\frac{1}{k}\,. \tag{13.108}$$

The nonintegrable singularity of $\widehat{f}(k)$ at $k = 0$ is indicative of the fact that the sign function does *not* decay as $|\, x \,| \to \infty$. In this case, neither the Fourier transform integral nor its inverse are well-defined as standard (Riemann, or even Lebesgue, [**153**]) integrals. Nevertheless, it is possible to rigorously justify these results within the framework of generalized functions.

More interesting are the even pulse functions $f_e(x)$, which, in the limit $a \to 0$, become the constant function

$$f(x) \equiv 1. \tag{13.109}$$

The limit of the Fourier transform (13.98) is

$$\lim_{a \to 0} \sqrt{\frac{2}{\pi}}\,\frac{2a}{k^2 + a^2} = \begin{cases} 0, & k \neq 0, \\ \infty, & k = 0. \end{cases} \tag{13.110}$$

This limiting behavior should remind the reader of our construction (11.31) of the delta function as the limit of the functions

$$\delta(x) = \lim_{n \to \infty} \frac{n}{\pi\,(1 + n^2\,x^2)} = \lim_{a \to 0} \frac{a}{\pi\,(a^2 + x^2)}\,,$$

Comparing with (13.110), we conclude that the Fourier transform of the constant function (13.109) is a multiple of the delta function in the frequency variable:

$$\widehat{f}(k) = \sqrt{2\,\pi}\;\delta(k). \tag{13.111}$$

The direct transform integral

$$\delta(k) = \frac{1}{2\,\pi}\int_{-\infty}^{\infty} e^{-\,\mathrm{i}\,k\,x}\,dx$$

is, strictly speaking, not defined because the infinite integral of the oscillatory sine and cosine functions doesn't converge! However, this identity can be validly interpreted within the framework of weak convergence and generalized functions. On the other hand, the inverse transform formula (13.86) yields

$$\int_{-\infty}^{\infty} \delta(k)\,e^{\,\mathrm{i}\,k\,x}\,dk = e^{\,\mathrm{i}\,k\,0} = 1,$$

which is in accord with the basic definition (11.36) of the delta function. As in the previous case, the delta function singularity at $k = 0$ manifests the lack of decay of the constant function.

Conversely, the delta function $\delta(x)$ has constant Fourier transform

$$\widehat{\delta}(k) = \frac{1}{\sqrt{2\,\pi}}\int_{-\infty}^{\infty} \delta(x)\,e^{-\,\mathrm{i}\,k\,x}\,dx = \frac{e^{-\,\mathrm{i}\,k0}}{\sqrt{2\,\pi}} \equiv \frac{1}{\sqrt{2\,\pi}}\,. \tag{13.112}$$

a result that also follows from the symmerty principle of Theorem 13.18. To determine the Fourier transform of a delta spike $\delta_y(x) = \delta(x - y)$ concentrated at position $x = y$, we compute

$$\widehat{\delta}_y(k) = \frac{1}{\sqrt{2\,\pi}}\int_{-\infty}^{\infty} \delta(x-y)\,e^{-\,\mathrm{i}\,k\,x}\,dx == \frac{e^{-\,\mathrm{i}\,k\,y}}{\sqrt{2\,\pi}}\,. \tag{13.113}$$

The result is a pure exponential in frequency space. Applying the inverse Fourier transform (13.86) leads, formally, to the remarkable identity

$$\delta_y(x) = \delta(x-y) = \frac{1}{2\,\pi}\int_{-\infty}^{\infty} e^{-\,\mathrm{i}\,k(x-y)}\,dk = \frac{1}{2\,\pi}\,\langle\, e^{\,\mathrm{i}\,k\,y}\,, e^{\,\mathrm{i}\,k\,x}\,\rangle\,, \tag{13.114}$$

where $\langle\,\cdot\,,\cdot\,\rangle$ denotes the usual L^2 inner product on $\mathbb{R}$. Since the delta function vanishes for $x \neq y$, this identity is telling us that complex exponentials of differing frequencies are mutually orthogonal. However, this statement must be taken with a grain of salt, since the integral does not converge in the normal (Riemann or Lebesgue) sense. But it is possible to make sense of this identity within the language of generalized functions. Indeed, multiplying both sides by $f(x)$, and then integrating with respect to x, we find

$$f(y) = \frac{1}{2\,\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x)\,e^{-\,\mathrm{i}\,k(x-y)}\,dx\,dk. \tag{13.115}$$

This *is* a perfectly valid formula, being a restatement (or, rather, combination) of the basic formulae (13.83, 86) connecting the direct and inverse Fourier transforms of the function $f(x)$.

Conversely, the symmetry principle tells us that the Fourier transform of a pure exponential $e^{\mathrm{i}\,l\,x}$ will be a shifted delta spike $\sqrt{2\,\pi}\,\delta(k-l)$, concentrated in frequency space. Both results are particular cases of the general Shift Theorem, whose proof is left as an exercise for the reader.

Proposition 13.19. *If $f(x)$ has Fourier transform $\widehat{f}(k)$, then the Fourier transform of the shifted function $f(x-y)$ is $e^{-\mathrm{i}\,k\,y}\,\widehat{f}(k)$. Similarly, the transform of the product function $e^{\mathrm{i}\,l\,x}\,f(x)$ is the shifted transform $\widehat{f}(k-l)$.*

Since the Fourier transform uniquely associates a function $\widehat{f}(k)$ on frequency space with each (reasonable) function $f(x)$ on physical space, one can characterize functions by their transforms. Practical applications rely on tables (or, even better, computer algebra systems) that recognize a wide variety of transforms of basic functions of importance in applications. The accompanying table lists some of the most important examples of functions and their Fourier transforms. Note that, according to the symmetry principle of Theorem 13.18, each tabular entry can be used to deduce two different Fourier transforms. A more extensive collection of Fourier transforms can be found in [**138**].

Derivatives and Integrals

One of the most remarkable and important properties of the Fourier transform is that it converts calculus into algebra! More specifically, the two basic operations in calculus — differentiation and integration of functions — are realized as algebraic operations on their Fourier transforms. (The downside is that algebraic operations become more complicated in the transform domain.)

Let us begin with derivatives. If we differentiate[†] the basic inverse Fourier transform formula

$$f(x) \;\sim\; \frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{\infty} \widehat{f}(k)\, e^{\mathrm{i}\,k\,x}\, dk.$$

with respect to x, we obtain

$$f'(x) \;\sim\; \frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{\infty} \mathrm{i}\,k\,\widehat{f}(k)\, e^{\mathrm{i}\,k\,x}\, dk. \tag{13.116}$$

The resulting integral is itself in the form of an inverse Fourier transform, namely of $2\,\pi\,\mathrm{i}\,k\,\widehat{f}(k)$ which immediately implies the following basic result.

Proposition 13.20. *The Fourier transform of the derivative $f'(x)$ of a function is obtained by multiplication of its Fourier transform by $\mathrm{i}\,k$:*

$$\widehat{f}'(k) = \mathrm{i}\,k\,\widehat{f}(k). \tag{13.117}$$

[†] We are assuming the integrand is sufficiently nice so that we can bring the derivative under the integral sign; see [**62**, **193**] for a fully rigorous derivation.

Table of Fourier Transforms

$f(x)$	$\widehat{f}(k)$
1	$\sqrt{2\,\pi}\,\delta(k)$
$\delta(x)$	$\dfrac{1}{\sqrt{2\,\pi}}$
$\sigma(x)$	$\sqrt{\dfrac{\pi}{2}}\,\delta(k) - \dfrac{\mathrm{i}}{\sqrt{2\,\pi}\,k}$
$\operatorname{sign} x$	$-\,\mathrm{i}\,\sqrt{\dfrac{2}{\pi}}\,\dfrac{1}{k}$
$\sigma(x+a) - \sigma(x-a)$	$\sqrt{\dfrac{2}{\pi}}\,\dfrac{\sin k\,a}{k}$
$e^{-a\,x}\,\sigma(x)$	$\dfrac{1}{\sqrt{2\,\pi}\,(a + \mathrm{i}\,k)}$
$e^{-a\,\lvert\,x\,\rvert}$	$\sqrt{\dfrac{2}{\pi}}\,\dfrac{a}{k^2 + a^2}$
$e^{-a\,x^2}$	$\dfrac{e^{-k^2/4a}}{\sqrt{2\,a}}$
$\tan^{-1} x$	$-\,\mathrm{i}\,\sqrt{\dfrac{\pi}{2}}\,\dfrac{e^{-\lvert\,k\,\rvert}}{k} + \dfrac{\pi^{3/2}}{\sqrt{2}}\,\delta(k)$
$f(c\,x + d)$	$\dfrac{e^{\mathrm{i}\,k\,d/c}}{\lvert\,c\,\rvert}\,\widehat{f}\left(\dfrac{k}{c}\right)$
$\overline{f(x)}$	$\overline{\widehat{f}(-k)}$
$\widehat{f}(x)$	$f(-k)$
$f'(x)$	$\mathrm{i}\,k\,\widehat{f}(k)$
$f^{(n)}(x)$	$(\mathrm{i}\,k)^n\,\widehat{f}(k)$
$x^n\,f(x)$	$\left(\mathrm{i}\,\dfrac{d}{dk}\right)^n\,\widehat{f}(k)$
$f * g(x)$	$\sqrt{2\,\pi}\,\widehat{f}(k)\,\widehat{g}(k)$

Note: The parameter $a > 0$ is always real and positive, while $c \neq 0$ is any nonzero complex number.

Example 13.21. The derivative of the even exponential pulse $f_e(x) = e^{-a|x|}$ is a multiple of the odd exponential pulse $f_o(x) = (\operatorname{sign} x)\, e^{-a|x|}$:

$$f_e'(x) = -a\,(\operatorname{sign} x)\, e^{-a|x|} = -a f_o(x).$$

Proposition 13.20 says that their Fourier transforms are related by

$$\widehat{f}_o(k) = -\frac{\mathrm{i}\,k}{a}\,\widehat{f}_e(k) = -\,\mathrm{i}\sqrt{\frac{2}{\pi}}\,\frac{k}{k^2+a^2} = -\,\frac{\mathrm{i}\,k}{a}\,\widehat{f}_o(k),$$

as previously noted in (13.98, 101). On the other hand, since the odd exponential pulse has a jump discontinuity of magnitude 2 at $x = 0$, its derivative contains a delta function, and is equal to

$$f_o'(x) = -\,a\,e^{-a|x|} + 2\,\delta(x) = -\,a\,f_e(x) + 2\,\delta(x).$$

This is reflected in the relation between their Fourier transforms. If we multiply (13.101) by $\mathrm{i}\,k$ we obtain

$$\mathrm{i}\,k\,\widehat{f}_o(k) = \sqrt{\frac{2}{\pi}}\,\frac{k^2}{k^2+a^2} = \sqrt{\frac{2}{\pi}} \;-\; \sqrt{\frac{2}{\pi}}\,\frac{a^2}{k^2+a^2} = 2\,\widehat{\delta}(k) - a\,\widehat{f}_e(k).$$

The Fourier transform, just like Fourier series, is completely compatible with the calculus of generalized functions.

Higher order derivatives are handled by iterating formula (13.117), and so:

Corollary 13.22. *The Fourier transform of $f^{(n)}(x)$ is $(\mathrm{i}\,k)^n\,\widehat{f}(k)$.*

This result has an important consequence: the smoothness of $f(x)$ is manifested in the rate of decay of its Fourier transform $\widehat{f}(k)$. We already noted that the Fourier transform of a (nice) function must decay to zero at large frequencies: $\widehat{f}(k) \to 0$ as $|\,k\,| \to \infty$. If the n^{th} derivative $f^{(n)}(x)$ is also a reasonable function, then its Fourier transform $\widehat{f^{(n)}}(k) = (\mathrm{i}\,k)^n\,\widehat{f}(k)$ must go to zero as $|\,k\,| \to \infty$. This requires that $\widehat{f}(k)$ go to zero more rapidly than $|\,k\,|^{-n}$. Thus, the smoother $f(x)$, the more rapid the decay of its Fourier transform. As a general rule of thumb, local features of $f(x)$, such as smoothness, are manifested by global features of $\widehat{f}(k)$, such as decay for large $|\,k\,|$. The symmetry principle implies that reverse is also true: global features of $f(x)$ correspond to local features of $\widehat{f}(k)$. This local-global duality is one of the major themes of Fourier theory.

Integration of functions is the inverse operation to differentiation, and so should correspond to division by $2\pi\,\mathrm{i}\,k$ in frequency space. As with Fourier series, this is not completely correct; there is an extra constant involved, and this contributes an extra delta function in frequency space.

Proposition 13.23. *If $f(x)$ has Fourier transform $\widehat{f}(k)$, then the Fourier transform of its integral $g(x) = \displaystyle\int_{-\infty}^{x} f(y)\,dy$ is*

$$\widehat{g}(k) = \frac{\widehat{f}(k)}{\mathrm{i}\,k} \;+\; c\sqrt{\frac{\pi}{2}}\,\delta(k), \qquad \text{where} \qquad c = \int_{-\infty}^{\infty} f(x)\,dx. \tag{13.118}$$

Proof: First notice that

$$\lim_{x \to -\infty} g(x) = 0, \qquad \lim_{x \to +\infty} g(x) = c = \int_{-\infty}^{\infty} f(x)\, dx.$$

Therefore, by subtracting a suitable multiple of the step function from the integral, the resulting function

$$h(x) = g(x) - c\,\sigma(x)$$

decays to 0 at both $\pm\infty$. Consulting our table of Fourier transforms, we find

$$\widehat{h}(k) = \widehat{g}(k) - c\sqrt{\frac{\pi}{2}}\,\delta(k) + \frac{\mathrm{i}\,c}{\sqrt{2\,\pi}\,k}\,. \tag{13.119}$$

On the other hand,

$$h'(x) = f(x) - c\,\delta(x).$$

Since $h(x) \to 0$ as $|\,x\,| \to \infty$, we can apply our differentiation rule (13.117), and conclude that

$$\mathrm{i}\,k\,\widehat{h}(k) = \widehat{f}(k) - \frac{c}{\sqrt{2\,\pi}}\,. \tag{13.120}$$

Combining (13.119) and (13.120) establishes the desired formula (13.118). *Q.E.D.*

Example 13.24. The Fourier transform of the inverse tangent function

$$f(x) = \tan^{-1} x = \int_0^x \frac{dy}{1+y^2} = \int_{-\infty}^x \frac{dy}{1+y^2} \;-\; \frac{\pi}{2}$$

can be computed by combining Proposition 13.23 with (13.105). Since $\displaystyle\int_{-\infty}^{\infty} \tan^{-1} x\, dx = \pi$, we find

$$\widehat{f}(k) = -\,\mathrm{i}\sqrt{\frac{\pi}{2}}\;\frac{e^{-|\,k\,|}}{k} \;+\; \frac{\pi^{3/2}}{\sqrt{2}}\,\delta(k).$$

Applications to Differential Equations

The fact that the Fourier transform changes differentiation in the physical domain into multiplication in the frequency domain is one of its most compelling features. A particularly important consequence is that the Fourier transform effectively converts differential equations into algebraic equations, and thereby opens the door to their solution by elementary algebra! One begins by applying the Fourier transform to both sides of the differential equation under consideration. Solving the resulting algebraic equation will produce a formula for the Fourier transform of the desired solution, which can then be immediately reconstructed via the inverse Fourier transform.

The Fourier transform is particularly well adapted to boundary value problems on the entire real line. In place of the boundary conditions used on finite intervals, we look for solutions that decay to zero sufficiently rapidly as $|\,x\,| \to \infty$ — in order that their Fourier transform be well-defined (in the context of ordinary functions). In quantum mechanics, [**127**], these solutions are known as the *bound states* of the system, and correspond to subatomic particles that are trapped or localized in a region of space by some sort of

force field. For example, the bound electrons in an atom are localized by the electrostatic attraction of the nucleus.

As a specific example, consider the boundary value problem

$$-\frac{d^2u}{dx^2} + \omega^2 u = h(x), \qquad -\infty < x < \infty, \tag{13.121}$$

where $\omega > 0$ is a positive constant. In lieu of boundary conditions, we require that the solution $u(x) \to 0$ as $|\,x\,| \to \infty$. We will solve this problem by applying the Fourier transform to both sides of the differential equation. Taking Corollary 13.22 into account, the result is a linear algebraic equation

$$k^2\,\widehat{u}(k) + \omega^2\,\widehat{u}(k) = \widehat{h}(k)$$

relating the Fourier transforms of u and h. Unlike the differential equation, the transformed equation can be immediately solved for

$$\widehat{u}(k) = \frac{\widehat{h}(k)}{k^2 + \omega^2} \tag{13.122}$$

Therefore, we can reconstruct the solution by applying the inverse Fourier transform formula (13.86):

$$u(x) = \frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{\infty} \frac{\widehat{h}(k)\,e^{\,\mathrm{i}\,k\,x}}{k^2 + \omega^2}\,dk. \tag{13.123}$$

For example, if the forcing function is an even exponential pulse,

$$h(x) = e^{-|\,x\,|} \qquad \text{with} \qquad \widehat{h}(k) = \sqrt{\frac{2}{\pi}}\,\frac{1}{k^2+1}\,,$$

then (13.123) writes the solution as a Fourier integral:

$$u(x) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{e^{\,\mathrm{i}\,k\,x}}{(k^2+\omega^2)(k^2+1)}\,dk = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\cos k\,x}{(k^2+\omega^2)(k^2+1)}\,dk\,.$$

noting that the imaginary part of the complex integral vanishes because the integrand is an odd function. (Indeed, if the forcing function is real, the solution must also be real.) The Fourier integral can be explicitly evaluated by using partial fractions to rewrite

$$\widehat{u}(k) = \sqrt{\frac{2}{\pi}}\,\frac{1}{(k^2+\omega^2)(k^2+1)} = \sqrt{\frac{2}{\pi}}\,\frac{1}{\omega^2-1}\left(\frac{1}{k^2+1} - \frac{1}{k^2+\omega^2}\right), \qquad \omega^2 \neq 1,$$

Thus, according to our Fourier Transform Table, the solution to this boundary value problem is

$$u(x) = \frac{e^{-|\,x\,|} - \dfrac{1}{\omega}\,e^{-\omega\,|\,x\,|}}{\omega^2 - 1} \qquad \text{when} \qquad \omega^2 \neq 1. \tag{13.124}$$

The reader may wish to verify that this function is indeed a solution, meaning that it is twice continuously differentiable (which is not so immediately apparent from the formula), decays to 0 as $|\,x\,| \to \infty$, and satisfies the differential equation everywhere. The "resonant" case $\omega^2 = 1$ is left as an exercise.

Remark: The method of partial fractions that you learned in first year calculus is often an effective tool for evaluating (inverse) Fourier transforms of rational functions.

A particularly important case is when the forcing function $h(x) = \delta_y(x) = \delta(x - y)$ represents a unit impulse concentrated at $x = y$. The resulting square-integrable solution is the Green's function $G(x, y)$ for the boundary value problem. According to (13.122), its Fourier transform with respect to x is

$$\widehat{G}(k, y) = \frac{1}{\sqrt{2\,\pi}}\, \frac{e^{-\,\mathrm{i}\,k\,y}}{k^2 + \omega^2}\,.$$

which is the product of an exponential factor $e^{-\,\mathrm{i}\,k\,y}$, representing the Fourier transform of $\delta_y(x)$, times a multiple of the Fourier transform of the even exponential pulse $e^{-\,\omega\,|\,x\,|}$. We apply Proposition 13.19, and conclude that the Green's function for this boundary value problem is an exponential pulse centered at y, namely

$$G(x, y) = \frac{1}{2\,\omega}\, e^{-\,\omega\,|\,x - y\,|}\,.$$

Observe that, as with other self-adjoint boundary value problems, the Green's function is symmetric under interchange of x and y. As a function of x, it satisfies the homogeneous differential equation $-u'' + \omega^2\, u = 0$, except at the point $x = y$ when its derivative has a jump discontinuity of unit magnitude. It also decays as $|\,x\,| \to \infty$, as required by the boundary conditions. The Green's function superposition principle tells us that the solution to the inhomogeneous boundary value problem (13.121) under a general forcing can be represented in the integral form

$$u(x) = \int_{-\infty}^{\infty} G(x, y)\, h(y)\, dy = \frac{1}{2\,\omega} \int_{-\infty}^{\infty} e^{-\,\omega\,|\,x - y\,|}\, h(y)\, dy. \tag{13.125}$$

The reader may enjoy recovering the particular exponential solution (13.124) from this integral formula.

Convolution

The final Green's function formula (13.125) is indicative of a general property of Fourier transforms. The right hand side has the form of a *convolution product* between functions.

Definition 13.25. The *convolution* of scalar functions $f(x)$ and $g(x)$ is the scalar function $h = f * g$ defined by the formula

$$h(x) = f(x) * g(x) = \int_{-\infty}^{\infty} f(x - y)\, g(y)\, dy. \tag{13.126}$$

We record the basic properties of the convolution product, leaving their verification as exercises for the reader. All of these assume that the implied convolution integrals converge.

(a) *Symmetry*: $\qquad f * g = g * f,$

(b) *Bilinearity*: $\begin{cases} f * (a\,g + b\,h) = a\,(f * g) + b\,(f * h), \\ (a\,f + b\,g) * h = a\,(f * h) + b\,(g * h), \end{cases}$ $a, b \in \mathbb{C}$,

(c) *Associativity*: $f * (g * h) = (f * g) * h$,

(d) *Zero function*: $f * 0 = 0$,

(e) *Delta function*: $f * \delta = f$.

One tricky feature is that the constant function 1 is *not* a unit for the convolution product; indeed,

$$f * 1 = \int_{-\infty}^{\infty} f(y)\, dy$$

is a constant function — the total integral of f — not the original function $f(x)$. In fact, according to the last property, the delta function plays the role of the "convolution unit":

$$f(x) * \delta(x) = \int_{-\infty}^{\infty} f(x - y)\, \delta(y)\, dy = f(x),$$

which follows from the basic property (11.36) of the delta function.

In particular, our solution (13.125) has the form of a convolution product between an even exponential pulse $g(x) = \dfrac{1}{2\omega}\, e^{-\omega |\, x\,|}$ and the forcing function:

$$u(x) = g(x) * h(x).$$

On the other hand, its Fourier transform (13.122) is the ordinary multiplicative product

$$\widehat{u}(k) = \widehat{g}(k)\, \widehat{h}(k)$$

of the Fourier transforms of g and h. In fact, this is a general property of the Fourier transform: convolution in the physical domain corresponds to multiplication in the frequency domain, and conversely.

Theorem 13.26. *The Fourier transform of the convolution $u(x) = f * g(x)$ of two functions is a multiple of the product of their Fourier transforms*:

$$\widehat{u}(k) = \sqrt{2\,\pi}\, \widehat{f}(k)\, \widehat{g}(k).$$

Vice versa, the Fourier transform of their product $h(x) = f(x)\, g(x)$ is, up to multiple, the convolution of their Fourier transforms:

$$\widehat{h}(k) = \frac{1}{\sqrt{2\,\pi}}\, \widehat{f} * \widehat{g}(k).$$

Proof: Combining the definition of the Fourier transform with the convolution formula (13.126), we find

$$\widehat{u}(k) = \frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{\infty} h(x)\, e^{-\,\mathrm{i}\,k\,x}\, dx = \frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x - y)\, g(y)\, e^{-\,\mathrm{i}\,k\,x}\, dx\, dy.$$

where we are assuming that the integrands are sufficiently nice to allow us to interchange the order of integration, [**9**]. Applying the change of variables $z = x - y$ in the inner integral produces

$$\begin{aligned}\widehat{u}(k) &= \frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(z)\, g(y)\, e^{-\,\mathrm{i}\, k(y+z)}\, dy\, dz \\ &= \sqrt{2\,\pi}\, \left(\frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{\infty} f(z)\, e^{-\,\mathrm{i}\, k\, z}\, dz \right) \left(\frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{\infty} g(y)\, e^{-\,\mathrm{i}\, k\, y}\, dy \right) = \sqrt{2\,\pi}\, \widehat{f}(k)\, \widehat{g}(k).\end{aligned}$$

The second statement follows directly from the symmetry principle of Theorem 13.18. *Q.E.D.*

Example 13.27. We already know, (13.106), that the Fourier transform of

$$f(x) = \frac{\sin x}{x}$$

is the box function

$$\widehat{f}(k) = \sqrt{\frac{\pi}{2}}\, \big[\, \sigma(k+1) - \sigma(k-1)\, \big] = \begin{cases} \sqrt{\dfrac{\pi}{2}}\,, & -1 < k < 1, \\ 0, & |\, k\,| > 1, \end{cases}$$

We also know that the Fourier transform of

$$g(x) = \frac{1}{x} \qquad \text{is} \qquad \widehat{g}(k) = -\,\mathrm{i}\, \sqrt{\frac{\pi}{2}}\, \operatorname{sign} k.$$

Therefore, the Fourier transform of their product

$$h(x) = f(x)\, g(x) = \frac{\sin x}{x^2}$$

can be obtained by convolution:

$$\begin{aligned}\widehat{h}(k) &= \frac{1}{\sqrt{2\,\pi}}\, \widehat{f} * \widehat{g}(k) = \frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{\infty} \widehat{f}(l)\, \widehat{g}(k-l)\, dl \\ &= -\,\mathrm{i}\, \sqrt{\frac{\pi}{8}} \int_{-1}^{1} \operatorname{sign}(k-l)\, dl = \begin{cases} \mathrm{i}\, \sqrt{\dfrac{\pi}{2}}, & k < -1, \\ -\,\mathrm{i}\, \sqrt{\dfrac{\pi}{2}}\, k, & -1 < k < 1, \\ -\,\mathrm{i}\, \sqrt{\dfrac{\pi}{2}}, & k > 1. \end{cases}\end{aligned}$$

A graph of $\widehat{h}(k)$ appears in Figure 13.15.

The Fourier Transform on Hilbert Space

While we do not have the space to embark on a fully rigorous treatment of the theory underlying the Fourier transform, it is worth outlining a few of the most basic features.

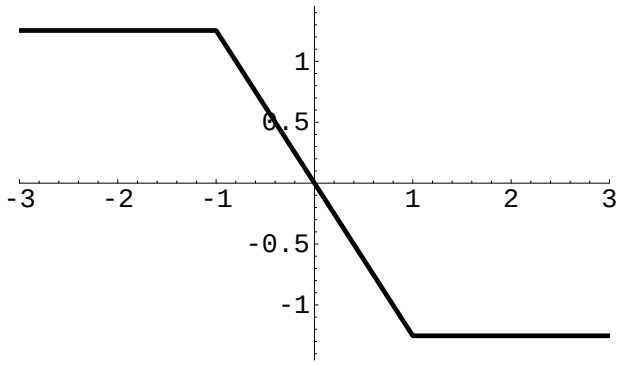


Figure 13.15. The Fourier transform of $\dfrac{\sin x}{x^2}$.

We have already noted that the Fourier transform, when defined, is a linear map, taking functions $f(x)$ on physical space to functions $\widehat{f}(k)$ on frequency space. A critical question is precisely which function space should the theory be applied to. Not every function admits a Fourier transform in the classical sense[†] — the Fourier integral (13.83) is required to converge, and this places restrictions on the function and its asymptotics at large distances.

It turns out the proper setting for the rigorous theory is the *Hilbert space* of complex-valued square-integrable functions — the same infinite-dimensional vector space that lies at the heart of modern quantum mechanics. In Section 12.5, we already introduced the Hilbert space $L2[-\pi, \pi]$ on a finite interval; here we adapt Definition 12.33 to the entire real line. Thus, the Hilbert space $\mathrm{L}^2 = \mathrm{L}^2(\mathbb{R})$ is the infinite-dimensional vector space consisting of all complex-valued functions $f(x)$ which are defined for all $x \in \mathbb{R}$ and have finite L^2 norm:

$$\| f \|^2 = \int_{-\infty}^{\infty} | f(x) |^2 \, dx < \infty. \tag{13.127}$$

For example, any piecewise continuous function that satisfies the decay criterion

$$| f(x) | \leq \frac{M}{| x |^{1/2+\delta}}, \qquad \text{for all sufficiently large} \qquad | x | \gg 0, \tag{13.128}$$

for some $M > 0$ and $\delta > 0$, belongs to L^2. However, as in Section 12.5, Hilbert space contains many more functions, and the precise definitions and identification of its elements is quite subtle. The inner product on the Hilbert space L^2 is prescribed in the usual manner,

$$\langle f , g \rangle = \int_{-\infty}^{\infty} f(x) \, \overline{g(x)} \, dx.$$

The Cauchy–Schwarz inequality

$$| \langle f , g \rangle | \leq \| f \| \, \| g \|$$

ensures that the inner product integral is finite whenever $f, g \in \mathrm{L}^2$.

[†] We leave aside the more advanced issues involving generalized functions in this subsection.

Let us state the fundamental theorem governing the effect of the Fourier transform on functions in Hilbert space. It can be regarded as a direct analog of the pointwise convergence Theorem 12.7 for Fourier series. A fully rigorous proof can be found in [**173**, **193**].

Theorem 13.28. *If $f(x) \in \mathrm{L}^2$ is square-integrable, then its Fourier transform $\widehat{f}(k) \in \mathrm{L}^2$ is a well-defined, square-integrable function of the frequency variable k. If $f(x)$ is continuously differentiable at a point x, then its inverse Fourier transform* (13.86) *equals its value $f(x)$. More generally, if the right and left hand limits $f(x^-)$, $f'(x^-)$, and $f(x^+)$, $f'(x^+)$ exist, then the inverse Fourier transform integral converges to the average value $\frac{1}{2}\big[\, f(x^-) + f(x^+) \,\big]$.*

Thus, the Fourier transform $\widehat{f} = \mathcal{F}[\, f\,]$ defines a linear transformation from L^2 functions of x to L^2 functions of k. In fact, the Fourier transform preserves inner products. This important result is known as *Parseval's formula*, whose Fourier series counterpart appeared in (12.116).

Theorem 13.29. *If $\widehat{f}(k) = \mathcal{F}[\, f(x)\,]$ and $\widehat{g}(k) = \mathcal{F}[\, g(x)\,]$, then $\langle\, f\,, g\,\rangle = \langle\, \widehat{f}\,, \widehat{g}\,\rangle$, i.e.,*

$$\int_{-\infty}^{\infty} f(x)\, \overline{g(x)}\, dx = \int_{-\infty}^{\infty} \widehat{f}(k)\, \overline{\widehat{g}(k)}\, dk. \tag{13.129}$$

Proof: Let us sketch a formal proof that serves to motivate why this result is valid. We use the definition (13.83) of the Fourier transform to evaluate

$$\begin{aligned} \int_{-\infty}^{\infty} \widehat{f}(k)\, \overline{\widehat{g}(k)}\, dk &= \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{\infty} f(x)\, e^{-\,\mathrm{i}\,k\,x}\, dx \right) \left(\frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{\infty} \overline{g(y)}\, e^{+\,\mathrm{i}\,k\,y}\, dy \right) dk \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x)\, \overline{g(y)} \left(\frac{1}{2\,\pi} \int_{-\infty}^{\infty} e^{-\,\mathrm{i}\,k(x-y)}\, dk \right) dx\, dy. \end{aligned}$$

Now according to (13.114), the inner k integral can be replaced by a delta function $\delta(x-y)$, and hence

$$\int_{-\infty}^{\infty} \widehat{f}(k)\, \overline{\widehat{g}(k)}\, dk = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x)\, \overline{g(y)}\, \delta(x-y)\, dx\, dy = \int_{-\infty}^{\infty} f(x)\, \overline{g(x)}\, dx.$$

This completes our "proof"; see [**173**, **193**] for a rigorous version. *Q.E.D.*

In particular, orthogonal functions, $\langle\, f\,, g\,\rangle = 0$, will have orthogonal Fourier transforms, $\langle\, \widehat{f}\,, \widehat{g}\,\rangle = 0$. Choosing $f = g$ in (13.129) results in the *Plancherel formula* $\| f \|^2 = \| \widehat{f}\, \|^2$, or, explicitly,

$$\int_{-\infty}^{\infty} |\, f(x\,|^2\, dx = \int_{-\infty}^{\infty} |\, \widehat{f}(k)\,|^2\, dk. \tag{13.130}$$

We conclude that the Fourier transform map $\mathcal{F}\colon \mathrm{L}^2 \to \mathrm{L}^2$ defines a *unitary* or norm-preserving linear transformation on Hilbert space, mapping L^2 functions of the physical variable x to L^2 functions of the frequency variable k.

Quantum Mechanics and the Uncertainty Principle

In quantum mechanics, the wave functions or probability densities of a quantum system are characterized as unit elements, $\| \varphi \| = 1$ of the underlying state space, which, in a one-dimensional model of a single particle, is the Hilbert space $\mathrm{L}^2 = \mathrm{L}^2(\mathbb{R})$. All measurable physical quantities are represented as linear operators $A \colon \mathrm{L}^2 \to \mathrm{L}^2$ acting on (a suitable dense subspace of) the state space. These are obtained by the tahter myusterious process of "quantizing" their classical counterparts, which are ordinary functions. In particular, the particle's *position*, usually denoted by Q, is represented by the operation of multiplication by x, so $Q[\varphi] = x\,\varphi(x)$, whereas its *momentum*, denoted by P for historical reasons, is represented by the differentiation operator $P = d/dx$, so that $P[\varphi] = \varphi'(x)$.

In its popularized form, Heisenberg's Uncertainty Principle is a by now familiar philosophical concept. The principle, first formulated by the twentieth century German physicist Werner Heisenberg, one of the founders of modern quantum mechanics, states that, in a physical system, certain quantities cannot be simultaneously measured with complete accuracy. For instance, the more precisely one measures the position of a particle, the less accuracy there will be in the measurement of its momentum; vice versa, the greater the accuracy in the momentum, the less certainty in its position. A similar uncertainty couples energy and time. Experimental verification of the uncertainty principle can be found even in fairly simple situations. Consider a light beam passing through a small hole. The position of the photons is constrained by the hole; the effect of their momenta is in the pattern of light diffused on a screen placed beyond the hole. The smaller the hole, the more constrained the position, and the wider the image on the screen, meaning the less certainty there is in the momentum.

This is not the place to discuss the philosophical and experimental consequences of Heisenberg's principle. What we will show is that the Uncertainty Principle is, in fact, a rigorous theorem concerning the Fourier transform! In quantum theory, each of the paired quantities, e.g., position and momentum, are interrelated by the Fourier transform. Indeed, Proposition 13.20 says that the Fourier transform of the differentiation operator representing monmentum is a multiplication operator representing position on the frequency space and vice versa. This Fourier transform-based duality between position and momentum, or multiplication and differentiation, lies at the heart of the Uncertainty Principle.

In general, suppose A is a linear operator on Hilbert space representing a physical quantity. If the quantum system is in a state represented by a particular wave function φ, then the *localization* of the quantity A is measured by the norm $\| A[\varphi] \|$. The smaller this norm, the more accurate the measurement. Thus, $\| Q[\varphi] \| = \| x\,\varphi(x) \|$ measures the localization of the position of the particle represented by φ; the smaller $\| Q[\varphi] \|$, the more concentrated the probability of finding the particle near[†] $x = 0$ and hence the smaller the error in the measurement of its position. Similarly, by Plancherel's formula (13.130),

$$\| P[\varphi] \| = \| \varphi'(x) \| = \| \mathrm{i}\, k\, \widehat{\varphi}(k) \| = \| k\, \widehat{\varphi}(k) \|,$$

[†] At another position a, one replaces x by $x - a$.

measures the localization in the momentum of the particle, which is small if and only if its Fourier transform is concentrated near $k = 0$, and hence the smaller the error in its measured momentum. With this interpretation, the Uncertainty Principle states that these two quantities cannot simultaneously be arbitrarily small.

Theorem 13.30. *If $\varphi(x)$ is a wave function, so $\| \varphi \| = 1$, then*

$$\| Q[\varphi] \| \, \| P[\varphi] \| \geq \tfrac{1}{2}. \tag{13.131}$$

Proof: The proof rests on the Cauchy–Schwarz inequality

$$\Big| \langle \, x \, \varphi(x) \, , \varphi'(x) \, \rangle \Big| \leq \| \, x \, \varphi(x) \, \| \, \| \, \varphi'(x) \, \| = \| \, Q[\varphi] \, \| \, \| \, P[\varphi] \, \|. \tag{13.132}$$

On the other hand, writing out the inner product term

$$\langle \, x \, \varphi(x) \, , \varphi'(x) \, \rangle = \int_{-\infty}^{\infty} x \, \varphi(x) \, \varphi'(x) \, dx.$$

Let us integrate by parts, using the fact that

$$\varphi(x) \, \varphi'(x) = \frac{d}{dx} \left[\, \tfrac{1}{2} \, \varphi(x)^2 \, \right].$$

Since $\varphi(x) \to 0$ as $| \, x \, | \to \infty$, the boundary terms vanish, and hence

$$\langle \, x \, \varphi(x) \, , \varphi'(x) \, \rangle = \int_{-\infty}^{\infty} x \, \varphi(x) \, \varphi'(x) \, dx = - \int_{-\infty}^{\infty} \tfrac{1}{2} \, \varphi(x)^2 \, dx = - \tfrac{1}{2},$$

since $\| \varphi \| = 1$. Substituting back into (13.132) completes the proof. *Q.E.D.*

The inequality (13.131) quantifies the statement that the more accurately we measure the momentum Q, the less accurately we are able to measure the position P, and vice versa. For more details and physical consequences, you should consult an introductory text on mathematical quantum mechanics, e.g., [**122**, **127**].

13.4. The Laplace Transform.

In engineering applications, the Fourier transform is often overshadowed by a close relative. The Laplace transform plays an essential role in control theory, linear systems analysis, electronics, and many other fields of practical engineering and science. However, the Laplace transform is most properly interpreted as a particular real form of the more fundamental Fourier transform. When the Fourier transform is evaluated along the imaginary axis, the complex exponential factor becomes real, and the result is the Laplace transform, which maps real-valued functions to real-valued functions. Since it is so closely allied to the Fourier transform, the Laplace transform enjoys many of its featured properties, including linearity. Moreover, derivatives are transformed into algebraic operations, which underlies its applications to solving differential equations. The Laplace transform is one-sided; it only looks forward in time and prefers functions that decay — transients. The Fourier transform looks in both directions and prefers oscillatory functions. For this

reason, while the Fourier transform is used to solve boundary value problems on the real line, the Laplace transform is much better adapted to initial value problems.

Since we will be applying the Laplace transform to initial value problems, we switch our variable from x to t to emphasize this fact. Suppose $f(t)$ is a (reasonable) function which vanishes on the negative axis, so $f(t) = 0$ for all $t < 0$. The Fourier transform of f is

$$\widehat{f}(k) = \frac{1}{\sqrt{2\,\pi}} \int_0^\infty f(t)\, e^{-\,\mathrm{i}\,k\,t}\, dt,$$

since, by our assumption, its negative t values make no contribution to the integral. The *Laplace transform* of such a function is obtained by replacing $\mathrm{i}\,k$ by a real[†] variable s, leading to

$$F(s) = \mathcal{L}[\,f(t)\,] = \int_0^\infty f(t)\, e^{-s\,t}\, dt, \tag{13.133}$$

where, in accordance with the standard convention, the factor of $\sqrt{2\,\pi}$ has been omitted. By allowing complex values of the Fourier frequency variable k, we may identify the Laplace transform with $\sqrt{2\,\pi}$ times the evaluation of the Fourier transform for values of $k = -\,\mathrm{i}\,s$ on the imaginary axis:

$$F(s) = \sqrt{2\,\pi}\; \widehat{f}(-\,\mathrm{i}\,s). \tag{13.134}$$

Since the exponential factor in the integral has become real, the Laplace transform $\mathcal{L}$ takes real functions to real functions. Moreover, since the integral kernel $e^{-s\,t}$ is exponentially decaying for $s > 0$, we are no longer required to restrict our attention to functions that decay to zero as $t \to \infty$.

Example 13.31. Consider an exponential function $f(t) = e^{\alpha\,t}$, where the exponent α is allowed to be complex. Its Laplace transform is

$$F(s) = \int_0^\infty e^{(\alpha - s)\,t}\, dt = \frac{1}{s - \alpha}\,. \tag{13.135}$$

Note that the integrand is exponentially decaying, and hence the integral converges, if and only if $\mathrm{Re}\,(\alpha - s) < 0$. Therefore, the Laplace transform (13.135) is, strictly speaking, only defined at sufficiently large $s > \mathrm{Re}\,\alpha$. In particular, for an oscillatory exponential,

$$\mathcal{L}[\,e^{\,\mathrm{i}\,\omega\,t}\,] = \frac{1}{s - \mathrm{i}\,\omega} = \frac{s + \mathrm{i}\,\omega}{s^2 + \omega^2} \qquad \text{provided} \qquad s > 0.$$

Taking real and imaginary parts of this identity, we discover the formulae for the Laplace transforms of the simplest trigonometric functions:

$$\mathcal{L}[\cos\omega\,t\,] \;=\; \frac{s}{s^2 + \omega^2}\,, \qquad\qquad \mathcal{L}[\sin\omega\,t\,] \;=\; \frac{\omega}{s^2 + \omega^2}\,. \tag{13.136}$$

[†] One can also define the Laplace transform at complex values of s, but this will not be required in the applications discussed here.

Two additional important transforms are

$$\mathcal{L}[1] = \int_0^\infty e^{-st}\,dt = \frac{1}{s}\,, \qquad \mathcal{L}[t] = \int_0^\infty t\,e^{-st}\,dt = \frac{1}{s}\int_0^\infty e^{-st}\,dt = \frac{1}{s^2}\,. \tag{13.137}$$

The second computation relies on an integration by parts, making sure that the boundary terms at $s = 0, \infty$ vanish.

Remark: In every case, we really mean the Laplace transform of the function whose values are given for $t > 0$ and is equal to 0 for all negative t. Therefore, the function 1 in reality signifies the step function

$$\sigma(t) = \begin{cases} 1, & t > 0, \\ 0, & t < 0, \end{cases} \tag{13.138}$$

and so the first formula in (13.137) should more properly be written

$$\mathcal{L}[\sigma(t)] = \frac{1}{s}\,. \tag{13.139}$$

However, in the traditional approach to the Laplace transform, one only considers the functions on the positive t axis, and so the step function and the constant function are, from this viewpoint, indistinguishable. However, once one moves beyond a purely mechanistic approach, any deeper understanding of the properties of the Laplace transform requires keeping this distinction firmly in mind.

Let us now pin down the precise class of functions to which the Laplace transform can be applied.

Definition 13.32. A function $f(t)$ is said to have *exponential growth* of *order* a if

$$|\,f(t)\,| < M\,e^{a\,t}, \qquad \text{for all} \qquad t > t_0, \tag{13.140}$$

for some $M > 0$ and $t_0 > 0$.

Note that the exponential growth condition only depends upon the function's behavior for large values of t. If $a < 0$, then f is, in fact, exponentially decaying as $x \to \infty$. Since $e^{a\,t} < e^{b\,t}$ for $a < b$ and all $t > 0$, if $f(t)$ has exponential growth of order a, it automatically has exponential growth of any higher order $b > a$. All polynomial, trigonometric, and exponential functions (with linear argument) have exponential growth. The simplest example of a function that does not satisfy any exponential growth bound is $f(t) = e^{t^2}$, since it grows faster than any simple exponential $e^{a\,t}$.

The following result guarantees the existence of the Laplace transform, at least for sufficiently large values of the transform variable s, for a rather broad class of functions that includes almost all of the functions that arise in applications.

Theorem 13.33. *If $f(t)$ is piecewise continuous and has exponential growth of order a, then its Laplace transform $F(s) = \mathcal{L}[\,f(t)\,]$ is defined for all $s > a$.*

Table of Laplace Transforms

$f(t)$	$F(s)$
1	$\dfrac{1}{s}$
t	$\dfrac{1}{s^2}$
t^n	$\dfrac{n!}{s^{n+1}}$
$\delta(t-c)$	$e^{-s\,c}$
$\sigma(t-c)$	$\dfrac{e^{-s\,c}}{s}$
$e^{\alpha\,t}$	$\dfrac{1}{s-\alpha}$
$\cos\omega\,t$	$\dfrac{s}{s^2+\omega^2}$
$\sin\omega\,t$	$\dfrac{\omega}{s^2+\omega^2}$
$e^{c\,t}\,f(t)$	$F(s-c)$
$\dfrac{f(t)}{t}$	$\displaystyle\int_s^\infty F(r)\,dr$
$f(t-c)$	$e^{-s\,c}\,F(s)$
$f'(t)$	$s\,F(s)-f(0)$
$f^{(n)}(t)$	$s^n\,F(s)-s^{n-1}\,f(0)-$ $-s^{n-2}\,f'(0)-\cdots-f^{(n-1)}(0)$
$f(t)*g(t)$	$F(s)\,G(s)$

Proof: The exponential growth inequality (13.140) implies that we can bound the integrand in (13.133) by $| f(t)\, e^{-s\,t} | < M\, e^{(a-s)\,t}$. Therefore, as soon as $s > a$, the integrand is exponentially decaying as $t \to \infty$, and this suffices to ensure the convergence of the Laplace transform integral. *Q.E.D.*

Theorem 13.33 is an existential result, and of course, in practice, we may not be able to explicitly evaluate the Laplace transform integral. Nevertheless, the Laplace transforms of most common functions are not hard to find, and extensive lists have been tabulated, [**139**]. An abbreviated table of Laplace transforms can be found on the following page. Nowadays, the most convenient sources of transform formulas are computer algebra packages, including MATHEMATICA and MAPLE.

According to Theorem 13.28, when it exists, the Fourier transform uniquely specifies the function, except possibly at jump discontinuities where the limiting value must be half way in between. An analogous result can be established for the Laplace transform.

Lemma 13.34. *If $f(t)$ and $g(t)$ are piecewise continuous functions that are of exponential growth, and $\mathcal{L}[\,f(t)\,] = \mathcal{L}[\,g(t)\,]$ for all s sufficiently large, then $f(t) = g(t)$ at all points of continuity $t > 0$.*

In fact, there is an explicit formula for the inverse Laplace transform, which follows from its identification, (13.134), with the Fourier transform along the imaginary axis. Under suitable hypotheses, a given function $F(s)$ is the Laplace transform of the function $f(t)$ determined by the complex integral formula[†]

$$f(t) = \frac{1}{2\,\pi\,\mathrm{i}} \int_{-\,\mathrm{i}\,\infty}^{\mathrm{i}\,\infty} F(s)\, e^{s\,t}\, ds, \qquad\qquad t > 0. \tag{13.141}$$

In practice, one hardly ever uses this complicated formula to compute the inverse Laplace transform. Rather, one simply relies on tables of known Laplace transforms, coupled with a few basic rules that will be covered in the following subsection.

The Laplace Transform Calculus

The first observation is that the Laplace transform is a linear operator, and so

$$\mathcal{L}[\,f + g\,] = \mathcal{L}[\,f\,] + \mathcal{L}[\,g\,], \qquad\qquad \mathcal{L}[\,c\,f\,] = c\,\mathcal{L}[\,f\,], \tag{13.142}$$

for any constant c. Moreover, just like its Fourier progenitor, the Laplace transform converts calculus into algebra. In particular, differentiation turns into multiplication by the transform variable, but with one additional term that depends upon the value of the function at $t = 0$.

[†] See Section 16.5 for details on complex integration. The stated formula doesn't apply to all functions of exponential growth. A more universally valid inverse Laplace transform formula is obtained by shifting the complex contour to run from $b - \mathrm{i}\,\infty$ to $b + \mathrm{i}\,\infty$ for some $b > a$, the order of exponential growth of f.

Theorem 13.35. *Let $f(t)$ be continuously differentiable for $t > 0$ and have exponential growth of order a. If $\mathcal{L}[\,f(t)\,] = F(s)$ then, for $s > a$,*

$$\mathcal{L}[\,f'(t)\,] = s\,F(s) - f(0). \tag{13.143}$$

Proof: The proof relies on an integration by parts:

$$\begin{aligned}\mathcal{L}[\,f'(t)(t)\,] &= \int_0^\infty f'(t)\,e^{-\,s\,t}\,dt = \,f(t)\,e^{-\,s\,t}\,\Big|_{t=0}^{\infty} + s\int_0^\infty f(t)\,e^{-\,s\,t}\,dt\\ &= \lim_{t\,\to\,\infty}\,f(t)\,e^{-\,s\,t} - f(0) + s\,F(s).\end{aligned}$$

The exponential growth inequality (13.140) implies that first term vanishes for $s > a$, and the remaining terms agree with (13.143). *Q.E.D.*

Example 13.36. According to Example 13.31, the Laplace transform of the function $\sin\omega\,t$ is

$$\mathcal{L}[\sin\omega\,t] = \frac{\omega}{s^2+\omega^2}\,.$$

Its derivative is

$$\frac{d}{dt}\sin\omega\,t = \omega\cos\omega\,t,$$

and therefore

$$\mathcal{L}[\omega\cos\omega\,t] = s\,\mathcal{L}[\sin\omega\,t] = \frac{\omega\,s}{s^2+\omega^2}\,,$$

since $\sin\omega\,t$ vanishes at $t = 0$. The result agrees with (13.136). On the other hand,

$$\frac{d}{dt}\cos\omega\,t = -\,\omega\sin\omega\,t,$$

and so

$$\mathcal{L}[-\,\omega\sin\omega\,t\,] \;=\; s\,\mathcal{L}[\cos\omega\,t] - 1 = \frac{s^2}{s^2+\omega^2} - 1 = -\,\frac{\omega^2}{s^2+\omega^2}\,,$$

again in agreement with the known formula.

Remark: The final term $-\,f(0)$ in (13.143) is a manifestation of the discontinuity in $f(t)$ at $t = 0$. Keep in mind that the Laplace transform only applies to functions with $f(t) = 0$ for all $t < 0$, and so if $f(0) \neq 0$, then $f(t)$ has a jump discontinuity of magnitude $f(0)$ at $t = 0$. Therefore, by the calculus of generalized functions, its derivative $f'(t)$ should include a delta function term, namely $f(0)\,\delta(0)$, which accounts for the additional constant term in its transform. In the practical approach to the Laplace transform calculus, one suppresses the delta function when computing the derivative $f'(t)$. However, its effect must reappear on the other side of the differentiation formula (13.143), and the upshot is the extra term $-\,f(0)$.

Laplace transforms of higher order derivatives are found by iterating the first order formula (13.143). For example, if $f \in \mathrm{C}^2$, then

$$\mathcal{L}[\, f''(t)\,] = s\,\mathcal{L}[\, f'(t)\,] - f'(0) = s^2\, F(s) - s\, f(0) - f'(0). \tag{13.144}$$

In general, for an n times continuously differentiable function,

$$\mathcal{L}\big[\, f^{(n)}\,\big] = s^n F(s) - s^{n-1} f(0) - s^{n-2} f'(0) - \,\cdots\, - f^{(n-1)}(0). \tag{13.145}$$

Conversely, integration corresponds to dividing the Laplace transform by s, so

$$\mathcal{L}\left[\, \int_0^t f(\tau)\, d\tau \,\right] = \frac{F(s)}{s}\,. \tag{13.146}$$

Unlike the Fourier transform, there are no additional terms in the integration formula as long as we start the integral at $t = 0$. For instance,

$$\mathcal{L}[\, t^2\,] = \frac{1}{s}\,\mathcal{L}[\, 2\,t\,] = \frac{2}{s^3}\,, \qquad \text{and, more generally,} \qquad \mathcal{L}[\, t^n\,] = \frac{n!}{s^{n+1}}\,. \tag{13.147}$$

There is also a shift formula, analogous to Proposition 13.19 for Fourier transforms, but with one important caveat. Since all functions must vanish for $t < 0$, we are only allowed to shift them to the right, a shift to the left would produce nonzero function values for some $t < 0$. In general, the Laplace transform of the function $f(t - c)$ shifted to the right by an amount $c > 0$, is

$$\begin{aligned} \mathcal{L}[\, f(t-c)\,] &= \int_0^\infty f(t-c)\, e^{-\,s\,t}\, dt = \int_{-\,c}^\infty f(t)\, e^{-\,s\,(t+c)}\, dt \\ &= \int_{-\,c}^0 f(t)\, e^{-\,s\,(t+c)}\, dt + \int_0^\infty f(t)\, e^{-\,s\,(t+c)}\, dt = e^{-\,s\,c} \int_0^\infty f(t)\, e^{-\,s\,t}\, dt = e^{-\,s\,c}\, F(s). \end{aligned} \tag{13.148}$$

In this computation, we first used a change of variables in the integral, replacing $t - c$ by t; then, the fact that $f(t) \equiv 0$ for $t < 0$ was used to eliminate the integral from $-c$ to 0. When using the shift formula in practice, it is important to keep in mind that $c > 0$ and the function $f(t - c)$ vanishes for all $t < c$.

Example 13.37. Consider the square wave pulse $f(t) = \begin{cases} 1, & b < t < c, \\ 0, & \text{otherwise,} \end{cases}$ for some $0 < b < c$. To compute its Laplace transform, we write it as the difference

$$f(t) = \sigma(t - b) - \sigma(t - c)$$

of shifted versions of the step function (13.138). Combining the shift formula (13.148) and the formula (13.139) for the Laplace transform of the step function, we find

$$\mathcal{L}[\, f(t)\,] = \mathcal{L}[\, \sigma(t-b)\,] - \mathcal{L}[\, \sigma(t-c)\,] = \frac{e^{-\,s\,b} - e^{-\,s\,c}}{s}\,. \tag{13.149}$$

We already noted that the Fourier transform of the convolution product of two functions is realized as the ordinary product of their individual transforms. A similar result

holds for the Laplace transform. Let $f(t), g(t)$ be given functions. Since we are implicitly assuming that the functions vanish at all negative values of t, their convolution product (13.126) reduces to a finite integral

$$h(t) = f(t) * g(t) = \int_0^t f(t-\tau)\, g(\tau)\, d\tau. \tag{13.150}$$

In particular $h(t) = 0$ for all $t < 0$ also. Further, it is not hard to show that the convolution of two functions of exponential growth also has exponential growth.

Theorem 13.38. *If $\mathcal{L}[f(t)] = F(s)$ and $\mathcal{L}[g(t)] = G(s)$, then the convolution $h(t) = f(t) * g(t)$ has Laplace transform given by the product $H(s) = F(s)\, G(s)$.*

The proof of the convolution theorem for the Laplace transform proceeds along the same lines as its Fourier transform version Theorem 13.26, and is left as Exercise ∎ for the reader.

Applications to Initial Value Problems

The key application of the Laplace transform is to facilitate the solution of initial value problems for linear, constant coefficient ordinary differential equations. As a prototypical example, consider the second order initial value problem

$$a\,\frac{d^2u}{dt^2} + b\,\frac{du}{dt} + c\,u = f(t), \qquad u(0) = \alpha, \qquad \frac{du}{dt}\,(0) = \beta, \tag{13.151}$$

in which a, b, c are constant. We will solve the initial value problem by applying the Laplace transform to both sides of the differential equation. In view of the differentiation formulae (13.143, 144),

$$a\big(\, s^2\mathcal{L}[\,u(t)\,] - s\,u(0) - \dot u(0)\,\big) + b\big(\, s\,\mathcal{L}[\,u(t)\,] - u(0)\,\big) + c\,\mathcal{L}[\,u(t)\,] = \mathcal{L}[\,f(t)\,].$$

Setting $\mathcal{L}[\,u(t)\,] = U(s)$ and $\mathcal{L}[\,f(t)\,] = F(s)$, and making use of the initial conditions, the preceding equation takes the form

$$(a\,s^2 + b\,s + c)\,U(s) = F(s) + (a\,s + b)\,\alpha + a\,\beta. \tag{13.152}$$

Thus, by applying the Laplace transform, we have effectively reduced the entire initial value problem to a single elementary algebraic equation! Solving for

$$U(s) = \frac{F(s) + (a\,s + b)\,\alpha + a\,\beta}{a\,s^2 + b\,s + c}\,, \tag{13.153}$$

we then recover solution $u(t)$ to the initial value problem by finding the inverse Laplace transform of $U(s)$. As noted earlier, in practice the inverse transform is found by suitably massaging the answer (13.153) to be a combination of known transforms.

Example 13.39. Consider the initial value problem

$$\ddot u + u = 10\, e^{-3\,t}, \qquad u(0) = 1, \qquad \dot u(0) = 2.$$

Taking the Laplace transform of the differential equation as above, we find

$$(s^2+1)\,U(s) - s - 2 = \frac{10}{s+3}\,, \qquad \text{and so} \qquad U(s) = \frac{s+2}{s^2+1} + \frac{10}{(s+3)(s^2+1)}.$$

The second summand does not directly correspond to any of the entries in our table of Laplace transforms. However, we can use the method of partial fractions to write it as a sum

$$U(s) = \frac{s+2}{s^2+1} + \frac{1}{s+3} + \frac{3-s}{s^2+1} = \frac{1}{s+3} + \frac{5}{s^2+1}$$

of terms appearing in the table. We conclude that the solution to our initial value problem is

$$u(t) = e^{-3\,t} + 5\sin t.$$

Of course, the last example is a problem that you can easily solve directly. The standard method learned in your first course on differential equations is just as effective in finding the final solution, and does not require all the extra Laplace transform machinery! The Laplace transform method is, however, particularly effective when dealing with complications that arise in cases of discontinuous forcing functions.

Example 13.40. Consider a mass vibrating on a spring with fixed stiffness $c = 4$. Assume that the mass starts at rest, is then subjected to a unit force over time interval $\frac{1}{2}\,\pi < t < 2\,\pi$, after which it left to vibrate on its own. The initial value problem is

$$\frac{d^2u}{dt^2} + 4\,u = \begin{cases} 1, & \frac{1}{2}\,\pi < t < 2\,\pi, \\ 0, & \text{otherwise}, \end{cases} \qquad\qquad u(0) = \dot{u}(0) = 0.$$

Taking the Laplace transform of the differential equation, and using (13.149), we find

$$(s^2+4)\,U(s) = \frac{e^{-\pi s/2} - e^{-2\,\pi s}}{s}\,, \qquad \text{and so} \qquad U(s) = \frac{e^{-\pi s/2} - e^{-2\,\pi s}}{s(s^2+4)}\,.$$

Therefore, by the shift formula (13.148)

$$u(t) = h\big(t - \tfrac{1}{2}\,\pi\big) - h(t - 2\,\pi),$$

where $h(t)$ is the function with Laplace transform

$$\mathcal{L}[\,h(t)\,] = H(s) = \frac{1}{s(s^2+4)} = \frac{1}{4}\left(\frac{1}{s} - \frac{s}{s^2+4}\right),$$

which has been conveniently rewritten using partial fractions. Referring to our table of Laplace transforms,

$$h(t) = \begin{cases} \frac{1}{4} - \frac{1}{4}\cos 2t, & t > 0, \\ 0, & t < 0. \end{cases}$$

Therefore, our desired solution is

$$u(t) = \begin{cases} 0, & 0 \le t \le \frac{1}{2}\,\pi, \\ \frac{1}{4} + \frac{1}{4}\cos 2t, & \frac{1}{2}\,\pi \le t \le 2\,\pi, \\ \frac{1}{2}\cos 2t, & 2\,\pi \le t. \end{cases}$$

Note that the solution $u(t)$ is only C^1 at the points of discontinuity of the forcing function.

Remark: A direct solution of this problem would proceed as follows. One solves the differential equation on each interval of continuity of the forcing function, leading to a solution on that interval depending upon two integration constants. The integration constants are then adjusted so that the solution satisfies the initial conditions and is continuously differentiable at each point of discontinuity of the forcing function. The details are straightforward, but messy. The Laplace transform method successfully bypasses these intervening manipulations.

Example 13.41. Consider the initial value problem

$$\frac{d^2u}{dt^2} + \omega^2\, u = f(t), \qquad\qquad u(0) = \dot{u}(0) = 0,$$

involving a general forcing function $f(t)$. Applying the Laplace transform to the differential equation and using the initial conditions,

$$(s^2 + \omega^2)\, U(s) = F(s), \qquad \text{and hence} \qquad U(s) = \frac{F(s)}{s^2 + \omega^2}\,.$$

The right hand side is the product of the Laplace transform of the forcing function $f(t)$ and that of the trigonometric function $k(t) = \dfrac{\sin \omega t}{\omega}$. Therefore, Theorem 13.38 implies that the solution can be written as their convolution

$$u(t) = f * k(t) = \int_0^t k(t-\tau)\, f(\tau)\, d\tau = \int_0^t \frac{\sin \omega(t-\tau)}{\omega}\, f(\tau)\, d\tau. \tag{13.154}$$

where

$$k(t) = \begin{cases} \dfrac{\sin \omega t}{\omega}\,, & t > 0, \\ 0, & t < 0. \end{cases}$$

The integral kernel $k(t-\tau)$ is known as the *fundamental solution* to the initial value problem, and prescribes the response of the system to a unit impulse force that is applied instantaneously at the time $t = \tau$. Note particularly that (unlike boundary value problems) the impulse only affect the solutions at later times $t > \tau$. For initial value problems, the fundamental solution plays a role similar to that of a Green's function in a boundary value problem. The convolution formula (13.154) can be viewed as a linear superposition of fundamental solution responses induced by expressing the forcing function

$$f(t) = \int_0^\infty f(\tau)\, \delta(t\, d\tau - \tau)$$

as a superposition of concentrated impulses over time.

This concludes our brief introduction to the Laplace transform and a few of its many applications to physical problems. More details can be found in almost all applied texts on mechanics, electronic circuits, signal processing, control theory, and many other areas, [**Ltr**].

Chapter 14

Vibration and Diffusion in One–Dimensional Media

In this chapter, we study the solutions, both analytical and numerical, to the two most important equations of one-dimensional continuum dynamics. The *heat equation* models the diffusion of thermal energy in a body; here, we analyze the case of a one-dimensional bar. The *wave equation* describes vibrations and waves in continuous media, including sound waves, water waves, elastic waves, electromagnetic waves, and so on. Again, we restrict our attention here to the case of waves in a one-dimensional medium, e.g., a string, or a bar, or a column of air. The two- and three-dimensional versions of these fundamental equations will be analyzed in the later Chapters 17 and 18.

As we saw in Section 12.1, the basic solution technique is inspired by our eigenvalue-based methods for solving linear systems of ordinary differential equations. Substituting the appropriate exponential or trigonometric ansatz will effectively reduce the partial differential equation to a one-dimensional boundary value problem. The linear superposition principle implies that general solution can then be expressed as a infinite series in the resulting eigenfunction solutions. In both cases considered here, the eigenfunctions of the one-dimensional boundary value problem are trigonometric, and so the solution to the partial differential equation takes the form of a time-dependent Fourier series. Although we cannot, in general, analytically sum the Fourier series to produce a simpler formula for the solution, there are a number of useful observations that can be gleaned from it.

In the case of the heat equation, the solutions decay exponentially fast to thermal equilibrium, at a rate governed by the smallest positive eigenvalue of the associated boundary value problem. The higher order Fourier modes damp out very rapidly, and so the heat equation can be used to automatically smooth and denoise signals and images. It also implies that the heat equation cannot be run backwards in time — determining the initial temperature profile from a later measurement is an ill-posed problem. The response to a concentrated unit impulse leads to the fundamental solution, which can then be used to construct integral representations of the solution to the inhomogeneous heat equation. We will also explain how to exploit the symmetry properties of the differential equation in order to construct new solutions from known solutions.

In the case of the wave equation, each Fourier mode vibrates with its natural frequency. In a stable situation, the full solution is a linear combination of these fundamental vibrational modes, while an instability induces an extra linearly growing mode. For one-dimensional media, the natural frequencies are integral multiples of a single lowest frequency, and hence the solution is periodic, which, in particular, explains the tonal qualities

of string and wind instruments. The one-dimensional wave equation admits an alternative explicit solution formula, due to d'Alembert, which points out the role of characteristics in signal propagation and the behavior of solutions. The analytic and series solution methods serve to shed complementary lights on the physical phenomena of waves and vibration in continuous media.

Following our analytical study, we introduce several basic numerical solution methods for both the heat and the wave equations. We begin with a general discussion of finite difference formulae for numerically approximating derivatives of functions. The basic *finite difference scheme* is obtained by replacing the derivatives in the equation by the appropriate numerical differentiation formulae. However, there is no guarantee that the resulting numerical scheme will accurately approximate the true solution, and further analysis is required to elicit bona fide, convergent numerical algorithms. In dynamical problems, the finite difference schemes replace the partial differential equation by an iterative linear matrix system, and the analysis of convergence relies on the methods covered in Chapter 10.

In preparation for our treatment of partial differential equations in higher dimensions, the final section introduces a general framework for dynamics that incorporates both discrete dynamics, modeled by linear systems of ordinary differential equations, and continuum dynamics, modeled by the heat and wave equations, their multidimensional counterparts, as well as the equilibrium boundary value problems governing bars, beams, membranes and solid bodies. Common features, based on eigenvalues, are discussed.

14.1. The Diffusion and Heat Equations.

Let us begin with a physical derivation of the heat equation from first principles of thermodynamics. The reader solely interested in the mathematical developments can skip ahead to the following subsection. However, physical insight can often play an critical role in understanding the underlying mathematics, and is neglected at one's peril.

We consider a bar — meaning a thin, heat-conducting body of length ℓ. "Thin" means that we can regard the bar as a one-dimensional continuum with no significant transverse temperature variation. We use $0 \le x \le \ell$ to denote the position along the bar. Our goal is to find the temperature $u(t,x)$ of the bar at position x and time t. The dynamical equations governing the temperature are based on three fundamental physical laws.

The first law is that, in the absence of external sources, thermal energy can only enter the bar through its ends. In physical terms, we are assuming that the bar is fully insulated along its length. Let $\varepsilon(t,x)$ to denote the thermal energy in the bar at position x and time t. Consider a small section of the bar lying between x and $x+\Delta x$. The total amount of heat energy contained in this section is obtained by integrating (summing): $\displaystyle\int_x^{x+\Delta x} \varepsilon(t,y)\,dy$. Further, let $w(t,x)$ denote the *heat flux*, i.e., the rate of flow of thermal energy along the bar. We use the convention that $w(t,x) > 0$ means that the energy is moving to the right, while $w(t,x) < 0$ if it moves to the left. The first law implies that the rate of change in the thermal energy in any section of the bar is equal to the total heat flux, namely the amount of the heat passing through its ends. Therefore, in view of our sign convention on

the flux,

$$\frac{\partial}{\partial t}\int_x^{x+\Delta x} \varepsilon(t,y)\,dy = -\,w(t,x+\Delta x) + w(t,x),$$

the latter two terms denoting the respective flux of heat *into* the section of the bar at its right and left ends. Assuming sufficient regularity of the integrand, we are permitted to bring the derivative inside the integral. Thus, dividing both sides of the resulting equation by Δx,

$$\frac{1}{\Delta x}\int_x^{x+\Delta x} \frac{\partial \varepsilon}{\partial t}\,(t,y)\,dy = -\,\frac{w(t,x+\Delta x) - w(t,x)}{\Delta x}\,.$$

In the limit as the length $\Delta x \to 0$, the right hand side of this equation converges to minus the x derivative of $w(t,x)$, while, by the Fundamental Theorem of Calculus, the left hand side converges to the integrand $\partial\varepsilon/\partial t$ at the point x; the net result is the fundamental differential equation

$$\frac{\partial \varepsilon}{\partial t} = -\,\frac{\partial w}{\partial x} \tag{14.1}$$

relating thermal energy ε and heat flux w. A partial differential equation of this particular form is known as a *conservation law*, and, in this instance, formulates the law of conservation of thermal energy. See Exercise ■ for details.

The second physical law is a *constitutive assumption*, based on experimental evidence. In most physical materials, thermal energy is found to be proportional to temperature,

$$\varepsilon(t,x) = \sigma(x)\,u(t,x). \tag{14.2}$$

The factor

$$\sigma(x) = \rho(x)\,\chi(x) > 0$$

is the product of the *density* ρ of the material and its *specific heat* χ, which is the amount of heat energy required to raise the temperature of a unit mass of the material by one unit. Note that we are assuming the bar is not changing in time, and so physical quantities such as density and specific heat depend only on position x. We also assume, perhaps with less physical justification, that the material properties do not depend upon the temperature; otherwise, we would be led to a much more difficult nonlinear diffusion equation.

The third physical law relates the heat flux to the temperature. Physical experiments in a wide variety of materials indicate that the heat energy moves from hot to cold at a rate that is in direct proportion to the rate of change — meaning the derivative — of the temperature. The resulting linear constitutive relation

$$w(t,x) = -\,\kappa(x)\,\frac{\partial u}{\partial x} \tag{14.3}$$

is known as *Fourier's Law of Cooling*. The proportionality factor $\kappa(x) > 0$ is called the *thermal conductivity* of the bar at position x. A good heat conductor, e.g., silver, will have high conductivity, while a poor conductor, e.g., glass, will have low conductivity. The minus sign tells us that heat energy moves from hot to cold; if $\dfrac{\partial u}{\partial x}(t,x) > 0$ the

temperature is increasing from left to right, and so the heat energy moves back to the left, with consequent flux $w(t,x) < 0$.

Combining the three laws (14.1), (14.2) and (14.3) produces the basic partial differential equation

$$\frac{\partial}{\partial t}\big[\,\sigma(x)\,u\,\big] = \frac{\partial}{\partial x}\left(\kappa(x)\,\frac{\partial u}{\partial x}\right), \qquad 0 < x < \ell, \tag{14.4}$$

governing the diffusion of heat in a non-uniform bar. The resulting *linear diffusion equation* is used to model a variety of diffusive processes, including heat flow, chemical diffusion, population dispersion, and the spread of infectious diseases. If, in addition, we allow external heat sources $h(t,x)$, then the linear diffusion equation acquires an inhomogeneous term:

$$\frac{\partial}{\partial t}\big[\,\sigma(x)\,u\,\big] = \frac{\partial}{\partial x}\left(\kappa(x)\,\frac{\partial u}{\partial x}\right) + h(t,x), \qquad 0 < x < \ell. \tag{14.5}$$

In order to uniquely prescribe the solution $u(t,x)$ to the diffusion equation, we need to specify the initial temperature distribution

$$u(t_0,x) = f(x), \qquad 0 \le x \le \ell, \tag{14.6}$$

along the bar at an initial time t_0. In addition, we must impose suitable boundary conditions at the two ends of the bar. As with the equilibrium equations discussed in Chapter 11, there are three common physical types. The first is a *Dirichlet boundary condition*, where an end of the bar is held at prescribed temperature. Thus, the boundary condition

$$u(t,0) = \alpha(t) \tag{14.7}$$

fixes the temperature at the left hand end of the bar. Alternatively, the *Neumann boundary condition*

$$\frac{\partial u}{\partial x}\,(t,0) = \xi(t) \tag{14.8}$$

prescribes the heat flux $w(t,0) = -\kappa(0)\,\dfrac{\partial u}{\partial x}\,(t,0)$ at the left hand end. In particular, the homogeneous Neumann condition with $\xi(t) \equiv 0$ corresponds to an insulated end, where no heat can flow in or out. Each end of the bar should have one or the other of these boundary conditions. For example, a bar with both ends having prescribed temperatures is governed by the pair of Dirichlet boundary conditions

$$u(t,0) = \alpha(t), \qquad u(t,\ell) = \beta(t), \tag{14.9}$$

whereas a bar with two insulated ends requires two homogeneous Neumann boundary conditions

$$\frac{\partial u}{\partial x}\,(t,0) = 0, \qquad \frac{\partial u}{\partial x}\,(t,\ell) = 0. \tag{14.10}$$

The mixed case, with one end fixed and the other insulated, is similarly formulated. Finally, the *periodic boundary conditions*

$$u(t,0) = u(t,\ell), \qquad \frac{\partial u}{\partial x}\,(t,0) = \frac{\partial u}{\partial x}\,(t,\ell), \tag{14.11}$$

correspond to a circular *ring* of length ℓ. As before, we are assuming the heat is only allowed to flow around the ring — insulation prevents any radiation of heat from one side of the ring to the other.

The Heat Equation

In this book, we will retain the term "heat equation" to refer to the homogeneous case, in which the bar is made of a uniform material, and so its density ρ, conductivity κ, and specific heat χ are all positive constants. Under these assumptions, the homogeneous diffusion equation (14.4) reduces to the *heat equation*

$$\frac{\partial u}{\partial t} = \gamma \, \frac{\partial^2 u}{\partial x^2} \tag{14.12}$$

for the temperature $u(t,x)$ in the bar at time t and position x. The constant

$$\gamma = \frac{\kappa}{\sigma} = \frac{\kappa}{\rho \, \chi} \tag{14.13}$$

is called the *thermal diffusivity* of the bar, and incorporates all of its relevant physical properties. The solution $u(t,x)$ will be uniquely prescribed once we specify initial conditions (14.6) and a suitable pair of boundary conditions at the ends of the bar.

As we learned in Section 12.1, the elementary, separable solutions to the heat equation are based on the exponential ansatz

$$u(t,x) = e^{-\lambda t} \, v(x), \tag{14.14}$$

where $v(x)$ is a time-independent function. Substituting the solution formula (14.14) into (14.12) and canceling the common exponential factors, we find that $v(x)$ must solve the ordinary differential equation

$$-\gamma \, v'' = \lambda \, v.$$

In other words, v is an *eigenfunction* with *eigenvalue* λ, for the second derivative operator $K = -\gamma \, D^2$. Once we determine the eigenvalues and eigenfunctions, we will be able to reconstruct the solution $u(t,x)$ as a linear combination, or, rather, infinite series in the corresponding separable eigenfunction solutions.

Let us consider the simplest case of a uniform bar held at zero temperature at each end. For simplicity, we take the initial time to be $t_0 = 0$, and so the initial and boundary conditions are

$$\begin{aligned} u(t,0) = 0, \qquad u(t,\ell) = 0, \qquad & t \geq 0, \\ u(0,x) = f(x), \qquad & 0 < x < \ell. \end{aligned} \tag{14.15}$$

According to the general prescription, we need to solve the eigenvalue problem

$$\gamma \, \frac{d^2 v}{dx^2} + \lambda \, v = 0, \qquad v(0) = 0, \qquad v(\ell) = 0. \tag{14.16}$$

As noted in Exercise ■, positive definiteness of the underlying differential operator $K = -\gamma \, D^2$ when subject to Dirichlet boundary conditions implies that we need only look for positive eigenvalues: $\lambda > 0$. In Exercises ■–■, the skeptical reader is asked to check

explicitly that if $\lambda \leq 0$ or λ is complex, then the boundary value problem (14.16) admits only the trivial solution $v(x) \equiv 0$.

Setting $\lambda = \gamma\,\omega^2$ with $\omega > 0$, the general solution to the differential equation is a trigonometric function

$$v(x) = a\,\cos \omega\, x + b\,\sin \omega\, x,$$

where a, b are constants whose values will be fixed by the boundary conditions. The boundary condition at $x = 0$ requires $a = 0$. The second boundary condition requires

$$v(\ell) = b\,\sin \omega\,\ell = 0.$$

Therefore, assuming $b \neq 0$, as otherwise the solution is trivial, $\omega\,\ell$ must be an integer multiple of π, and so

$$\omega \;=\; \frac{\pi}{\ell}\,, \qquad \frac{2\pi}{\ell}\,, \qquad \frac{3\pi}{\ell}\,, \qquad \dots\ .$$

We conclude that the eigenvalues and eigenfunctions of the boundary value problem (14.16) are

$$\lambda_n = \gamma\,\left(\frac{n\,\pi}{\ell}\right)^2, \qquad\qquad v_n(x) = \sin\frac{n\,\pi\,x}{\ell}\,, \qquad\qquad n = 1, 2, 3, \dots. \tag{14.17}$$

The corresponding separable solutions (14.14) to the heat equation with the given boundary conditions are

$$u_n(t,x) = \exp\left(-\,\frac{\gamma\,n^2\,\pi^2\,t}{\ell^2}\right)\,\sin\frac{n\,\pi\,x}{\ell}, \qquad\qquad n = 1, 2, 3, \dots\ . \tag{14.18}$$

Each represents a trigonometrically oscillating temperature profile that maintains its form while decaying to zero at an exponential rate. The first of these,

$$u_1(t,x) = \exp\left(-\,\frac{\gamma\,\pi^2\,t}{\ell^2}\right)\,\sin\frac{\pi\,x}{\ell}\,,$$

experiences the slowest decay. The higher "frequency" modes $u_n(t,x)$, $n \geq 2$, all go to zero at a faster rate, with those having a highly oscillatory temperature profile, where $n \gg 0$, effectively disappearing almost instantaneously. Thus, small scale temperature fluctuations tend to rapidly cancel out through diffusion of heat energy.

Linear superposition is used to assemble the general series solution

$$u(t,x) = \sum_{n=1}^{\infty}\, b_n\,u_n(t,x) = \sum_{n=1}^{\infty}\, b_n\,\exp\left(-\,\frac{\gamma\,n^2\,\pi^2\,t}{\ell^2}\right)\,\sin\frac{n\,\pi\,x}{\ell} \tag{14.19}$$

as a combination of the separable solutions. Assuming that the series converges, the initial temperature profile is

$$u(0,x) = \sum_{n=1}^{\infty}\, b_n\,\sin\frac{n\,\pi\,x}{\ell} = f(x). \tag{14.20}$$

This has the form of a Fourier sine series (12.44) on the interval $[0, \ell]$. By orthogonality of the eigenfunctions — which is a direct consequence of the self-adjointness of the underlying

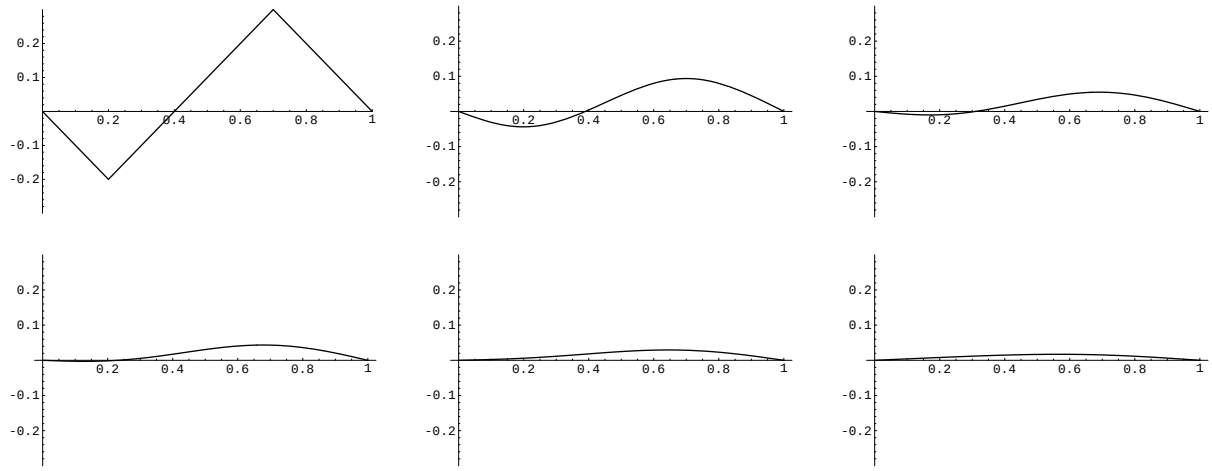


Figure 14.1. A Solution to the Heat Equation.

boundary value problem (14.16) — the coefficients are determined by the inner product formulae (12.45), and so

$$b_n = \frac{2}{\ell} \int_0^\ell f(x) \sin \frac{n\pi x}{\ell}\, dx, \qquad n = 1, 2, 3, \dots . \tag{14.21}$$

The resulting solution (14.19) describes the Fourier sine series for the temperature $u(t, x)$ of the bar at each later time $t \geq 0$. It can be rigorously proved that, for quite general initial conditions, the Fourier series does indeed converge to a solution to the initial-boundary value problem, [**181**].

Example 14.1. Consider the initial temperature profile

$$u(0, x) = f(x) = \begin{cases} -x, & 0 \leq x \leq \frac{1}{5}, \\ x - \frac{2}{5}, & \frac{1}{5} \leq x \leq \frac{7}{10}, \\ 1 - x, & \frac{7}{10} \leq x \leq 1, \end{cases} \tag{14.22}$$

on a bar of length 1, plotted in the first graph in Figure 14.1. Using (14.21), the first few Fourier coefficients of $f(x)$ are computed as

$$\begin{aligned} &b_1 = .0448\dots, \qquad b_2 = -.096\dots, \qquad b_3 = -.0145\dots, \qquad b_4 = 0, \\ &b_5 = -.0081\dots, \qquad b_6 = .0066\dots, \qquad b_7 = .0052\dots, \qquad b_8 = 0, \qquad \dots . \end{aligned}$$

Setting $\gamma = 1$, the resulting Fourier series solution to the heat equation is

$$\begin{aligned} u(t, x) &= \sum_{n=1}^{\infty} b_n\, u_n(t, x) = \sum_{n=1}^{\infty} b_n\, e^{-n^2 \pi^2 t} \sin n\pi x \\ &= .0448\, e^{-\pi^2 t} \sin \pi x - .096\, e^{-4\pi^2 t} \sin 2\pi x - .0145\, e^{-9\pi^2 t} \sin 3\pi x - \cdots . \end{aligned}$$

In Figure 14.1, the solution is plotted at the successive times $t = ., .02, .04, \dots, .1$. Observe that the corners in the initial data are immediately smoothed out. As time progresses, the solution decays, at an exponential rate of $\pi^2 \approx 9.87$, to a uniform, zero temperature,

which is the equilibrium temperature distribution for the homogeneous Dirichlet boundary conditions. As the solution decays to thermal equilibrium, it also assumes the progressively more symmetric shape of a single sine arc, of exponentially decreasing amplitude.

Smoothing and Long Time Behavior

The fact that we can write the solution to an initial-boundary value problem in the form of an infinite series is progress of a sort. However, because it cannot be summed in closed form, this "solution" is much less satisfying than a direct, explicit formula. Nevertheless, there are important qualitative and quantitative features of the solution that can be easily gleaned from such series expansions.

If the initial data $f(x)$ is piecewise continuous, then its Fourier coefficients are uniformly bounded; indeed, for any $n \geq 1$,

$$| b_n | \;\leq\; \frac{2}{\ell} \int_0^\ell \left| f(x) \sin \frac{n \pi x}{\ell} \right| dx \;\leq\; \frac{2}{\ell} \int_0^\ell | f(x) | \, dx \equiv M. \tag{14.23}$$

This property holds even for quite irregular data; for instance, the Fourier coefficients (12.59) of the delta function are also uniformly bounded. Under these conditions, each term in the series solution (14.19) is bounded by an exponentially decaying function

$$\left| \; b_n \exp\left(- \frac{\gamma \, n^2 \, \pi^2}{\ell^2} \, t \right) \sin \frac{n \pi x}{\ell} \; \right| \;\leq\; M \exp\left(- \frac{\gamma \, n^2 \, \pi^2}{\ell^2} \, t \right).$$

This means that, as soon as $t > 0$, most of the high frequency terms, $n \gg 0$, will be extremely small. Only the first few terms will be at all noticeable, and so the solution essentially degenerates into a finite sum over the first few Fourier modes. As time increases, more and more of the Fourier modes will become negligible, and the sum further degenerates into progressively fewer significant terms. Eventually, as $t \to \infty$, *all* of the Fourier modes will decay to zero. Therefore, the solution will converge exponentially fast to a zero temperature profile: $u(t, x) \to 0$ as $t \to \infty$, representing the bar in its final uniform thermal equilibrium. The fact that its equilibrium temperature is zero is the result of holding both ends of the bar fixed at zero temperature, and any initial heat energy will eventually be dissipated away through the ends. The last term to disappear is the one with the slowest decay, namely

$$u(t, x) \;\approx\; b_1 \exp\left(- \frac{\gamma \, \pi^2}{\ell^2} \, t \right) \sin \frac{\pi x}{\ell}, \qquad \text{where} \qquad b_1 = \frac{1}{\pi} \int_0^\pi f(x) \sin x \, dx. \tag{14.24}$$

Generically, $b_1 \neq 0$, and the solution approaches thermal equilibrium exponentially fast with rate equal to the smallest eigenvalue, $\lambda_1 = \gamma \, \pi^2 / \ell^2$, which is proportional to the thermal diffusivity divided by the square of the length of the bar. The longer the bar, or the smaller the diffusivity, the longer it takes for the effect of holding the ends at zero temperature to propagate along the entire bar. Also, again provided $b_1 \neq 0$, the asymptotic shape of the temperature profile is a small sine arc, just as we observed in Example 14.1. In exceptional situations, namely when $b_1 = 0$, the solution decays even faster, at a rate

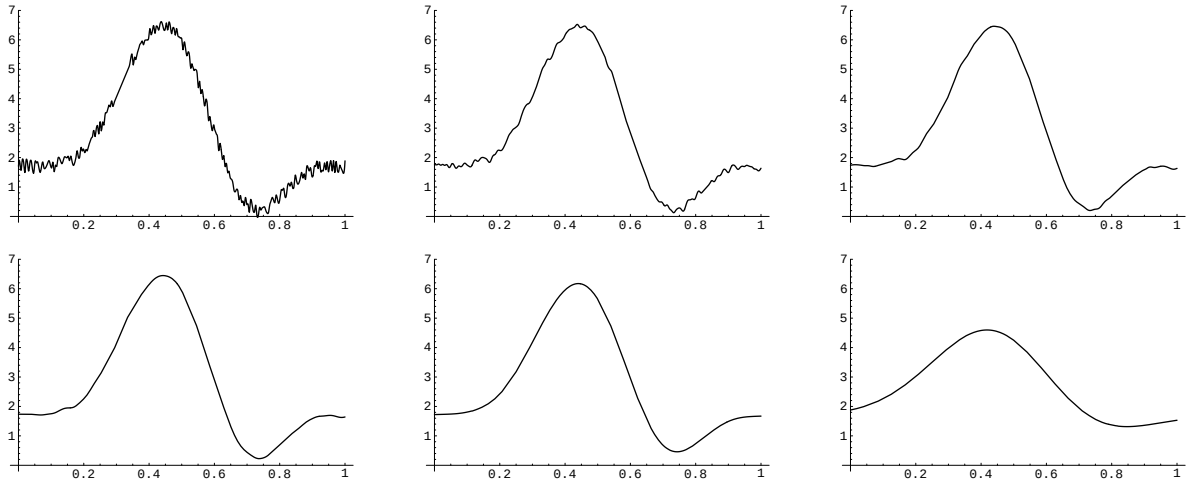


Figure 14.2. Denoising a Signal with the Heat Equation.

equal to the eigenvalue $\lambda_k = \gamma\, k^2\, \pi^2/\ell^2$ corresponding to the first nonzero term, $b_k \neq 0$, in the series; its asymptotic shape now oscillates k times over the interval.

The heat equation's smoothing effect on irregular initial data by fast damping of the high frequency modes underlies its effectiveness for smoothing out and denoising signals. We take the initial data $u(0,x) = f(x)$ to be a noisy signal, and then evolve the heat equation forward to a prescribed time $t^\star > 0$. The resulting function $g(x) = u(t^\star, x)$ will be a smoothed version of the original signal $f(x)$ in which most of the high frequency noise has been eliminated. Of course, if we run the heat flow for too long, all of the low frequency features will be also be smoothed out and the result will be a uniform, constant signal. Thus, the choice of stopping time $t^\star$ is crucial to the success of this method. Figure 14.2 shows the effect running the heat equation[†], with $\gamma = 1$, to times $t = 0., .00001, .00005, .0001, .001, .01$ on the same signal from Figure 13.5. Observe how quickly the noise is removed. By the final time, the overall smoothing effect of the heat flow has caused significant degradation (blurring) of the original signal. The heat equation approach to denoising has the advantage that no Fourier coefficients need be explicitly computed, nor does one need to reconstruct the smoothed signal from its remaining Fourier coefficients. The final section discusses some numerical methods that can be used to solve the heat equation directly.

Another, closely related observation is that, for any fixed time $t > 0$ after the initial moment, the coefficients in the Fourier series (14.19) decay exponentially fast as $n \to \infty$. According to the discussion at the end of Section 12.3, this implies that the solution $u(t,x)$ is a very smooth, infinitely differentiable function of x at each positive time t, *no matter how unsmooth the initial temperature profile*. We have discovered the basic smoothing property of heat flow.

Theorem 14.2. *If $u(t,x)$ is a solution to the heat equation with piecewise continuous*

[†] To be honest, we are using periodic boundary conditions in the figures, although the Dirichlet version leads to similar results

initial data $f(x) = u(0,x)$, or, more generally, initial data satisfying (14.23), *then, for any $t > 0$, the solution $u(t,x)$ is an infinitely differentiable function of x.*

After even a very short amount of time, the heat equation smoothes out most, and, eventually, all of the fluctuations in the initial temperature profile. As a consequence, it becomes impossible to reconstruct the initial temperature $u(0,x) = f(x)$ by measuring the temperature distribution $h(x) = u(t,x)$ at a later time $t > 0$. *Diffusion is irreversible* — we cannot run the heat equation backwards in time! Indeed, if the initial data $u(0,x) = f(x)$ is not smooth, there is *no* function $u(t,x)$ for $t < 0$ that could possibly yield such an initial distribution because all corners and singularities are smoothed out by the diffusion process as t goes forward! Or, to put it another way, the Fourier coefficients (14.21) of any purported solution will be exponentially growing when $t < 0$, and so high frequency noise will completely overwhelm the solution. For this reason, the backwards heat equation is said to be *ill-posed*.

On the other hand, the unsmoothing effect of the backwards heat equation does have potential benefits. For example, in image processing, diffusion will gradually blur an image. Image enhancement is the reverse process, and requires running the heat flow backwards in some stable manner. One option is to restrict to the backwards evolution to the first few Fourier modes, which prevents the small scale fluctuations from overwhelming the computation. Similar issues arise in the reconstruction of subterranean profiles from seismic data, a problem of great concern in the oil and gas industry. In forensics, determining the time of death based on the current temperature of a corpse also requires running the equations governing the dissipation of body heat backwards in time. For these and other applications, a key issue in contemporary research is how to cleverly circumventing the ill-posedness of the backwards heat flow.

Remark: The irreversibility of the heat equation points out a crucial distinction between partial differential equations and ordinary differential equations. Ordinary differential equations are always reversible — unlike the heat equation, existence, uniqueness and continuous dependence properties of solutions are all equally valid in reverse time (although the detailed qualitative and quantitative properties of solutions can very well depend upon whether time is running forwards or backwards). The irreversibility of partial differential equations modeling the diffusive processes in our universe may well be why, in our experience, Time's Arrow points exclusively to the future.

The Heated Ring

Let us next consider the periodic boundary value problem modeling heat flow in an insulated circular ring. Let us fix the length of the ring to be $\ell = 2\pi$, with $-\pi < x < \pi$ representing "angular" coordinate around the ring. For simplicity, we also choose units in which the thermal diffusivity is $\gamma = 1$. Thus, we seek to solve the heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}\,, \qquad -\pi < x < \pi, \qquad t > 0, \tag{14.25}$$

subject to periodic boundary conditions

$$u(t,-\pi) = u(t,\pi), \qquad \frac{\partial u}{\partial x}\,(t,-\pi) = \frac{\partial u}{\partial x}\,(t,\pi), \qquad t \geq 0. \tag{14.26}$$

The initial temperature distribution is

$$u(0,x) = f(x), \qquad -\pi < x < \pi. \tag{14.27}$$

The resulting temperature $u(t,x)$ will be a periodic function in x of period 2π.

Substituting the separable solution ansatz $u(t,x) = e^{-\lambda t} v(x)$ into the heat equation and the boundary conditions leads to the periodic eigenvalue problem

$$\frac{d^2v}{dx^2} + \lambda\, v = 0, \qquad v(-\pi) = v(\pi), \qquad v(-\pi) = v(\pi). \tag{14.28}$$

As we know, in this case the eigenvalues are $\lambda_n = n^2$ where $n = 0, 1, 2, \dots$ is a non-negative integer, and the corresponding eigenfunction solutions are the trigonometric functions

$$v_n(x) = \cos n\,x, \qquad \widetilde{v}_n(x) = \sin n\,x, \qquad n = 0, 1, 2, \dots\ .$$

Note that $\lambda_0 = 0$ is a simple eigenvalue, with constant eigenfunction $\cos 0x = 1$ — the sine solution $\sin 0x \equiv 0$ is trivial — while the positive eigenvalues are, in fact, double, each possessing two linearly independent eigenfunctions. The corresponding separable solutions to the heated ring equation are

$$u_n(t,x) = e^{-n^2 t} \cos n\,x, \qquad \widetilde{u}_n(t,x) = e^{-n^2 t} \sin n\,x, \qquad n = 0, 1, 2, 3, \dots\ .$$

The resulting infinite series solution is

$$u(t,x) = \tfrac{1}{2}\, a_0 + \sum_{n=1}^{\infty} \big(a_n\, e^{-n^2 t} \cos n\,x + b_n\, e^{-n^2 t} \sin n\,x \big). \tag{14.29}$$

The initial conditions require

$$u(0,x) = \tfrac{1}{2}\, a_0 + \sum_{n=1}^{\infty} \big(a_n \cos n\,x + b_n \sin n\,x \big) = f(x), \tag{14.30}$$

which is precisely the Fourier series of the initial temperature profile $f(x)$. Consequently,

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos n\,x\, dx, \qquad b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin n\,x\, dx, \tag{14.31}$$

are the usual Fourier coefficients of $f(x)$.

As in the Dirichlet problem, after the initial instant, the high frequency terms in the series (14.29) become extremely small, since $e^{-n^2 t} \ll 1$ for $n \gg 0$. Therefore, as soon as $t > 0$, the solution essentially degenerates into a finite sum over the first few Fourier modes. Moreover, as $t \to \infty$, *all* of the Fourier modes will decay to zero with the exception of the constant one, with null eigenvalue $\lambda_0 = 0$. Therefore, the solution will converge exponentially fast to a constant temperature profile:

$$u(t,x) \;\longrightarrow\; \frac{1}{2}\, a_0 \;=\; \frac{1}{2\,\pi} \int_{-\pi}^{\pi} f(x)\, dx,$$

which equals the *average* of the initial temperature profile. Physically, we observe that the heat energy is redistributed so that the ring achieves a uniform constant temperature and is in thermal equilibrium. Indeed, the total heat

$$H(t) = \int_{-\pi}^{\pi} u(t,x)\,dx = \text{constant} \tag{14.32}$$

is conserved, meaning constant, for all time; the proof of this fact is left as an Exercise . (On the other hand, the Dirichlet boundary value problem does not conserve hear energy.)

Prior to equilibrium, only the lowest frequency Fourier modes will still be noticeable, and so the solution will asymptotically look like

$$u(t,x) \approx \tfrac{1}{2}\,a_0 + e^{-t}(a_1 \cos x + b_1 \sin x) = \tfrac{1}{2}\,a_0 + r_1\, e^{-t} \cos(x + \delta_1), \tag{14.33}$$

where

$$a_1 = r_1 \cos\delta_1 = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\cos x\,dx, \qquad b_1 = r_1 \sin\delta_1 = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\sin x\,dx.$$

Thus, for most initial data, the solution approaches thermal equilibrium exponentially fast, at a unit rate. The exceptions are when $r_1 = \sqrt{a_1^2 + b_1^2} = 0$, for which the rate of convergence is even faster, namely at a rate $e^{-k^2 t}$ where k is the smallest integer such that $r_k = \sqrt{a_k^2 + b_k^2} \neq 0$.

Inhomogeneous Boundary Conditions

So far, we have concentrated our attention on homogeneous boundary conditions. There is a simple trick that will convert a boundary value problem with inhomogeneous but constant Dirichlet boundary conditions,

$$u(t,0) = \alpha, \qquad u(t,\ell) = \beta, \qquad t \geq 0, \tag{14.34}$$

into a homogeneous Dirichlet problem. We begin by solving for the equilibrium temperature profile, which is the affine function that satisfies the two boundary conditions, namely

$$u^\star(x) = \alpha + \frac{\beta - \alpha}{\ell}\,x. \tag{14.35}$$

The difference

$$\widetilde{u}(t,x) = u(t,x) - u^\star(x) = u(t,x) - \alpha - \frac{\beta - \alpha}{\ell}\,x \tag{14.36}$$

measures the deviation of the solution from equilibrium. It clearly satisfies the homogeneous boundary conditions at both ends:

$$\widetilde{u}(t,0) = 0 = \widetilde{u}(t,\ell).$$

Moreover, by linearity, since both $u(t,x)$ and $u^\star(x)$ are solutions to the heat equation, so is $\widetilde{u}(t,x)$. The initial data must be similarly adapted:

$$\widetilde{u}(0,x) = \widetilde{f}(x) = f(x) - u^\star(x) = f(x) - \alpha - \frac{\beta - \alpha}{\ell}\,x.$$

Solving the resulting homogeneous initial value problem, we write $\widetilde{u}(t,x)$ in Fourier series form (14.19), where the Fourier coefficients are computed from the modified initial data $\widetilde{f}(x)$. The solution to the inhomogeneous boundary value problem thus has the series form

$$u(t,x) = \alpha + \frac{\beta - \alpha}{\ell}\, x \;+\; \sum_{n=1}^{\infty} \widetilde{b}_n \exp\left(-\frac{\gamma\, n^2\, \pi^2}{\ell^2}\, t \right) \sin \frac{n\,\pi\, x}{\ell}\,, \tag{14.37}$$

where

$$\widetilde{b}_n = \frac{2}{\ell} \int_0^{\ell} \widetilde{f}(x) \sin \frac{n\,\pi\, x}{\ell}\, dx, \qquad\qquad n = 1, 2, 3, \dots\ . \tag{14.38}$$

Since, for any reasonable initial data, $\widetilde{u}(t,0)$ will decay to zero at an exponential rate as $t \to \infty$, the actual temperature profile (14.37) will asymptotically decay to the equilibrium profile,

$$u(t,x) \quad \longrightarrow \quad u^\star(x) = \alpha \;+\; \frac{\beta - \alpha}{\ell}\, x$$

at the same exponentially fast rate.

This method does not apply when the boundary conditions are time-dependent: $u(t,0) = \alpha(t),\ \ u(t,\ell) = \beta(t)$. Attempting to mimic the preceding technique, we discover that the deviation

$$\widetilde{u}(t,x) = u(t,x) - u^\star(t,x), \qquad \text{where} \qquad u^\star(t,x) = \alpha(t) + \frac{\beta(t) - \alpha(t)}{\ell}\, x, \tag{14.39}$$

does satisfy the homogeneous boundary conditions, but now solves an inhomogeneous version of the heat equation:

$$\frac{\partial \widetilde{u}}{\partial t} = \frac{\partial^2 \widetilde{u}}{\partial x^2} - h(t,x), \qquad \text{where} \qquad h(t,x) = \frac{\partial u^\star}{\partial t}\,(t,x). \tag{14.40}$$

Solution techniques in this case will be discussed below.

14.2. Symmetry and the Maximum Principle.

So far we have relied on the method of separation of variables to construct explicit solutions to partial differential equations. A second useful solution technique relies on exploiting inherent symmetry properties of the differential equation. Unlike separation of variables[†], symmetry methods can be also successfully applied to produce solutions to a broad range of nonlinear partial differential equations; examples can be found in Chapter 22. While we do not have the space or required mathematical tools to develop the full apparatus of symmetry techniques, we can introduce the important concept of a *similarity solution*, applied in the particular context of the heat equation.

In general, by a *symmetry* of an equation, we mean a transformation, either linear (as in Section 7.2), affine (as in Section 7.3), or even nonlinear, that takes solutions to solutions. Thus, if we have a symmetry, and know one solution, then we can construct a

[†] This is not quite fair: separation of variables can be applied to some special nonlinear partial differential equations such as Hamilton–Jacobi equations, [**128**].

second solution by applying the symmetry. And, possibly, a third solution by applying the symmetry yet again. And so on. If we know lots of symmetries, then we can produce lots and lots of solutions by this simple device.

Remark: General symmetry techniques are founded on the theory of Lie groups, named after the influential nineteenth century Norwegian mathematician Sophus Lie (pronounced "Lee"). Lie's theory provides an algorithm for completely determining all the symmetries of a given differential equation, but this is beyond the scope of this introductory text. However, direct inspection and/or physical intuition will often detect the most important symmetries without appealing to such a sophisticated theory. Modern applications of Lie's symmetry methods to partial differential equations arising in physics and engineering can be traced back to the influential book of G. Birkhoff, [**18**], on hydrodynamics. A complete and comprehensive treatment of symmetry methods can be found in the first author's book [**141**], and, at a more introductory level, in the recent books by Hydon, [**104**], and Cantwell, [**38**], with particular emphasis on fluid mechanics.

The heat equation serves as an excellent testing ground for the general symmetry methodology, as it admits a rich variety of symmetry transformations that take solutions to solutions. The simplest are the translations. Moving the space and time coordinates by a fixed amount,

$$t \longmapsto t - a, \qquad x \longmapsto x - b, \tag{14.41}$$

where a, b are constants, changes the function $u(t, x)$ into the translated function

$$U(t, x) = u(t - a, x - b). \tag{14.42}$$

A simple application of the chain rule proves that the partial derivatives of U with respect to t and x agree with the corresponding partial derivatives of u, so

$$\frac{\partial U}{\partial t} = \frac{\partial u}{\partial t}, \qquad \frac{\partial U}{\partial x} = \frac{\partial u}{\partial x}, \qquad \frac{\partial^2 U}{\partial x^2} = \frac{\partial^2 u}{\partial x^2},$$

and so on. In particular, the function $U(t, x)$ is a solution to the heat equation $U_t = \gamma\, U_{xx}$ whenever $u(t, x)$ also solves $u_t = \gamma\, u_{xx}$. Physically, the translation symmetries formalize the property that the heat equation models a homogeneous medium, and hence the solution does not depend on the choice of reference point or origin of our coordinate system.

As a consequence, each solution to the heat equation will produce an infinite family of translated solutions. For example, starting with the separable solution

$$u(t, x) = e^{-\gamma t} \sin x,$$

we immediately produce the additional solutions

$$u(t, x) = e^{-\gamma (t-a)} \sin \pi (x - b),$$

valid for any choice of constants a, b.

Warning: Typically, the symmetries of a differential equation do not respect initial or boundary conditions. For instance, if $u(t, x)$ is defined for $t \geq 0$ and in the domain $0 \leq x \leq \ell$, then its translated version $U(t, x)$ is defined for $t \geq a$ and in the translated domain $b \leq x \leq \ell + b$, and so will solve a translated initial-boundary value problem.

A second, even more important class of symmetries are the scaling invariances. We already know that if $u(t,x)$ is a solution, so is any scalar multiple $c\,u(t,x)$; this is a simple consequence of linearity of the heat equation. We can also add an arbitrary constant to the temperature, noting that

$$U(t,x) = c\,u(t,x) + k \tag{14.43}$$

is a solution for any choice of constants c, k. Physically, the transformation (14.43) amounts to a change in the scale for measuring temperature. For instance, if u is measured degrees Celsius, and we set $c = \frac{9}{5}$ and $k = 32$, then $U = \frac{9}{5}\,u + 32$ will be measured in degrees Fahrenheit. Thus, reassuringly, the physical processes described by the heat equation do not depend upon our choice of thermometer.

More interestingly, suppose we rescale the space and time variables:

$$t \;\longmapsto\; \alpha\,t, \qquad\qquad x \;\longmapsto\; \beta\,x, \tag{14.44}$$

where $\alpha, \beta > 0$ are positive constants. The effect of such a scaling transformation is to convert $u(t,x)$ into a rescaled function

$$U(t,x) = u(\alpha\,t, \beta\,x). \tag{14.45}$$

The derivatives of U are related to those of u according to the following formulae, which are direct consequences of the multi-variable chain rule:

$$\frac{\partial U}{\partial t} = \alpha\,\frac{\partial u}{\partial t}\,, \qquad\qquad \frac{\partial U}{\partial x} = \beta\,\frac{\partial u}{\partial x}\,, \qquad\qquad \frac{\partial^2 U}{\partial x^2} = \beta^2\,\frac{\partial^2 u}{\partial x^2}\,.$$

Therefore, if u satisfies the heat equation $u_t = \gamma\,u_{xx}$, then U satisfies the rescaled heat equation

$$U_t = \alpha\,u_t = \alpha\,\gamma\,u_{xx} = \frac{\alpha\,\gamma}{\beta^2}\,U_{xx},$$

which we rewrite as

$$U_t = \Gamma\,U_{xx}, \qquad\qquad \text{where} \qquad\qquad \Gamma = \frac{\gamma\,\alpha}{\beta^2}\,. \tag{14.46}$$

Thus, the net effect of scaling space and time is merely to rescale the diffusion coefficient in the heat equation. Physically, the scaling symmetry (14.44) corresponds to a change in the physical units used to measure time and distance. For instance, to change from seconds to minutes, set $\alpha = 60$, and from meters to yards, set $\beta = 1.0936$. The net effect (14.46) on the diffusion coefficient is a reflection of its physical units, namely distance^2/time.

In particular, if we choose

$$\alpha = \frac{1}{\gamma}\,, \qquad\qquad \beta = 1,$$

then the rescaled diffusion coefficient becomes $\Gamma = 1$. This observation has the following important consequence. If $U(t,x)$ solves the heat equation for a unit diffusivity, $\Gamma = 1$, then

$$u(t,x) = U(\gamma\,t, x) \tag{14.47}$$

solves the heat equation for the diffusivity γ. Thus, the only effect of the diffusion coefficient γ is to speed up or slow down time! A body with diffusivity $\gamma = 2$ will cool down twice as fast as a body (of the same shape subject to the same boundary conditions and initial conditions) with diffusivity $\gamma = 1$. Note that this particular rescaling has not altered the space coordinates, and so $U(t,x)$ is defined on the same domain as $u(t,x)$.

On the other hand, if we set $\alpha = \beta^2$, then the rescaled diffusion coefficient is exactly the same as the original: $\Gamma = \gamma$. Thus, the transformation

$$t \longmapsto \beta^2 t, \qquad\qquad x \longmapsto \beta x, \tag{14.48}$$

does not alter the equation, and hence defines a *scaling symmetry*, also known as a *similarity transformation*, for the heat equation. Combining (14.48) with the linear rescaling $u \mapsto c\,u$, we make the elementary, but important observation that if $u(t,x)$ is any solution to the heat equation, then so is the function

$$U(t,x) = c\,u(\beta^2 t, \beta x), \tag{14.49}$$

for the *same* diffusion coefficient γ. For example, rescaling the solution $u(t,x) = e^{-\gamma t^2}\cos x$ leads to the solution $U(t,x) = c\,e^{-\gamma \beta^2 t^2}\cos \beta x$.

Warning: As in the case of translations, rescaling space by a factor $\beta \neq 1$ will alter the domain of definition of the solution. If $u(t,x)$ is defined for $0 \leq x \leq \ell$, then $U(t,x)$ is defined for $0 \leq x \leq \ell/\beta$.

Suppose that we have solved the heat equation for the temperature $u(t,x)$ on a bar of length 1, subject to certain initial and boundary conditions. We are then given a bar composed of the same material of length 2. Since the diffusivity coefficient has not changed, and we can directly construct the new solution $U(t,x)$ by rescaling. Setting $\beta = \frac{1}{2}$ will serve to double the length. If we also rescale time by a factor $\alpha = \beta^2 = \frac{1}{4}$, then the rescaled function $U(t,x) = u\big(\frac{1}{4}t, \frac{1}{2}x\big)$ will be a solution of the heat equation on the longer bar with the same diffusivity constant. The net effect is that the rescaled solution will be evolving four times as slowly as the original. Thus, it effectively takes a bar that is twice the length four times as long to cool down.

The Maximum Principle

The *Maximum Principle* is the mathematical formulation of the thermodynamical law that heat cannot, in the absence of external sources, achieve a value at a point that strictly larger than its surroundings.

We will help in the proof to formulate it for the more general case when the heat equation is subject to external forcing.

Theorem 14.3. *Let $\gamma > 0$. Suppose $u(t,x)$ is a solution to the forced heat equation*

$$\frac{\partial u}{\partial t} = \gamma\,\frac{\partial^2 u}{\partial x^2} + F(t,x), \quad \text{on the rectangular domain} \quad R = \{\, a \leq x \leq b,\ 0 \leq t \leq c. \,\}$$

Suppose $F(t,x) \leq 0$ for all $(t,x) \in R$. Then the global maximum of $u(t,x)$ on the domain R occurs either at $t = 0$ or at $x = a$ or $x = b$.

In other words, if we are no;t adding in any external heat, the maximum temperature ion the bar occurs either at the initial time, or at one of the endpoints.

Proof: First let us prove the result under the assumption that $F(t,x) < 0$, and hence

$$\frac{\partial u}{\partial t} < \gamma \, \frac{\partial^2 u}{\partial x^2} \tag{14.50}$$

everywhere in R. Suppose $u(t,x)$ has a (local) maximum at a point in the interior of R. Then, by multivariable calculus (see also Theorem 19.42), its gradient must vanish, so $u_t = u_x = 0$, and its Hessian matrix must be positive definite, which, in particular, requires that $u_{xx} \leq 0$. But these contradict the inequality (14.50). Thus, the solution cannot have a local interior maximum. If the maximum were to occur on the top of the rectangle, at a point (t,x) where $t = c$, then we would necessarily have $u_t \geq 0$ there, as otherwise $u(t,x)$ would be decreasing as a function of t and hence (c,x) could not be a local maximum on R, and also $u_{xx} \leq 0$, again contradicting (14.50).

To generalize to the case when $F(t,x) \leq 0$ — which includes the heat equation when $F(t,x) \equiv 0$, requires a little trick. We set

$$v(t,x) = u(t,x) + \varepsilon \, x^2, \qquad \text{where} \qquad \varepsilon > 0.$$

Then,

$$\frac{\partial v}{\partial t} = \gamma \, \frac{\partial^2 v}{\partial x^2} - 2\gamma\,\varepsilon + F(t,x) = \gamma \, \frac{\partial^2 v}{\partial x \, \partial +} \, \widetilde{F}(t,x),$$

where

$$\widetilde{F}(t,x) = F(t,x) - 2\gamma\,\varepsilon < 0$$

everywhere in R. Thus, by the previous paragraph, the maximum of v occurs on either the bottom or sides of the rectangle. Now we let $\varepsilon \to 0$ and conclude the same for u. More precisely, let $u(t,x) \leq M$ on $t = 0$ or $x = a$ or $x = b$. Then

$$v(t,x) \leq M + \varepsilon \max\{\, a^2, b^2 \,\}$$

and hence, on all of R,

$$u(t,x) \leq v(t,x) \leq M + \varepsilon \max\{\, a^2, b^2 \,\}.$$

letting $\varepsilon \to 0$ proves that $u(t,x) \leq M$ everywhere, which completes the proof. *Q.E.D.*

Thus, solution $u(t,x)$ to the heat equation (14.12) can only achieve a maximum on the boundary of its domain of definition. In other words, a solution cannot have a maximum at any point (t,x) that lies in the interior of the domain and after the initial time $t > 0$. Physically, if the temperature in a fully insulated bar starts out everywhere above freezing, then, in the absence of external heat sources, it can never dip below freezing at any later time.

Let us apply the Maximum Principle to prove uniqueness of solutions to the heat equation.

Theorem 14.4. *There is at most one solution to the boundary value problem.*

Proof: Suppose u and $\widetilde{u}$ are any two solutions. Then their difference $v = u - \widetilde{u}$ solves the homogeneous boundary value problem. Thus, by the maximu principle $v(t,x) \leq 0$ at all points of$R$. But $-v = \widetilde{u} - u$ also solves the homogeneous boundary value problem, and hence $-v \leq 0$ too. This implies $v(t,x) \equiv 0$ and hence $u \equiv \widetilde{u}$ everywhbere, proving uniqueness. *Q.E.D.*

Remark: Existence follows from the Fourier series solution — assuming the initial and boundary data and the forcing function are sufficiently nice.

14.3. The Fundamental Solution.

One disadvantage of the Fourier series solution to the heat equation is that it is not nearly as explicit as one might desire for either practical applications, numerical computations, or even further theoretical investigations and developments. An alternative, general approach is based on the idea of the *fundamental solution*, which derives its inspiration from the Green's function method for solving boundary value problems. For the heat equation, the fundamental solution measures the effect of a concentrated heat source.

Let us restrict our attention to homogeneous boundary conditions. The idea is to analyze the case when the initial data $u(0,x) = \delta_y(x) = \delta(x-y)$ is a delta function, which we can interpret as a highly concentrated unit heat source, e.g., a soldering iron or laser beam, that is instantaneously applied at a position y along the bar. The heat will diffuse away from its initial concentration, and the resulting *fundamental solution* is denoted by

$$u(t,x) = F(t,x;y), \qquad \text{with} \qquad F(0,x;y) = \delta(x-y). \tag{14.51}$$

For each fixed y, the fundamental solution, as a function of $t > 0$ and x, must satisfy the differential equation as well as the specified homogeneous boundary conditions.

Once we have determined the fundamental solution, we can then use linear superposition to reconstruct the general solution to the initial-boundary value problem. Namely, we first write the initial data

$$u(0,x) = f(x) = \int_0^\ell \delta(x-y)\, f(y)\, dy \tag{14.52}$$

as a superposition of delta functions, as in (11.36). Linearity implies that the solution is then the corresponding superposition of the responses to those concentrated delta profiles:

$$u(t,x) = \int_0^\ell F(t,x;y)\, f(y)\, dy. \tag{14.53}$$

Assuming that we can differentiate under the integral sign, the fact that $F(t,x;y)$ satisfies the differential equation and the homogeneous boundary conditions for each fixed y immediately implies that the integral (14.53) is also a solution with the correct initial and (homogeneous) boundary conditions.

Unfortunately, most boundary value problems do not have fundamental solutions that can be written down in closed form. An important exception is the case of an infinitely

long homogeneous bar, which requires solving the heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}\,, \qquad \text{for} \qquad -\infty < x < \infty, \qquad t > 0. \tag{14.54}$$

For simplicity, we have chosen units in which the thermal diffusivity is $\gamma = 1$. The solution $u(t,x)$ is defined for all $x \in \mathbb{R}$, and has initial conditions

$$u(0,x) = f(x) \qquad \text{for} \qquad -\infty < x < \infty.$$

In order to specify the solution uniquely, we shall require that the temperature be square-integrable at all times, so that

$$\int_{-\infty}^{\infty} |\, u(t,x)\,|^2\, dx < \infty \qquad \text{for all} \qquad t \geq 0. \tag{14.55}$$

Roughly speaking, we are requiring that the temperature be small at large distances, which are the relevant boundary conditions for this situation.

On an infinite interval, the Fourier series solution to the heat equation becomes a Fourier integral. We write the initial temperature distribution as a superposition

$$f(x) = \int_{-\infty}^{\infty} \widehat{f}(k)\, e^{2\pi\,\mathrm{i}\,k\,x}\, dk$$

of complex exponentials $e^{2\pi\,\mathrm{i}\,k\,x}$, where $\widehat{f}(k)$ is the Fourier transform (13.83) of $f(x)$. The corresponding separable solutions to the heat equation are

$$u_k(t,x) = e^{-4\pi^2 k^2 t}\, e^{2\pi\,\mathrm{i}\,k\,x} = e^{-4\pi^2 k^2 t}\,\big(\cos 2\pi\,\mathrm{i}\,k\,x + \mathrm{i}\,\sin 2\pi\,\mathrm{i}\,k\,x\,\big), \tag{14.56}$$

where the frequency variable k is allowed to assume any real value. We invoke linear superposition to combine these complex solutions into a Fourier integral

$$u(t,x) = \int_{-\infty}^{\infty} e^{-4\pi^2 k^2 t}\, e^{2\pi\,\mathrm{i}\,k\,x}\, \widehat{f}(k)\, dk \tag{14.57}$$

to form the solution to the initial value problem for the heat equation.

In particular, to recover the fundamental solution, we take the initial temperature profile to be a delta function $\delta_y(x) = \delta(x-y)$ concentrated at $x = y$. According to (13.113), its Fourier transform is

$$\widehat{\delta}_y(k) = e^{-2\pi\,\mathrm{i}\,k\,y}.$$

Plugging this into (14.57), and then referring to our table of Fourier transforms, we find the following explicit formula for the fundamental solution:

$$F(t,x;y) = \int_{-\infty}^{\infty} e^{-4\pi^2 k^2 t}\, e^{2\pi\,\mathrm{i}\,k\,(x-y)}\, dk = \frac{1}{2\,\sqrt{\pi\,t}}\, e^{-(x-y)^2/(4t)}. \tag{14.58}$$

As you can verify, for each fixed y, the function $F(t,x;y)$ is indeed a solution to the heat equation for all $t > 0$. In addition,

$$\lim_{t\,\to\,0^+} F(t,x;y) = \begin{cases} 0, & x \neq y, \\ \infty, & x = y. \end{cases}$$

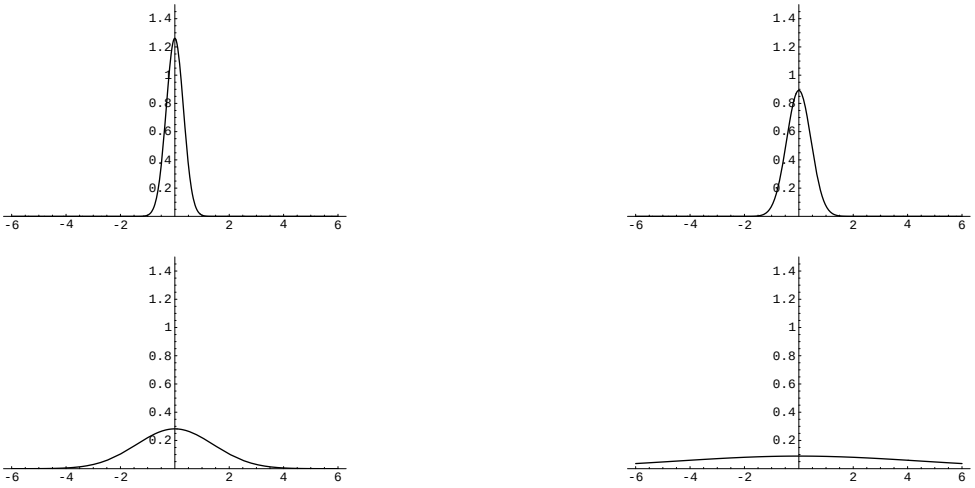


Figure 14.3. The Fundamental Solution to the Heat Equation.

Furthermore, its integral

$$\int_{-\infty}^{\infty} F(t,x;y)\,dx = 1, \tag{14.59}$$

which represents the total heat energy, is constant — in accordance with the law of conservation of energy; see Exercise ∎. Therefore, as $t \to 0^+$, the fundamental solution satisfies the original limiting definition (11.29–30) of the delta function, and so $F(0,x;y) = \delta_y(x)$ has the desired initial temperature profile. In Figure 14.3 we graph $F(t,x;0)$ at times $t = .05, .1, 1., 10.$. It starts out as a delta spike concentrated at the origin and then immediately smoothes out into a tall and narrow bell-shaped curve, centered at $x = 0$. As time increases, the solution shrinks and widens, decaying everywhere to zero. Its maximal amplitude is proportional to $t^{-1/2}$, while its overall width is proportional to $t^{1/2}$. The total heat energy (14.59), which is the area under the graph, remains fixed while gradually spreading out over the entire real line.

Remark: In probability, these exponentially bell-shaped curves are known as *normal* or *Gaussian distributions*. The width of the bell curve corresponds to the *standard deviation*. For this reason, the fundamental solution to the heat equation sometimes referred to as a "Gaussian filter".

Remark: One of the non-physical artifacts of the heat equation is that the heat energy propagates with infinite speed. Indeed, the effect of any initial concentration of heat energy will immediately be felt along the entire length of an infinite bar, because, at any $t > 0$, the fundamental solution is nonzero for all x. (The graphs in Figure 14.3 are a little misleading because they fail to show the extremely small, but still positive, exponentially decreasing tails.) This effect, while more or less negligible at large distances, is nevertheless in clear violation of physical intuition — not to mention relativity that postulates that signals cannot propagate faster than the speed of light. Despite this non-physical property, the heat equation remains an extremely accurate model for heat propagation and similar diffusive phenomena.

With the fundamental solution in hand, we can then adapt the linear superposition formula (14.53) to reconstruct the general solution

$$u(t,x) = \frac{1}{2\sqrt{\pi\, t}} \int_{-\infty}^{\infty} e^{-(x-y)^2/(4t)}\, f(y)\, dy \tag{14.60}$$

to our initial value problem (14.54). Comparing with (13.126), we see that the solutions are obtained by convolution,

$$u(t,x) = g(t,x) * f(x), \qquad \text{where} \qquad g(t,x) = F(t,x;0) = e^{-x^2/(4t)},$$

of the initial data with a one-parameter family of progressively wider and shorter Gaussian filters. Since $u(t,x)$ solves the heat equation, we conclude that Gaussian filter convolution has the same smoothing effect on the initial signal $f(x)$. Indeed, the convolution integral (14.60) serves to replace each initial value $f(x)$ by a weighted average of nearby values, the weight being determined by the Gaussian distribution. Such a weighted averaging has the effect of smoothing out high frequency variations in the signal, and, consequently, the Gaussian convolution formula (14.60) provides an effective method for denoising signals and images. In fact, for practical reasons, the graphs displayed earlier in Figure 14.2 were computed by using a standard numerical integration routine to evaluate the convolution integral (14.60), rather than by a numerical solution scheme for the heat equation.

Example 14.5. An infinite bar is initially heated to unit temperature along a finite interval. This corresponds to an initial temperature profile

$$u(0,x) = f(x) = \sigma(x-a) - \sigma(x-b) = \begin{cases} 1, & a < x < b, \\ 0, & \text{otherwise.} \end{cases}$$

The corresponding solution to the heat equation is obtained by the integral formula (14.60), producing

$$u(t,x) = \frac{1}{2\sqrt{\pi\, t}} \int_a^b e^{-(x-y)^2/(4t)}\, dy = \frac{1}{2}\left[\operatorname{erf}\left(\frac{x-a}{2\sqrt{t}} \right) - \operatorname{erf}\left(\frac{x-b}{2\sqrt{t}} \right) \right], \tag{14.61}$$

where

$$\operatorname{erf} x = \frac{2}{\sqrt{\pi}} \int_0^x e^{-z^2}\, dz \tag{14.62}$$

is known as the *error function* due to its applications in probability and statistics, [**65**]. A graph appears in Figure 14.4. The error function integral cannot be written in terms of elementary functions. Nevertheless, its importance in various applications means that its properties have been well studied, and its values tabulated, [**58**]. In particular, it has asymptotic values

$$\lim_{x \to \infty} \operatorname{erf} x = 1, \qquad \lim_{x \to -\infty} \operatorname{erf} x = -1. \tag{14.63}$$

A graph of the heat equation solution (14.61) when $a = -5$, $b = 5$, at successive times $t = 0., .1, 1, 5, 30, 300$, is displayed in Figure 14.5. Note the initial smoothing or blurring of the sharp interface, followed by a gradual decay to thermal equilibrium.

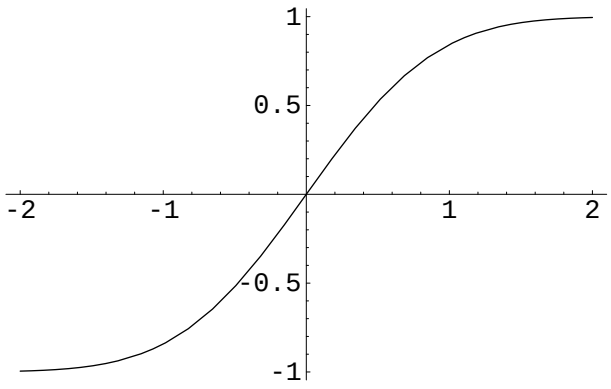


Figure 14.4. The Error Function.

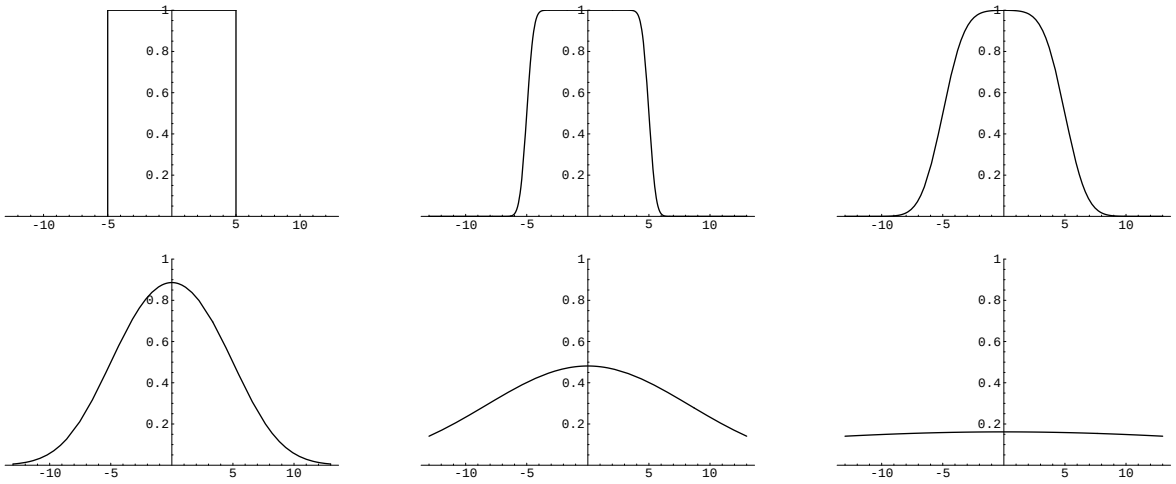


Figure 14.5. Error Function Solution to the Heat Equation.

The Inhomogeneous Heat Equation

The fundamental solution can be also used to solve the inhomogeneous heat equation

$$u_t = u_{xx} + h(t, x), \tag{14.64}$$

that models a bar under an external heat source $h(t, x)$, that might depend upon both position and time. We begin by solving the particular case

$$u_t = u_{xx} + \delta(t - s)\, \delta(x - y), \tag{14.65}$$

whose inhomogeneity represents a heat source of unit magnitude that is concentrated at a position $0 < y < \ell$ and applied instantaneously at a single time $t = s > 0$. Physically, we apply a soldering iron or laser beam to a single spot on the bar for a brief moment. Let us also impose homogeneous initial conditions

$$u(0, x) = 0 \tag{14.66}$$

as well as homogeneous boundary conditions of one of our standard types. The resulting solution

$$u(t, x) = G(t, x; s, y) \tag{14.67}$$

will be referred to as the *general fundamental solution* to the heat equation. Since a heat source which is applied at time s will only affect the solution at later times $t \geq s$, we expect that

$$G(t, x; s, y) = 0 \qquad \text{for all} \qquad t < s. \tag{14.68}$$

Indeed, since $u(t, x)$ solves the unforced heat equation at all times $t < s$ subject to homogeneous boundary conditions and has zero initial temperature, this follows immediately from the uniqueness of the solution to the initial-boundary value problem.

Once we know the general fundamental solution (14.67), we are able to solve the problem for a general external heat source (14.64) by appealing to linearity. We first write the forcing as a superposition

$$h(t, x) = \int_0^\infty \int_0^\ell h(s, y)\, \delta(t - s)\, \delta(x - y)\, dy\, ds \tag{14.69}$$

of concentrated instantaneous heat sources. Linearity allows us to conclude that the solution is given by the self-same superposition formula

$$u(t, x) = \int_0^t \int_0^\ell h(s, y)\, G(t, x; s, y)\, dy\, ds. \tag{14.70}$$

The fact that we only need to integrate over times $0 \leq s \leq t$ follows from (14.68).

Remark: If we have a nonzero initial condition, $u(0, x) = f(x)$, then we appeal to linear superposition to write the solution

$$u(t, x) = \int_0^\ell F(t, x; y)\, f(y)\, dy + \int_0^t \int_0^\ell h(s, y)\, G(t, x; s, y)\, dy\, ds \tag{14.71}$$

as a combination of (a) the solution with no external heat source, but nonzero initial conditions, plus (b) the solution with homogeneous initial conditions but nonzero heat source.

Let us solve the forced heat equation in the case of a infinite bar, so $-\infty < x < \infty$. We begin by computing the general fundamental solution to (14.65), (14.66). As before, we take the Fourier transform of both sides of the partial differential equation with respect to x. In view of (13.113), (13.117), we find

$$\frac{\partial \widehat{u}}{\partial t} + 4\pi^2 k^2\, \widehat{u} = e^{-2\pi \mathrm{i}\, k\, y}\, \delta(t - s), \tag{14.72}$$

which is an inhomogeneous first order ordinary differential equation for the Fourier transform $\widehat{u}(t, k)$ of $u(t, x)$. Assuming $s > 0$, by (14.68), the initial condition is

$$\widehat{u}(0, k) = 0. \tag{14.73}$$

We solve the initial value problem (14.72–73) by the usual method, [**26**]. Multiplying the differential equation by the integrating factor $e^{4\pi^2 k^2 t}$ yields

$$\frac{\partial}{\partial t}\left(e^{4\pi^2 k^2 t}\, \widehat{u} \right) = e^{4\pi^2 k^2 t - 2\pi \mathrm{i}\, k\, y}\, \delta(t - s).$$

Integrating both sides from 0 to t and using the initial condition, we find

$$\widehat{u}(t,k) = e^{4\pi^2 k^2 (s-t) - 2\pi \mathrm{i}\, k y}\, \sigma(t-s),$$

where $\sigma(t)$ is the usual step function (11.42). Finally, we apply the inverse Fourier transform formula (13.86) and then (14.58), we deduce that

$$\begin{aligned} G(t,x;s,y) = u(t,x) &= \sigma(t-s) \int_{-\infty}^{\infty} e^{4\pi^2 k^2 (s-t) + 2\pi \mathrm{i}\, k (x-y)}\, dk \\ &= \frac{\sigma(t-s)}{2\sqrt{\pi(t-s)}}\, \exp\left[-\frac{(x-y)^2}{4(t-s)} \right] = \sigma(t-s)\, F(t-s,x;y)\,. \end{aligned}$$

Thus, the general fundamental solution is obtained by translating the fundamental solution $F(t,x;y)$ for the initial value problem to a starting time of $t = s$ instead of $t = 0$. Thus, an initial condition has the same aftereffect on the temperature as an instantaneous applied heat source of the same magnitude. Finally, the superposition principle (14.70) produces the solution

$$u(t,x) = \int_0^t \int_{-\infty}^{\infty} \frac{h(s,y)}{2\sqrt{\pi(t-s)}}\, \exp\left[-\frac{(x-y)^2}{4(t-s)} \right] dy\, ds. \tag{14.74}$$

to the heat equation with source term on an infinite bar.

The Root Cellar Problem

As a final example, we discuss a problem that involves analysis of the heat equation on a semi-infinite interval. The question is: how deep should you dig a root cellar? In the prerefrigeration era, a root cellar was used to keep food cool in the summer, but not freeze in the winter. We assume that the temperature in the earth only depends on the depth and the time of year. Let $u(t,x)$ denote the deviation in the temperature in the earth, from its annual mean, at depth $x > 0$ and time t. We shall assume that the temperature at the earth's surface, $x = 0$, fluctuates in a periodic manner; specifically, we set

$$u(t,0) = a\, \cos \omega\, t, \tag{14.75}$$

where the oscillatory frequency

$$\omega = \frac{2\pi}{365.25 \text{ days}} = 2.0 \times 10^{-7} \text{sec}^{-1} \tag{14.76}$$

refers to yearly temperature variations. In this model, we shall ignore daily temperature fluctuations as their effect is not significant below a very thin surface layer. At large depth the temperature is assumed to be unvarying:

$$u(t,x) \;\longrightarrow\; 0 \qquad \text{as} \qquad x \;\longrightarrow\; \infty, \tag{14.77}$$

where 0 refers to the mean temperature.

Thus, we must solve the heat equation on a semi-infinite bar $0 < x < \infty$, with time-dependent boundary conditions (14.75), (14.77) at the ends. The analysis will be simplified

a little if we replace the cosine by a complex exponential, and so look for a complex solution with boundary conditions

$$u(t,0) = a\, e^{\,\mathrm{i}\,\omega\, t}, \qquad \lim_{x \to \infty} u(t,x) = 0. \tag{14.78}$$

Let us try a separable solution of the form

$$u(t,x) = v(x)\, e^{\,\mathrm{i}\,\omega\, t}. \tag{14.79}$$

Substituting this expression into the heat equation $u_t = \gamma\, u_{xx}$ leads to

$$\mathrm{i}\,\omega\, v(x)\, e^{\,\mathrm{i}\,\omega\, t} = \gamma\, v''(x)\, e^{\,\mathrm{i}\,\omega\, t}.$$

Canceling the common exponential factors, we conclude that $v(x)$ should solve the boundary value problem

$$\gamma\, v''(x) = \,\mathrm{i}\,\omega\, v, \qquad v(0) = a, \qquad \lim_{x \to \infty} v(x) = 0.$$

The solutions to the ordinary differential equation are

$$v_1(x) = e^{\sqrt{\,\mathrm{i}\,\omega/\gamma}\,x} = e^{\sqrt{\omega/2\gamma}\,(1+\mathrm{i}\,)\,x}, \qquad v_2(x) = e^{-\sqrt{\,\mathrm{i}\,\omega/\gamma}\,x} = e^{-\sqrt{\omega/2\gamma}\,(1+\mathrm{i}\,)\,x}.$$

The first solution is exponentially growing as $x \to \infty$, and so not appropriate to our problem. The solution to the boundary value problem must therefore be a multiple,

$$v(x) = a\, e^{-\sqrt{\omega/2\gamma}\,(1+\mathrm{i}\,)\,x}$$

of the exponentially decaying solution. Substituting back into (14.79), we find the (complex) solution to the root cellar problem to be

$$u(t,x) = a\, e^{-x\sqrt{\omega/2\gamma}}\, e^{\,\mathrm{i}\,\omega\left(t - \sqrt{\omega/2\gamma}\right)x}. \tag{14.80}$$

The corresponding real solution is obtained by taking the real part,

$$u(t,x) = a\, e^{-x\sqrt{\omega/2\gamma}}\, \cos\left(\omega\, t - \sqrt{\frac{\omega}{2\gamma}}\, x\right). \tag{14.81}$$

The first term in (14.81) is exponentially decaying as a function of the depth. Thus, the further down one goes, the less noticeable the effect of the surface temperature fluctuations. The second term is periodic with the same annual frequency ω. The interesting feature is the phase lag in the response. The temperature at depth x is out of phase with respect to the surface temperature fluctuations, having an overall phase lag

$$\delta = \sqrt{\frac{\omega}{2\gamma}}\, x$$

that depends linearly on depth. In particular, a cellar built at a depth where δ is an odd multiple of π will be completely out of phase, being hottest in the winter, and coldest in

the summer. Thus, the (shallowest) ideal depth at which to build a root cellar would take $\delta = \pi$, corresponding to a depth of

$$x = \pi \sqrt{\frac{2\gamma}{\omega}}\,.$$

For typical soils in the earth, $\gamma \approx 10^{-6}$ meters^2 sec^{-1}, and hence, by (14.76), $x \approx 9.9$ meters. However, at this depth, the relative amplitude of the oscillations is

$$e^{-x\sqrt{\omega/2\gamma}} = e^{-\pi} = .04$$

and hence there is only a 4% temperature fluctuation. In Minnesota, the temperature varies, roughly, from $-40°$C to $+40°$C, and hence our 10 meter deep root cellar would experience only a 3.2°C annual temperature deviation from the winter, when it is the warmest, to the summer, where it is the coldest. Building the cellar twice as deep would lead to a temperature fluctuation of .2%, now in phase with the surface variations, which means that the cellar is, for all practical purposes, at constant temperature year round.

14.4. The Wave Equation.

The second important class of dynamical partial differential equations are those modeling vibrations of continuous media. As we saw in Chapter 9, Newton's Law implies that the free vibrations of a discrete mechanical system are governed by a second order system of ordinary differential equations of the form

$$M\,\frac{d^2\mathbf{u}}{dt^2} = -K\,\mathbf{u},$$

in which M is the positive definite, diagonal mass matrix, while $K = A^*A = A^TCA$ is the positive definite (or semi-definite in the case of an unstable system) stiffness matrix.

The corresponding dynamical equations describing the small vibrations of continuous media take an entirely analogous form

$$\rho\,\frac{\partial^2 u}{\partial t^2} = -K[u]. \tag{14.82}$$

In this framework, ρ describes the density, while $K = L^* \circ L$ is the same self-adjoint differential operator, with appropriate boundary conditions, that appears in the equilibrium equations (11.89). For one-dimensional media, such as a vibrating bar or string, we are led to a partial differential equation in the particular form

$$\rho(x)\,\frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial x}\left(\kappa(x)\,\frac{\partial u}{\partial x}\right), \qquad 0 < x < \ell, \tag{14.83}$$

where $\rho(x)$ is the density of the bar or string at position x, while $\kappa(x) > 0$ denotes its stiffness or tension. The second order partial differential equation (14.83) models the dynamics of vibrations and waves in a broad range of applications, including elastic vibrations of a bar, sound vibrations in a column of air, e.g., inside a wind instrument, and also transverse

vibrations of a string, e.g., a violin string. (However, bending vibrations of a beam lead to a fourth order partial differential equation; see Exercise ■.) The wave equation is also used to model small amplitude water waves, electromagnetic waves, including light, radio and microwaves, gravitational waves, and many others. A detailed derivation of the model from first principles in the case of a vibrating string can be found in [**181**].

To specify the solution, we must impose suitable boundary conditions. The usual suspects — Dirichlet, Neumann, mixed, and periodic boundary conditions — continue to play a central role, and have immediate physical interpretations. Tying down an end of the string imposes a Dirichlet condition $u(t,0) = \alpha$, while a free end is prescribed by a homogeneous Neumann boundary condition $u_x(t,0) = 0$. Periodic boundary conditions, as in (14.11), correspond to the vibrations of a circular ring. As with all second order Newtonian systems of ordinary differential equations, the solution to the full boundary value problem for the second order partial differential equation will then be uniquely specified by its initial displacement and initial velocity:

$$u(0,x) = f(x), \qquad \frac{\partial u}{\partial t}\,(0,x) = g(x). \tag{14.84}$$

The simplest situation occurs when the medium is homogeneous, and so both its density and stiffness are constant. Then the general vibration equation (14.83) reduces to the one-dimensional *wave equation*

$$\frac{\partial^2 u}{\partial t^2} = c^2\,\frac{\partial^2 u}{\partial x^2}\ . \tag{14.85}$$

The constant

$$c = \sqrt{\frac{\kappa}{\rho}}\ > \ 0 \tag{14.86}$$

is known as the *wave speed*, for reasons that will soon become apparent.

The method for solving such second order systems is motivated by our solution in the discrete case discussed in Section 9.5. To keep matters simple, we shall concentrate on the wave equation (14.85), although the method is easily extended to the general homogeneous Newtonian system (14.83). We will try a separable solution with trigonometric time dependence:

$$u(t,x) = \cos(\omega t)\,v(x), \tag{14.87}$$

in which both the frequency ω and profile $v(x)$ are to be determined. Differentiating (14.87), we find

$$\frac{\partial^2 u}{\partial t^2} = -\omega^2\cos(\omega t)\,v(x), \qquad \frac{\partial^2 u}{\partial x^2} = \cos(\omega t)\,v''(x).$$

Substituting into the wave equation (14.85) and canceling the common cosine factors, we deduce that $v(x)$ must satisfy the ordinary differential equation

$$c^2\ \frac{d^2 v}{dx^2} + \omega^2\,v = 0. \tag{14.88}$$

Thus $\omega^2 = \lambda$ can be viewed as an eigenvalue with eigenfunction $v(x)$ for the second order differential operator $K = -c^2\,D^2$. Ignoring the boundary conditions for the moment, if

$\omega > 0$, the solutions are the trigonometric functions $\cos \frac{\omega x}{c}$, $\sin \frac{\omega x}{c}$. Thus, (14.87) leads to the pair of explicit solutions

$$\cos \omega t \ \cos \frac{\omega x}{c}, \qquad \cos \omega t \ \sin \frac{\omega x}{c},$$

to the wave equation. Now, the computation will work just as well with a sine function, which yields two additional solutions

$$\sin \omega t \ \cos \frac{\omega x}{c}, \qquad \sin \omega t \ \sin \frac{\omega x}{c}.$$

Each of these four solutions represents a spatially periodic standing wave form of period $2\pi c/\omega$, that is vibrating with frequency ω. Observe that the smaller scale waves vibrate faster.

On the other hand, if $\omega = 0$, then (14.88) has the solution $v = \alpha x + \beta$, leading to the solutions

$$u(t,x) = 1, \qquad \text{and} \qquad u(t,x) = x. \tag{14.89}$$

The first is a constant, nonvibrating solution, while the second is also constant in time, but will typically not satisfy the boundary conditions and so can be discarded. As we learned in Chapter 9, the existence of a zero eigenvalue corresponds to an unstable mode in the physical system, in which the displacement grows linearly in time. In the present situation, these correspond to the two additional solutions

$$u(t,x) = t, \qquad \text{and} \qquad u(t,x) = x\,t, \tag{14.90}$$

of the wave equation. Again, the second solution will typically not satisfy the homogeneous boundary conditions, and can usually be ignored. Such null eigenfunction modes will only arise in unstable situations.

The boundary conditions will distinguish the particular eigenvalues and natural frequencies of vibration. Consider first the case of a string of length ℓ with two fixed ends, and thus subject to homogeneous Dirichlet boundary conditions

$$u(t,0) = 0 = u(t,\ell).$$

This forms a positive definite boundary value problem, and so there is no unstable mode. Indeed, the eigenfunctions of the boundary value problem (14.88) with Dirichlet boundary conditions $v(0) = 0 = v(\ell)$ were found in (14.17):

$$v_n(x) = \sin \frac{n\pi x}{\ell}, \qquad \omega_n = \frac{n\pi c}{\ell}, \qquad n = 1, 2, 3, \ldots .$$

Therefore, we can write the general solution as a Fourier sine series

$$\begin{aligned} u(t,x) &= \sum_{n=1}^{\infty} \left[b_n \cos \frac{n\pi c t}{\ell} \ \sin \frac{n\pi x}{\ell} + d_n \sin \frac{n\pi c t}{\ell} \ \sin \frac{n\pi x}{\ell} \right] \\ &= \sum_{n=1}^{\infty} r_n \cos \left(\frac{n\pi c t}{\ell} + \delta_n \right) \sin \frac{n\pi x}{\ell}. \end{aligned} \tag{14.91}$$

The solution is thus a linear combination of the natural Fourier modes vibrating with frequencies

$$\omega_n = \frac{n\pi c}{\ell} = \frac{n\pi}{\ell}\sqrt{\frac{\kappa}{\rho}}, \qquad n = 1, 2, 3, \dots . \tag{14.92}$$

The longer the length ℓ of the string, or the higher its density ρ, the slower the vibrations; whereas increasing its stiffness or tension κ speeds them up — in exact accordance with our physical intuition.

The Fourier coefficients b_n and d_n in (14.91) will be uniquely determined by the initial conditions (14.84). Differentiating the series term by term, we discover that we must represent the initial displacement and velocity as Fourier sine series

$$u(0,x) = \sum_{n=1}^{\infty} b_n \sin\frac{n\pi x}{\ell} = f(x), \qquad \frac{\partial u}{\partial t}(0,x) = \sum_{n=1}^{\infty} d_n \frac{n\pi c}{\ell}\sin\frac{n\pi x}{\ell} = g(x).$$

Therefore,

$$b_n = \frac{2}{\ell}\int_0^{\ell} f(x)\sin\frac{n\pi x}{\ell}\,dx, \qquad n = 1, 2, 3, \dots .$$

are the Fourier sine coefficients (12.83) of the initial displacement $f(x)$, while

$$d_n = \frac{2}{n\pi c}\int_0^{\ell} g(x)\sin\frac{n\pi x}{\ell}\,dx, \qquad n = 1, 2, 3, \dots .$$

are rescaled versions of the Fourier sine coefficients of the initial velocity $g(x)$.

Example 14.6. A string of unit length is held taut in the center and then released. Our goal is to describe the ensuing vibrations. Let us assume the physical units are chosen so that $c^2 = 1$, and so we are asked to solve the initial-boundary value problem

$$u_{tt} = u_{xx}, \qquad u(0,x) = f(x), \qquad u_t(0,x) = 0, \qquad u(t,0) = u(t,1) = 0. \tag{14.93}$$

To be specific, we assume that the center of the string has been displaced by half a unit, and so the initial displacement is

$$f(x) = \begin{cases} x, & 0 \le x \le \frac{1}{2}, \\ 1 - x, & \frac{1}{2} \le x \le 1. \end{cases}$$

The vibrational frequencies $\omega_n = n\pi$ are the integral multiples of π, and so the natural modes of vibration are

$$\cos n\pi t\, \sin n\pi x \qquad \text{and} \qquad \sin n\pi t\, \sin n\pi x \qquad \text{for} \qquad n = 1, 2, \dots .$$

Consequently, the general solution to the boundary value problem is

$$u(t,x) = \sum_{n=1}^{\infty} \big[\, b_n \cos n\pi t\, \sin n\pi x + d_n \sin n\pi t\, \sin n\pi x \,\big],$$

where

$$b_n = 2\int_0^1 f(x)\sin n\pi x\,dx = \begin{cases} 4\displaystyle\int_0^{1/2} x\sin n\pi x\,dx = \frac{4\,(-1)^k}{(2k+1)^2\pi^2}, & n = 2k+1, \\ 0, & n = 2k, \end{cases}$$

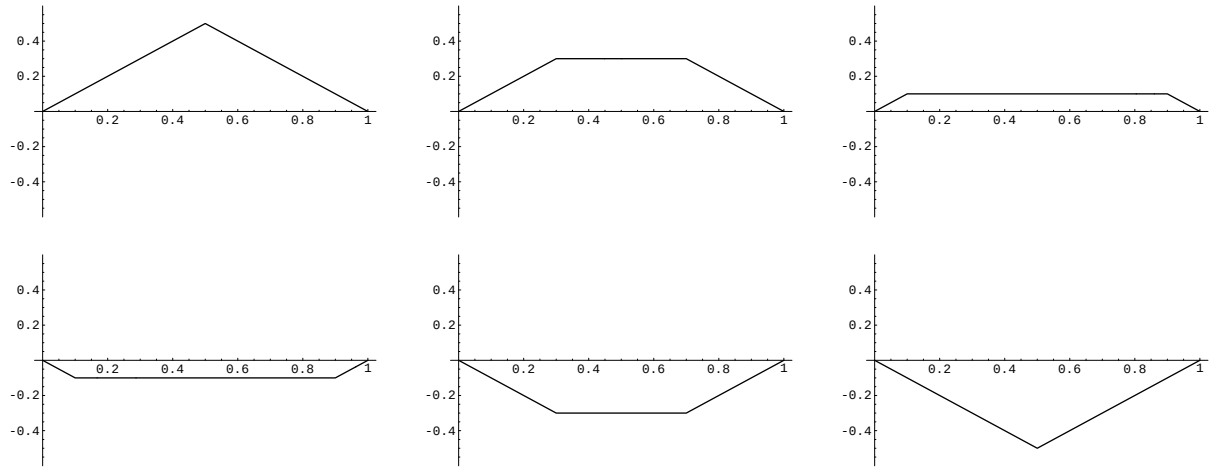


Figure 14.6. Plucked String Solution of the Wave Equation.

are the Fourier sine coefficients of the initial displacement, while $d_n = 0$ are the Fourier sine coefficients of the initial velocity. Therefore, the solution is the Fourier sine series

$$u(t,x) = 4 \sum_{k=0}^{\infty} (-1)^k \, \frac{\cos(2k+1)\pi t \, \sin(2k+1)\pi x}{(2k+1)^2 \pi^2} \,, \tag{14.94}$$

whose graph at times $t = 0, .2, .4, .6, .8, 1.$ is depicted in Figure 14.6. At time $t = 1$, the original displacement is reproduced exactly, but upside down. The subsequent dynamics proceeds as before, but in mirror image form. The original displacement reappears at time $t = 2$, after which time the motion is periodically repeated. Interestingly, at times $t_k = .5, 1.5, 2.5, \ldots$, the displacement is identically zero: $u(t_k, x) \equiv 0$, although the velocity $u_t(t_k, x) \not\equiv 0$. The solution appears to be piecewise affine, i.e., its graph is a collection of straight lines. This fact will be verified in Exercise ■, where you are asked to construct an exact analytical formula for this solution. Unlike the heat equation, the wave equation does *not* smooth out discontinuities and corners in the initial data.

While the series form (14.91) of the solution is not entirely satisfying, we can still use it to deduce important qualitative properties. First of all, since each term is periodic in t with period $2\,\ell/c$, the entire solution is time periodic with that period: $u(t + 2\,\ell/c, x) = u(t, x)$. In fact, after half the period, at time $t = \ell/c$, the solution reduces to

$$u\left(\frac{\ell}{c}\,, x\right) = \sum_{n=1}^{\infty} (-1)^n \, b_n \, \sin \frac{n\pi x}{\ell} = - \sum_{n=1}^{\infty} b_n \, \sin \frac{n\pi\,(\ell - x)}{\ell} = - u(0, \ell - x) = - f(\ell - x).$$

In general,

$$u\left(t + \frac{\ell}{c}\,,\; x\right) = - u(t, \ell - x), \qquad\qquad u\left(t + \frac{2\,\ell}{c}\,,\; x\right) = u(t, x). \tag{14.95}$$

Therefore, the initial wave form is reproduced, first as an upside down mirror image of itself at time $t = \ell/c$, and then in its original form at time $t = 2\,\ell/c$. This has the important consequence that vibrations of (homogeneous) one-dimensional media are purely periodic

phenomena! There is no quasi-periodicity because the fundamental frequencies are all integer multiples of each other.

Remark: The preceding analysis has important musical consequences. To the human ear, sonic vibrations that are integral multiples of a single frequency are harmonic, whereas those that admit quasi-periodic vibrations, with irrationally related frequencies, sound percussive. This is why most tonal instruments rely on vibrations in one dimension, be it a violin string, a column of air in a wind instrument (flute, clarinet, trumpet or saxophone), a xylophone bar or a triangle. On the other hand, most percussion instruments rely on the vibrations of two-dimensional media, e.g., drums and cymbals, or three-dimensional solid bodies, e.g., blocks. As we shall see in Chapters 17 and 18, the frequency ratios of the latter are typically irrational.

A bar with both ends left free, and so subject to the Neumann boundary conditions

$$\frac{\partial u}{\partial x}\,(t,0) = 0 = \frac{\partial u}{\partial x}\,(t,\ell), \tag{14.96}$$

will have a slightly different behavior, owing to the instability of the underlying equilibrium equations. The eigenfunctions of (14.88) with Neumann boundary conditions $v'(0) = 0 = v'(\ell)$ are now

$$v_n(x) = \cos\frac{n\,\pi\,x}{\ell} \qquad \text{with} \qquad \omega_n = \frac{n\,\pi\,c}{\ell}, \qquad n = 0, 1, 2, 3, \ldots\,.$$

The resulting solution takes the form of a Fourier cosine series

$$u(t,x) = a_0 + c_0\,t + \sum_{n=1}^{\infty} \left(a_n \cos\frac{n\,\pi\,c\,t}{\ell}\,\cos\frac{n\,\pi\,x}{\ell} + c_n \sin\frac{n\,\pi\,c\,t}{\ell}\,\cos\frac{n\,\pi\,x}{\ell} \right). \tag{14.97}$$

In accordance with (14.89), the first two terms come from the null eigenfunction $v_0(x) = 1$ with $\omega_0 = 0$. The bar vibrates with the same fundamental frequencies (14.92) as in the fixed end case, but there is now an additional unstable mode $c_0\,t$ that is no longer periodic, but grows linearly in time.

Substituting (14.97) into the initial conditions (14.84), we find the Fourier coefficients are prescribed, as before, by the initial displacement and velocity,

$$a_n = \frac{2}{\ell}\int_0^{\ell} f(x)\,\cos\frac{n\,\pi\,x}{\ell}\,dx, \qquad c_n = \frac{2}{n\,\pi\,c}\int_0^{\ell} g(x)\,\cos\frac{n\,\pi\,x}{\ell}\,dx, \qquad n = 1, 2, 3, \ldots\,.$$

The order zero coefficients[†],

$$a_0 = \frac{1}{\ell}\int_0^{\ell} f(x)\,dx, \qquad\qquad c_0 = \frac{1}{\ell}\int_0^{\ell} g(x)\,dx,$$

[†] Note that, we have not included the usual $\frac{1}{2}$ factor in the constant terms in the Fourier series (14.97).

are equal to the average initial displacement and average initial velocity of the bar. In particular, when $c_0 = 0$ there is no net initial velocity, and the unstable mode is not excited. In this case, the solution is time-periodic, oscillating around the position given by the average initial displacement. On the other hand, if $c_0 \neq 0$, the bar will move off with constant average speed c_0, all the while vibrating at the same fundamental frequencies.

Similar considerations apply to the periodic boundary value problem for the wave equation on a circular ring. The details are left as Exercise ■ for the reader.

Forcing and Resonance

In Section 9.6, we learned that periodically forcing an undamped mechanical structure (or a resistanceless electrical circuit) at a frequency that is distinct from its natural vibrational frequencies leads, in general, to a quasi-periodic response. The solution is a sum of the unforced vibrations superimposed with an additional vibrational mode at the forcing frequency. However, if forced at one of its natural frequencies, the system may go into a catastrophic resonance.

The same type of quasi-periodic/resonant response is also observed in the partial differential equations governing periodic vibrations of continuous media. To keep the analysis as simple as possible, we restrict our attention to the forced wave equation for a homogeneous bar

$$\frac{\partial^2 u}{\partial t^2} = c^2 \, \frac{\partial^2 u}{\partial x^2} + F(t,x), \tag{14.98}$$

subject to specified homogeneous boundary conditions. The external forcing function $F(t,x)$ may depend upon both time t and position x. We will be particularly interested in a periodically varying external force of the form

$$F(t,x) = \cos(\omega\, t)\, h(x), \tag{14.99}$$

where the function $h(x)$ is fixed.

As always, the solution to an inhomogeneous linear equation can be written as a combination,

$$u(t,x) = u_\star(t,x) + z(t,x) \tag{14.100}$$

of a particular solution $u_\star(t,x)$ plus the general solution $z(t,x)$ to the homogeneous equation, namely

$$\frac{\partial^2 z}{\partial t^2} = c^2 \, \frac{\partial^2 z}{\partial x^2}\,. \tag{14.101}$$

The boundary and initial conditions will serve to uniquely prescribe the solution $u(t,x)$, but there is some flexibility in its two constituents (14.100). For instance, we may ask that the particular solution $u_\star$ satisfy the homogeneous boundary conditions along with zero (homogeneous) initial conditions, and thus represents the pure response of the system to the forcing. The homogeneous solution $z(t,x)$ will then reflect the effect of the initial and boundary conditions unadulterated by the external forcing. The final solution will equal the sum of the two individual responses.

In the case of periodic forcing (14.99), we look for a particular solution

$$u_\star(t,x) = \cos(\omega\, t)\, v_\star(x) \tag{14.102}$$

that vibrates at the forcing frequency. Substituting the ansatz (14.102) into the equation (14.98), and canceling the common cosine factors, we discover that $v_\star(x)$ must satisfy the boundary value problem prescribed by

$$-c^2\, v_\star'' - \omega^2\, v_\star = h(x), \tag{14.103}$$

supplemented by the relevant homogeneous boundary conditions — Dirichlet, Neumann, mixed, or periodic.

At this juncture, there are two possibilities. If the unforced, homogeneous boundary value problem has only the trivial solution $v \equiv 0$, then there is a solution to the forced boundary value problem for any form of the forcing function $h(x)$. On the other hand, the homogeneous boundary value problem has a nontrivial solution $v(x)$ when $\omega^2 = \lambda$ is an eigenvalue, and so ω is a natural frequency of vibration to the homogeneous problem; the solution $v(x)$ is the corresponding eigenfunction appearing in the solution series (14.91). In this case, according to the Fredholm alternative, Theorem 5.55, the boundary value problem (14.103) has a solution if and only if the forcing function $h(x)$ is orthogonal to the eigenfunction(s):

$$\langle\, h\,, v\,\rangle = 0. \tag{14.104}$$

See Example 11.3 and Exercise 11.5.7 for details. If we force in a resonant manner — meaning that (14.104) is not satisfied — then the solution will be a resonantly growing vibration

$$u_\star(t,x) = t\, \sin(\omega\, t)\, v_\star(x).$$

In a real-world situation, such large resonant (or near resonant) vibrations will either cause a catastrophic breakdown, e.g., the bar breaks or the string snaps, or will send the system into a different, nonlinear regime that helps mitigate the resonant effects, but is no longer modeled by the simple linear wave equation.

Example 14.7. As a specific example, consider the initial-boundary value problem modeling the forced vibrations of a uniform bar of unit length and fixed at both ends:

$$\begin{aligned} u_{tt} &= c^2\, u_{xx} + \cos(\omega\, t)\, h(x), \\ u(t,0) = 0 = u(t,1), \qquad u(0,x) &= f(x), \qquad u_t(0,x) = g(x). \end{aligned} \tag{14.105}$$

The particular solution will have the nonresonant form (14.102) provided there exists a solution $v_\star(x)$ to the boundary value problem

$$c^2\, v_\star'' + \omega^2\, v_\star = -\,h(x), \qquad\qquad v_\star(0) = 0 = v_\star(1). \tag{14.106}$$

The natural frequencies and associated eigenfunctions are

$$\omega_n = n\, c\, \pi, \qquad\qquad v_n(x) = \sin n\,\pi\, x, \qquad\qquad n = 1, 2, 3, \dots\ .$$

The boundary value problem (14.106) will have a solution, and hence the forcing is not resonant, provided either $\omega \neq \omega_n$ is not a natural frequency, or $\omega = \omega_n$, but

$$0 = \langle\, h\,, v_n\,\rangle = \int_0^1 h(x) \sin n\,\pi\, x\, dx \tag{14.107}$$

is orthogonal to the associated eigenfunction. Otherwise, the forcing profile will induce a resonant response.

For example, under periodic forcing of frequency ω with trigonometric profile $h(x) \equiv \sin k\pi x$, the particular solution to (14.106) is

$$v_\star(x) = \frac{\sin k\pi x}{\omega^2 - k^2\pi^2 c^2}, \qquad \text{so that} \qquad u_\star(t,x) = \frac{\cos \omega t \ \sin k\pi x}{\omega^2 - k^2\pi^2 c^2}, \tag{14.108}$$

which is a valid solution as long as $\omega \neq \omega_k = k\pi c$. Note that we may allow the forcing frequency $\omega = \omega_n$ to coincide with any other resonant forcing frequency, $n \neq k$, because the sine profiles are mutually orthogonal and so the nonresonance condition (14.107) holds. On the other hand, if $\omega = \omega_k = k\pi c$, then the particular solution

$$u_\star(t,x) = \frac{t \sin k\pi c t \ \sin k\pi x}{2k\pi c}, \tag{14.109}$$

is resonant, and grows linearly in time.

To obtain the full solution to the initial-boundary value problem, we write $u = u_\star + z$ where $z(t,x)$ must satisfy

$$z_{tt} - c^2 z_{xx} = 0, \qquad z(t,0) = 0 = z(t,1),$$

along with the modified initial conditions

$$z(0,x) = f(x) \ - \ \frac{\sin k\pi x}{\omega^2 - k^2\pi^2 c^2}, \qquad \frac{\partial u}{\partial x}(0,x) = g(x),$$

stemming from the fact that the particular solution (14.108) has non-trivial initial data. (In the resonant case (14.109), there is no extra term in the initial data.) Note that, the closer ω is to the resonant frequency, the larger the modification of the initial data, and hence the larger the response of the system to the periodic forcing. As before, the solution $z(t,x)$ to the homogeneous equation can be written as a Fourier sine series (14.91). The final formulae are left to the reader to write out.

14.5. d'Alembert's Solution.

The one-dimensional wave equation is distinguished by admitting an explicit solution formula, originally discovered by the eighteenth century French mathematician Jean d'Alembert, that entirely avoids the complicated Fourier series form. D'Alembert's solution provides additional valuable insight into the behavior of the solutions. Unfortunately, unlike series methods that have a very broad range of applicability, d'Alembert's method only succeeds in this one very special situation: the homogeneous wave equation in a single space variable.

The starting point is to write the wave equation (14.85) in the suggestive form

$$\Box u = (\partial_t^2 - c^2\,\partial_x^2)\,u = u_{tt} - c^2\,u_{xx} = 0. \tag{14.110}$$

Here $\Box = \partial_t^2 - c^2\,\partial_x^2$ is a common mathematical notation for the *wave operator*, while ∂_t, ∂_x are convenient shorthands for the partial derivative operators with respect to t and

x. We note that $\Box$ is a linear, second order partial differential operator. In analogy with the elementary polynomial factorization

$$t^2 - c^2\, x^2 = (t - c\, x)(t + c\, x),$$

we shall factor the wave operator into a product of two first order partial differential operators:

$$\Box = \partial_t^2 - c^2\, \partial_x^2 = (\partial_t - c\, \partial_x)\, (\partial_t + c\, \partial_x). \tag{14.111}$$

Now, if the second factor annihilates the function $u(t, x)$, meaning

$$(\partial_t + c\, \partial_x)\, u = u_t + c\, u_x = 0, \tag{14.112}$$

then u is automatically a solution to the wave equation, since

$$\Box\, u = (\partial_t - c\, \partial_x)\, (\partial_t + c\, \partial_x)\, u = (\partial_t - c\, \partial_x)\, 0 = 0.$$

In other words, every solution to the simpler first order partial differential equation (14.112) is a solution to the wave equation (14.85). (The converse is, of course, not true.)

It is relatively easy to solve linear[†] first order partial differential equations.

Proposition 14.8. *Every solution $u(t, x)$ to the partial differential equation*

$$\frac{\partial u}{\partial t} + c\, \frac{\partial u}{\partial x} = 0 \tag{14.113}$$

has the form

$$u(t, x) = p(x - c\, t), \tag{14.114}$$

where $p(\xi)$ is an arbitrary function of the characteristic variable $\xi = x - c\, t$.

Proof: We adopt a linear change of variables to rewrite the solution

$$u(t, x) = p(t, x - c\, t) = p(t, \xi)$$

as a function of the characteristic variable ξ and the time t. Applying the chain rule, we express the derivatives of u in terms of the derivatives of p as follows:

$$\frac{\partial u}{\partial t} = \frac{\partial p}{\partial t} - c\, \frac{\partial p}{\partial \xi}\,, \qquad\qquad \frac{\partial u}{\partial x} = \frac{\partial p}{\partial \xi}\,,$$

and hence

$$\frac{\partial u}{\partial t} + c\, \frac{\partial u}{\partial x} = \frac{\partial p}{\partial t} - c\, \frac{\partial p}{\partial \xi} + c\, \frac{\partial p}{\partial \xi} = \frac{\partial p}{\partial t}\,.$$

Therefore, u is a solution to (14.113) if and only if $p(t, \xi)$ is a solution to the very simple partial differential equation

$$\frac{\partial p}{\partial t} = 0.$$

[†] See Chapter 22 for more details, including extensions to first order nonlinear partial differential equations.

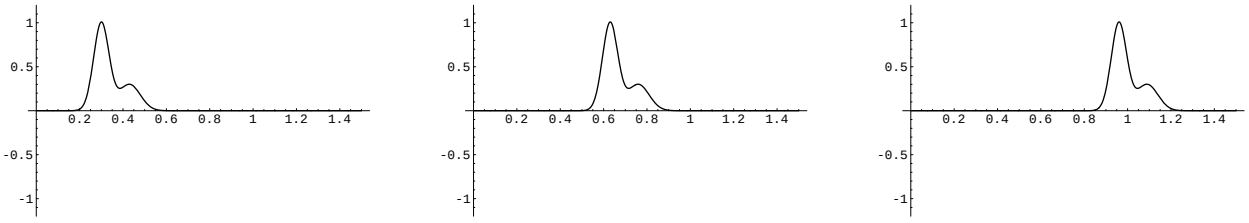


Figure 14.7. Traveling Wave.

This clearly[‡] implies that $p = p(\xi)$ does not depend on the variable t, and hence

$$u = p(\xi) = p(x - c\,t)$$

is of the desired form. *Q.E.D.*

Therefore, *any* function of the characteristic variable, e.g., $\xi^2 + 1$, or $\cos \xi$, or e^{ξ}, will produce a corresponding solution, $(x - c\,t)^2 + 1$, or $\cos(x - c\,t)$, or $e^{x - c\,t}$, to the first order partial differential equation (14.113), and hence a solution to the wave equation (14.85). The functions of the form $u(t, x) = p(x - c\,t)$ are known as *traveling waves*. At $t = 0$ the wave has the initial profile $u(0, x) = p(x)$. As t progresses, the wave moves to the *right*, with constant speed $c > 0$ and unchanged in form; see Figure 14.7. For this reason, (14.113) is sometimes referred to as the *one-way* or *unidirectional wave equation*. Proposition 14.8 tells us that every traveling wave with wave speed c is a solution to the full wave equation (14.85). But keep in mind that such solutions do not necessarily respect the boundary conditions, which, when present, will affect their ultimate behavior.

Now, since c is constant, the factorization (14.111) can be written equally well in the reverse order:

$$\Box = \partial_t^2 - c^2\,\partial_x^2 = (\partial_t + c\,\partial_x)\,(\partial_t - c\,\partial_x). \tag{14.115}$$

The same argument tells us that any solution to the alternative first order partial differential equation

$$\frac{\partial u}{\partial t} - c\,\frac{\partial u}{\partial x} = 0, \tag{14.116}$$

also provides a solution to the wave equation. This is also a unidirectional wave equation, but with the opposite wave speed $-c$. Applying Proposition 14.8, now with c replaced by $-c$, we conclude that the general solution to (14.116) has the form

$$u(t, x) = q(x + c\,t) \tag{14.117}$$

where $q(\eta)$ is an arbitrary differentiable function of the alternate characteristic variable $\eta = x + c\,t$. The solutions (14.117) represent traveling waves moving to the *left* with constant speed $c > 0$ and unchanged in form.

[‡] More rigorously, one should also assume that, at each time t, the domain of definition of $p(\xi)$ is a connected interval. A similar technical restriction should be imposed upon the solutions in the statement of Proposition 14.8. See Exercise ■ for a detailed example.

The wave equation (14.110) is *bidirectional* and admits both left and right traveling wave solutions. Linearity of the wave equation implies that the sum of solutions is again a solution. In this way, we can produce solutions which are superpositions of left and right traveling waves. The remarkable fact is that *every* solution to the wave equation can be so represented.

Theorem 14.9. *Every solution to the wave equation* (14.85) *can be written as a combination*

$$u(t,x) = p(\xi) + q(\eta) = p(x - c\,t) + q(x + c\,t) \tag{14.118}$$

of right and left traveling waves, each depending on its characteristic variable

$$\xi = x - c\,t, \qquad \eta = x + c\,t. \tag{14.119}$$

Proof: The key is to use a linear changes of variables to rewrite the wave equation entirely in terms of the characteristic variables. We set

$$u(t,x) = w(x - c\,t, x + c\,t) = w(\xi,\eta), \qquad \text{whereby} \qquad w(\xi,\eta) = u\left(\frac{\eta - \xi}{2\,c}\,, \frac{\xi + \eta}{2}\right).$$

Then, using the chain rule to compute the partial derivatives,

$$\frac{\partial u}{\partial t} = c\left(\frac{\partial w}{\partial \xi} - \frac{\partial w}{\partial \eta}\right), \qquad \frac{\partial u}{\partial x} = \frac{\partial w}{\partial \xi} + \frac{\partial w}{\partial \eta}\,.$$

and hence

$$\frac{\partial^2 u}{\partial t^2} = c^2\left(\frac{\partial^2 w}{\partial \xi^2} - 2\,\frac{\partial^2 w}{\partial \xi\,\partial \eta} + \frac{\partial^2 w}{\partial \eta^2}\right), \qquad \frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 w}{\partial \xi^2} + 2\,\frac{\partial^2 w}{\partial \xi\,\partial \eta} + \frac{\partial^2 w}{\partial \eta^2}\,.$$

Therefore

$$\Box\, u = \frac{\partial^2 u}{\partial t^2} - c^2\,\frac{\partial^2 u}{\partial x^2} = -\,4\,c^2\,\frac{\partial^2 w}{\partial \xi\,\partial \eta}\,.$$

We conclude that $u(t,x)$ solves the wave equation $\Box\, u = 0$ if and only if $w(\xi,\eta)$ solves the second order partial differential equation

$$\frac{\partial^2 w}{\partial \xi\,\partial \eta} = 0, \qquad \text{which we write in the form} \qquad \frac{\partial}{\partial \xi}\left(\frac{\partial w}{\partial \eta}\right) = 0.$$

This partial differential equation can be integrated once with respect to ξ, resulting in

$$\frac{\partial w}{\partial \eta} = r(\eta),$$

where r is an arbitrary function of the characteristic variable η. Integrating both sides of the latter partial differential equation with respect to η, we find

$$w(\xi,\eta) = p(\xi) + q(\eta), \qquad \text{where} \qquad q'(\eta) = r(\eta),$$

while $p(\xi)$ represents the integration "constant". Replacing the characteristic variables by their formulae in terms of t and x completes the proof. *Q.E.D.*

Remark: As noted above, we have been a little cavalier with our specification of the domain of definition of the functions and the differentiability assumptions required. Sorting out the precise technical details is left to the meticulous reader.

Remark: As we know, the general solution to a second order *ordinary* differential equation depends on two arbitrary constants. Here we observe that the general solution to a second order *partial* differential equation depends on two arbitrary functions — in this case $p(\xi)$ and $q(\eta)$. This serves as a useful rule of thumb, but should not be interpreted too literally.

Let us now see how this new form of solution to the wave equation can be used to effectively solve initial value problems. The simplest case is that of a bar or string of infinite length, in which case we have a pure initial value problem

$$\frac{\partial^2 u}{\partial t^2} = c^2 \, \frac{\partial^2 u}{\partial x^2}\,, \qquad u(0,x) = f(x), \qquad \frac{\partial u}{\partial t}\,(0,x) = g(x), \qquad \text{for} \quad -\infty < x < \infty.$$

Substituting the solution formula (14.118) into the initial conditions, we find

$$u(0,x) = p(x) + q(x) = f(x), \qquad \frac{\partial u}{\partial t}\,(0,x) = -\,c\,p'(x) + c\,q'(x) = g(x).$$

To solve this pair of linear equations for p and q, we differentiate the first equation:

$$p'(x) + q'(x) = f'(x).$$

Subtracting the second equation divided by c, we find

$$2\,p'(x) = f'(x) - \frac{1}{c}\,g(x).$$

Therefore,

$$p(x) = \frac{1}{2}\,f(x) - \frac{1}{2\,c}\int_0^x g(z)\,dz + a,$$

where a is an integration constant. The first equation then yields

$$q(x) = f(x) - p(x) = \frac{1}{2}\,f(x) + \frac{1}{2\,c}\int_0^x g(z)\,dz - a.$$

Substituting these two expressions back into (14.118), we find

$$\begin{aligned} u(t,x) = p(\xi) + q(\eta) &= \frac{f(\xi) + f(\eta)}{2} + \frac{1}{2\,c}\left[-\int_0^\xi + \int_0^\eta \right] g(z)\,dz \\ &= \frac{f(\xi) + f(\eta)}{2} + \frac{1}{2\,c}\int_\xi^\eta g(z)\,dz, \end{aligned}$$

where ξ, η are the characteristic variables (14.119). In this fashion, we have derived *d'Alembert's solution* to the wave equation on the entire line $-\infty < x < \infty$.

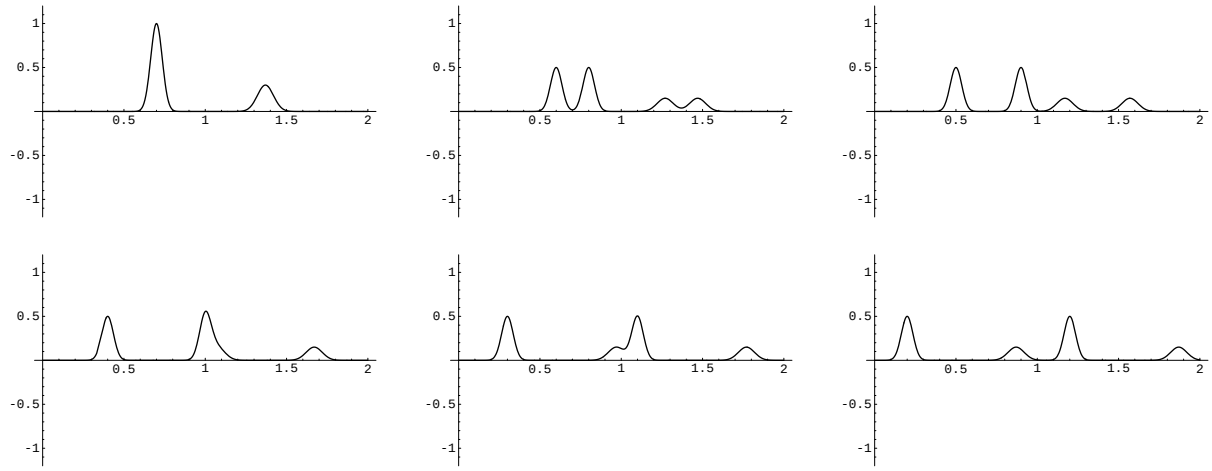


Figure 14.8. Interaction of Waves.

Theorem 14.10. *The solution to the initial value problem*

$$\frac{\partial^2 u}{\partial t^2} = c^2 \, \frac{\partial^2 u}{\partial x^2}\,, \qquad u(0,x) = f(x), \qquad \frac{\partial u}{\partial t}\,(0,x) = g(x), \qquad -\infty < x < \infty. \tag{14.120}$$

is given by

$$u(t,x) = \frac{f(x - c\,t) + f(x + c\,t)}{2} + \frac{1}{2\,c} \int_{x - c\,t}^{x + c\,t} g(z)\, dz. \tag{14.121}$$

.

Let us investigate the implications of d'Alembert's solution formula (14.121). First, suppose there is no initial velocity, so $g(x) \equiv 0$, and the motion is purely the result of the initial displacement $u(0,x) = f(x)$. In this case, the solution (14.121) reduces to

$$u(t,x) = \tfrac{1}{2}\, f(x - c\,t) + \tfrac{1}{2}\, f(x + c\,t).$$

The effect is that the initial displacement $f(x)$ splits into two waves, one traveling to the right and one traveling to the left, each of constant speed c, and each of exactly the same shape as the initial displacement $f(x)$ but only half as tall. For example, if the initial displacement is a localized pulse centered at the origin, say

$$u(0,x) = e^{-\,x^2}, \qquad\qquad \frac{\partial u}{\partial t}\,(0,x) = 0,$$

then the solution

$$u(t,x) = \tfrac{1}{2}\, e^{-\,(x - c\,t)^2} + \tfrac{1}{2}\, e^{-\,(x + c\,t)^2}$$

consists of two half size pulses running away from the origin in opposite directions with equal speed c. If we take two separated pulses, say

$$u(0,x) = e^{-\,x^2} + 2\, e^{-\,(x-1)^2}, \qquad\qquad \frac{\partial u}{\partial x}\,(0,x) = 0,$$

centered at $x = 0$ and $x = 1$, then the solution

$$u(t,x) = \tfrac{1}{2}\, e^{-\,(x - c\,t)^2} + e^{-\,(x - 1 - c\,t)^2} + \tfrac{1}{2}\, e^{-\,(x + c\,t)^2} + e^{-\,(x - 1 + c\,t)^2}$$

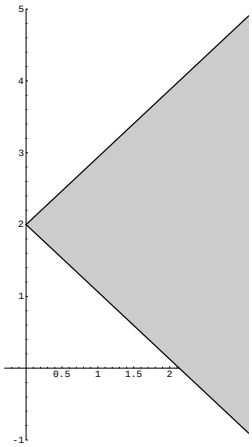


Figure 14.9. Characteristic Lines for the Wave Equation.

will consist of four pulses, two moving to the right and two to the left, all with the same speed, as pictured in Figure 14.8.

An important observation is that when a right-moving pulse collides with a left-moving pulse, they emerge from the collision unchanged — a consequence of the inherent linearity of the wave equation. The first picture shows the initial displacement. In the second and third pictures, the two localized bumps have each split into two copies moving in opposite directions. In the fourth and fifth, the larger right moving bump is in the process of interacting with the smaller left moving bump. Finally, in the last picture the interaction is complete, and the two left moving bumps and two right moving bumps travel in tandem with no further collisions.

Remark: If the initial displacement has bounded support, and so $f(x) = 0$ if $x < a$ or $x > b$ for some $a < b$, then after a finite time the right and left-moving waves will be completely separated, and the observer will see two exact half size replicas running away, with speed c, in opposite directions. If the displacement is not localized, then the left and right traveling waves will never fully disengage, and one might be hard pressed (just as in our earlier discussion of quasi-periodic phenomena) in recognizing that a complicated solution pattern is, in reality, just the superposition of two very simple traveling waves. For example, using a trigonometric identity, any separable solution, e.g.,

$$\sin c\,t \sin x = \tfrac{1}{2}\cos(x - c\,t) - \tfrac{1}{2}\cos(x + c\,t),$$

can be rewritten in d'Alembert form. The interpretation of the two solution formulas is quite different. Most observers will see a standing sinusoidal wave, vibrating with frequency c, as represented on the left side of the equation. However, the right hand side says that this is the same as the difference of a right and left traveling cosine wave. The interactions of their peaks and troughs reproduces the standing wave. Thus, the same solution can be interpreted in seemingly incompatible ways!

In general, the lines of slope $\pm c$ in the (t, x)–plane, where the characteristic variables are constant,

$$\xi = x - c\,t = a, \qquad \eta = x + c\,t = b, \tag{14.122}$$

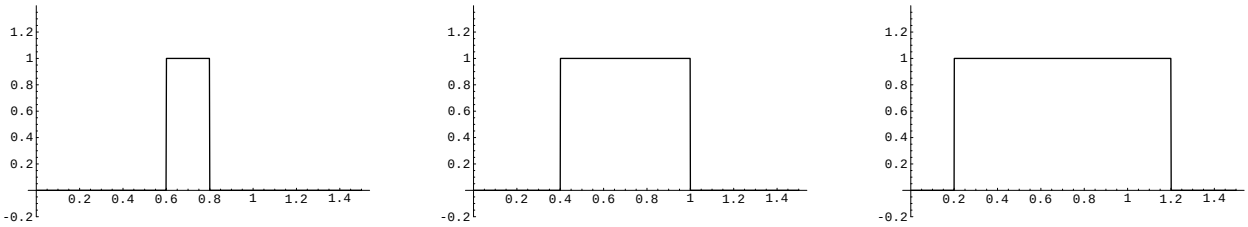


Figure 14.10. Concentrated Initial Velocity for Wave Equation.

are known as the *characteristics* of the wave equation. The two characteristics going through a point on the x axis, where the initial data is prescribed, are illustrated in Figure 14.9. The reader should note that, in this figure, the t axis is horizontal, while the x axis is vertical.

In general, signals propagate along characteristics. More specifically, if we start out with an initial displacement concentrated very close to a point $x = a$, $t = 0$, then the solution will be concentrated along the two characteristic lines emanating from the point, namely $x - c\,t = a$ and $x + c\,t = a$. In the limit, a unit impulse or delta function displacement at $x = a$, corresponding to the initial condition

$$u(0, x) = \delta(x - a), \qquad \frac{\partial u}{\partial t}\,(0, x) = 0, \tag{14.123}$$

will result in a solution

$$u(t, x) = \tfrac{1}{2}\,\delta(x - c\,t - a) + \tfrac{1}{2}\,\delta(x + c\,t - a) \tag{14.124}$$

consisting of two half-strength delta spikes traveling away from the starting position concentrated on the two characteristic lines.

Let us return to the general initial value problem (14.120). Suppose now that there is no initial displacement, $u(0, x) = f(x) \equiv 0$, but rather a concentrated initial velocity, say a delta function

$$\frac{\partial u}{\partial t}\,(0, x) = \delta_a(x) = \delta(x - a).$$

Physically, this would correspond to striking the string with a highly concentrated blow at the point $x = a$. The d'Alembert solution (14.121) to this "hammer blow" problem is

$$u(t, x) = \frac{1}{2\,c} \int_{x - c\,t}^{x + c\,t} \delta_a(z)\, dz = \begin{cases} \dfrac{1}{2\,c}\,, & x - c\,t < a < x + c\,t, \\ 0, & \text{otherwise}, \end{cases} \tag{14.125}$$

and consists of a constant displacement, of magnitude $1/(2\,c)$, between the two characteristic lines $x - c\,t = a = x + c\,t$ based at $x = a$, $t = 0$ — the shaded region of Figure 14.9. This region is known as the *domain of influence* of the point $(a, 0)$ since, in general, the value of the initial data at that point will only affect the solution values in the region. The particular solution, which is plotted in Figure 14.10, has two jump discontinuities between the undisturbed state and the displaced state, each propagating along its characteristic line with speed c, but in opposite directions. Note that, unlike a concentrated initial displacement, where the signal remains concentrated and each point along the bar is temporarily

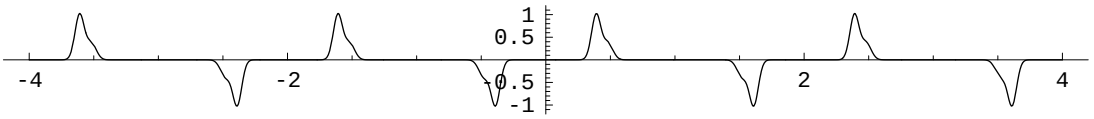


Figure 14.11. Odd Periodic Extension of a Concentrated Pulse.

displaced, eventually returning to its undisturbed state, a concentrated initial velocity has a lasting effect, and the bar remains permanently deformed by an amount $1/(2\,c)$.

Solutions on Bounded Intervals

So far, we have been using d'Alembert formula to solve the wave equation on an infinite interval. The formula can still be used on bounded intervals, but in a suitably modified format so as to respect the boundary conditions. The easiest to deal with is the periodic problem on $0 \le x \le \ell$, with boundary conditions

$$u(t,0) = u(t,\ell), \qquad u_x(t,0) = u_x(t,\ell). \tag{14.126}$$

If we extend the initial displacement $f(x)$ and velocity $g(x)$ to be periodic functions of period ℓ, so $f(x+\ell) = f(x)$ and $g(x+\ell) = g(x)$ for all $x \in \mathbb{R}$, then the resulting d'Alembert solution (14.121) will also be periodic in x, so $u(t, x+\ell) = u(t,x)$. In particular, it satisfies the boundary conditions (14.126) and so coincides with the desired solution. Details can be found in Exercises ■–■.

Next, suppose we have fixed (Dirichlet) boundary conditions

$$u(t,0) = 0, \qquad u(t,\ell) = 0. \tag{14.127}$$

The resulting solution can be written as a Fourier sine series (14.91), and hence is both odd and $2\,\ell$ periodic in x. Therefore, to write the solution in d'Alembert form, we extend the initial displacement $f(x)$ and velocity $g(x)$ to be odd, periodic functions of period $2\,\ell$:

$$f(-x) = -f(x), \qquad f(x+2\,\ell) = f(x), \qquad g(-x) = -g(x), \qquad g(x+2\,\ell) = g(x).$$

This will ensure that the d'Alembert solution also remains odd and periodic. As a result, it satisfies the boundary conditions (14.127) for all t. Keep in mind that, while the solution $u(t,x)$ is defined for all x, the only physically relevant values occur on the interval $0 \le x \le \ell$. Nevertheless, the effects of displacements in the unphysical regime will eventually be "felt" as the propagating waves pass through the physical interval.

For example, consider an initial displacement which is concentrated near $x = a$ for some $0 < a < \ell$. Its odd, periodic extension consists of two sets of replicas: those of the same form occurring at positions $a \pm 2\,\ell, a \pm 4\,\ell, \ldots$, and their mirror images at the intermediate positions $-a, -a \pm 2\,\ell, -a \pm 4\,\ell, \ldots$; Figure 14.11 shows a representative example. The resulting solution begins with each of the pulses, both positive and negative, splitting into two half-size replicas that propagate with speed c in opposite directions. As the individual pulses meet, they interact as they pass unaltered through each other. The process repeats periodically, with an infinite row of half-size pulses moving to the right kaleidoscopically interacting with an infinite row moving to the left.

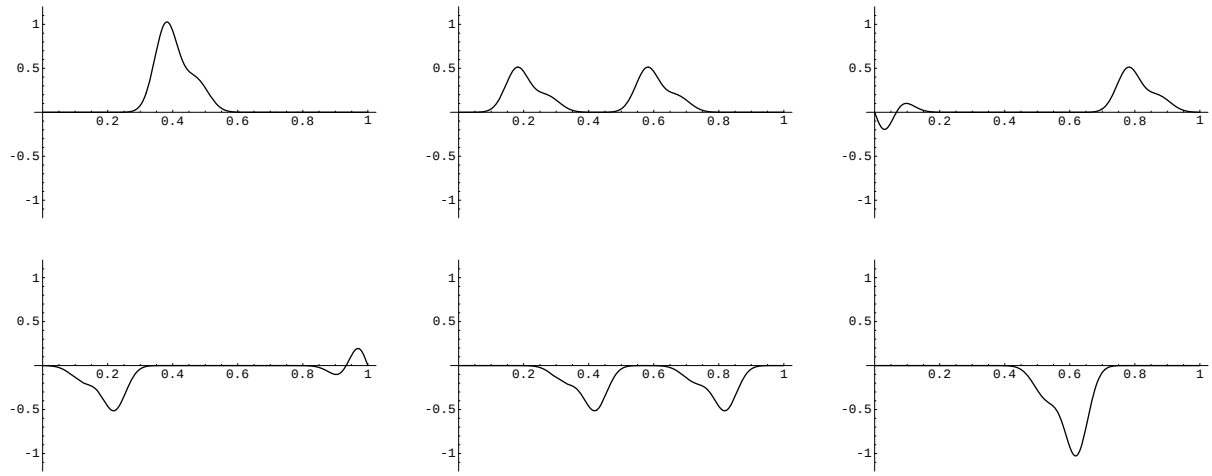


Figure 14.12. Solution to Wave Equation with Fixed Ends.

However, only the part of this solution that lies on $0 \le x \le \ell$ is actually realized on the physical bar. The net effect is as if we were forced to view the complete solution as it passes by a window of length ℓ that blocks out all other regions of the real axis. What the viewer effectively sees assumes a somewhat different interpretation. To wit, the original pulse at position $0 < a < \ell$ splits up into two half size replicas that move off in opposite directions. As each half-size pulse reaches an end of the bar, it meets a mirror image pulse that has been propagating in the opposite direction from the non-physical regime. The pulse appears to be reflected at the end of the interval, and changes into an upside down mirror image moving in the opposite direction. The original positive pulse has moved off the end of the bar just as its mirror image has moved into the physical regime. (A common physical illustration is a pulse propagating down a jump rope that is held fixed at its end; the reflected pulse returns upside down.) A similar reflection occurs as the other half-size pulse hits the other end of the physical interval, after which the solution consists of two upside down half-size pulses moving back towards each other. At time $t = \ell/c$ they recombine at the point $\ell - a$ to instantaneously form a full-sized, but upside-down mirror image of the original disturbance — in accordance with (14.95). The recombined pulse in turn splits apart into two upside down half-size pulses that, when each collides with the end, reflects and returns to its original upright form. At time $t = 2\,\ell/c$, the pulses recombine to exactly reproduce the original displacement. The process then repeats, and the solution is periodic in time with period $2\,\ell/c$.

In Figure 14.12, the first picture displays the initial displacement. In the second, it has splits into left and right moving, half-size clones. In the third picture, the left moving bump is in the process of colliding with the left end of the bar. In the fourth picture, it has emerged from the collision, and is now upside down, reflected, and moving to the right. Meanwhile, the right moving pulse is starting to collide with the right end. In the fifth picture, both pulses have completed their collisions and are now moving back towards each other, where, in the last picture, they recombine into an upside-down mirror image of the original pulse. The process then repeats itself, in mirror image, finally recombining to the original pulse, at which point the entire process starts over.

The Neumann (free) boundary value problem

$$\frac{\partial u}{\partial x}(t,0) = 0, \qquad \frac{\partial u}{\partial x}(t,\ell) = 0, \tag{14.128}$$

is handled similarly. Since the solution has the form of a Fourier cosine series in x, we extend the initial conditions to be *even*, $2\,\ell$ periodic functions

$$f(-x) = f(x), \qquad f(x + 2\,\ell) = f(x), \qquad g(-x) = g(x), \qquad g(x + 2\,\ell) = g(x).$$

The resulting d'Alembert solution is also even and $2\,\ell$ periodic in x, and hence satisfies the boundary conditions. In this case, when a pulse hits one of the ends, its reflection remains upright, but becomes a mirror image of the original; a familiar physical illustration is a water wave that reflects off a solid wall. Further details are left to the reader in Exercise ■

In summary, we have now learned two different ways to solve the one-dimensional wave equation. The first, based on Fourier analysis, emphasizes the vibrational or wave character of the solutions, while the second, based on the d'Alembert formula, emphasizes their particle aspects, where individual wave packets collide with each other, or reflect at the boundary, all the while maintaining their overall form. Some solutions look like vibrating waves, while others seem much more like interacting particles. But, like the blind men describing the elephant, these are merely two facets of the *same* solution. The Fourier series formula shows how every particle-like solution can be decomposed into its constituent vibrational modes, while the d'Alembert formula demonstrates how vibrating waves combine as moving particles.

The coexistence of particle and wave features is reminiscent of the long running historical debate over the nature of light. In the beginning, Newton and his disciples advocated a particle basis, in the form of photons. However, until the beginning of the twentieth century, most physicists advocated a wave or vibrational viewpoint. Einstein's explanation of the photoelectric effect in 1905 served to resurrect the particle interpretation. Only with the establishment of quantum mechanics was the debate resolved — light, and, indeed, all subatomic particles are *both*, manifesting particle and wave features, depending upon the experiment and the physical situation. But the theoretical evidence for wave-particle duality already existed in the competing solution formulae of the classical wave equation!

14.6. Numerical Methods.

As you know, most differential equations are much too complicated to be solved analytically. Thus, to obtain quantitative results, one is forced to construct a sufficiently accurate numerical approximation to the solution. Even in cases, such as the heat and wave equations, where explicit solution formulas (either closed form or infinite series) exist, numerical methods still can be profitably employed. Moreover, justification of the numerical algorithm is facilitated by the ability compare it with an exact solution. Moreover, the lessons learned in the design of numerical algorithms for "solved" problems prove to be of immense value when one is confronted with more complicated problems for which solution formulas no longer exist.

In this final section, we present some of the most basic numerical solution techniques for the heat and wave equations. We will only introduce the most basic algorithms, leaving

One-Sided Difference Central Difference

Figure 14.13. Finite Difference Approximations.

more sophisticated variations and extensions to a more thorough treatment, which can be found in numerical analysis texts, e.g., [**27**, **34**, **107**].

Finite Differences

Numerical solution methods for differential equations can be partitioned into two principal classes. (In this oversimplified presentation, we are ignoring more specialized methods of less general applicability.) The first category, already introduced in Section 11.6, are the *finite element methods*. Finite elements are designed for the differential equations describing equilibrium configurations, since they rely on minimizing a functional. A competitive alternative is to directly approximate the derivatives appearing in the differential equation, which requires knowing appropriate numerical differentiation formulae.

In general, to approximate the derivative of a function at a point, say $f'(x)$ or $f''(x)$, one constructs a suitable combination of sampled function values at nearby points. The underlying formalism used to construct these approximation formulae is known as the *calculus of finite differences*. Its development has a long and influential history, dating back to Newton. The resulting *finite difference numerical methods* for solving differential equations have extremely broad applicability, and can, with proper care, be adapted to most problems that arise in mathematics and its many applications.

The simplest finite difference approximation is the ordinary *difference quotient*

$$\frac{u(x+h) - u(x)}{h} \approx u'(x), \tag{14.129}$$

used to approximate the first derivative of the function $u(x)$. Indeed, if u is differentiable at x, then $u'(x)$ is, by definition, the limit, as $h \to 0$ of the finite difference quotients. Geometrically, the difference quotient equals the slope of the secant line through the two points $\big(x, u(x)\big)$ and $\big(x+h, u(x+h)\big)$ on the graph of the function. For small h, this should be a reasonably good approximation to the slope of the tangent line, $u'(x)$, as illustrated in the first picture in Figure 14.13.

How close an approximation is the difference quotient? To answer this question, we assume that $u(x)$ is at least twice continuously differentiable, and examine the first order Taylor expansion

$$u(x+h) = u(x) + u'(x)\, h + \tfrac{1}{2}\, u''(\xi)\, h^2. \tag{14.130}$$

We have used the Cauchy form (C.8) for the remainder term, in which ξ represents some point lying between x and $x + h$. The *error* or difference between the finite difference formula and the derivative being approximated is given by

$$\frac{u(x+h) - u(x)}{h} \;-\; u'(x) = \tfrac{1}{2}\, u''(\xi)\, h. \tag{14.131}$$

Since the error is proportional to h, we say that the finite difference quotient (14.131) is a *first order* approximation. When the precise formula for the error is not so important, we will write

$$u'(x) = \frac{u(x+h) - u(x)}{h} + \mathrm{O}(h). \tag{14.132}$$

The "big Oh" notation $\mathrm{O}(h)$ refers to a term that is proportional to h, or, more rigorously, bounded by a constant multiple of h as $h \to 0$.

Example 14.11. Let $u(x) = \sin x$. Let us try to approximate $u'(1) = \cos 1 = 0.5403023\ldots$ by computing finite difference quotients

$$\cos 1 \;\approx\; \frac{\sin(1+h) - \sin 1}{h}\,.$$

The result for different values of h is listed in the following table.

h	1	.1	.01	.001	.0001
approximation	0.067826	0.497364	0.536086	0.539881	0.540260
error	−0.472476	−0.042939	−0.004216	−0.000421	−0.000042

We observe that reducing the step size by a factor of $\frac{1}{10}$ reduces the size of the error by approximately the same factor. Thus, to obtain 10 decimal digits of accuracy, we anticipate needing a step size of about $h = 10^{-11}$. The fact that the error is more of less proportional to the step size confirms that we are dealing with a first order numerical approximation.

To approximate higher order derivatives, we need to evaluate the function at more than two points. In general, an approximation to the n^{th} order derivative $u^{(n)}(x)$ requires at least $n+1$ distinct sample points. For simplicity, we shall only use equally spaced points, leaving the general case to the exercises.

For example, let us try to approximate $u''(x)$ by sampling u at the particular points x, $x + h$ and $x - h$. Which combination of the function values $u(x-h), u(x), u(x+h)$ should be used? The answer to such a question can be found by consideration of the relevant Taylor expansions

$$\begin{aligned} u(x+h) &= u(x) + u'(x)\, h + u''(x)\, \frac{h^2}{2} + u'''(x)\, \frac{h^3}{6} + \mathrm{O}(h^4), \\ u(x-h) &= u(x) - u'(x)\, h + u''(x)\, \frac{h^2}{2} - u'''(x)\, \frac{h^3}{6} + \mathrm{O}(h^4), \end{aligned} \tag{14.133}$$

where the error terms are proportional to h^4. Adding the two formulae together gives

$$u(x+h) + u(x-h) = 2\, u(x) + u''(x)\, h^2 + \mathrm{O}(h^4).$$

Rearranging terms, we conclude that

$$u''(x) = \frac{u(x+h) - 2\,u(x) + u(x-h)}{h^2} + \mathrm{O}(h^2), \tag{14.134}$$

The result is known as the *centered finite difference approximation* to the second derivative of a function. Since the error is proportional to h^2, this is a second order approximation.

Example 14.12. Let $u(x) = e^{x^2}$, with $u''(x) = (4\,x^2 + 2)\,e^{x^2}$. Let us approximate $u''(1) = 6\,e = 16.30969097\ \ldots$ by using the finite difference quotient (14.134):

$$6\,e \;\approx\; \frac{e^{(1+h)^2} - 2\,e + e^{(1-h)^2}}{h^2}\,.$$

The results are listed in the following table.

h	1	.1	.01	.001	.0001
approximation	50.16158638	16.48289823	16.31141265	16.30970819	16.30969115
error	33.85189541	0.17320726	0.00172168	0.00001722	0.00000018

Each reduction in step size by a factor of $\frac{1}{10}$ reduces the size of the error by a factor of $\frac{1}{100}$ and results in a gain of two new decimal digits of accuracy, confirming that the finite difference approximation is of second order.

However, this prediction is not completely borne out in practice. If we take[†] $h = .00001$ then the formula produces the approximation 16.3097002570, with an error of 0.0000092863 — which is *less* accurate that the approximation with $h = .0001$. The problem is that round-off errors have now begun to affect the computation, and underscores the difficulty with numerical differentiation. Finite difference formulae involve dividing very small quantities, which can induce large numerical errors due to round-off. As a result, while they typically produce reasonably good approximations to the derivatives for moderately small step sizes, to achieve high accuracy, one must switch to a higher precision. In fact, a similar comment applied to the previous Example 14.11, and our expectations about the error were not, in fact, fully justified as you may have discovered if you tried an extremely small step size.

Another way to improve the order of accuracy of finite difference approximations is to employ more sample points. For instance, if the first order approximation (14.132) to the first derivative based on the two points x and $x+h$ is not sufficiently accurate, one can try combining the function values at three points x, $x+h$ and $x-h$. To find the appropriate combination of $u(x-h), u(x), u(x+h)$, we return to the Taylor expansions (14.133). To solve for $u'(x)$, we subtract[‡] the two formulae, and so

$$u(x+h) - u(x-h) = 2\,u'(x)\,h + u'''(x)\,\frac{h^3}{3} + \mathrm{O}(h^4).$$

[†] This next computation depends upon the computer's precision; here we used single precision in MATLAB.

[‡] The terms $\mathrm{O}(h^4)$ do *not* cancel, since they represent potentially different multiples of h^4.

Rearranging the terms, we are led to the well-known *centered difference formula*

$$u'(x) = \frac{u(x+h) - u(x-h)}{2\,h} + \mathrm{O}(h^2), \tag{14.135}$$

which is a second order approximation to the first derivative. Geometrically, the centered difference quotient represents the slope of the secant line through the two points $\big(x-h, u(x-h)\big)$ and $\big(x+h, u(x+h)\big)$ on the graph of u centered symmetrically about the point x. Figure 14.13 illustrates the two approximations; the advantages in accuracy in the centered difference version are graphically evident. Higher order approximations can be found by evaluating the function at yet more sample points, including, say, $x+2\,h, x-2\,h$, etc.

Example 14.13. Return to the function $u(x) = \sin x$ considered in Example 14.11. The centered difference approximation to its derivative $u'(1) = \cos 1 = 0.5403023\ \ldots$ is

$$\cos 1 \approx \frac{\sin(1+h) - \sin(1-h)}{2\,h}\,.$$

The results are tabulated as follows:

h	.1	.01	.001	.0001
approximation	0.53940225217	0.54029330087	0.54030221582	0.54030230497
error	−0.00090005370	−0.00000900499	−0.00000009005	−0.00000000090

As advertised, the results are much more accurate than the one-sided finite difference approximation used in Example 14.11 at the same step size. Since it is a second order approximation, each reduction in the step size by a factor of $\frac{1}{10}$ results in two more decimal places of accuracy.

Many additional finite difference approximations can be constructed by similar manipulations of Taylor expansions, but these few very basic ones will suffice for our subsequent purposes. In the following subsection, we apply the finite difference formulae to develop numerical solution schemes for the heat and wave equations.

Numerical Algorithms for the Heat Equation

Consider the heat equation

$$\frac{\partial u}{\partial t} = \gamma\,\frac{\partial^2 u}{\partial x^2}\,, \qquad 0 < x < \ell, \qquad t \geq 0, \tag{14.136}$$

representing a bar of length ℓ and thermal diffusivity $\gamma > 0$, which is assumed to be constant. To be concrete, we impose time-dependent Dirichlet boundary conditions

$$u(t,0) = \alpha(t), \qquad u(t,\ell) = \beta(t), \qquad t \geq 0, \tag{14.137}$$

specifying the temperature at the ends of the bar, along with the initial conditions

$$u(0,x) = f(x), \qquad 0 \leq x \leq \ell, \tag{14.138}$$

specifying the bar's initial temperature distribution. In order to effect a numerical approximation to the solution to this initial-boundary value problem, we begin by introducing a *rectangular mesh* consisting of points (t_i, x_j) with

$$0 = x_0 < x_1 < \cdots < x_n = \ell \qquad \text{and} \qquad 0 = t_0 < t_1 < t_2 < \cdots .$$

For simplicity, we maintain a uniform mesh spacing in both directions, with

$$h = x_{j+1} - x_j = \frac{\ell}{n}, \qquad\qquad k = t_{i+1} - t_i,$$

representing, respectively, the spatial mesh size and the time step size. It will be essential that we do *not* a priori require the two to be the same. We shall use the notation

$$u_{i,j} \approx u(t_i, x_j) \qquad \text{where} \qquad t_i = i\,k, \qquad x_j = j\,h, \tag{14.139}$$

to denote the numerical approximation to the solution value at the indicated mesh point.

As a first attempt at designing a numerical method, we shall use the simplest finite difference approximations to the derivatives. The second order space derivative is approximated by (14.134), and hence

$$\begin{aligned} \frac{\partial^2 u}{\partial x^2}\,(t_i, x_j) &\approx \frac{u(t_i, x_{j+1}) - 2\,u(t_i, x_j) + u(t_i, x_{j-1})}{h^2} + \mathrm{O}(h^2) \\ &\approx \frac{u_{i,j+1} - 2\,u_{i,j} + u_{i,j-1}}{h^2} + \mathrm{O}(h^2), \end{aligned} \tag{14.140}$$

where the error in the approximation is proportional to h^2. Similarly, the one-sided finite difference approximation (14.132) is used for the time derivative, and so

$$\frac{\partial u}{\partial t}\,(t_i, x_j) \approx \frac{u(t_{i+1}, x_j) - u(t_i, x_j)}{k} + \mathrm{O}(k) \approx \frac{u_{i+1,j} - u_{i,j}}{k} + \mathrm{O}(k), \tag{14.141}$$

where the error is proportion to k. In practice, one should try to ensure that the approximations have similar orders of accuracy, which leads us to choose

$$k \approx h^2.$$

Assuming $h < 1$, this requirement has the important consequence that the time steps must be *much* smaller than the space mesh size.

Remark: At this stage, the reader might be tempted to replace (14.141) by the second order central difference approximation (14.135). However, this produces significant complications, and the resulting numerical scheme is not practical.

Replacing the derivatives in the heat equation (14.142) by their finite difference approximations (14.140), (14.141), and rearranging terms, we end up with the linear system

$$u_{i+1,j} = \mu\,u_{i,j+1} + (1 - 2\,\mu)u_{i,j} + \mu\,u_{i,j-1}, \qquad \begin{array}{l} i = 0, 1, 2, \dots , \\ j = 1, \dots, n-1, \end{array} \tag{14.142}$$

in which

$$\mu = \frac{\gamma\, k}{h^2}\,. \tag{14.143}$$

The resulting numerical scheme takes the form of an iterative linear system for the solution values $u_{i,j} \approx u(t_i, x_j)$, $j = 1, \dots, n-1$, at each time step t_i.

The initial condition (14.138) means that we should initialize our numerical data by sampling the initial temperature at the mesh points:

$$u_{0,j} = f_j = f(x_j), \qquad\qquad j = 1, \dots, n-1. \tag{14.144}$$

Similarly, the boundary conditions (14.137) require that

$$u_{i,0} = \alpha_i = \alpha(t_i), \qquad\qquad u_{i,n} = \beta_i = \beta(t_i), \qquad\qquad i = 0, 1, 2, \dots\,. \tag{14.145}$$

For consistency, we should assume that the initial and boundary conditions agree at the corners of the domain:

$$f_0 = f(0) = u(0,0) = \alpha(0) = \alpha_0, \qquad\qquad f_n = f(\ell) = u(0,\ell) = \beta(0) = \beta_0.$$

The three equations (14.142–145) completely prescribe the numerical approximation algorithm for solving the initial-boundary value problem (14.136–138).

Let us rewrite the scheme in a more transparent matrix form. First, let

$$\mathbf{u}^{(i)} = \big(\, u_{i,1}, u_{i,2}, \dots, u_{i,n-1} \,\big)^T \approx \big(\, u(t_i, x_1), u(t_i, x_2), \dots, u(t_i, x_{n-1}) \,\big)^T \tag{14.146}$$

be the vector whose entries are the numerical approximations to the solution values at time t_i at the *interior* nodes. We omit the boundary nodes $x_0 = 0$, $x_n = \ell$, since those values are fixed by the boundary conditions (14.137). Then (14.142) assumes the compact vectorial form

$$\mathbf{u}^{(i+1)} = A\,\mathbf{u}^{(i)} + \mathbf{b}^{(i)}, \tag{14.147}$$

where

$$A = \begin{pmatrix} 1-2\,\mu & \mu & & & & \\ \mu & 1-2\,\mu & \mu & & & \\ & \mu & 1-2\,\mu & \mu & & \\ & & \mu & \ddots & \ddots & \\ & & & \ddots & \ddots & \mu \\ & & & & \mu & 1-2\,\mu \end{pmatrix}, \qquad \mathbf{b}^{(i)} = \begin{pmatrix} \mu\,\alpha_i \\ 0 \\ 0 \\ \vdots \\ 0 \\ \mu\,\beta_i \end{pmatrix}. \tag{14.148}$$

The coefficient matrix A is symmetric and tridiagonal. The contributions (14.145) of the boundary nodes appear in the vector $\mathbf{b}^{(i)}$. This numerical method is known as an *explicit scheme* since each iterate is computed directly without relying on solving an auxiliary equation — unlike the implicit schemes to be discussed below.

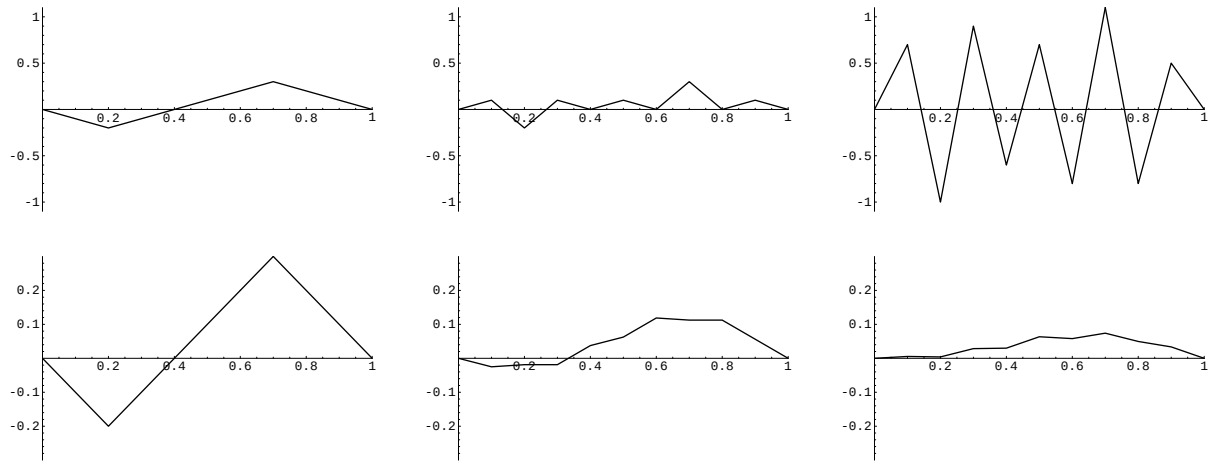


Figure 14.14. Numerical Solutions for the Heat Equation Based on the Explicit Scheme.

Example 14.14. Let us fix the diffusivity $\gamma = 1$ and the bar length $\ell = 1$. For illustrative purposes, we take a spatial step size of $h = .1$. In Figure 14.14 we compare two (slightly) different time step sizes on the same initial data as used in (14.22). The first sequence uses the time step $k = h^2 = .01$ and plots the solution at times $t = 0., .02, .04$. The solution is already starting to show signs of instability, and indeed soon thereafter becomes completely wild. The second sequence takes $k = .005$ and plots the solution at times $t = 0., .025, .05$. (Note that the two sequences of plots have different vertical scales.) Even though we are employing a rather coarse mesh, the numerical solution is not too far away from the true solution to the initial value problem, which can be found in Figure 14.1.

In light of this calculation, we need to understand why our scheme sometimes gives reasonable answers but at other times utterly fails. To this end, let us specialize to homogeneous boundary conditions

$$u(t,0) = 0 = u(t,\ell), \qquad \text{whereby} \qquad \alpha_i = \beta_i = 0 \qquad \text{for all} \qquad i = 0, 1, 2, 3, \dots , \tag{14.149}$$

and so (14.147) reduces to a homogeneous, linear iterative system

$$\mathbf{u}^{(i+1)} = A\,\mathbf{u}^{(i)}. \tag{14.150}$$

According to Proposition 10.11, all solutions will converge to zero, $\mathbf{u}^{(i)} \to \mathbf{0}$ — as they are supposed to (why?) — if and only if A is a convergent matrix. But convergence depends upon the step sizes. Example 14.14 is indicating that for mesh size $h = .1$, the time step $k = .01$ yields a non-convergent matrix, while $k = .005$ leads to a convergent matrix and a valid numerical scheme.

As we learned in Theorem 10.14, the convergence property of a matrix is fixed by its spectral radius, i.e., its largest eigenvalue in magnitude. There is, in fact, an explicit formula for the eigenvalues of the particular tridiagonal matrix in our numerical scheme, which follows from the following general result, which solves Exercise 8.2.48. It is a direct consequence of Exercise 8.2.47, which contains the explicit formulae for the eigenvectors.

Lemma 14.15. *The eigenvalues of an $(n-1) \times (n-1)$ tridiagonal matrix all of whose diagonal entries are equal to a and all of whose sub- and super-diagonal entries are equal to b are*

$$\lambda_k = a + 2\,b \cos \frac{\pi\, k}{n}\,, \qquad\qquad k = 1, \dots, n-1. \tag{14.151}$$

In our particular case, $a = 1 - 2\,\mu$ and $b = \mu$, and hence the eigenvalues of the matrix A given by (14.148) are

$$\lambda_k = 1 - 2\,\mu + 2\,\mu \cos \frac{\pi\, k}{n}\,, \qquad\qquad k = 1, \dots, n-1.$$

Since the cosine term ranges between -1 and $+1$, the eigenvalues satisfy

$$1 - 4\,\mu < \lambda_k < 1.$$

Thus, assuming that $0 < \mu \leq \frac{1}{2}$ guarantees that all $|\,\lambda_k\,| < 1$, and hence A is a convergent matrix. In this way, we have deduced the basic stability criterion

$$\mu = \frac{\gamma\, k}{h^2} \leq \frac{1}{2}\,, \qquad \text{or} \qquad k \leq \frac{h^2}{2\,\gamma}\,. \tag{14.152}$$

With some additional analytical work, [**107**], it can be shown that this is sufficient to conclude that the numerical scheme (14.142–145) converges to the true solution to the initial-boundary value problem for the heat equation.

Since not all choices of space and time steps lead to a convergent scheme, the numerical method is called *conditionally stable*. The convergence criterion (14.152) places a severe restriction on the time step size. For instance, if we have $h = .01$, and $\gamma = 1$, then we can only use a time step size $k \leq .00005$, which is minuscule. It would take an inordinately large number of time steps to compute the value of the solution at even a moderate times, e.g., $t = 1$. Moreover, owing to the limited accuracy of computers, the propagation of round-off errors might then cause a significant reduction in the overall accuracy of the final solution values.

An unconditionally stable method — one that does not restrict the time step — can be constructed by using the backwards difference formula

$$\frac{\partial u}{\partial t}\,(t_i, x_j) \approx \frac{u(t_i, x_j) - u(t_{i-1}, x_j)}{k} + \mathrm{O}(h^k) \tag{14.153}$$

to approximate the temporal derivative. Substituting (14.153) and the same approximation (14.140) for u_{xx} into the heat equation, and then replacing i by $i+1$, leads to the iterative system

$$u_{i+1,j} - \mu\,\big(\,u_{i+1,j+1} - 2\,u_{i+1,j} + u_{i+1,j-1}\,\big) = u_{i,j}, \qquad \begin{array}{l} i = 0, 1, 2, \dots\,, \\ j = 1, \dots, n-1, \end{array} \tag{14.154}$$

where the parameter $\mu = \gamma\, k/h^2$ is as above. The initial and boundary conditions also have the same form (14.144, 145). The system can be written in the matrix form

$$\widehat{A}\,\mathbf{u}^{(i+1)} = \mathbf{u}^{(i)} + \mathbf{b}^{(i+1)}, \tag{14.155}$$

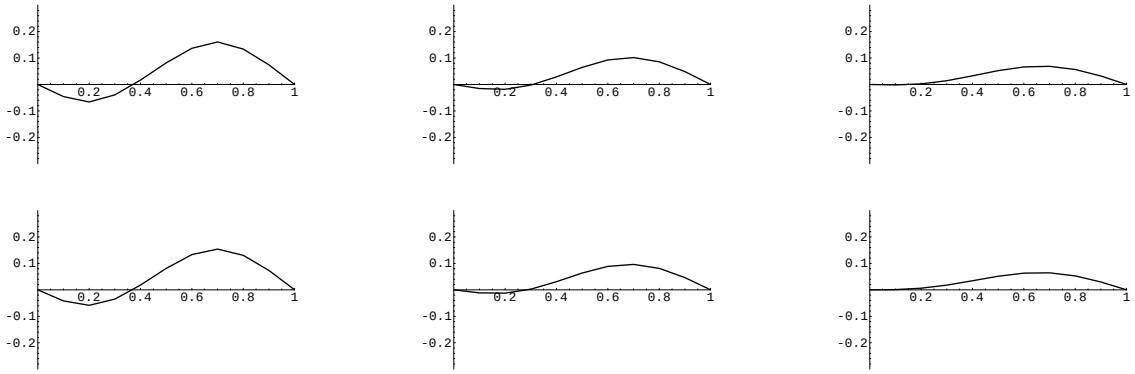


Figure 14.15. Numerical Solutions for the Heat Equation Based on the Implicit Scheme.

where $\widehat{A}$ is obtained from the matrix A in (14.148) by replacing μ by $-\mu$. This defines an *implicit method* since we have to solve a tridiagonal linear system at each step in order to compute the next iterate $\mathbf{u}^{(i+1)}$. However, as we learned in Section 1.7, tridiagonal systems can be solved very rapidly, and so speed does not become a significant issue in the practical implementation of this implicit scheme.

Let us look at the convergence properties of the implicit scheme. For homogeneous Dirichlet boundary conditions (14.149), the system takes the form

$$\mathbf{u}^{(i+1)} = \widehat{A}^{-1}\,\mathbf{u}^{(i)},$$

and the convergence is now governed by the eigenvalues of $\widehat{A}^{-1}$. Lemma 14.15 tells us that the eigenvalues of $\widehat{A}$ are

$$\lambda_k = 1 + 2\,\mu - 2\,\mu\cos\frac{\pi\,k}{n}\,, \qquad\qquad k = 1, \dots, n-1.$$

As a result, its inverse $\widehat{A}^{-1}$ has eigenvalues

$$\frac{1}{\lambda_k} = \frac{1}{1 + 2\,\mu\left(1 - \cos\dfrac{\pi\,k}{n}\right)}\,, \qquad\qquad k = 1, \dots, n-1.$$

Since $\mu > 0$, the latter are *always* less than 1 in absolute value, and so $\widehat{A}$ is always a convergent matrix. The implicit scheme (14.155) is convergent for any choice of step sizes h, k, and hence *unconditionally stable*.

Example 14.16. Consider the same initial-boundary value problem considered in Example 14.14. In Figure 14.15, we plot the numerical solutions obtained using the implicit scheme. The initial data is not displayed, but we graph the numerical solutions at times $t = .2, .4, .6$ with a mesh size of $h = .1$. On the top line, we use a time step of $k = .01$, while on the bottom $k = .005$. Unlike the explicit scheme, there is very little difference between the two — both come much closer to the actual solution than the explicit scheme. Indeed, even significantly larger time steps give reasonable numerical approximations to the solution.

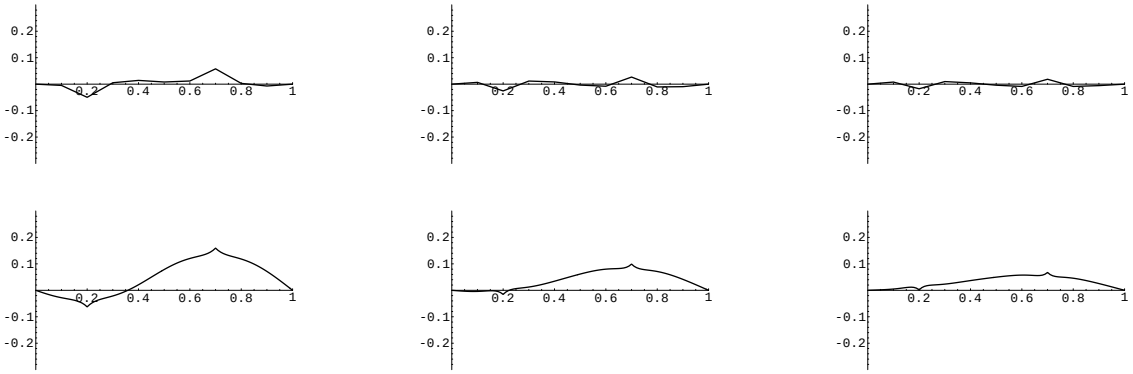


Figure 14.16. Numerical Solutions for the Heat Equation Based on the Crank–Nicolson Scheme.

Another popular numerical scheme is the *Crank–Nicolson method*

$$u_{i+1,j} - u_{i,j} = \frac{\mu}{2}\left(u_{i+1,j+1} - 2\,u_{i+1,j} + u_{i+1,j-1} + u_{i,j+1} - 2\,u_{i,j} + u_{i,j-1} \right). \tag{14.156}$$

which can be obtained by averaging the explicit and implicit schemes (14.142, 154). We can write the iterative system in matrix form

$$B\,\mathbf{u}^{(i+1)} = C\,\mathbf{u}^{(i)} + \tfrac{1}{2}\big(\,\mathbf{b}^{(i)} + \mathbf{b}^{(i+1)}\,\big),$$

where

$$B = \begin{pmatrix} 1+\mu & -\frac{1}{2}\,\mu & & \\ -\frac{1}{2}\,\mu & 1+\mu & -\frac{1}{2}\,\mu & \\ & -\frac{1}{2}\,\mu & \ddots & \ddots \\ & & \ddots & \ddots \end{pmatrix}, \qquad C = \begin{pmatrix} 1-\mu & \frac{1}{2}\,\mu & & \\ \frac{1}{2}\,\mu & 1-\mu & \frac{1}{2}\,\mu & \\ & \frac{1}{2}\,\mu & \ddots & \ddots \\ & & \ddots & \ddots \end{pmatrix}. \tag{14.157}$$

Convergence is governed by the generalized eigenvalues of the tridiagonal matrix pair B, C, or, equivalently, the eigenvalues of the product $B^{-1}C$, cf. Exercise 9.5.33. According to Exercise ■, these are

$$\lambda_k = \frac{1 - \mu\left(1 - \cos\dfrac{\pi\,k}{n}\right)}{1 + \mu\left(1 - \cos\dfrac{\pi\,k}{n}\right)}, \qquad k = 1, \ldots, n-1. \tag{14.158}$$

Since $\mu > 0$, all of the eigenvalues are strictly less than 1 in absolute value, and so the Crank–Nicolson scheme is also unconditionally stable. A detailed analysis will show that the errors are of the order of k^2 and h^2, and so it is reasonable to choose the time step to have the same order of magnitude as the space step, $k \approx h$. This gives the Crank–Nicolson scheme one advantage over the previous two methods. However, applying it to the initial value problem considered earlier points out a significant weakness. Figure 14.16 shows the result of running the scheme on the initial data (14.22). The top row has space and time

step sizes $h = k = .1$, and does a rather poor job replicating the solution. The second row uses $h = k = .01$, and performs better except near the corners where an annoying and incorrect local time oscillation persists as the solution decays. Indeed, since most of its eigenvalues are near -1, the Crank–Nicolson scheme does not do a good job of damping out the high frequency modes that arise from small scale features, including discontinuities and corners in the initial data. On the other hand, most of the eigenvalues of the fully implicit scheme are near zero, and it tends to handle the high frequency modes better, losing out to Crank–Nicolson when the data is smooth. Thus, a good strategy is to first evolve using the implicit scheme until the small scale noise is dissipated away, and then switch to Crank–Nicolson to use a much larger time step for final the large scale changes.

Numerical Solution Methods for the Wave Equation

Let us now look at some numerical solution techniques for the wave equation. Although this is in a sense unnecessary, owing to the explicit d'Alembert solution formula (14.121), the experience we gain in designing workable schemes will serve us well in more complicated situations, including inhomogeneous media, and higher dimensional problems, when analytic solution formulas are no longer available.

Consider the wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \, \frac{\partial^2 u}{\partial x^2} \, , \qquad 0 < x < \ell, \qquad t \geq 0, \tag{14.159}$$

modeling vibrations of a homogeneous bar of length ℓ with constant wave speed $c > 0$. We impose Dirichlet boundary conditions

$$u(t,0) = \alpha(t), \qquad u(t,\ell) = \beta(t), \qquad t \geq 0. \tag{14.160}$$

and initial conditions

$$u(0,x) = f(x), \qquad \frac{\partial u}{\partial t}\,(0,x) = g(x), \qquad 0 \leq x \leq \ell. \tag{14.161}$$

We adopt the same uniformly spaced mesh

$$t_i = i\,k, \qquad x_j = j\,h, \qquad \text{where} \qquad h = \frac{\ell}{n}\,.$$

In order to discretize the wave equation, we replace the second order derivatives by their standard finite difference approximations (14.134), namely

$$\begin{aligned} \frac{\partial^2 u}{\partial t^2}\,(t_i,x_j) &\approx \frac{u(t_{i+1},x_j) - 2\,u(t_i,x_j) + u(t_{i-1},x_j)}{k^2} + \mathrm{O}(h^2), \\ \frac{\partial^2 u}{\partial x^2}\,(t_i,x_j) &\approx \frac{u(t_i,x_{j+1}) - 2\,u(t_i,x_j) + u(t_i,x_{j-1})}{h^2} + \mathrm{O}(k^2), \end{aligned} \tag{14.162}$$

Since the errors are of orders of k^2 and h^2, we anticipate to be able to choose the space and time step sizes of comparable magnitude:

$$k \approx h.$$

Substituting the finite difference formulae (14.162) into the partial differential equation (14.159), and rearranging terms, we are led to the iterative system

$$u_{i+1,j} = \sigma^2\, u_{i,j+1} + 2\,(1-\sigma^2)\, u_{i,j} + \sigma^2\, u_{i,j-1} - u_{i-1,j}, \qquad \begin{array}{l} i = 1, 2, \dots, \\ j = 1, \dots, n-1, \end{array} \tag{14.163}$$

for the numerical approximations $u_{i,j} \approx u(t_i, x_j)$ to the solution values at the mesh points. The positive parameter

$$\sigma = \frac{c\,k}{h} > 0 \tag{14.164}$$

depends upon the wave speed and the ratio of space and time step sizes. The boundary conditions (14.160) require that

$$u_{i,0} = \alpha_i = \alpha(t_i), \qquad u_{i,n} = \beta_i = \beta(t_i), \qquad i = 0, 1, 2, \dots. \tag{14.165}$$

This allows us to rewrite the system in matrix form

$$\mathbf{u}^{(i+1)} = B\,\mathbf{u}^{(i)} - \mathbf{u}^{(i-1)} + \mathbf{b}^{(i)}, \tag{14.166}$$

where

$$B = \begin{pmatrix} 2\,(1-\sigma^2) & \sigma^2 & & & \\ \sigma^2 & 2\,(1-\sigma^2) & \sigma^2 & & \\ & \sigma^2 & \ddots & \ddots & \\ & & \ddots & \ddots & \sigma^2 \\ & & & \sigma^2 & 2\,(1-\sigma^2) \end{pmatrix}, \quad \mathbf{u}^{(j)} = \begin{pmatrix} u_{1,j} \\ u_{2,j} \\ \vdots \\ u_{n-2,j} \\ u_{n-1,j} \end{pmatrix}, \quad \mathbf{b}^{(j)} = \begin{pmatrix} \sigma^2\,\alpha_j \\ 0 \\ \vdots \\ 0 \\ \sigma^2\,\beta_j \end{pmatrix}. \tag{14.167}$$

The entries of $\mathbf{u}^{(i)}$ are, as in (14.146), the numerical approximations to the solution values at the *interior* nodes. Note that the system (14.166) is a second order iterative scheme, since computing the next iterate $\mathbf{u}^{(i+1)}$ requires the value of the preceding two, $\mathbf{u}^{(i)}$ and $\mathbf{u}^{(i-1)}$.

The one difficulty is getting the method started. We know $\mathbf{u}^{(0)}$ since $u_{0,j} = f_j = f(x_j)$ is determined by the initial position. However, we also need to find $\mathbf{u}^{(1)}$ with entries $u_{1,j} \approx u(k, x_j)$ at time $t_1 = k$ in order launch the iteration, but the initial velocity $u_t(0,x) = g(x)$ prescribes the derivatives $u_t(0, x_j) = g_j = g(x_j)$ at time $t_0 = 0$ instead. One way to resolve this difficult would be to utilize the finite difference approximation

$$g_j = \frac{\partial u}{\partial t}\,(0, x_j) \approx \frac{u(k, x_j) - u(0, x_j)}{k} \approx \frac{u_{1,j} - g_j}{k} \tag{14.168}$$

to compute the required values

$$u_{1,j} = f_j + k\,g_j.$$

However, the approximation (14.168) is only accurate to order k, whereas the rest of the scheme has error proportional to k^2. Therefore, we would introduce an unacceptably large error at the initial step.

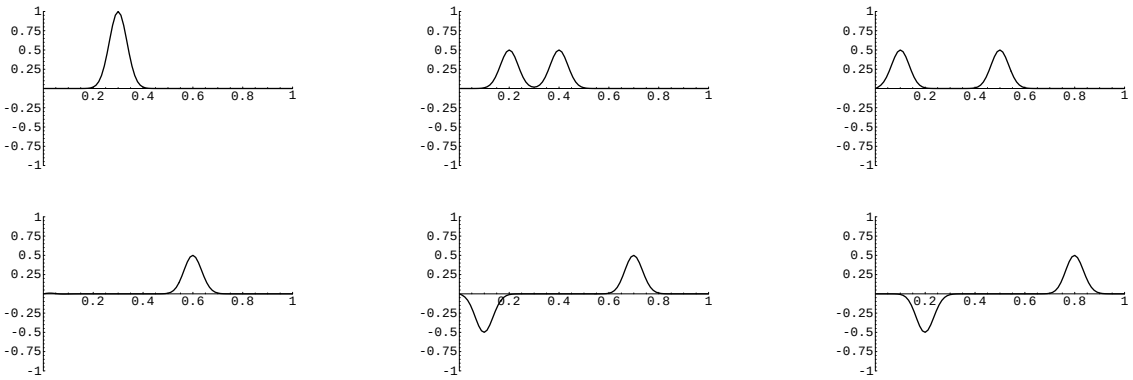


Figure 14.17. Numerically Stable Waves.

To construct an initial approximation to $\mathbf{u}^{(1)}$ with error on the order of k^2, we need to analyze the local error in more detail. Note that, by Taylor's theorem,

$$\frac{u(k,x_j)-u(0,x_j)}{k} \approx \frac{\partial u}{\partial t}\,(0,x_j) + \frac{k}{2}\,\frac{\partial^2 u}{\partial t^2}\,(0,x_j) = \frac{\partial u}{\partial t}\,(0,x_j) + \frac{c^2\,k}{2}\,\frac{\partial^2 u}{\partial x^2}\,(0,x_j)\,,$$

where the error is now of order k^2, and, in the final equality, we have used the fact that u is a solution to the wave equation. Therefore, we find

$$\begin{aligned} u(k,x_j) &\approx u(0,x_j) + k\,\frac{\partial u}{\partial t}\,(0,x_j) + \frac{c^2\,k^2}{2}\,\frac{\partial^2 u}{\partial x^2}\,(0,x_j) \\ &= f(x_j) + k\,g(x_j) + \frac{c^2\,k^2}{2}\,f''(x_j) \approx f_j + k\,g_j + \frac{c^2\,k^2}{2\,h^2}(f_{j+1} - 2\,f_j + f_{j-1})\,, \end{aligned}$$

where we can use the finite difference approximation (14.134) for the second derivative of $f(x)$ if no explicit formula is known. Therefore, when we initiate the scheme by setting

$$u_{1,j} = \tfrac{1}{2}\,\sigma^2\,f_{j+1} + (1-\sigma^2)f_j + \tfrac{1}{2}\,\sigma^2\,f_{j-1} + k\,g_j, \tag{14.169}$$

or, in matrix form,

$$\mathbf{u}^{(0)} = \mathbf{f}, \qquad\qquad \mathbf{u}^{(1)} = \tfrac{1}{2}\,B\,\mathbf{u}^{(0)} + k\,\mathbf{g} + \tfrac{1}{2}\,\mathbf{b}^{(0)}, \tag{14.170}$$

we will have maintained the desired order k^2 (and h^2) accuracy.

Example 14.17. Consider the particular initial value problem

$$u_{tt} = u_{xx}, \qquad \begin{aligned} u(0,x) &= e^{-400\,(x-.3)^2}, \qquad u_t(0,x) = 0, \qquad & 0 \le x \le 1, \\ u(t,0) &= u(1,0) = 0, & t \ge 0, \end{aligned}$$

subject to homogeneous Dirichlet boundary conditions on the interval $[0,1]$. The initial data is a fairly concentrated single hump centered at $x = .3$, and we expect it to split into two half sized humps, which then collide with the ends. Let us choose a space discretization consisting of 90 equally spaced points, and so $h = \frac{1}{90} = .0111\ldots$. If we choose a time step of $k = .01$, whereby $\sigma = .9$, then we get reasonably accurate solution over a fairly long time range, as plotted in Figure 14.17 at times $t = 0, .1, .2, \ldots, .5$. On the other hand,

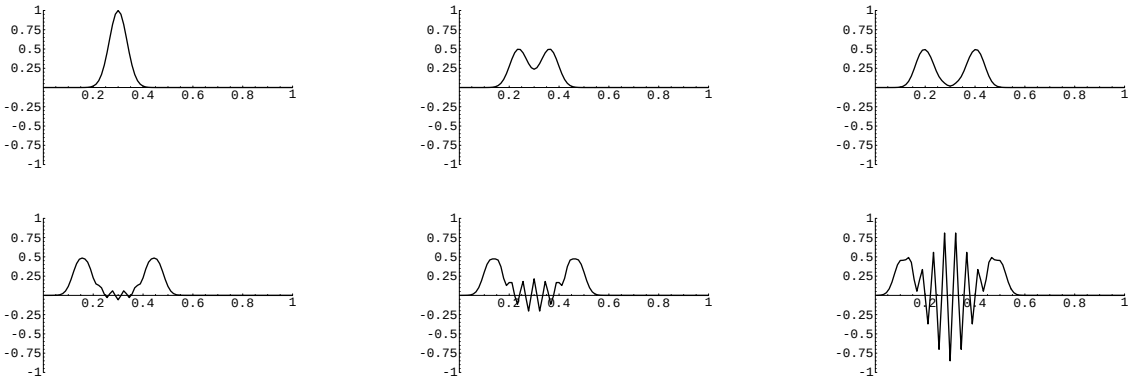


Figure 14.18. Numerically Unstable Waves.

if we double the time step, setting $k = .02$, so $\sigma = 1.8$, then, as plotted in Figure 14.18 at times $t = 0, .05, .1, .14, .16, .18$, we observe an instability eventually creeping into the picture that eventually overwhelms the numerical solution. Thus, the numerical scheme appears to only be conditionally stable.

The stability analysis of this numerical scheme proceeds as follows. We first need to recast the second order iterative system (14.166) into a first order system. In analogy with Example 10.6, this is accomplished by introducing the vector $\mathbf{z}^{(i)} = \begin{pmatrix} \mathbf{u}^{(i)} \\ \mathbf{u}^{(i-1)} \end{pmatrix} \in \mathbb{R}^{2n-2}$. Then

$$\mathbf{z}^{(i+1)} = C\,\mathbf{z}^{(i)} + \mathbf{c}^{(i)}, \qquad \text{where} \qquad C = \begin{pmatrix} B & -\mathrm{I} \\ \mathrm{I} & \mathrm{O} \end{pmatrix}. \tag{14.171}$$

Therefore, the stability of the method will be determined by the eigenvalues of the coefficient matrix C. The eigenvector equation $C\,\mathbf{z} = \lambda\,\mathbf{z}$, where $\mathbf{z} = \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix}$, can be written out in its individual components:

$$B\,\mathbf{u} - \mathbf{v} = \lambda\,\mathbf{u}, \qquad \mathbf{u} = \lambda\,\mathbf{v}.$$

Substituting the second equation into the first, we find

$$(\lambda\,B - \lambda^2 - 1)\,\mathbf{v} = \mathbf{0}, \qquad \text{or} \qquad B\,\mathbf{v} = \left(\lambda + \frac{1}{\lambda}\right)\,\mathbf{v}.$$

The latter equation implies that $\mathbf{v}$ is an eigenvector of B with $\lambda + \lambda^{-1}$ the corresponding eigenvalue. The eigenvalues of the tridiagonal matrix B are governed by Lemma 14.15, in which $a = 2(1 - \sigma^2)$ and $b = \sigma^2$, and hence are

$$\lambda + \frac{1}{\lambda} = 2\left(1 - \sigma^2 + \sigma^2 \cos\frac{\pi\,k}{n}\right), \qquad k = 1, \dots, n-1.$$

Multiplying both sides by λ leads to a quadratic equation for the eigenvalues,

$$\lambda^2 - 2\,a_k\,\lambda + 1 = 0, \qquad \text{where} \qquad 1 - 2\sigma^2 < a_k = 1 - \sigma^2 + \sigma^2 \cos\frac{\pi\,k}{n} < 1. \tag{14.172}$$

Figure 14.19. The Courant Condition.

Each pair of solutions to these $n-1$ quadratic equations, namely

$$\lambda_k^{\pm} = a_k \pm \sqrt{a_k^2 - 1}\,, \tag{14.173}$$

yields two eigenvalues of the matrix C. If $a_k > 1$, then one of the two eigenvalues will be larger than one in magnitude, which means that the linear iterative system has an exponentially growing mode, and so $\| \mathbf{u}^{(i)} \| \to \infty$ as $i \to \infty$ for almost all choices of initial data. This is clearly incompatible with the wave equation solution that we are trying to approximate, which is periodic and hence remains bounded. On the other hand, if $| a_k | < 1$, then the eigenvalues (14.173) are complex numbers of modulus 1, indicated stability (but not convergence) of the matrix C. Therefore, in view of (14.172), we should require that

$$\sigma = \frac{c\,k}{h} < 1, \qquad \text{or} \qquad k < \frac{h}{c}\,, \tag{14.174}$$

which places a restriction on the relative sizes of the time and space steps. We conclude that the numerical scheme is conditionally stable.

The stability criterion (14.174) is known as the *Courant condition*, and can be assigned a simple geometric interpretation. Recall that the wave speed c is the slope of the characteristic lines for the wave equation. The Courant condition requires that the *mesh slope*, which is defined to be the ratio of the space step size to the time step size, namely h/k, must be strictly greater than the characteristic slope c. A signal starting at a mesh point (t_i, x_j) will reach positions $x_j \pm k/c$ at the next time $t_{i+1} = t_i + k$, which are still between the mesh points x_{j-1} and x_{j+1}. Thus, characteristic lines that start at a mesh point are not allowed to reach beyond the neighboring mesh points at the next time step.

For instance, in Figure 14.19, the wave speed is $c = 1.25$. The first figure has equal mesh spacing $k = h$, and does not satisfy the Courant condition (14.174), whereas the second figure has $k = \frac{1}{2}\,h$, which does. Note how the characteristic lines starting at a given mesh point have progressed beyond the neighboring mesh points after one time step in the first case, but not in the second.

14.7. A General Framework for Dynamics.

According to Section 12.1, the one-dimensional heat and wave equations are individual instances of two broad classes of dynamical systems that include, in a common framework, both discrete dynamics modeled by systems of ordinary differential equations and continuum systems modeled by (systems of) partial differential equations. In preparation for their multi-dimensional generalizations, it will be useful to summarize the general mathematical framework, which can be regarded as the dynamical counterpart to the framework for equilibrium developed in Sections 7.5 and 15.4, which you are advised to review before going through this section. Readers more attuned to concrete examples, on the other hand, might prefer to skip this material entirely, referring back as needed.

In all situations, the starting point is a certain linear function

$$L\colon U \longrightarrow V \tag{14.175}$$

that maps a vector space U to another vector space V. In mechanics, the elements of U represent displacements, while the elements of V represent strains (elongations). In electromagnetism and gravitation, elements of U represent potentials and elements of V electric or magnetic or gravitational fields. In thermodynamics, elements of U represent temperature distributions, and elements of V temperature gradients. In fluid mechanics, U contains potential functions and V is fluid velocities. And so on.

In the discrete, finite-dimensional setting, when $U = \mathbb{R}^n$ and $V = \mathbb{R}^m$, the linear function $L[\mathbf{u}] = A\,\mathbf{u}$ is represented by multiplication by an $m \times n$ matrix A — the incidence matrix and its generalizations. In the infinite-dimensional function space situation presented in Chapter 11 and earlier in this chapter, the linear map L is a differential operator; in the case of one-dimensional elastic bars, it is the first order derivative $L = D_x$, while for beams it is a second order derivative $L = D_x^2$. In the multi-dimensional physics treated in subsequent chapters, the most important example is the gradient operator, $L = \nabla$, that maps scalar potentials to vector fields.

The vector spaces U and V are further each endowed with inner products, that encapsulate the consitutive assumptions underlying the physical problem. Definition 7.53 explains how to construct the resulting adjoint map

$$L^*\colon V \longrightarrow U \tag{14.176}$$

which goess in the *reverse* direction. In finite dimensions, when $U = \mathbb{R}^n$ and $V = \mathbb{R}^m$ are both equipped with the Euclidean dot product, the adjoint corresponds to the simple transpose operation. Modifications under more general weighted inner products were discussed in Section 7.5. In infinite-dimensional contexts, the computation of the adjoint presented in Section 11.3 relied on an integration by parts argument, supplemented by suitable boundary conditions. In the next chapter, we will adapt this argument to determine adjoints of multi-dimensional differential operators.

The crucial operator underlying the equilibrium and dynamical equations for a remarkably broad range of physics is the self-adjoint combination

$$K = L^* \circ L\colon U \longrightarrow U. \tag{14.177}$$

According to Theorem 7.60, the operator K is *positive semi-definite* in all situations, and *positive definite* if and only if $\ker L = \{0\}$. In the finite-dimensional context, K is represented by the symmetric positive (semi-)definite Gram matrix $A^T A$ when both $U = \mathbb{R}^n$ and $V = \mathbb{R}^m$ have the dot product, by the symmetric combination $A^T C\, A$ when V has a weighted inner product represented by the positive definite matrix $C > 0$, and the more general self-adjoint form $M^{-1} A^T C\, A$ when U also is endowed with a weighted inner product represented by $M > 0$. In one-dimensional bars, K is represented by a self-adjoint second order differential operator, whose form depends upon the underlying inner products., while in beams it becomes a fourth order differential operator. The definiteness of K (or lack thereof) depends upon the imposed boundary conditions. In higher dimensions, as we discuss in Section 15.4, K becomes the Laplacian or a more general elliptic partial differential operator.

With this set-up, the basic *equilibrium equation* has the form

$$K[u] = f, \tag{14.178}$$

where f represents an external forcing function. In finite dimensions, this is a linear system consisting of n equations in n unknowns with positive (semi-)definite coefficient matrix. In function space, it becomes a self-adjoint boundary value problem for the unknown function u. If K is positive definite, the solution is unique. (But rigorously proving the existence of a solution is not trivial, requiring serious analysis beyond the scope of this text.) Theorem 7.61 says that the solutions are characterized by a minimization principle, as the minimizers of the quadratic function(al)

$$p[u] = \tfrac{1}{2} \,\|\, L[u] \,\|^2 - \langle\, u \,,\, f \,\rangle, \tag{14.179}$$

which typically represents the potential energy in the system. If K is only positive semi-definite, existence of a solution requires that the forcing function satisfy the Fredholm condition that it be orthogonal to the the unstable modes, that is the elements of $\ker K = \ker L$

With the equilibrium operator K in hand, there are two basic types of dynamical systems of importance in physical models. Unforced *diffusion processes* are modeled by a dynamical system of the form

$$u_t = -\,K[u], \qquad \text{where} \qquad K = L^* \circ L \tag{14.180}$$

is the standard self-adjoint combination of a linear operator L. In the discrete case, K is a matrix and so this represents a first order system of ordinary differential equations, that has the form of a linear *gradient flow* (9.18), so named because it decreases the energy function

$$q[u] = \|\, L[u] \,\|^2, \tag{14.181}$$

which is (14.179) when $f = 0$, as rapidly as possible. In the continuous case, K is a differential operator, and (14.180) represents a partial differential equation for the time-varying function $u = u(t, \mathbf{x})$.

The solution to the general diffusion equation (14.180) mimics earlier our separation of variables method for the one-dimensional heat equation, as well as the original solution

technique for linear systems of ordinary differential equations, as developed in Chapter 9. The separable solutions are of exponential form

$$u(t) = e^{-\lambda t}\, v, \tag{14.182}$$

where $v \in U$ is a fixed element of the domain space — i.e., a vector in the discrete context, or a function $v = v(x)$ that only depends on the spatial variables in the continuum versions. Since the operator K is linear and does not involve t differentiation, we find

$$\frac{\partial u}{\partial t} = -\lambda\, e^{-\lambda t}\, v, \qquad \text{while} \qquad K[u] = e^{-\lambda t}\, K[v].$$

Substituting back into (14.180) and canceling the common exponential factors, we are led to the eigenvalue problem

$$K[v] = \lambda\, v. \tag{14.183}$$

Thus, v must be an eigenvector/eigenfunction for the linear operator K, with λ the corresponding eigenvalue.

Generalizing our earlier observations on the eigenvalues of positive definite matrices and the boundary value problems associated with the one-dimensional heat equation, let us establish the positivity of eigenvalues of such general self-adjoint, positive (semi-)definite linear operators.

Theorem 14.18. *All eigenvalues of the linear operator $K = L^* \circ L$ are real and non-negative*: $\lambda \geq 0$. *If, moreover, K is positive definite, or, equivalently,* $\ker K = \ker L = 0$, *then all eigenvalues are strictly positive*: $\lambda > 0$.

Proof: Suppose $K[u] = \lambda\, u$ with $u \neq 0$. Then

$$\lambda \,\|\, u \,\|^2 = \lambda \,\langle\, u\,, u \,\rangle = \langle\, K[u]\,, u \,\rangle = \langle\, L^* \circ L[u]\,, u \,\rangle = \langle\, L[u]\,, L[u] \,\rangle = \|\, L[u] \,\|^2 \geq 0,$$

by the defining equation (7.74) of the adjoint operator. Since $\| u \|^2 > 0$, this immediately implies that $\lambda \geq 0$. Furthermore, in the positive definite case $\ker L = \{0\}$, and so $L[u] \neq 0$. Thus, $\| L[u] \|^2 > 0$, proving that $\lambda > 0$. *Q.E.D.*

We index the eigenvalues in increasing order:

$$0 < \lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \ \cdots\ . \tag{14.184}$$

where the eigenvalues are repeated according to their multiplicities, and $\lambda_0 = 0$ is an eigenvalue only in the positive semi-definite case. Each eigenvalue induces a separable *eigensolution*

$$u_k(t) = e^{-\lambda_k t}\, v_k \tag{14.185}$$

to the diffusion equation (14.180). Those associated with strictly positive definite eigenvalues are exponentially decaying, at a rate equal to the eigenvalue $\lambda_k > 0$, while any null eigenvalue modes correspond to constant solutions $u_0(t) \equiv v_0$ where v_0 is any null eigenvector/eigenfunction, i.e., element of $\ker K = \ker L$. The general solution is built up as a linear combination

$$u(t) = \sum_k\, c_k\, u_k(t) = \sum_k\, c_k\, e^{-\lambda_k t}\, v_k \tag{14.186}$$

of the eigensolutions. Thus, all solutions decay exponentially fast as $t \to \infty$ to an equilibrium solution, which is 0 in the positive definite case, or, more generally, the null eigenspace component $c_0 v_0$. The overall rate of decay is, generally, prescribed by the smallest positive eigenvalue $\lambda_1 > 0$.

In the discrete version, the summation (14.186) has only finitely many terms, corresponding to the n eigenvalues of the matrix representing K. Moreover, thanks to the Spectral Theorem 8.26, the eigensolutions form a basis for the solution space to the diffusion equation. In infinite-dimensional function space, there are, in many instances, an infinite number of eigenvalues, with $\lambda_k \to \infty$ as $k \to \infty$. Completeness of the resulting eigensolutions is a more delicate issue. Often, as in the one-dimensional heat equation, the eigenfunctions are complete and the (Fourier) series (14.186) converges and represents the general solution. But in situations involving unbounded domains, like the hydrogen atom to be discussed in Section 18.7, there are additional separable solutions corresponding to the so-called *continuous spectrum* that are not represented in terms of eigenfunctions, and the analysis is considerably more involved, requiring analogs of the Fourier transform. A full discussion of completeness and convergence of eigenfunction expansions must be relegated to an advanced course in analysis, [**46**, **149**].

Assuming completeness, the eigensolution coefficients c_k in (14.186) are prescribed by the initial conditions, which require

$$\sum_k c_k \, v_k = h, \tag{14.187}$$

where $h = u(0)$ represents the initial data. To compute the coefficients c_k in the eigenfunction expansion (14.187), we appeal, as in the case of ordinary Fourier series, to orthogonality of the eigenvectors/eigenfunctions v_k. Orthogonality is proved by a straightforward adaptation of our earlier proof of part (b) of Theorem 8.20, guaranteeing the orthogonality of the eigenvectors of a symmetric matrix.

Theorem 14.19. *Two eigenvectors/eigenfunctions u, v of the self-adjoint linear operator $K = L^* \circ L$ that are associated with distinct eigenvalues $\lambda \neq \mu$ are orthogonal.*

Proof: Self-adjointness of the operator K means that

$$\langle\, K[u]\,, v\,\rangle = \langle\, L[u]\,, L[v]\,\rangle = \langle\, u\,, K[v]\,\rangle$$

for any for any u, v. In particular, if $K[u] = \lambda\, u$, $K[v] = \mu\, v$, are eigenfunctions, then

$$\lambda\,\langle\, u\,, v\,\rangle = \langle\, K[u]\,, v\,\rangle = \langle\, u\,, K[v]\,\rangle = \mu\,\langle\, u\,, v\,\rangle,$$

which, assuming $\lambda \neq \mu$, immediately implies orthogonality: $\langle\, u\,, v\,\rangle = 0$. *Q.E.D.*

If an eigenvalue λ admits more than one independent eigenvector/eigenfunction, we can apply the Gram–Schmidt process to produce an orthogonal basis of its eigenspace[†]

[†] This assumes that the eigenspace V_λ is finite-dimensional — which is assured whenever the eigenvalues $\lambda_k \to \infty$ as $k \to \infty$.

$V_\lambda = \ker(K - \lambda\, \mathrm{I}\,)$. In this way, the entire set of eigenvectors/eigenfunctions can be assumed to be mutually orthogonal:

$$\langle\, v_j\,, v_k\,\rangle = 0, \qquad\qquad j \neq k.$$

As a consequence, taking the inner product of both sides of (14.187) with the eigenfunction v_k leads to the equation

$$c_k \,\|\, v_k \,\|^2 = \langle\, h\,, v_k\,\rangle \qquad \text{and hence} \qquad c_k = \frac{\langle\, h\,, v_k\,\rangle}{\|\, v_k\,\|^2}\,. \tag{14.188}$$

In this manner we recover our standard orthogonality formula (5.7) for expressing elements of a vector space in terms of an orthogonal basis.

The second important class of dynamical systems consists of second order (in time) vibration equations

$$u_{tt} = -\,K[u]. \tag{14.189}$$

Newton's equations of motion in the absence of non-conservative frictional forces, the propagation of waves in fluids and solids, as well as electromagnetic waves, and many other related physical problems lead to such vibrational systems. In this case, the separable solutions are of trigonometric form

$$u(t,x,y) = \cos(\omega\, t)\, v \qquad \text{or} \qquad \sin(\omega\, t)\, v. \tag{14.190}$$

Substituting this ansatz back into the vibration equation (14.189) results in the same eigenvalue problem (14.183) with eigenvalue $\lambda = \omega^2$ equal to the square of the vibrational frequency. We conclude that the *normal mode* or separable eigensolutions take the form

$$u_k(t) = \cos(\omega_k\, t)\, v_k, \qquad \widetilde{u}_k(t) = \sin(\omega_k\, t)\, v_k, \qquad \text{provided} \qquad \lambda_k = \omega_k^2 > 0$$

is a non-zero eigenvalue. In the stable, positive definite case, there are no zero eigenvalues, and so the general solution is built up as a (quasi-)periodic[†] combination

$$u(t) = \sum_k \big[\, c_k u_k(t) + d_k \widetilde{u}_k(t)\,\big] = \sum_k r_k \cos(\omega_k\, t + \delta_k)\, v_k, \tag{14.191}$$

of the eigenmodes. The initial conditions

$$g = u(0) = \sum_k c_k v_k, \qquad\qquad h = u_t(0) = \sum_k d_k \omega_k v_k, \tag{14.192}$$

are used to specify the coefficients c_k, d_k, using the same orthogonality formula (14.188):

$$c_k = \frac{\langle\, f\,, v_k\,\rangle}{\|\, v_k\,\|^2}\,, \qquad\qquad d_k = \frac{\langle\, f\,, v_k\,\rangle}{\omega_k \,\|\, v_k\,\|^2}\,. \tag{14.193}$$

[†] The solution is periodic if and only if the frequencies appearing in the sum are all integer multiples of a common frequency: $\omega_k = n_k \omega_\star$ for $n_k \in \mathbb{N}$.

In the unstable, positive semi-definite cases, the null eigensolutions have the form

$$u_0(t) = v_0, \qquad \widetilde{u}_0(t) = t\, v_0,$$

where $v_0 \in \ker K = \ker L$, and must be appended to the series solution. The unstable mode $\widetilde{u}_0(t)$ is excited if and only if the initial velocity is not orthogonal to the kernel element: $\langle\, h\,, v_0\,\rangle \neq 0$.

In classical mechanics, the diffusion and vibration equations and their variants (see, for instance, Exercises ■, ■ for versions with external forcing and Exercise ■ for vibrations with frictional effects) are the most important classes of dynamical systems. In quantum mechanics, the basic system for quantum dynamics is the *Schrödinger equation*, first written down by the the German physicist Erwin Schrödinger, one of the founders of quantum mechanics. The abstract form of the Schrödinger equation is

$$\mathrm{i}\,\hbar\, u_t = K[u]. \tag{14.194}$$

Here $i = \sqrt{-1}$, while

$$\hbar = \frac{h}{2\,\pi} \approx 1.055 \times 10^{-34} \quad \text{Joule seconds} \tag{14.195}$$

is *Planck's constant*, whose value governs the quantization of all physical quantities. At each time t, the solution $u(t, \mathbf{x})$ to the Schrödinger equation represents the wave function of the quantum system, and so is a complex-valued square integrable function of constant L^2 norm: $\| u \| = 1$. (See Sections 12.5 and 13.3 for the basics of quantum mechanics and Hilbert space.) As usual, we interpret the wave fuction as a probability density on the possible quantum states, and so the Schrödinger equation governs the dynamical evolution of quantum probabilities. The operator $K = L^* \circ L$ is known as the *Hamiltonian* for the quantum mechanical system governed by (14.194), and, typically represents the quantum energy operator. For physical systems such as atoms and nuclei, the relevant Hamiltonian operator is constructed from the classical energy through the rather mysterious process of "quantization". The interested reader should consult a basic text on quantum mechanics, e.g., [**122**, **127**], for full details on both the physics and underlying mathematics.

Proposition 14.20. *If $u(t)$ is a solution to the Schrödinger equation, its Hermitian L^2 norm $\| u \|$ is constant.*

Proof: Since the solution is complex-valued, we use the sesquilinearity of the underlying Hermitian inner product, as in (3.92), to compute

$$\begin{aligned}\frac{d}{dt}\,\| u \|^2 &= \langle\, u_t\,, u\,\rangle + \langle\, u\,, u_t\,\rangle \\ &= \left\langle\, -\frac{\mathrm{i}}{\hbar}\, K[u]\,,\, u\,\right\rangle + \left\langle\, u\,, -\frac{\mathrm{i}}{\hbar}\, K[u]\,\right\rangle = -\,\frac{\mathrm{i}}{\hbar}\,\langle\, K[u]\,, u\,\rangle + \frac{\mathrm{i}}{\hbar}\,\langle\, u\,, K[u]\,\rangle = 0,\end{aligned}$$

which vanishes since K is self-adjoint. Since its derivative vanishes everywhere, this implies that $\| u \|^2$ is constant. *Q.E.D.*

As a result, if the initial data $u(t_0) = h$ is a quantum mechanical wave function, meaning that $\| h \| = 1$, then, at each time t, the solution to the Schrödinger equation also has norm 1, and hence remains a wave function for all t.

Apart from the extra factor of $\mathrm{i}\,\hbar$, the Schrödinger equation looks like a diffusion equation (14.180). (*Warning*: Despite this superficial similarity, their solutions have radically different behavior.) This inspires us to seek separable solutions with an exponential ansatz:

$$u(t,x) = e^{\alpha\, t}\, v(x).$$

Substituting this expression into the Schrödinger equation (14.194) and canceling the common exponential factors reduces us to the usual eigenvalue problem

$$K[v] = \lambda\, v, \qquad \text{with eigenvalue} \qquad \lambda = -\,\mathrm{i}\,\hbar\,\alpha.$$

Let $v_k(x)$ denote the normalized eigenfunction associated with the k^{th} eigenvalue λ_k. Th corresponding separable solution of the Schrödinger equation is the complex wave functions

$$u_k(t,x) = e^{\,\mathrm{i}\,\lambda_k t/\hbar}\, v_k(x).$$

Observe that, in contrast to the exponentially decaying solutions to the diffusion equation, the eigensolutions to the Schrödinger equation are periodic, of frequencies proportional to the eigenvalues: $\omega_k = \lambda_k/\hbar$. (Along with constant solutions corresponding to the null eigenmodes, if any.) The general solution is a (quasi-)periodic series in the fundamental eigensolutions. The periodicity of the summands has the additional implication that, again unlike the diffusion equation, the Schrödinger equation can be run backwards in time. So, we can figure out both the past and future behavior of a quantum system from its present configuration.

Example 14.21. In a single space dimension, the simplest version of the Schrödinger equation is based on the derivative operator $L = D_x$, for which, assuming appropriate boundary conditions, the self-adjoint combination $K = L^* \circ L = -D_x^2$. Thus, (14.194) reduces to the second order partial differential equation

$$\mathrm{i}\,\hbar\, u_t = -\,u_{xx}. \tag{14.196}$$

Imposing the Dirichlet boundary conditions

$$u(t,0) = u(t,\ell) = 0,$$

the Schrödinger equation governs the dynamics of a quantum particle confined to the interval $0 < x < \ell$; the boundary conditions imply that there is zero probability of the particle escaping from the interval.

According to Section 14.1, the eigenfunctions of the Dirichlet eigenvalue problem

$$v_{xx} + \lambda\, v = 0, \qquad v(0) = v(\ell) = 0,$$

are

$$v_k(x) = \sqrt{\frac{2}{\ell}}\, \sin\frac{k\,\pi}{\ell}\, x \qquad \text{for} \qquad k = 1, 2, \dots, \qquad \text{with eigenvalue} \qquad \lambda_k = \frac{k^2\pi^2}{\ell^2},$$

where the initial factor is ensures that v_k has unit L^2 norm, and hence is a bona fide wave function. The corresponding separable solutions or eigenmodes are

$$u_k(t,x) = \sqrt{\frac{2}{\ell}}\, \exp\left(\mathrm{i}\, \frac{k^2\pi^2}{\hbar\,\ell^2}\, t \right) \sin \frac{k\,\pi}{\ell}\, x.$$

The eigenvalues represent the energy levels of the particle, which can be observed from the spectral lines emitted by the system. For instance, when an electron jumps from one level to another, it conserves energy by emitting a photon, with energy equal to the difference in energy between the two quantum levels. These emitted photons define the observed electromagnetic spectral lines, hence the adoption of the physics term "spectrum" to describe the eigenvalues of the Hamiltonian operator K.

Chapter 15

The Planar Laplace Equation

The fundamental partial differential equations that govern the equilibrium mechanics of multi-dimensional media are the Laplace equation and its inhomogeneous counterpart, the Poisson equation. The Laplace equation is arguably the most important differential equation in all of applied mathematics. It arises in an astonishing variety of mathematical and physical systems, ranging through fluid mechanics, electromagnetism, potential theory, solid mechanics, heat conduction, geometry, probability, number theory, and on and on. The solutions to the Laplace equation are known as "harmonic functions", and the discovery of their many remarkable properties forms one of the most significant chapters in the history of mathematics.

In this chapter, we concentrate on the Laplace and Poisson equations in a two-dimensional (planar) domain. Their status as equilibrium equations implies that the solutions are determined by their values on the boundary of the domain. As in the one-dimensional equilibrium boundary value problems, the principal cases are Dirichlet or fixed, Neumann or free, and mixed boundary conditions arise. In the introductory section, we shall briefly survey the basic boundary value problems associated with the Laplace and Poisson equations. We also take the opportunity to summarize the crucially important tripartite classification of planar second order partial differential equations: *elliptic*, such as the Laplace equation; *parabolic*, such as the heat equation; and *hyperbolic*, such as the wave equation. Each species has quite distinct properties, both analytical and numerical, and each forms an essentially distinct discipline. Thus, by the conclusion of this chapter, you will have encountered all three of the most important genres of partial differential equations.

The most important general purpose method for constructing explicit solutions of linear partial differential equations is the method of separation of variables. The method will be applied to the Laplace and Poisson equations in the two most important coordinate systems — rectangular and polar. Linearity implies that we may combine the separable solutions, and the resulting infinite series expressions will play a similar role as for the heat and wave equations. In the polar coordinate case, we can, in fact, sum the infinite series in closed form, leading to the explicit Poisson integral formula for the solution. More sophisticated techniques, relying on complex analysis, but (unfortunately) only applicable to the two-dimensional case, will be deferred until Chapter 16.

Green's formula allows us to properly formulate the Laplace and Poisson equations in self-adjoint, positive definite form, and thereby characterize the solutions via a minimization principle, first proposed by the nineteenth century mathematician Lejeune Dirichlet, who also played a crucial role in putting Fourier analysis on a rigorous foundation. Minimization forms the basis of the most important numerical solution technique — the finite

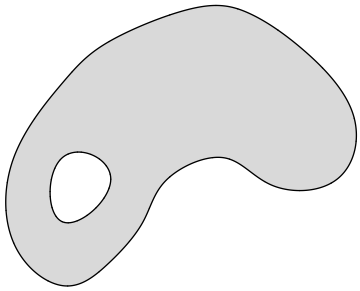


Figure 15.1. Planar Domain.

element method that we first encountered in Chapter 11. In the final section, we discuss numerical solution techniques based on finite element analysis for the Laplace and Poisson equations and their elliptic cousins, including the Helmholtz equation and more general positive definite boundary value problems.

15.1. The Planar Laplace Equation.

The two-dimensional *Laplace equation* is the second order linear partial differential equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0. \tag{15.1}$$

It is named in honor of the outstanding eighteenth century French mathematician Pierre–Simon Laplace. Along with the heat and wave equations, it completes the trinity of truly fundamental partial differential equations. A real-valued solution $u(x, y)$ to the Laplace equation is known as a *harmonic function*. The space of harmonic functions can thus be identified as the kernel of the second order linear partial differential operator

$$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\,, \tag{15.2}$$

known as the *Laplace operator*, or *Laplacian* for short. The inhomogeneous or forced version, namely

$$-\Delta[\,u\,] = -\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f(x, y) \tag{15.3}$$

is known as *Poisson's equation*, named for Siméon–Denis Poisson, who was taught by Laplace. Poisson's equation can be viewed as the higher dimensional analogue of the basic equilibrium equation (11.12) for a bar.

The Laplace and Poisson equations arise as the basic equilibrium equations in a remarkable variety of physical systems. For example, we may interpret $u(x, y)$ as the displacement of a *membrane*, e.g., a drum skin; the inhomogeneity $f(x, y)$ in the Poisson equation represents an external forcing. Another example is in the thermal equilibrium of flat plates; here $u(x, y)$ represents the temperature and $f(x, y)$ an external heat source. In fluid mechanics, $u(x, y)$ represents the potential function whose gradient $\mathbf{v} = \nabla u$ is the vleocity vector of a steady planar fluid flow. Similar considerations apply to two-dimensional electrostatic and gravitational potentials. The dynamical counterparts to the

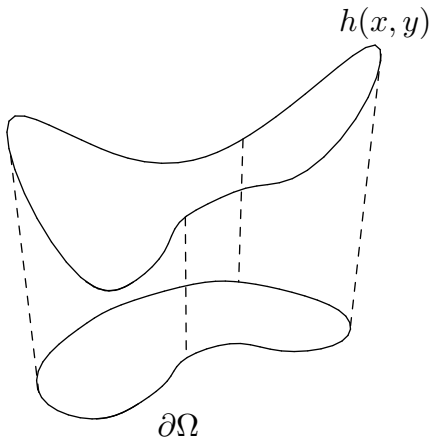


Figure 15.2. Dirichlet Boundary Conditions.

Laplace equation are the higher dimensional versions of the heat and wave equations, to be analyzed in Chapter 17.

Since both the Laplace and Poisson equations describe equilibrium configurations, they arise in applications in the context of boundary value problems. We seek a solution $u(x,y)$ to the partial differential equation defined on a fixed bounded, open domain[†] $(x,y) \in \Omega \subset \mathbb{R}^2$. The solution is required to satisfy suitable conditions on the boundary of the domain, denoted $\partial\Omega$, which will consist of one or more simple, closed curves, as illustrated in Figure 15.1. As in one-dimensional equilibria, there are three especially important types of boundary conditions.

The first are the *fixed* or *Dirichlet boundary conditions*, which specify the value of the function u on the boundary:

$$u(x,y) = h(x,y) \qquad \text{for} \qquad (x,y) \in \partial\Omega. \tag{15.4}$$

The Dirichlet conditions (15.4) serve to uniquely specify the solution $u(x,y)$ to the Laplace or the Poisson equation. Physically, in the case of a free or forced membrane, the Dirichlet boundary conditions correspond to gluing the edge of the membrane to a wire at height $h(x,y)$ over each boundary point $(x,y) \in \partial\Omega$, as illustrated in Figure 15.2. Uniqueness means that the shape of the boundary wire will unambiguously specify the vertical displacement of the membrane in equilibrium. Similarly, in the modeling of thermal equilibrium, a Dirichlet boundary condition represents the imposition of a prescribed temperature distribution, represented by the function h, along the boundary of the plate.

The second important class are the *Neumann boundary conditions*

$$\frac{\partial u}{\partial \mathbf{n}} = \nabla u \cdot \mathbf{n} = k(x,y) \qquad \text{on} \qquad \partial\Omega, \tag{15.5}$$

in which the normal derivative of the solution u on the boundary is prescribed. For example, in thermomechanics, a Neumann boundary condition specifies the heat flux into the

[†] See Appendix A for the precise definitions of the terms "domain", "bounded", "boundary", etc.

plate through its boundary. The "no-flux" or homogeneous Neumann boundary conditions, where $k(x, y) \equiv 0$, correspond to a fully insulated boundary. In the case of a membrane, homogeneous Neumann boundary conditions correspond to an unattached edge of the drum. In fluid mechanics, the no-flux conditions imply that the normal component of the velocity vector $\mathbf{v} = \nabla u$ vanishes on the boundary, and so no fluid is allowed to flow across the solid boundary.

Finally, one can mix the previous two sorts of boundary conditions, imposing Dirichlet conditions on part of the boundary, and Neumann on the complementary part. The general *mixed boundary value problem* has the form

$$-\Delta u = f \quad \text{in} \quad \Omega, \qquad u = h \quad \text{on} \quad D, \qquad \frac{\partial u}{\partial \mathbf{n}} = k \quad \text{on} \quad N, \tag{15.6}$$

with the boundary $\partial\Omega = D \cup N$ being the disjoint union of a "Dirichlet part", denoted by D, and a "Neumann part" N. For example, if u represents the equilibrium temperature in a plate, then the Dirichlet part of the boundary is where the temperature is fixed, while the Neumann part is insulated, or, more generally, has prescribed heat flux. Similarly, when modeling the displacement of a membrane, the Dirichlet part is where the edge of the drum is attached to a support, while the homogeneous Neumann part is where it is left hanging free.

Classification of Linear Partial Differential Equations in Two Variables

We have, at last, encountered all three of the fundamental linear, second order, partial differential equations for functions of two variables. The homogeneous versions of the trinity are

(*a*)	The wave equation:	$u_{tt} - c^2\, u_{xx} = 0,$	*hyperbolic*,
(*b*)	The heat equation:	$u_t - \gamma\, u_{xx} = 0,$	*parabolic*,
(*c*)	Laplace's equation:	$u_{xx} + u_{yy} = 0,$	*elliptic.*

The last column specifies the equations' *type*, in accordance with the standard taxonomy of partial differential equations. An explanation of the terminology will appear momentarily.

The wave, heat and Laplace equations are the prototypical representatives of the three fundamental genres of partial differential equations, each with its own intrinsic features and physical manifestations. Equations governing vibrations, such as the wave equation, are typically hyperbolic. Equations governing diffusion, such as the heat equation, are parabolic. Hyperbolic and parabolic equations both govern dynamical processes, and one of the variables is identified with the time. On the other hand, equations modeling equilibrium phenomena, including the Laplace and Poisson equations, are typically elliptic, and only involve spatial variables. Elliptic partial differential equations are associated with boundary value problems, whereas parabolic and hyperbolic equations require initial-boundary value problems, with, respectively, one or two required initial conditions. Furthermore, each type requires a fundamentally different kind of numerical solution algorithm.

While the initial tripartite classification first appears in partial differential equations in two variables, the terminology, underlying properties, and assocaited physical models

carry over to equations in higher dimensions. Most of the important partial differential equations arising in applications are of one of these three types, and it is fair to say that the field of partial differential equations breaks into three major, disjoint subfields. Or, rather four subfields, the last being all the equations, including higher order equations, that do not fit into this preliminary categorization.

The classification of linear, second order partial differential equations for a scalar-valued function $u(x, y)$ of two variables[†] proceeds as follows. The most general such equation has the form

$$L[u] = A\,u_{xx} + B\,u_{xy} + C\,u_{yy} + D\,u_x + E\,u_y + F\,u = f, \tag{15.7}$$

where the coefficients A, B, C, D, E, F are all allowed to be functions of (x, y), as is the inhomogeneity or forcing function $f = f(x, y)$. The equation is *homogeneous* if and only if $f \equiv 0$. We assume that at least one of the leading coefficients A, B, C is nonzero, as otherwise the equation degenerates to a first order equation.

The key quantity that determines the *type* of such a partial differential equation is its *discriminant*

$$\Delta = B^2 - 4\,AC. \tag{15.8}$$

This should (and for good reason) remind the reader of the discriminant of the quadratic equation

$$Q(\xi, \eta) = A\,\xi^2 + B\,\xi\,\eta + C\,\eta^2 + D\,\xi + E\,\eta + F = 0. \tag{15.9}$$

The solutions (ξ, η) describes a plane curve — namely, a conic section. In the nondegenerate cases, the discriminant determines its geometrical type; it is

- a hyperbola when $\Delta > 0$,
- a parabola when $\Delta = 0$, or
- an ellipse when $\Delta < 0$.

This tripartite classification provides the underlying motivation for the terminology used to classify second order partial differential equations.

Definition 15.1. At a point (x, y), the linear, second order partial differential equation (15.7) is called

(a) hyperbolic		$\Delta(x, y) > 0$,
(b) parabolic	if and only if	$\Delta(x, y) = 0$, but $A^2 + B^2 + C^2 \neq 0$,
(c) elliptic		$\Delta(x, y) < 0$,
(d) degenerate		$A = B = C = 0$.

In particular:

- The wave equation $u_{xx} - u_{yy} = 0$ has discriminant $\Delta = 4$, and is hyperbolic.
- The heat equation $u_{xx} - u_y = 0$ has discriminant $\Delta = 0$, and is parabolic.
- The Poisson equation $u_{xx} + u_{yy} = -f$ has discriminant $\Delta = -4$, and is elliptic.

[†] For dynamical equations, we will identify y as the time variable t.

Example 15.2. Since the coefficients in the partial differential equation are allowed to vary over the domain, the type of an equation may vary from point to point. Equations that change type are much less common, as well as being much harder to handle. One example arising in the theory of supersonic aerodynamics is the *Tricomi equation*

$$y\,u_{xx} - u_{yy} = 0. \tag{15.10}$$

Comparing with (15.7), we find that

$$A = y, \qquad C = -1, \qquad \text{and} \qquad B = D = E = F = f = 0.$$

The discriminant in this particular case is $\Delta = 4y$, and hence the equation is hyperbolic when $y > 0$, elliptic when $y < 0$, and parabolic on the transition line $y = 0$. The hyperbolic region corresponds to subsonic fluid flow, while the supersonic regions are of elliptic type. The transitional parabolic boundary represents the shock line between the sub- and supersonic regions.

Characteristics

In Section 14.5, we learned the importance of the characteristic lines in understanding the behavior of solutions to the wave equation. Characteristic curves play a similarly fundamental role in the study of more general linear hyperbolic partial differential equations. Indeed, characteristics are another means of distinguishing between the three classes of second order partial differential equations.

Definition 15.3. A smooth curve $\mathbf{x}(t) \subset \mathbb{R}^2$ is called a *characteristic curve* for the second order partial differential equation (15.7) if its tangent vector $\dot{\mathbf{x}} = (\dot{x}, \dot{y})^T$ satisfies the quadratic *characteristic equation*

$$C(x,y)\,\dot{x}^2 - B(x,y)\,\dot{x}\,\dot{y} + A(x,y)\,\dot{y}^2 = 0. \tag{15.11}$$

Pay careful attention to the form of the characteristic equation — the positions of A and C are the opposite of what you might expect, while a minus sign appears in front of B. Furthermore, the first and zero$^{\text{th}}$ order terms in the original partial differential equation play no role.

For example, consider the hyperbolic wave equation[†]

$$c^2\,u_{xx} + u_{yy} = 0.$$

In this case, $A = -c^2$, $B = 0$, $C = 1$, and so (15.11) takes the form

$$\dot{x}^2 - c^2\,\dot{y}^2 = 0, \qquad \text{which implies that} \qquad \dot{x} = \pm\, c\,\dot{y}.$$

All solutions to the latter ordinary differential equations are straight lines

$$x = \pm\, c\,y + k, \tag{15.12}$$

[†] *Warning*: We regard y as the "time" variable in the differential equation, rather than t, which assumes the role of the curve parameter.

where k is an integration constant. Therefore, the wave equation has two characteristic curves passing through each point (a, b), namely the straight lines (15.12) of slope $\pm 1/c$, in accordance with our earlier definition of characteristics. In general, a linear partial differential equation is hyperbolic at a point (x, y) if and only if there are two characteristic curves passing through it. Moreover, as with the wave equation, disturbances that are concentrated near the point will tend to propagate along the characteristic curves. This fact lies at the foundation of geometric optics. Light rays move along characteristic curves, and are thereby subject to the optical phenomena of refraction and focusing.

On the other hand, the elliptic Laplace equation

$$u_{xx} + u_{yy} = 0$$

has no (real) characteristic curves since the characteristic equation (15.11) reduces to

$$\dot{x}^2 + \dot{y}^2 = 0.$$

Elliptic equations have no characteristics, and as a consequence, do not admit propagating signals; the effect of a localized disturbance, say on a membrane, is immediately felt everywhere.

Finally, for the parabolic heat equation

$$u_{xx} - u_y = 0,$$

the characteristic equation is simply

$$\dot{y}^2 = 0,$$

and so there is only one characteristic curve through each point (a, b), namely the horizontal line $y = b$. Indeed, our observation that the effect of an initial concentrated heat source is immediately felt all along the bar is in accordance with propagation of localized disturbances along the characteristics.

In this manner, elliptic, parabolic, and hyperbolic partial differential equations are distinguished by the number of (real) characteristic curves passing through a point — namely, zero, one and two, respectively. Further discussion of characteristics and their applications to solving both linear and nonlinear partial differential equations can be found in Section 22.1.

15.2. Separation of Variables.

One of the oldest — and stillmost widely used — techniques for constructing explicit analytical solutions to partial differential equations is the method of *separation of variables*. We have, in fact, already used separation of variables to construct particular solutions to the heat and wave equations. In each case, we sought a solution in the form of a product, $u(t, x) = h(t)\, v(x)$, of scalar functions of each individual variable. For the heat and similar parabolic equations, $h(t)$ was an exponential, while the wave equation chose a trigonometric function. In more general situations, we might not know in advance which function $h(t)$ is appropriate. When the method succeeds (which is not guaranteed in advance), both factors are found as solutions to certain ordinary differential equations.

Turning to the Laplace equation, the solution depends on x and y, and so the multiplicative separation of variables ansatz has the form

$$u(x,y) = v(x)\, w(y). \tag{15.13}$$

Let us see whether such a function can solve the Laplace equation by direct substitution. First of all,

$$\frac{\partial^2 u}{\partial x^2} = v''(x)\, w(y), \qquad \frac{\partial^2 u}{\partial y^2} = v(x)\, w''(y),$$

where the primes indicate ordinary derivatives, and so

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = v''(x)\, w(y) + v(x)\, w''(y) = 0.$$

The method will succeed if we are able to separate the variables by placing all of the terms involving x on one side of the equation and all the terms involving y on the other. Here, we first write the preceding equation in the form

$$v''(x)\, w(y) = -\, v(x)\, w''(y).$$

Dividing both sides by $v(x)\, w(y)$ (which we assume is not identically zero as otherwise the solution would be trivial) yields

$$\frac{v''(x)}{v(x)} = -\,\frac{w''(y)}{w(y)}\,, \tag{15.14}$$

which effectively "separates" the x and y variables on each side of the equation. Now, how could a function of x alone be equal to a function of y alone? A moment's reflection should convince the reader that this can happen if and only if the two functions are constant[†], so

$$\frac{v''(x)}{v(x)} = -\,\frac{w''(y)}{w(y)} = \lambda,$$

where we use λ to indicate the common *separation constant*. Thus, the individual factors $v(x)$ and $w(y)$ satisfy ordinary differential equations

$$v'' - \lambda\, v = 0, \qquad w'' + \lambda\, w = 0,$$

as promised.

We already know how to solve both of these ordinary differential equations by elementary techniques. There are three different cases, depending on the sign of the separation constant λ, each leading to four different solutions to the Laplace equation. We collect the entire family of separable hbarmonic functions together in the following table.

[†] Technical detail: one should assume that the underlying domain be connected for this to be valid; however, in practical analysis, this technicality is irrelevant.

λ	$v(x)$	$w(y)$	$u(x,y) = v(x)\,w(y)$
$\lambda = -\omega^2 < 0$	$\cos \omega x,\ \sin \omega x$	$e^{-\omega y},\ e^{\omega y},$	$e^{\omega y}\cos \omega x,\quad e^{\omega y}\sin \omega x,$ $e^{-\omega y}\cos \omega x,\quad e^{-\omega y}\sin \omega x$
$\lambda = 0$	$1,\ x$	$1,\ y$	$1,\ x,\ y,\ x\,y$
$\lambda = \omega^2 > 0$	$e^{-\omega x},\ e^{\omega x}$	$\cos \omega y,\ \sin \omega y$	$e^{\omega x}\cos \omega y,\quad e^{\omega x}\sin \omega y,$ $e^{-\omega x}\cos \omega y,\quad e^{-\omega x}\sin \omega y$

Since Laplace's equation is a homogeneous linear system, any linear combination of solutions is also a solution. Thus, we can try to build general solutions as finite linear combinations, or, provided we pay proper attention to convergence issues, infinite series in the separable solutions. To solve boundary value problems, one must ensure that the resulting combination satisfies the boundary conditions. This is not easy, unless the underlying domain has a rather specific geomdetry.

In fact, the only domains for which we can explicitly solve boundary value problems using the separable solutions constructed above are rectangles. In this manner, we are led to consider boundary value problems for Laplace's equation

$$\Delta u = 0 \qquad \text{on a rectangle} \qquad R = \{\, 0 < x < a,\ \ 0 < y < b \,\}. \tag{15.15}$$

To be completely specific, we will focus on the following Dirichlet boundary conditions:

$$u(x,0) = f(x), \qquad u(x,b) = 0, \qquad u(0,y) = 0, \qquad u(a,y) = 0. \tag{15.16}$$

It will be important to only allow a nonzero boundary condition on one of the four sides of the rectangle. Once we know how to solve this type of problem, we can emplloy linear superposition to solve the general Dirichlet boundary value problem on a rectangle; see Exercise ■ for details. Other boundary conditions can be treated in a similar fashion — with the proviso that the condition on each side of the rectangle is either entirely Dirichlet or entirely Neumann.

We will ensure that the series solution we construct satisfies the three homogeneous boundary conditions by only using separable solutions that satisfy them. The remaining nonzero boundary condition will then specify the coefficients of the individual summands. The function $u(x,y) = v(x)\,w(y)$ will vanish on the top, right and left sides of the rectangle provided

$$v(0) = v(a) = 0, \qquad \text{and} \qquad w(b) = 0.$$

Referring to the preceding table, the first condition $v(0) = 0$ requires

$$v(x) = \begin{cases} \sin \omega x, & \lambda = \omega^2 > 0, \\ x, & \lambda = 0, \\ \sinh \omega x, & \lambda = -\omega^2 < 0, \end{cases}$$

where $\sinh z = \frac{1}{2}(e^z - e^{-z})$ is the usual hyperbolic sine function. However, the second and third cases cannot satisfy the second boundary condition $v(a) = 0$, and so we discard them. The first case leads to the condition

$$v(a) = \sin \omega\, a = 0, \qquad \text{and hence} \qquad \omega\, a = \pi, 2\pi, 3\pi, \dots$$

is an integral multiple of π. Therefore, the separation constant

$$\lambda = \omega^2 = \frac{n^2\, \pi^2}{a^2}\,, \qquad \text{where} \qquad n = 1, 2, 3, \dots\,, \tag{15.17}$$

and the corresponding functions are

$$v(x) = \sin \frac{n\, \pi\, x}{a}\,, \qquad\qquad n = 1, 2, 3, \dots\,. \tag{15.18}$$

Note: We have merely recomputed the known eigenvalues and eigenfunctions of the familiar boundary value problem $v'' + \lambda\, v = 0,\ \ v(0) = v(a) = 0$.

Since $\lambda = \omega^2 > 0$, the third boundary condition $w(b) = 0$ requires that, up to constant multiple,

$$w(y) = \sinh\, \omega\, (b - y) = \sinh \frac{n\, \pi\, (b - y)}{a}\,. \tag{15.19}$$

Therefore, each of the separable solutions

$$u_n(x, y) = \sin \frac{n\, \pi\, x}{a}\ \sinh \frac{n\, \pi\, (b - y)}{a}\,, \qquad\qquad n = 1, 2, 3, \dots\,, \tag{15.20}$$

satisfies the three homogeneous boundary conditions. It remains to analyze the inhomogeneous boundary condition along the bottom edge of the rectangle. To this end, let us try a linear superposition of the separable solutions in the form of an infinite series

$$u(x, y) = \sum_{n=1}^{\infty}\, c_n\, u_n(x, y) = \sum_{n=1}^{\infty}\, c_n\, \sin \frac{n\, \pi\, x}{a}\ \sinh \frac{n\, \pi\, (b - y)}{a}\,,$$

whose coefficients $c_1, c_2, \dots$ are to be prescribed by the remaining boundary condition. At the bottom edge, $y = 0$, we find

$$u(x, 0) = \sum_{n=1}^{\infty}\, c_n\, \sinh \frac{n\, \pi\, b}{a}\ \sin \frac{n\, \pi\, x}{a} = f(x), \qquad 0 \le x \le a, \tag{15.21}$$

which takes the form of a Fourier sine series for the function $f(x)$. According to (12.83), the coefficients b_n of the Fourier sine series

$$f(x) = \sum_{n=1}^{\infty}\, b_n\, \sin \frac{n\, \pi\, x}{a} \qquad \text{are given by} \qquad b_n = \frac{2}{a} \int_0^a f(x) \sin \frac{n\, \pi\, x}{a}\, dx. \tag{15.22}$$

Comparing (15.21, 22), we discover that

$$c_n \sinh \frac{n\, \pi b}{a} = b_n \qquad \text{or} \qquad c_n = \frac{b_n}{\sinh \dfrac{n\, \pi\, b}{a}} \ = \ \frac{2}{a \sinh \dfrac{n\, \pi\, b}{a}} \int_0^a f(x) \sin \frac{n\, \pi\, x}{a}\, dx.$$

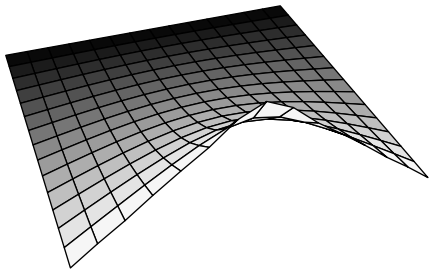


Figure 15.3. Square Membrane on a Wire.

Therefore, the solution to the boundary value problem takes the form of an infinite series

$$u(x,y) = \sum_{n=1}^{\infty} b_n \sin \frac{n\pi x}{a} \, \frac{\sinh \dfrac{n\pi(b-y)}{a}}{\sinh \dfrac{n\pi b}{a}}, \tag{15.23}$$

where b_n are the Fourier sine coefficients (15.22) of $f(x)$.

Does this series actually converge to the solution to the boundary value problem? Fourier analysis says that, under very mild conditions on the boundary function $f(x)$, the answer is "yes". Suppose that its Fourier coefficients are uniformly bounded,

$$| b_n | \le M \qquad \text{for all} \qquad n \ge 1, \tag{15.24}$$

which, according to (14.23) is true whenever $f(x)$ is piecewise continuous or, more generally, integrable: $\displaystyle\int_0^a | f(x) | \, dx < \infty$. Boundedness is also satisfied by many generalized functions, such as the delta function. In this case, as you are asked to prove in Exercise ■, the coefficients of the Fourier sine series (15.23)

$$B_n \;=\; \frac{\sinh \dfrac{n\pi(b-y)}{a}}{\sinh \dfrac{n\pi b}{a}} \, b_n \quad \longrightarrow \quad 0 \qquad \text{as} \qquad n \longrightarrow \infty \tag{15.25}$$

exponentially fast for all $0 < y \le b$. Thus, according to Section 12.3, the solution $u(x,y)$ is an infinitely differentiable function of x at each point in the rectangle, and can be well approximated by partial summation. The solution is also infinitely differentiable with respect to y; see Exercise ■. In fact, as we shall see, the solutions to the Laplace equation are *always analytic* functions inside their domain of definition — even when their boundary values are rather rough.

Example 15.4. A membrane is stretched over a wire in the shape of a unit square

with one side bent in half, as graphed in Figure 15.3. The precise boundary conditions are

$$u(x,y) = \begin{cases} x, & 0 \le x \le \frac{1}{2}, & y = 0, \\ 1 - x, & \frac{1}{2} \le x \le 1, & y = 0, \\ 0, & 0 \le x \le 1, & y = 1, \\ 0, & x = 0, & 0 \le y \le 1, \\ 0, & x = 1, & 0 \le y \le 1. \end{cases}$$

The Fourier sine series of the inhomogeneous boundary function is readily computed:

$$f(x) = \begin{cases} x, & 0 \le x \le \frac{1}{2}, \\ 1 - x, & \frac{1}{2} \le x \le 1, \end{cases}$$
$$= \frac{4}{\pi^2}\left(\sin \pi x - \frac{\sin 3\pi x}{9} + \frac{\sin 5\pi x}{25} - \cdots\right) = \frac{4}{\pi^2}\sum_{m=0}^{\infty} (-1)^m \frac{\sin(2m+1)\pi x}{(2m+1)^2}.$$

Specializing (15.23) when $a = b = 1$, we conclude that the solution to the boundary value problem is given by the Fourier series

$$u(x,y) = \frac{4}{\pi^2}\sum_{m=0}^{\infty} (-1)^m \frac{\sin(2m+1)\pi x \, \sinh(2m+1)\pi(1-y)}{(2m+1)^2 \sinh(2m+1)\pi}.$$

In Figure 15.3 we plot the sum of the first 10 terms in the series, which gives is a reasonably good approximation to the actual solutio n, except when we are very close to the raised corner of the boundary wire — which is the point of maximal displacement of the membrane.

Polar Coordinates

The method of separation of variables can be successfully exploited in certain other very special geometries. One particularly important case is a circular disk. To be specific, let us take the disk to have radius 1 and centered at the origin. Consider the Dirichlet boundary value problem

$$\Delta u = 0, \qquad x^2 + y^2 < 1, \qquad \text{and} \qquad u = h, \qquad x^2 + y^2 = 1, \tag{15.26}$$

so that the function $u(x,y)$ satisfies the Laplace equation on the unit disk and satisfies the specified Dirichlet boundary conditions on the unit circle. For example, $u(x,y)$ might represent the displacement of a circular drum that is attached to a wire of height

$$h(x,y) = h(\cos\theta, \sin\theta) \equiv h(\theta), \qquad 0 \le \theta \le 2\pi, \tag{15.27}$$

above each point $(x,y) = (\cos\theta, \sin\theta)$ on the unit circle.

The rectangular separable solutions are not particularly helpful in this situation. The fact that we are dealing with a circular geometry inspires us to adopt polar coordinates

$$x = r\cos\theta, \qquad y = r\sin\theta, \qquad \text{or} \qquad r = \sqrt{x^2 + y^2}, \qquad \theta = \tan^{-1}\frac{y}{x},$$

and write the solution $u(r,\theta)$ as a function thereof.

Warning: We will retain the same symbol, e.g., u, when rewriting a function in a different coordinate system. This is the convention of tensor analysis and differential geometry, [**2**], that treats the function or tensor as an intrinsic object, which is concretely realized through its formula in any chosen coordinate system. For instance, if $u(x,y) = x^2 + 2y$ in rectangular coordinates, then $u(r,\theta) = r^2\cos\theta + 2r\sin\theta$ — and *not* $r^2 + 2\theta$ — is its expression in polar coordinates. This convention avoids introducing new symbols when changing coordinates.

We also need to relate derivatives with respect to x and y to those with respect to r and θ. Performing a standard chain rule computation, we find

$$\begin{aligned} \frac{\partial}{\partial r} &= \cos\theta\,\frac{\partial}{\partial x} + \sin\theta\,\frac{\partial}{\partial y}\,, \\ \frac{\partial}{\partial \theta} &= -r\sin\theta\,\frac{\partial}{\partial x} + r\cos\theta\,\frac{\partial}{\partial y}\,, \end{aligned} \qquad \text{so} \qquad \begin{aligned} \frac{\partial}{\partial x} &= \cos\theta\,\frac{\partial}{\partial r} - \frac{\sin\theta}{r}\,\frac{\partial}{\partial \theta}\,, \\ \frac{\partial}{\partial y} &= \sin\theta\,\frac{\partial}{\partial r} + \frac{\cos\theta}{r}\,\frac{\partial}{\partial \theta}\,. \end{aligned} \tag{15.28}$$

These formulae allow us to rewrite the Laplace equation in polar coordinates; after some calculation in which many of the terms cancel, we find

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\,\frac{\partial u}{\partial r} + \frac{1}{r^2}\,\frac{\partial^2 u}{\partial \theta^2} = 0. \tag{15.29}$$

The boundary conditions are imposed on the unit circle $r = 1$, and so, by (15.27), take the form

$$u(1,\theta) = h(\theta). \tag{15.30}$$

Keep in mind that, in order to be single-valued functions of x, y, the solution $u(r,\theta)$ and its boundary values $h(\theta)$ must both be 2π periodic functions of the angular coordinate:

$$u(r,\theta + 2\pi) = u(r,\theta), \qquad\qquad h(\theta + 2\pi) = h(\theta). \tag{15.31}$$

Polar separation of variables is based on the ansatz

$$u(r,\theta) = v(r)\,w(\theta) \tag{15.32}$$

that assumes that the solution is a product of functions of the individual polar variables. Substituting (15.32) into the polar form (15.29) of Laplace's equation, we find

$$v''(r)\,w(\theta) + \frac{1}{r}\,v'(r)\,w(\theta) + \frac{1}{r^2}\,v(r)\,w''(\theta) = 0.$$

We now separate variables by moving all the terms involving r onto one side of the equation and all the terms involving θ onto the other. This is accomplished by first multiplying the equation by $r^2/v(r)\,w(\theta)$, and then moving the last term to the right hand side:

$$\frac{r^2\,v''(r) + r\,v'(r)}{v(r)} = -\,\frac{w''(\theta)}{w(\theta)} = \lambda.$$

As in the rectangular case, a function of r can equal a function of θ if and only if both are equal to a common separation constant, which we call λ. The partial differential equation thus splits into a pair of ordinary differential equations

$$r^2\, v'' + r\, v' - \lambda\, r = 0, \qquad\qquad w'' + \lambda\, w = 0, \tag{15.33}$$

that will prescribe the separable solution (15.32). Observe that both have the form of eigenfunction equations in which the separation constant λ plays the role of the eigenvalue, and we are only interested in nonzero solutions or eigenfunctions.

We have already solved the eigenvalue problem for $w(\theta)$. According to (15.31), $w(\theta + 2\,\pi) = w(\theta)$ must be a $2\,\pi$ periodic function. Therefore, according the discussion in Section 12.1, this periodic boundary value problem has the nonzero eigenfunctions

$$1, \qquad \sin n\,\theta, \qquad \cos n\,\theta, \qquad \text{for} \qquad n = 1, 2, \dots . \tag{15.34}$$

corresponding to the eigenvalues (separation constants) $\lambda = n^2$, where $n = 0, 1, 2, \dots$. Fixing the value of λ, the remaining ordinary differential equation

$$r^2\, v'' + r\, v' - n^2\, r = 0. \tag{15.35}$$

has the form of a second order Euler equation for the radial component $v(r)$. As discussed in Example 7.35, its solutions are obtained by substituting the power ansatz $v(r) = r^k$. We discover that this is a solution if and only if

$$k^2 - n^2 = 0, \qquad \text{and hence} \qquad k = \pm\, n.$$

Therefore, for $n \neq 0$, we find two linearly independent solutions,

$$v_1(r) = r^n, \qquad v_2(r) = r^{-n}, \qquad n = 1, 2, \dots . \tag{15.36}$$

If $n = 0$, there is an additional logarithmic solution

$$v_1(r) = 1, \qquad v_2(r) = \log r, \qquad n = 0. \tag{15.37}$$

Combining (15.34) and (15.36–37), we produce a complete list of separable polar coordinate solutions to the Laplace equation:

$$\begin{array}{lll} 1, & r^n \cos n\,\theta, & r^n \sin n\,\theta, \\ \log r, & r^{-n} \cos n\,\theta, & r^{-n} \sin n\,\theta, \end{array} \qquad n = 1, 2, 3, \dots . \tag{15.38}$$

Now, the solutions in the top row of (15.38) are continuous (in fact analytic) at the origin, whereas the solutions in the bottom row have singularities as $r \to 0$. The latter are not relevant since we require the solution u to remain bounded and smooth — even at the center of the disk. Thus, we should only use the former to concoct a candidate series solution

$$u(r, \theta) \;=\; \frac{a_0}{2} \;+\; \sum_{n=1}^{\infty} \big(a_n\, r^n \cos n\,\theta + b_n\, r^n \sin n\,\theta \big) \tag{15.39}$$

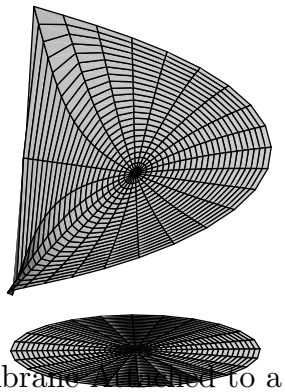


Figure 15.4. Membrane Attached to a Helical Wire.

to the Dirichlet boundary value problem. The coefficients a_n, b_n will be prescribed by the boundary conditions (15.30). Substituting $r = 1$, we find

$$u(1,\theta) \;=\; \frac{a_0}{2} \;+\; \sum_{n=1}^{\infty} \big(a_n \cos n\theta + b_n \sin n\theta \big) = h(\theta).$$

We recognize this as a standard Fourier series for the 2π periodic function $h(\theta)$. Therefore,

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} h(\theta)\cos n\theta\, d\theta, \qquad\qquad b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} h(\theta)\sin n\theta\, d\theta, \tag{15.40}$$

are precisely its Fourier coefficients, cf. (12.28).

Remark: Introducing the complex variable $z = r\,e^{\,\mathrm{i}\,\theta} = x + \mathrm{i}\,y$ allows us to write

$$z^n = r^n\, e^{\,\mathrm{i}\,n\theta} = r^n \cos n\theta + \mathrm{i}\, r^n \sin n\theta. \tag{15.41}$$

Therefore, the non-singular separable solutions are nothing but the harmonic polynomial solutions we first found in Example 7.52, namely

$$r^n \cos n\theta = \operatorname{Re} z^n, \qquad\qquad r^n \sin n\theta = \operatorname{Im} z^n. \tag{15.42}$$

Exploitation of the remarkable connections between the solutions to the Laplace equation and complex functions will form the focus of Chapter 16.

In view of (15.42), the n^{th} order term in the series solution (15.39),

$$a_n\, r^n \cos n\theta + b_n\, r^n \sin n\theta = a_n \operatorname{Re} z^n + b_n \operatorname{Im} z^n = \operatorname{Re}\big[\,(a_n - \mathrm{i}\, b_n)\, z^n\,\big],$$

is, in fact, a homogeneous polynomial in (x, y) of degree n. This means that, when written in rectangular coordinates x and y, (15.39) is, in fact, a *power series* for the function $u(x, y)$. Proposition C.4 implies that the power series is, in fact, the *Taylor series* for $u(x, y)$ based at the origin, and so its coefficients are multiples of the derivatives of u at $x = y = 0$. Details are worked out in Exercise ■. Thus, the fact that $u(x, y)$ has a convergent Taylor series implies that it is an analytic function at the origin. Indeed, as we will see, analyticity holds at any point of the domain of definintion of a harmonic function .

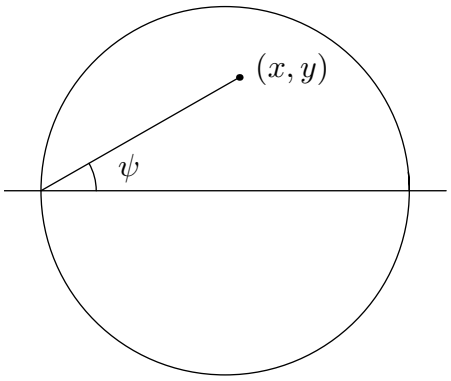


Figure 15.5. Geometrical Construction of the Solution.

Example 15.5. Consider the Dirichlet boundary value problem on the unit disk with

$$u(1,\theta) = \theta \qquad \text{for} \qquad -\pi < \theta < \pi. \tag{15.43}$$

The boundary data can be interpreted as a wire in the shape of a single turn of a spiral helix sitting over the unit circle, with a jump discontinuity, of magnitude 2π, at $(-1, 0)$. The required Fourier series

$$h(\theta) = \theta \;\sim\; 2\left(\sin\theta - \frac{\sin 2\theta}{2} + \frac{\sin 3\theta}{3} - \frac{\sin 4\theta}{4} + \cdots\right)$$

was computed in Example 12.2. Therefore, invoking our solution formula (15.39–40),

$$u(r,\theta) = 2\left(r\sin\theta - \frac{r^2\sin 2\theta}{2} + \frac{r^3\sin 3\theta}{3} - \frac{r^4\sin 4\theta}{4} + \cdots\right) \tag{15.44}$$

is the desired solution, and is plotted in Figure 15.4. In fact, this series can be explicitly summed. In view of (15.42),

$$u = 2\,\mathrm{Im}\left(z - \frac{z^2}{2} + \frac{z^3}{3} - \frac{z^4}{4} + \cdots\right) = 2\,\mathrm{Im}\;\log(1+z) = 2\,\mathrm{ph}\,(1+z) = 2\psi, \tag{15.45}$$

where

$$\psi = \tan^{-1}\frac{y}{1+x} \tag{15.46}$$

is the angle that the line passing through the two points (x, y) and $(-1, 0)$ makes with the x-axis, as sketched in Figure 15.5. You should try to convince yourself that, on the unit circle, $2\psi = \theta$ has the correct boundary values. Obserfve that, even though the boundary values are discontinuous, the solution is an analytic function inside the disk.

Unlike the rectangular series solution (15.23), the polar series solution (15.39) can, in fact, be summed in closed form! If we substitute the explicit Fourier formulae (15.40) into

(15.39) — remembering to change the integration variable to, say, ϕ to avoid a notational conflict — we find

$$\begin{aligned} u(r,\theta) &= \frac{a_0}{2} + \sum_{n=1}^{\infty} \big(a_n r^n \cos n\theta + b_n r^n \sin n\theta \big) \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} h(\phi)\, d\phi \\ &\qquad + \sum_{n=1}^{\infty} \left[\frac{r^n \cos n\theta}{\pi} \int_{-\pi}^{\pi} h(\phi) \cos n\phi \, d\phi + \frac{r^n \sin n\theta}{\pi} \int_{-\pi}^{\pi} h(\phi) \sin n\phi \, d\phi \right] \\ &= \frac{1}{\pi} \int_{-\pi}^{\pi} h(\phi) \left[\frac{1}{2} \ + \ \sum_{n=1}^{\infty} r^n \big(\cos n\theta \cos n\phi + \sin n\theta \sin n\phi \big) \right] d\phi \\ &= \frac{1}{\pi} \int_{-\pi}^{\pi} h(\phi) \left[\frac{1}{2} \ + \ \sum_{n=1}^{\infty} r^n \cos n(\theta - \phi) \right] d\phi. \end{aligned} \tag{15.47}$$

We next show how to sum the final series. Using (15.41), we can write it as the real part of a geometric series:

$$\begin{aligned} &\frac{1}{2} \ + \ \sum_{n=1}^{\infty} r^n \cos n\theta = \operatorname{Re} \left(\frac{1}{2} + \sum_{n=1}^{\infty} z^n \right) = \operatorname{Re} \left(\frac{1}{2} + \frac{z}{1-z} \right) = \operatorname{Re} \left(\frac{1+z}{2(1-z)} \right) \\ &= \operatorname{Re} \left(\frac{(1+z)(1-\overline{z})}{2\,|\,1-z\,|^2} \right) = \frac{\operatorname{Re}\,(1+z-\overline{z}-|\,z\,|^2)}{2\,|\,1-z\,|^2} = \frac{1-|\,z\,|^2}{2\,|\,1-z\,|^2} = \frac{1-r^2}{2\,(1+r^2-2r\cos\theta)}\,. \end{aligned}$$

Substituting back into (15.47) leads to the important *Poisson Integral Formula* for the solution to the boundary value problem.

Theorem 15.6. *The solution to the Laplace equation in the unit disk subject to Dirichlet boundary conditions $u(1,\theta) = h(\theta)$ is*

$$u(r,\theta) = \frac{1}{2\pi} \int_{-\pi}^{\pi} h(\phi)\, \frac{1-r^2}{1+r^2-2\,r\cos(\theta-\phi)}\, d\phi. \tag{15.48}$$

Example 15.7. A particularly important case is when the boundary value

$$h(\theta) = \delta(\theta - \phi)$$

is a delta function concentrated at the point $(\cos\phi, \sin\phi)$, $-\pi < \phi \le \pi$, on the unit circle. The solution to the resulting boundary value problem is the *Poisson integral kernel*

$$u(r,\theta) = \frac{1-r^2}{2\pi\big[\,1+r^2-2r\cos(\theta-\phi)\,\big]} = \frac{1-|\,z\,|^2}{2\pi\,|\,1-z\,e^{-\,\mathrm{i}\,\phi}\,|^2}\,. \tag{15.49}$$

The reader may enjoy verifying that this function does indeed, solve the Laplace equation and has the correct boundary values in the limit as $r \to 1$. Physically, if $u(r,\theta)$ represents the equilibrium temperature of the disk, then the delta function boundary data correspond

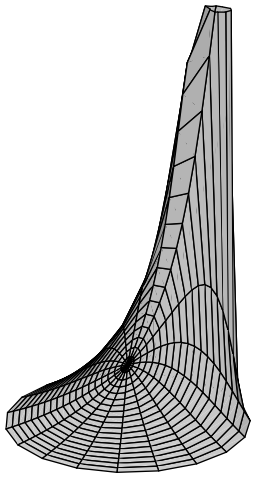


Figure 15.6. The Poisson Kernel.

to a concentrated unit heat source applied to a single point on the boundary. The resulting solution is sketched in Figure 15.6. OThus, the Poisson kernel plays the role of the fundamental solution for the boundary value problem. Indeed, Poisson integral formula (15.48) follows from our general superposition principle, writing the boundary data as a superposition of delta functions:

$$h(\theta) = \int_{-\pi}^{\pi} h(\phi)\, \delta(\phi - \theta)\, d\phi,$$

Averaging and the Maximum Principle

If we set $r = 0$ in the Poisson formula (15.48), then we obtain

$$u(0, \theta) = \frac{1}{2\,\pi} \int_{-\pi}^{\pi} h(\phi)\, d\phi. \tag{15.50}$$

The left hand side is the value of u at the origin — the center of the disk; the right hand side is the average of its boundary values around the unit circle. This is a particular instance of an important general fact.

Theorem 15.8. *Let $u(x, y)$ be harmonic inside a disk of radius a centered at a point (x_0, y_0) with piecewise continuous (or, more generally, integrable) boundary values on the circle $C = \{\, (x - x_0)^2 + (y - y_0)^2 = a^2 \,\}$. Then its value at the center of the disk is equal to the average of its values on the boundary circle*:

$$u(x_0, y_0) = \frac{1}{2\,\pi\, a} \oint_C u\, ds = \frac{1}{2\,\pi} \int_0^{2\,\pi} u(x_0 + a\cos\theta, y_0 + a\sin\theta)\, d\theta. \tag{15.51}$$

Proof: We use the scaling and translation symmetries of the Laplace equation to map the disk of radius r centered at (x_0, y_0) to the unit disk centered at the origin. Specifically, we set

$$U(x, y) = u(x_0 + a\,x, y_0 + a\,y). \tag{15.52}$$

An easy chain rule computation proves that $U(x,y)$ is harmonic on the unit disk, with boundary values

$$h(\theta) = U(\cos\theta, \sin\theta) = u(x_0 + a\cos\theta, y_0 + a\sin\theta).$$

Therefore, by (15.50) ,

$$U(0,0) = \frac{1}{2\,\pi}\int_{-\pi}^{\pi} h(\theta)\,d\theta = \frac{1}{2\,\pi}\int_{-\pi}^{\pi} U(\cos\theta,\sin\theta)\,d\theta.$$

Replacing U by its formula (15.52) produces the desired result. *Q.E.D.*

An important consequence of the integral formula (15.51) is the *Maximum Principle* for harmonic functions.

Theorem 15.9. *If u is a nonconstant harmonic function defined on a domain Ω, then u does not have a local maximum or local minimum at any interior point of Ω.*

Proof: The average of a continuous real function lies strictly between its maximum and minimum values — except in the trivial case when the function is constant. Since u is harmonic, it is continuous inside Ω. So Theorem 15.8 implies that the value of u at (x,y) lies strictly between its maximal and minimal values on any small circle centered at (x,y). This clearly excludes the possibility of u having a local maximum or minimum at (x,y). *Q.E.D.*

Thus, on a bounded domain, a harmonic function achieves its maximum and minimum values only at boundary points. Any interior critical point, where $\nabla u = \mathbf{0}$, must be a saddle point. Physically, if we interpret $u(x,y)$ as the vertical displacement of a membrane, then Theorem 15.9 says that, in the absence of external forcing, the membrane cannot have any internal bumps — its highest and lowest points are necessarily on the boundary of the domain. This reconfirms our physical intuition: the restoring force exerted by the stretched membrane will serve to flatten any bump, and hence a membrane with a local maximum or minimum cannot be in equilibrium. A similar interpretation holds for heat conduction. A body in thermal equilibrium can achieve its maximum and minimum temperature only on the boundary of the domain. Again, physically, heat energy would flow away from any internal maximum, or towards any local minimum, and so if the body contained a local maximum or minimum on its interior, it could not be in thermal equilibrium.

This concludes our discussion of separation of variables for the planar Laplace equation. The method works in a few other special coordinate systems. See Exercise ■ for one example, and [**128**, **131**, **133**] for a complete account, including connections with the underlying symmetries of the equation.

15.3. The Green's Function.

Now we turn to the Poisson equation (15.3), which is the inhomogeneous form of the Laplace equation. In Section 11.2, we learned how to solve one-dimensional inhomogeneous boundary value problems by contructing the associated Green's function. This important technique can be adapted to solve inhomogeneous boundary value problems for elliptic

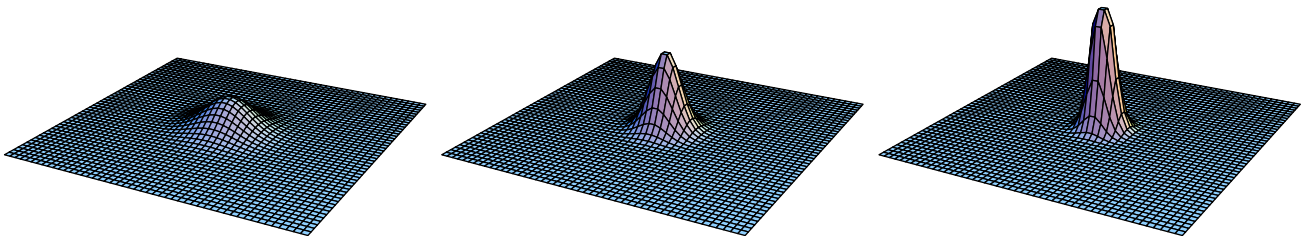


Figure 15.7. Gaussian Distributions Converging to the Delta Function.

partial differential equations in higher dimensions, including Poisson's equation. As before, the Green's function is characterized as the solution to the homogeneous boundary value problem in which the inhomogeneity is a concentrated unit impulse — a delta function. The solution to the general forced boundary value problem is then obtained via linear superposition, that is, as a convolution integral with the Green's function.

The first order of business is to establish the proper form for a unit impulse in our two-dimensional situation. We denote the *delta function* concentrated at position $\boldsymbol{\xi} = (\xi, \eta) \in \mathbb{R}^2$ by

$$\delta_{\boldsymbol{\xi}}(\mathbf{x}) = \delta_{(\xi,\eta)}(x, y) = \delta(\mathbf{x} - \boldsymbol{\xi}). \tag{15.53}$$

The delta function $\delta_{\mathbf{0}}(\mathbf{x}) = \delta(x, y)$ at the origin can be viewed as the limit, as $n \to \infty$, of a sequence of more and more highly concentrated functions $g_n(x, y)$, with

$$\lim_{n \to \infty} g_n(x, y) = 0, \quad \text{for} \quad (x, y) \neq (0, 0), \qquad \text{while} \qquad \iint_\Omega g_n(x, y)\, dx\, dy = 1.$$

A good example of a suitable sequence is provided by the *radial Gaussian distributions*

$$g_n(x, y) = \frac{n}{\pi}\, e^{-n\,(x^2+y^2)}\,, \tag{15.54}$$

which relies on the fact that

$$\iint_{\mathbb{R}^2} e^{-n\,(x^2+y^2)}\, dx\, dy = \frac{\pi}{n}\,,$$

established in Exercise ■. As plotted in Figure 15.7, as $n \to \infty$, the Gaussian profiles become more and more concentrated at the origin, while maintaining a unit volume underneath their graphs.

Alternatively, one can assign the delta function a dual interpretation as a linear functional on the vector space of continuous scalar-valued functions. We formally prescribe the delta function by the integral formula

$$\langle\, \delta_{(\xi,\eta)}\,, f\,\rangle = \iint_\Omega \delta_{(\xi,\eta)}(x, y) f(x, y)\, dx\, dy = \begin{cases} f(\xi, \eta), & (\xi, \eta) \in \Omega, \\ 0, & (\xi, \eta) \notin \overline{\Omega}, \end{cases} \tag{15.55}$$

which holds for any continuous function $f(x, y)$ and any domain $\Omega \subset \mathbb{R}^2$. As in the one-dimensional situation, we will avoid defining the integral when the delta function is concentrated at a boundary point, $(\xi, \eta) \in \partial\Omega$, of the integration domain.

Since double integrals can be evaluated as repeated one-dimensional integrals, we can conveniently view

$$\delta_{(\xi,\eta)}(x,y) = \delta_\xi(x)\,\delta_\eta(y) = \delta(x-\xi)\,\delta(y-\eta) \tag{15.56}$$

as the product of a pair of one-dimensional delta functions. Indeed, if

$$(\xi,\eta) \in R = \big\{\, a < x < b,\ c < y < d \,\big\} \subset \Omega$$

is contained in a rectangle inside the domain Ω, then

$$\begin{aligned}\iint_\Omega \delta_{(\xi,\eta)}(x,y) f(x,y)\,dx\,dy &= \iint_R \delta_{(\xi,\eta)}(x,y) f(x,y)\,dx\,dy \\ &= \int_a^b \int_a^b \delta(x-\xi)\,\delta(y-\eta)\,f(x,y)\,dy\,dx = \int_a^b \delta(x-\xi)\,f(x,\eta)\,dx = f(\xi,\eta).\end{aligned}$$

To find the Green's function, we must solve the equilibrium equation subject to a concentrated unit delta force at a prescribed point $\boldsymbol{\xi} = (\xi,\eta) \in \Omega$ inside the domain. In the case of Poisson's equation, the partial differential equation takes the form

$$-\Delta u = \delta_{\boldsymbol{\xi}}, \qquad \text{or} \qquad -\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = \delta(x-\xi)\,\delta(y-\eta), \qquad (x,y) \in \Omega, \tag{15.57}$$

and the solution is subject to homogeneous boundary conditions, either Dirichlet or mixed. (The nonuniqueness of solutions to the pure Neumann boundary value problem precludes the existence of a Green's function.) The resulting solution to the Poisson boundary value problem is denoted as

$$G(\mathbf{x};\boldsymbol{\xi}) = G(x,y;\xi,\eta), \tag{15.58}$$

and called the *Green's function*. Thus, the Green's function (15.58) measures the effect, at position $\mathbf{x} = (x,y)$, of a concentrated force applied at position $\boldsymbol{\xi} = (\xi,\eta)$.

Once we know the Green's function, the solution to the general Poisson boundary value problem

$$-\Delta u = f \qquad \text{in} \qquad \Omega, \qquad\qquad u = 0 \qquad \text{on} \qquad \partial\Omega \tag{15.59}$$

is reconstructed through a superposition principle. We regard the forcing function

$$f(x,y) = \iint_\Omega \delta(x-\xi)\,\delta(y-\eta) f(\xi,\eta)\,d\xi\,d\eta$$

as a superposition of delta impulses, whose strength at each point equals the value of f there. Linearity implies that the solution to the boundary value problem is the corresponding superposition of Green's function responses to each of the constituent impulses. The net result is the fundamental *superposition formula*

$$u(x,y) = \iint_\Omega G(x,y;\xi,\eta)\,f(\xi,\eta)\,d\xi\,d\eta \tag{15.60}$$

for the solution. This can be verified by direct evalution:

$$\begin{aligned} -\Delta u(x,y) &= \iint_\Omega \, [\, -\Delta G(x,y;\xi,\eta)\,]\, f(\xi,\eta)\, d\xi\, d\eta \\ &= \iint_\Omega \, \delta(x-\xi, y-\eta)\, f(\xi,\eta)\, d\xi\, d\eta = f(x,y), \end{aligned}$$

as claimed.

As in the one-dimensional situation, self-adjointness of the boundary value problem is manifested in the symmetry of the Green's function under interchange of its arguments:

$$G(\xi,\eta;x,y) = G(x,y;\xi,\eta). \tag{15.61}$$

The general proof of symmetry follows as in the one-dimensional version (11.90); see Exercise ∎. Symmetry has the following intriguing physical interpretation: Let $\mathbf{x}, \boldsymbol{\xi} \in \Omega$ be any pair of points in the domain. We apply a unit impulse to the membrane at the first point, and measure its deflection at the second; the result is exactly the same as if we apply the impulse at the second point, and measure the deflection at the first! (On the other hand, the deflections at other points in the domain will typically bear very little connection with each other.) Similarly, in electrostatics, the solution $u(x,y)$ is interpreted as the electrostatic potential for a system in equilibrium. A delta function corresponds to a point charge, e.g., an electron. The symmetry property says that the electrostatic potential at $\mathbf{x}$ due to a point charge placed at position $\boldsymbol{\xi}$ is exactly the same as the potential at $\boldsymbol{\xi}$ due to a point charge at $\mathbf{x}$. The reader may wish to meditate on the physical plausibility of these remarkable facts.

Unfortunately, most Green's functions — with a few notable exceptions — cannot be written down in closed form. However, their intrinsic form can be based on the following construction. Let us begin by considering the solution to the required Poisson equation

$$-\Delta u = \delta(x-\xi, y-\eta) \tag{15.62}$$

where $(\xi,\eta) \in \Omega$ is the point that the unit impulse force is being applied. As usual, the general solution to an inhomogeneous linear equation is a sum

$$u(x,y) = u_\star(x,y) + z(x,y) \tag{15.63}$$

of a particular solution $u_\star$ combined with the general solution z to the corresponding homogeneous equation, namely

$$-\Delta z = 0.$$

That is, $z(x,y)$ is an arbitrary harmonic function. We shall assume that the particular solution $u_\star(x,y)$ is due to the effect of the unit impulse, irrespective of any imposed boundary conditions. Once we have determined $u_\star$, we shall use the freedom inherent in the harmonic constituent $z(x,y)$ to ensure that the sum (15.63) satisfies the required boundary conditions.

One way to find a particular solution $u_\star$ is to appeal to physical intuition. First, since the delta function is concentrated at the point $\boldsymbol{\xi}$, the solution $u_\star$ must solve the homogeneous Laplace equation $\Delta u_\star = 0$ except at the point $\mathbf{x} = \boldsymbol{\xi}$, where we expect

it to have some sort of discontinuity. Second, since the Poisson equation is modeling a homogeneous, uniform medium (membrane, plate, gravitational potential in empty space, etc.), in the absence of boundary conditions, the effect of a unit impulse should only depend upon on the distance away from the source of the impulse. Therefore, we expect that the desired particular solution will depend only on the radial variable:

$$u_\star = u_\star(r), \qquad \text{where} \qquad r = \| \mathbf{x} - \boldsymbol{\xi} \| = \sqrt{(x-\xi)^2 + (y-\eta)^2}\,.$$

According to (15.37), the only radially symmetric solutions to the Laplace equation are

$$u(r) = a + b \log r, \tag{15.64}$$

where a and b are constants. The constant term a is smooth and harmonic everywhere, and so cannot contribute to a delta function singularity. Therefore, our only chance to produce a solution with such a singularity at the point $\boldsymbol{\xi}$ is to take a multiple of the logarithmic potential:

$$u_\star = b \, \log r.$$

We claim that, modulo the determination of b, this gives the correct formula, so

$$-\Delta u_\star = -b\,\Delta(\log r) = \delta(\mathbf{x} - \boldsymbol{\xi}), \qquad r = \| \mathbf{x} - \boldsymbol{\xi} \|. \tag{15.65}$$

is the delta function for an appropriate constant b.

To justify this claim, and so determine the proper value of b, we first note that, by construction, $\log r$ solves the Laplace equation everywhere except at $r = 0$, i.e., at $\mathbf{x} = \boldsymbol{\xi}$:

$$\Delta \log r = 0, \qquad r \neq 0. \tag{15.66}$$

Secondly, if $D_a = \big\{\, 0 \leq r \leq a \,\big\} = \big\{\, \| \mathbf{x} - \boldsymbol{\xi} \| \leq a \,\big\}$ is any disk centered at $\boldsymbol{\xi}$, then, by the divergence form (A.58) of Green's Theorem,

$$\begin{aligned} \iint_{D_a} \Delta(\log r)\, dx\, dy &= \iint_{D_a} \nabla \cdot \nabla(\log r)\, dx\, dy \\ &= \oint_{C_a} \frac{\partial(\log r)}{\partial \mathbf{n}}\, ds = \oint_{C_a} \frac{\partial(\log r)}{\partial r}\, ds = \oint_{C_a} \frac{1}{r}\, ds = \int_{-\pi}^{\pi} d\theta = 2\,\pi, \end{aligned}$$

where $C_a = \partial D_a = \big\{\, \| \mathbf{x} - \boldsymbol{\xi} \| = a \,\big\}$ is the boundary of the disk, i.e., the circle of radius a centered at $\boldsymbol{\xi}$. (The identification $\partial/\partial \mathbf{n} = \partial/\partial r$ on a circle can be found in Exercise ■.) Thus, if Ω is any domain, then

$$\iint_{\Omega} \Delta(\log r)\, dx\, dy = \begin{cases} 2\,\pi, & \boldsymbol{\xi} \in \Omega, \\ 0, & \boldsymbol{\xi} \notin \overline{\Omega}. \end{cases} \tag{15.67}$$

In the first case, when $\boldsymbol{\xi} \in \Omega$, (15.66) allows us replace the integral over Ω by an integral over a small disk centered at $\boldsymbol{\xi}$, and then apply the preceding identity; in the second case, (15.68) implies that the integrand vanishes on all of the domain, and so the integral is 0. Equations (15.66–67) are the defining properties for $2\,\pi$ times the delta function, so

$$\Delta(\log r) = 2\,\pi\,\delta(\mathbf{x} - \boldsymbol{\xi}). \tag{15.68}$$

Comparing (15.68) with (15.65), we conclude that

$$u_\star(x,y) = -\frac{1}{2\pi}\,\log r = -\frac{1}{2\pi}\log\|\,\mathbf{x}-\boldsymbol{\xi}\,\| = -\frac{1}{4\pi}\log\big[\,(x-\xi)^2+(y-\eta)^2\,\big] \qquad (15.69)$$

is a particular solution to the Poisson equation (15.62) with a unit impulse force.

The *logarithmic potential* (15.69) represents the gravitational potential in empty two-dimensional space due to a unit point mass at position $\boldsymbol{\xi}$, or, equivalently, the two-dimensional electrostatic potential due to a point charge at $\boldsymbol{\xi}$. The corresponding gravitational (electrostatic) force field is obtained by taking its gradient:

$$\mathbf{F} = \nabla\left(-\frac{1}{2\pi}\log\|\,\mathbf{x}-\boldsymbol{\xi}\,\|\right) = -\frac{\mathbf{x}-\boldsymbol{\xi}}{2\pi\,\|\,\mathbf{x}-\boldsymbol{\xi}\,\|^2}\,.$$

Note that $\|\,\mathbf{F}\,\| = 1/(2\pi\,\|\,\mathbf{x}-\boldsymbol{\xi}\,\|)$ is proportional to the inverse distance, which is the two-dimensional form of Newton's (Coulomb's) three-dimensional inverse square law. The gravitational potential due to a mass, e.g., a plate, in the shape of a domain $\Omega \subset \mathbb{R}^2$ can be obtained by superimposing delta function sources with strengths equalo to the density of the material at each point. The result is the potential function

$$u(x,y) = -\frac{1}{4\pi}\iint_\Omega \rho(\xi,\eta)\,\log\big[\,(x-\xi)^2+(y-\eta)^2\,\big]\,d\xi\,d\eta, \qquad (15.70)$$

in which $\rho(\xi,\eta)$ denotes the density of the body at position (ξ,η). For example, the gravitational potential due to the unit disk $D = \{\,x^2+y^2 \le 1\,\}$ with unit density $\rho \equiv 1$ is

$$u(x,y) = -\frac{1}{4\pi}\iint_D \log\big[\,(x-\xi)^2+(y-\eta)^2\,\big]\,d\xi\,d\eta.$$

Returning to our boundary value problem, the general solution to the Poisson equation (15.62) can, therefore, be written in the form

$$u(x,y) = -\frac{1}{2\pi}\,\log\|\,\mathbf{x}-\boldsymbol{\xi}\,\| + z(x,y), \qquad (15.71)$$

where $z(x,y)$ is an arbitrary harmonic function. To construct the Green's function for a prescribed domain, we need to choose the harmonic function $z(x,y)$ so that (15.71) satisfies the relevant homogeneous boundary conditions. Let us state this result for the Dirichlet problem.

Proposition 15.10. *The Green's function for the Dirichlet boundary value problem*

$$-\Delta u = f \qquad \text{on} \qquad \Omega, \qquad\qquad u = 0 \qquad \text{on} \qquad \partial\Omega,$$

has the form

$$G(x,y;\xi,\eta) = -\frac{1}{4\pi}\,\log\big[\,(x-\xi)^2+(y-\eta)^2\,\big] + z(x,y) \qquad (15.72)$$

where $z(x,y)$ is the harmonic function that has the same boundary values as the logarithmic potential function:

$$\Delta z = 0 \quad \text{on} \quad \Omega, \qquad\qquad z(x,y) = \frac{1}{4\pi}\,\log\big[\,(x-\xi)^2+(y-\eta)^2\,\big] \quad \text{for} \quad (x,y)\in\partial\Omega.$$

Let us conclude this subsection by summarizing the key properties of the Green's function $G(\mathbf{x}, \boldsymbol{\xi})$ for the two-dimensional Poisson equation., which

(a) Solves Laplace's equation, $\Delta G = 0$, for all $\mathbf{x} \neq \boldsymbol{\xi}$.

(b) Has a logarithmic singularity[†] at $\mathbf{x} = \boldsymbol{\xi}$.

(c) Satisfies the relevant homogeneous boundary conditions.

(d) Is symmetric: $G(\boldsymbol{\xi}, \mathbf{x}) = G(\mathbf{x}, \boldsymbol{\xi})$.

(e) Establishes the superposition formula (15.60) for a general forcing function.

The Method of Images

The preceding analysis exposes the underlying form of the Green's function, but we are still left with the determination of the harmonic component $z(x, y)$ required to match the logarithmic potential boundary values. There are three principal analytical techniques employed to produce explicit formulas. The first is an adaptation of the method of separation of variables, and leads to infinite series expressions, similar to those of the fundamental solution for the heat equation derived in Chapter 14. We will not dwell on this approach here, although a couple of the exercises ask the reader to fill in the details. The second is the *method of images* and will be developed in this section. The most powerful is based on the theory of conformal mappings, but must be deferred until we have learned the basics of complex analysis; the details can be found in Section 16.3. While the first two methods only apply to a fairly limited class of domains, they do adapt straightforwardly to higher dimensional problems, as well as certain other types of elliptic partial differential equations, whereas the method of conformal mapping is, unfortunately, restricted to two-dimensional problems involving the Laplace and Poisson equations.

We already know that the singular part of the Green's function for the two-dimensional Poisson equation is provided by a logarithmic potential. The problem, then, is to construct the harmonic part, called $z(x, y)$ in (15.72), so that the sum has the correct homogeneous boundary values, or, equivalently, that $z(x, y)$ has the same boundary values as the logarithmic potential. In certain cases, $z(x, y)$ can be thought of as the potential induced by one or more hypothetical electric charges (or, equivalently, gravitational point masses) that are located *outside* the domain Ω, arranged in such a manner that their combined electrostatic potential happens to coincide with the logarithmic potential on the boundary of the domain. The goal, then, is to place the image charges of suitable strength in the proper positions.

Here, we will only consider the case of a single image charge, located at a position $\boldsymbol{\eta} \notin \Omega$. We scale the logarithmic potential (15.69) by the charge strength, and, for added flexibility, include an additional constant — the charge's potential baseline:

$$z(x, y) = a \log \| \mathbf{x} - \boldsymbol{\eta} \| + b, \qquad \eta \in \mathbb{R}^2 \setminus \overline{\Omega}.$$

This function is harmonic inside Ω since the logarithmic potential is harmonic everywhere except at the singularity $\boldsymbol{\eta}$, which is assumed to lies outside the domain. For the Dirichlet

[†] Note that this is in contrast to the one-dimensional situation, where the Green's function is continuous at the impulse point.

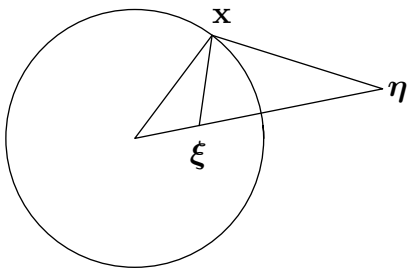


Figure 15.8. Method of Images for the Unit Disk.

boundary value problem, then, for each point $\boldsymbol{\xi} \in \Omega$, we must find a corresponding image point $\boldsymbol{\eta} \in \mathbb{R}^2 \setminus \overline{\Omega}$ and constants $a, b \in \mathbb{R}$, such that[‡]

$$\log \| \mathbf{x} - \boldsymbol{\xi} \| = a \log \| \mathbf{x} - \boldsymbol{\eta} \| + b \qquad \text{for all} \qquad \mathbf{x} \in \partial\Omega,$$

or, equivalently,

$$\| \mathbf{x} - \boldsymbol{\xi} \| = \lambda \| \mathbf{x} - \boldsymbol{\eta} \|^a \qquad \text{for all} \qquad \mathbf{x} \in \partial\Omega, \tag{15.73}$$

where $\lambda = \log b$. For each fixed $\boldsymbol{\xi}, \boldsymbol{\eta}, \lambda, a$, the equation in (15.73) will, typically, implicitly prescribe a plane curve, but it is not clear that one can always arrange that these curves all coincide with the boundary of our domain.

In order to make further progress, we appeal to a geometrical construction based upon similar triangles. We select $\boldsymbol{\eta} = c\,\boldsymbol{\xi}$ to be a point lying on the ray through $\boldsymbol{\xi}$. Its location is fixed so that the triangle with vertices $\mathbf{0}, \mathbf{x}, \boldsymbol{\eta}$ is similar to the triangle with vertices $\mathbf{0}, \boldsymbol{\xi}, \mathbf{x}$, noting that they have the same angle at the common vertex $\mathbf{0}$ — see Figure 15.8. Similarity requires that the triangles' corresponding sides have a common ratio, and so

$$\frac{\| \boldsymbol{\xi} \|}{\| \mathbf{x} \|} = \frac{\| \mathbf{x} \|}{\| \boldsymbol{\eta} \|} = \frac{\| \mathbf{x} - \boldsymbol{\xi} \|}{\| \mathbf{x} - \boldsymbol{\eta} \|} = \lambda. \tag{15.74}$$

The last equality implies that (15.73) holds with $a = 1$. Consequently, if we choose

$$\| \boldsymbol{\eta} \| = \frac{1}{\| \boldsymbol{\xi} \|}, \qquad \text{so that} \qquad \boldsymbol{\eta} = \frac{\boldsymbol{\xi}}{\| \boldsymbol{\xi} \|^2}, \tag{15.75}$$

then

$$\| \mathbf{x} \|^2 = \| \boldsymbol{\xi} \| \, \| \boldsymbol{\eta} \| = 1.$$

Thus $\mathbf{x}$ lies on the unit circle, and, as a result, $\lambda = \| \boldsymbol{\xi} \|$. The map taking a point $\boldsymbol{\xi}$ inside the disk to its image point $\boldsymbol{\eta}$ defined by (15.75) is known as *inversion* with respect to the unit circle.

[‡] To simplify the formulas, we have omitted the $1/(2\pi)$ factor, which can easily be reinstated at the end of the analysis.

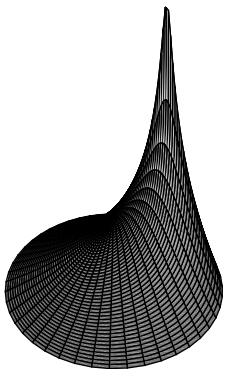


Figure 15.9. Green's Function for the Unit Disk.

We have now demonstrated that the functions

$$\frac{1}{2\pi}\,\log\|\,\mathbf{x}-\boldsymbol{\xi}\,\| = \frac{1}{2\pi}\,\log\big(\,\|\,\boldsymbol{\xi}\,\|\;\|\,\mathbf{x}-\boldsymbol{\eta}\,\|\,\big) = \frac{1}{2\pi}\,\log\frac{\|\,\|\,\boldsymbol{\xi}\,\|^2\,\mathbf{x}-\boldsymbol{\xi}\,\|}{\|\,\boldsymbol{\xi}\,\|} \quad \text{when} \quad \|\,\mathbf{x}\,\| = 1, \tag{15.76}$$

has the same boundary values on the unit circle. Consequently, their difference

$$G(\mathbf{x};\boldsymbol{\xi}) = -\,\frac{1}{2\pi}\log\|\,\mathbf{x}-\boldsymbol{\xi}\,\| + \frac{1}{2\pi}\log\frac{\|\,\|\,\boldsymbol{\xi}\,\|^2\,\mathbf{x}-\boldsymbol{\xi}\,\|}{\|\,\boldsymbol{\xi}\,\|} = \frac{1}{2\pi}\log\frac{\|\,\|\,\boldsymbol{\xi}\,\|^2\,\mathbf{x}-\boldsymbol{\xi}\,\|}{\|\,\boldsymbol{\xi}\,\|\;\|\,\mathbf{x}-\boldsymbol{\xi}\,\|} \tag{15.77}$$

has the required properties for the Green's function for the Dirichlet problem on the unit disk. In terms of polar coordinates

$$\mathbf{x} = (r\cos\theta, r\sin\theta), \qquad \boldsymbol{\xi} = (\rho\cos\varphi, \rho\sin\varphi),$$

applying the Law of Cosines to the triangles in Figure 15.8 leads to the explicit formula

$$G(r,\theta;\rho,\varphi) = \frac{1}{4\pi}\;\log\left(\,\frac{1+r^2\rho^2-2\,r\,\rho\cos(\theta-\varphi)}{r^2+\rho^2-2\,r\,\rho\cos(\theta-\varphi)}\,\right). \tag{15.78}$$

In Figure 15.9 we sketch the Green's function corresponding to a unit impulse being applied at a point half way between the center and the edge of the disk.

Remark: Unlike one-dimensional boundary value problems, the Green's function (15.78) has a singularity and is not continuous at the impulse point $\mathbf{x} = \boldsymbol{\xi}$.

Applying the general superposition rule (15.60), we arrive at a solution to the Dirichlet boundary value problem for the Poisson equation in the unit disk.

Theorem 15.11. *The solution to the homogeneous Dirichlet boundary value problem*

$$-\,\Delta u = f, \quad \textit{for} \quad r = \|\,\mathbf{x}\,\| < 1, \qquad\qquad u = 0, \quad \textit{for} \quad r = 1,$$

is, when expressed in polar coordinates,

$$u(r,\theta) = \frac{1}{4\pi}\;\int_0^{2\pi}\int_0^1 f(\rho,\varphi)\;\log\left(\,\frac{1+r^2\rho^2-2\,r\,\rho\cos(\theta-\varphi)}{r^2+\rho^2-2\,r\,\rho\cos(\theta-\varphi)}\,\right)\,\rho\,d\rho\,d\varphi. \tag{15.79}$$

The Green's function was originally designed for the homogeneous boundary value problem. Interestingly, it can also be used to handle inhomogeneous boundary conditions.

Theorem 15.12. *Let $G(\mathbf{x};\boldsymbol{\xi})$ denote the Green's function for the homogeneous Dirichlet boundary value problem for the Poisson equation on a domain $\Omega \subset \mathbb{R}^2$. Then the solution to the inhomogeneous Dirichlet problem*

$$-\Delta u = f, \qquad \mathbf{x} \in \Omega, \qquad\qquad u = h, \qquad \mathbf{x} \in \partial\Omega, \tag{15.80}$$

is given by

$$u(\mathbf{x}) = \iint_\Omega G(\mathbf{x};\boldsymbol{\xi})\, f(\boldsymbol{\xi})\, d\xi\, d\eta - \oint_{\partial\Omega} \frac{\partial G(\mathbf{x};\boldsymbol{\xi})}{\partial \mathbf{n}}\, h(\boldsymbol{\xi})\, ds. \tag{15.81}$$

For example, applying (15.81) to the Green's function (15.78) for the unit disk with $f \equiv 0$ recovers the Poisson integral formula (15.48).

Proof: Let $\psi(\mathbf{x})$ be any function such that

$$\psi = h \qquad \text{for} \qquad \mathbf{x} \in \partial\Omega.$$

Set $v = u - \psi$, so that v satisfies the homogeneous boundary value problem

$$-\Delta v = f + \Delta\psi \qquad \text{in} \qquad \Omega, \qquad\qquad v = 0 \qquad \text{on} \qquad \partial\Omega.$$

We can therefore express

$$v(\mathbf{x}) = \iint_\Omega G(\mathbf{x};\boldsymbol{\xi}) \big[\, f(\boldsymbol{\xi}) + \Delta\psi(\boldsymbol{\xi}) \,\big]\, d\xi\, d\eta.$$

Integration by parts, based on the second formula in Exercise ■, can be used to simplify the integral:

$$\begin{aligned} \iint_\Omega G(\mathbf{x};\boldsymbol{\xi})\, \Delta\psi(\boldsymbol{\xi})\, d\xi\, d\eta = \iint_\Omega \Delta G(\mathbf{x};\boldsymbol{\xi})\, \psi(\boldsymbol{\xi})\, d\xi\, d\eta + \\ + \oint_{\partial\Omega} \left(G(\mathbf{x};\boldsymbol{\xi})\, \frac{\partial\psi(\boldsymbol{\xi})}{\partial \mathbf{n}} - \frac{\partial G(\mathbf{x};\boldsymbol{\xi})}{\partial \mathbf{n}}\, \psi(\boldsymbol{\xi}) \right) ds. \end{aligned}$$

Since the Green's function solves $-\Delta G = \delta_{\boldsymbol{\xi}}$, the first term reproduces $-\psi(\mathbf{x})$. Moreover, $G = 0$ and $\psi = h$ on $\partial\Omega$, and so the right hand side reduces to (15.81). *Q.E.D.*

15.4. Adjoints and Minimum Principles.

In this section, we explain how the Laplace and Poisson equations fit into our universal self-adjoint equilibrium framework. The most important outcome will be to establish a very famous minimization principle characterizing the equilibrium solution, that we will exploit in the design of the finite element numerical solution method.

The one-dimensional version of the Poisson equation,

$$-\frac{d^2u}{dx^2} = f,$$

is the equilibrium equation for a uniform elastic bar. In Section 11.3, we wrote the underlying boundary value problems in self-adjoint form

$$K[u] = D^* \circ D[u] = f$$

based on the product of the derivative operator $Du = u'$ and its adjoint $D^* = -D$ with respect to the standard L^2 inner product.

For the two-dimensional Poisson equation

$$-\Delta[u] = -\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = f(x,y)$$

the role of the one-dimensional derivative D will be played by the *gradient* operator

$$\nabla u = \operatorname{grad} u = \begin{pmatrix} u_x \\ u_y \end{pmatrix}.$$

The gradient ∇ defines a linear map that takes a scalar-valued function $u(x,y)$ to the vector-valued function consisting of its two first order partial derivatives. Thus, its domain is the vector space $U = \mathrm{C}^1(\Omega, \mathbb{R})$ consisting of all continuously differentiable functions $u(x,y)$ defined for $(x,y) \in \Omega$. The target space $V = \mathrm{C}^0(\Omega, \mathbb{R}^2)$ consists of all continuous vector-valued functions $\mathbf{v}(x,y) = (\, v_1(x,y), v_2(x,y)\,)^T$, also known as *vector fields*. (By way of analogy, scalar-valued functions are sometimes referred to as *scalar fields*.) Indeed, if $u_1, u_2 \in U$ are any two scalar functions and $c_1, c_2 \in \mathbb{R}$ any constants, then

$$\nabla(c_1\, u_1 + c_2\, u_2) = c_1\, \nabla u_1 + c_2\, \nabla u_2,$$

which is the requirement for linearity as stated in Definition 7.1.

In accordance with the general Definition 7.53, the adjoint of the gradient must go in the reverse direction,

$$\nabla^* \colon V \longrightarrow U,$$

mapping vector fields $\mathbf{v}(x,y)$ to scalar functions $z(x,y) = \nabla^* \mathbf{v}$. The defining equation for the adjoint

$$\langle\!\langle\, \nabla u \,,\, \mathbf{v} \,\rangle\!\rangle = \langle\, u \,,\, \nabla^* \mathbf{v} \,\rangle \tag{15.82}$$

depends on the choice of inner products on the two vector spaces. The simplest inner product between real-valued scalar functions $u(x,y), z(x,y)$ defined on a domain $\Omega \subset \mathbb{R}^2$ is given by the double integral

$$\langle\, u \,,\, v \,\rangle = \iint_\Omega u(x,y)\, z(x,y)\, dx\, dy. \tag{15.83}$$

As in the one-dimensional case (3.12), this is often referred to as the L^2 *inner product* between scalar fields, with associated norm

$$\| u \|^2 = \langle\, u \,,\, u \,\rangle = \iint_\Omega u(x,y)^2\, dx\, dy.$$

Similarly, the L^2 inner product between vector-valued functions (vector fields) defined on Ω is obtained by integrating their usual dot product:

$$\langle\!\langle\, \mathbf{v}\,, \mathbf{w}\,\rangle\!\rangle = \iint_\Omega \mathbf{v}(x,y)\cdot\mathbf{w}(x,y)\,dx\,dy = \iint_\Omega \big[\, v_1(x,y)\,w_1(x,y) + v_2(x,y)\,w_2(x,y)\,\big]\,dx\,dy. \tag{15.84}$$

These form the two most basic inner products on the spaces of scalar and vector fields, and are the ones required to place the Laplace and Poisson equations in self-adjoint form.

The adjoint identity (15.82) is supposed to hold for all appropriate scalar fields u and vector fields $\mathbf{v}$. For the L^2 inner products (15.83, 84), the two sides of the identity read

$$\langle\!\langle\, \nabla u\,, \mathbf{v}\,\rangle\!\rangle = \iint_\Omega \nabla u\cdot\mathbf{v}\,dx\,dy = \iint_\Omega \left(\frac{\partial u}{\partial x}\,v_1 + \frac{\partial u}{\partial y}\,v_2\right) dx\,dy,$$
$$\langle\, u\,, \nabla^*\mathbf{v}\,\rangle = \iint_\Omega u\,\nabla^*\mathbf{v}\,dx\,dy.$$

Thus, to equate these two double integrals, we must soomehow remove the derivatives from the scalar field u. As in the one-dimensional computation (11.73), the secret is integration by parts.

For single integrals, the integration by parts formula is found by applying the Fundamental Theorem of Calculus to Leibniz's rule for the derivative of the product of two functions. According to Appendix A, Green's Theorem A.26 plays the role of the Fundamental Theorem when dealing with double integrals. We will find the divergence form

$$\iint_\Omega \nabla\cdot\mathbf{v}\,dx\,dy = \oint_{\partial\Omega} \mathbf{v}\cdot\mathbf{n}\,ds, \tag{15.85}$$

as in (A.58), the more convenient for the present purposes. Proceeding in analogy with the one-dimensional argument, we replace the vector field $\mathbf{v}$ by the product $u\,\mathbf{v}$ of a scalar field u and a vector field $\mathbf{v}$. An elementary computation proves that

$$\nabla\cdot(u\,\mathbf{v}) = u\,\nabla\cdot\mathbf{v} + \nabla u\cdot\mathbf{v}. \tag{15.86}$$

As a result, we deduce what is usually known as *Green's formula*

$$\iint_\Omega \big[\, u\,\nabla\cdot\mathbf{v} + \nabla u\cdot\mathbf{v}\,\big]\,dx\,dy = \oint_{\partial\Omega} u\,(\mathbf{v}\cdot\mathbf{n})\,ds, \tag{15.87}$$

which is valid for arbitrary bounded domains Ω, and arbitrary scalar and vector fields defined thereon. Rearranging the terms in this integral identity produces the required *integration by parts* formula for double integrals:

$$\iint_\Omega \nabla u\cdot\mathbf{v}\,dx\,dy = \oint_{\partial\Omega} u\,(\mathbf{v}\cdot\mathbf{n})\,ds - \iint_\Omega u\,\nabla\cdot\mathbf{v}\,dx\,dy. \tag{15.88}$$

The terms in this identity have direct counterparts in our one-dimensional integration by parts formula (11.76). The first term on the right hand side of this identity is a boundary term, just like the first terms on the right hand side of the one-dimensional formula (11.76). Moreover, the derivative operation has moved from a gradient on the scalar field in the

doubl.e integral on the left to a divergence on the vector field in the double integral on the right — even the minus sign is there!

Now, the left hand side in the integration by parts formula (15.88) is the same as the left hand side of (15.82). If the boundary integral vanishes,

$$\oint_{\partial\Omega} u\,\mathbf{v}\cdot\mathbf{n}\,ds = 0, \tag{15.89}$$

then the right hand side of formula (15.88) also reduces to an L^2 inner product

$$-\iint_\Omega u\,\nabla\cdot\mathbf{v}\,dx\,dy = \iint_\Omega u\,(-\nabla\cdot\mathbf{v})\,dx\,dy = \langle\, u\,, -\nabla\cdot\mathbf{v}\,\rangle$$

between the scalar field u and minus the divergence of the vector field $\mathbf{v}$. Therefore, subject to the boundary constraint (15.89), the integration by parts formula reduces to the inner product identity

$$\langle\!\langle\, \nabla u\,, \mathbf{v}\,\rangle\!\rangle = \langle\, u\,, -\nabla\cdot\mathbf{v}\,\rangle. \tag{15.90}$$

Comparing (15.82) with adjgrad, we conclude that

$$\nabla^*\mathbf{v} = -\nabla\cdot\mathbf{v},$$

and hence, when subject to the proper boundary conditions, the adjoint of the gradient operator is minus the divergence: $\nabla^* = -\nabla\cdot$. In this manner, we are able to write the two-dimensional Poisson equation in the standard self-adjoint form

$$-\Delta u = \nabla^*\circ\nabla u = -\nabla\cdot(\nabla u) = f \tag{15.91}$$

subject to an appropriate system of boundary conditions that justify (15.90).

The vanishing of the boundary integral (15.89) will be ensured by the imposition of suitable homogeneous boundary conditions on the scalar field u and/or the vector field $\mathbf{v}$. Clearly the line integral will vanish if either $u = 0$ or $\mathbf{v}\cdot\mathbf{n} = 0$ at each point on the boundary. These lead immediately to the three principle types of boundary conditions. The first are the fixed or *Dirichlet boundary conditions*, which require

$$u = 0 \qquad \text{on} \qquad \partial\Omega. \tag{15.92}$$

Alternatively, we can require

$$\mathbf{v}\cdot\mathbf{n} = 0 \qquad \text{on} \qquad \partial\Omega, \tag{15.93}$$

which requires that $\mathbf{v}$ be tangent to $\partial\Omega$ at each point, and so there is no net flux across the (solid) boundary. If we identify $\mathbf{v} = \nabla u$, then the no flux boundary condition (15.93) translates into the *Neumann boundary conditions*

$$\frac{\partial u}{\partial\mathbf{n}} = \nabla u\cdot\mathbf{n} = 0 \qquad \text{on} \qquad \partial\Omega. \tag{15.94}$$

One can evidently also mix the boundary conditions, imposing Dirichlet conditions on part of the boundary, and Neumann on the complementary part:

$$u = 0 \quad \text{on} \quad D, \qquad \frac{\partial u}{\partial\mathbf{n}} = 0 \quad \text{on} \quad N, \qquad \text{where} \qquad \partial\Omega = D\cup N \tag{15.95}$$

is the disjoint union of the Dirichlet and Neumann parts.

More generally, when modeling inhomogeneous membranes, heat flow through inhomogeneous media, and similar physical equilibria, we replace the L^2 inner product between vector fields by the weighted version

$$\langle\!\langle\, \mathbf{v}\,, \mathbf{w}\,\rangle\!\rangle = \iint_\Omega \big[\, p(x,y)\, v_1(x,y)\, w_1(x,y) + q(x,y)\, v_2(x,y)\, w_2(x,y)\,\big]\, dx\, dy, \tag{15.96}$$

in which $p(x,y), q(x,y) > 0$ are strictly positive functions on the domain $(x,y) \in \Omega$ and are prescribed by the physical properties of the membrane. Retaining the usual L^2 inner product (15.83) between scalar fields, let us compute the weighted adjoint of the gradient operator. Retaining the basic defining formula (15.82),

$$\langle\!\langle\, \nabla u\,, \mathbf{v}\,\rangle\!\rangle = \iint_\Omega \left(p\, v_1\, \frac{\partial u}{\partial x} + q\, v_2\, \frac{\partial u}{\partial y} \right) dx\, dy = \iint_\Omega \nabla u \cdot \mathbf{w}\, dx\, dy, \text{ where } \mathbf{w} = \begin{pmatrix} p\, v_1 \\ q\, v_2 \end{pmatrix}.$$

We then apply out integration by parts formula (15.88) to remove the derivatives from the scalar field u, leading to

$$\begin{aligned} \iint_\Omega \nabla u \cdot \mathbf{w}\, dx\, dy &= \oint_{\partial\Omega} -u(\mathbf{w}\cdot\mathbf{n})\, ds - \iint_\Omega u\, \nabla\cdot\mathbf{w}\, dx\, dy \\ &= \oint_{\partial\Omega} \big[\, -u\, q\, v_2\, dx + u\, p\, v_1\, dy\,\big] - \iint_\Omega u \left(\frac{\partial (p\, v_1)}{\partial x} + \frac{\partial (q\, v_2)}{\partial y} \right) dx\, dy. \end{aligned} \tag{15.97}$$

Equating this to the right hand side $\langle\, u\,, \nabla^* \mathbf{v}\,\rangle$, we deduce that, provided the boundary integral vanishes, the weighted adjoint of the gradient operator with respect to (15.96) is given by

$$\nabla^* \mathbf{v} = -\nabla \mathbf{w} = -\frac{\partial (p\, v_1)}{\partial x} - \frac{\partial (q\, v_2)}{\partial y} = -p\, \frac{\partial v_1}{\partial x} - q\, \frac{\partial v_2}{\partial y} - v_1\, \frac{\partial p}{\partial x} - v_2\, \frac{\partial q}{\partial y}\,. \tag{15.98}$$

The boundary integral in (15.97) vanishes provided either $u = 0$ or $\mathbf{v} = 0$ on $\partial\Omega$. Therefore, the same homogeneous boundary conditions — Dirichlet, Neumann or mixed — remain valid in this more general context.

The corresponding self-adjoint boundary value problem takes the form

$$\nabla^* \circ \nabla u = -\frac{\partial}{\partial x} \left(p(x,y)\, \frac{\partial u}{\partial x} \right) - \frac{\partial}{\partial x} \left(q(x,y)\, \frac{\partial u}{\partial x} \right) = f(x,y), \qquad (x,y) \in \Omega, \tag{15.99}$$

along with boundary conditions of either Dirichlet, Neumann or mixed type.

Remark: In electrostatics, the gradient equation $\mathbf{v} = \nabla u$ relates the voltage drop to the electrostatic potential u, and is the continuous analog of the circuit formula (6.18) relating potentials to voltages. The continuous version of Kirchhoff's Voltage Law (6.20) that the net voltage drop around any loop is zero is the fact that any gradient vector has zero curl, $\nabla \wedge \mathbf{v} = \mathbf{0}$, i.e., the flow is irrotational. Ohm's law (6.23) has the form $\mathbf{y} = C\,\mathbf{v}$ where the vector field $\mathbf{y}$ represents the current, while $C = \mathrm{diag}\,(p(x,y), q(x,y))$ represents th conductance of the medium; in the case of Laplace's equation, we are assuming a

uniform unit conductance. Finally, the equation $f = \nabla \cdot \mathbf{y} = \nabla^* \mathbf{v}$ relating current and external current sources forms the continuous analog of Kirchhoff's Current Law (6.26) — the transpose of the discrete incidence matrix translates into the adjoint of the gradient operator is the divergence. Thus, our discrete electro-mechanical analogy carries over, in the continuous realm, to a tripartite electro-mechanical-fluid analogy, with all three physical systems leading to the same very general mathematical structure.

Positive Definiteness and the Dirichlet Principle

In conclusion, as a result of the integration by parts calculation, we have formulated the Poisson and Laplace equations (as well as their weighted counterparts) in positive (semi-)definite, self-adjoint form

$$-\Delta u = \nabla^* \circ \nabla u = f,$$

when subject to the appropriate homogeneous boundary conditions: Dirichlet, Neumann, or mixed. A key benefit is, in the positive definite cases, the characterization of the solutions by a minimization principle.

According to Theorem 7.60, the self-adjoint operator $\nabla^* \circ \nabla$ is positive definite if and only if the kernel of the underlying gradient operator — restricted to the appropriate space of scalar fields — is trivial: $\ker \nabla = \{0\}$. The determination of the kernel of the gradient operator relies on the following elementary fact.

Lemma 15.13. *If $u(x, y)$ is a C^1 function defined on a connected domain Ω, then $\nabla u \equiv 0$ if and only if $u \equiv c$ is a constant.*

This result can be viewed as the multi-variable counterpart of the result that the only function with zero derivative is a constant. It is a simple consequence of Theorem A.20; see Exercise ∎. Therefore, the only functions which could show up in $\ker \nabla$, and thus prevent positive definiteness, are the constants. The boundary conditions will tell us whether or not this occurs. The only constant function that satisfies either homogeneous Dirichlet or homogeneous mixed boundary conditions is the zero function, and thus, just as in the one-dimensional case, the boundary value problem for the Poisson equation with Dirichlet or mixed boundary conditions is positive definite. On the other hand, any constant function satisfies the homogeneous Neumann boundary conditions $\partial u/\partial \mathbf{n} = 0$, and hence such boundary value problems are only positive semi-definite.

In the positive definite definite cases, the equilibrium solution is characterized by our basic minimization principle (7.81). For the Poisson equation, the result is the justly famous *Dirichlet minimization principle*.

Theorem 15.14. *The function $u(x, y)$ that minimizes the* Dirichlet integral

$$\tfrac{1}{2} \| \nabla u \|^2 - \langle u, f \rangle = \iint_\Omega \left(\tfrac{1}{2} u_x^2 + \tfrac{1}{2} u_y^2 - f u \right) dx\, dy \tag{15.100}$$

among all C^1 functions that satisfy the prescribed homogeneous Dirichlet or mixed boundary conditions is the solution to the corresponding boundary value problem for the Poisson equation $-\Delta u = f$.

In physical applications, the Dirichlet integral (15.100) represents the energy in the system. As always, Nature chooses the equilibrium configuration so as to minimize the energy. A key application of the Dirichlet minimum principle is the finite element numerical solution scheme, to be described in detail in Section 15.5.

Remark: The fact that a minimizer to the Dirichlet integral (15.100) satisfies the Poisson equation is an immediate consequence of our general Minimization Theorem 7.62. However, unlike the finite-dimensional situation, proving the *existence* of a minimizing function is a non-trivial issue. This was not immediatiely recognized: Dirichlet originally thought this to be self-evident, but it then took about 50 years until Hilbert supplied the first rigorous existence proof. In this introductory treatment, we adopt a pragmatic approach, concentrating on the computation of the solution — reassured, if necessary, by the theoreticians' efforts in establishing its existence.

The Dirichlet minimization principle (15.100) was derived under the assumption that the boundary conditions are homogeneous — either pure Dirichlet or mixed. As it turns out, the principle, as stated, also applies to inhomogeneous Dirichlet boundary conditions. However, if we have a mixed boundary value problem with inhomogeneous Neumann conditions on part of the boundary, then we must include an additional boundary term in the minimizing functional. The general result can be stated as follows:

Theorem 15.15. *The solution $u(x,y)$ to the boundary value problem*

$$-\Delta u = f \quad \text{in} \quad \Omega, \qquad u = h \quad \text{on} \quad D, \qquad \frac{\partial u}{\partial \mathbf{n}} = k \quad \text{on} \quad N,$$

with $\partial\Omega = D \cup N$, and $D \neq \varnothing$, is characterized as the unique function that minimizes the modified Dirichlet integral

$$\iint_\Omega \left(\tfrac{1}{2} \| \nabla u \|^2 - f\, u \right) dx\, dy \; + \; \int_N u\, k\, ds \tag{15.101}$$

among all C^1 *functions that satisfy the prescribed boundary conditions.*

The inhomogeneous Dirichlet problem has $N = \varnothing$ and $D = \partial\Omega$, in which case the boundary integral does not appear. An outline of the proof of this result appears in the exercises.

As we know, positive definiteness is directly related to the stability of the physical system. The Dirichlet and mixed boundary value problems are stable, and can support any imposed force. On the other hand, the pure Neumann boundary value problem is unstable, owing to the existence of a nontrivial kernel — the constant functions. Physically, the unstable mode represents a rigid translation of the entire membrane in the vertical direction. Indeed, the Neumann problem leaves the entire boundary of the membrane unattached to any support, and so the unforced membrane is free to move up or down without affecting its equilibrium status.

Furthermore, as in finite-dimensional linear systems, non-uniqueness and non-existence of solutions go hand in hand. As we learned in Section 11.3, the existence of a solution to a Neumann boundary value problem is subject to the *Fredholm alternative*, suitably

adapted to this multi-dimensional situation. A necessary condition for the existence of a solution is that the forcing function be orthogonal to the elements of the kernel of the underlying self-adjoint linear operator, which, in the present situation requires that f be orthogonal to the subspace consisting of all constant functions. In practical terms, we only need to check orthogonality with respect to a basis for the subspace, which in this situation consists of the constant function 1.

Theorem 15.16. *The Neumann boundary value problem*

$$-\Delta u = f, \quad \text{in} \quad \Omega, \qquad\qquad \frac{\partial u}{\partial \mathbf{n}} = 0, \quad \text{on} \quad \partial\Omega, \tag{15.102}$$

admits a solution $u(x,y)$ if and only if

$$\langle\, 1\,,\, f\,\rangle = \iint_{\Omega} f(x,y)\, dx\, dy = 0. \tag{15.103}$$

Moreover, when it exists, the solution is not unique since any function of the form $u(x,y)+c$, where $c \in \mathbb{R}$ is an arbitrary constant, is also a solution.

Forcing functions $f(x,y)$ which do not satisfy the orthogonality constraint (15.103) will excite the translational instability, and no equilibrium configuration is possible. For example, if we force a free membrane, (15.103) requires that the net force in the vertical direction be zero; otherwise, the membrane will start moving and cannot be in an equilibrium.

15.5. Finite Elements.

As the reader has no doubt already surmised, explicit solutions to boundary value problems for the Laplace and Poisson equations are few and far between. In most cases, exact solution formulae are not available, or are so complicated as to be of scant utility. To proceed further, one is forced to design suitable numerical approximation schemes that can accurately evaluate the desired solution.

An especially powerful class of numerical algorithms for solving elliptic boundary value problems are the finite element methods. We have already learned, in Section 11.6, the key underlying idea. One begins with a minimization principle, prescribed by a quadratic functional defined on a suitable vector space of functions U that serves to incorporate the (homogeneous) boundary conditions. The desired solution is characterized as the unique minimizer $u_\star \in U$. One then restricts the functional to a suitably chosen finite-dimensional subspace $W \subset U$, and seeks a minimizer $w_\star \in W$. Finite-dimensionality of W has the effect of reducing the infinite-dimensional minimization problem to a finite-dimensional problem, which can then be solved by numerical linear algebra. The resulting minimizer $w_\star$ will — provided the subspace W has been cleverly chosen — provide a good approximation to the true minimizer $u_\star$ on the entire domain. Here we concentrate on the practical design of the finite element procedure, and refer the reader to a more advanced text, e.g., [**168**], for the analytical details and proofs of convergence. Most of the multi-dimensional complications are not in the underlying theory, but rather in the realms of data management and organizational details.

In this section, we first concentrate on applying these ideas to the two-dimensional Poisson equation. For specificity, we concentrate on the homogeneous Dirichlet boundary value problem

$$-\Delta u = f \quad \text{in} \quad \Omega \qquad\qquad u = 0 \quad \text{on} \quad \partial\Omega. \tag{15.104}$$

According to Theorem 15.14, the solution $u = u_\star$ is characterized as the unique minimizing function for the Dirichlet functional (15.100) among all smooth functions $u(x, y)$ that satisfy the prescribed boundary conditions. In the finite element approximation, we restrict the Dirichlet functional to a suitably chosen finite-dimensional subspace. As in the one-dimensional situation, the most convenient finite-dimensional subspaces consist of functions that may lack the requisite degree of smoothness that qualifies them as possible solutions to the partial differential equation. Nevertheless, they do provide good approximations to the actual solution. An important practical consideration, impacting the speed of the calculation, is to employ functions with small support, as in Definition 13.5. The resulting finite element matrix will then be sparse and the solution to the linear system can be relatively rapidly calculate, usually by application of an iterative numerical scheme such as the Gauss–Seidel or SOR methods discussed in Chapter 10.

Finite Elements and Triangulation

For one-dimensional boundary value problems, the finite element construction rests on the introduction of a mesh $a = x_0 < x_1 < \cdots < x_n = b$ on the interval of definition. The mesh nodes x_k break the interval into a collection of small subintervals. In two-dimensional problems, a *mesh* consists of a finite number of points $\mathbf{x}_k = (x_k, y_k)$, $k = 1, \dots, m$, known as *nodes*, usually lying inside the domain $\Omega \subset \mathbb{R}^2$. As such, there is considerable freedom in the choice of mesh nodes, and completely uniform spacing is often not possible. We regard the nodes as forming the vertices of a *triangulation* of the domain Ω, consisting of a finite number of small triangles, which we denote by $T_1, \dots, T_N$. The nodes are split into two categories — *interior nodes* and *boundary nodes*, the latter lying on or close to the boundary of the domain. A curved boundary is approximated by the polygon through the boundary nodes formed by the sides of the triangles lying on the edge of the domain; see Figure 15.10 for a typical example. Thus, in computer implementations of the finite element method, the first module is a routine that will automatically triangulate a specified domain in some reasonable manner; see below for details on what "reasonable" entails.

As in our one-dimensional finite element construction, the functions $w(x, y)$ in the finite-dimensional subspace W will be continuous and *piecewise affine*. "Piecewise affine" means that, on each triangle, the graph of w is flat, and so has the formula[†]

$$w(x, y) = \alpha^\nu + \beta^\nu x + \gamma^\nu y, \qquad \text{for} \qquad (x, y) \in T_\nu. \tag{15.105}$$

Continuity of w requires that its values on a common edge between two triangles must agree, and this will impose certain compatibility conditions on the coefficients $\alpha^\mu, \beta^\mu, \gamma^\mu$ and $\alpha^\nu, \beta^\nu, \gamma^\nu$ associated with adjacent pairs of triangles T_μ, T_ν. The graph of $z = w(x, y)$

[†] Here and subsequently, the index ν is a superscript, not a power!

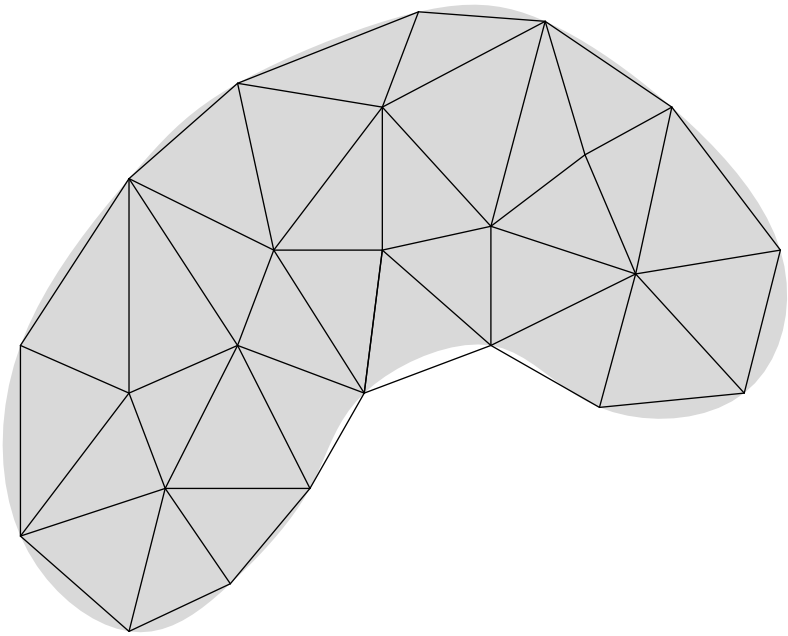


Figure 15.10. Triangulation of a Planar Domain.

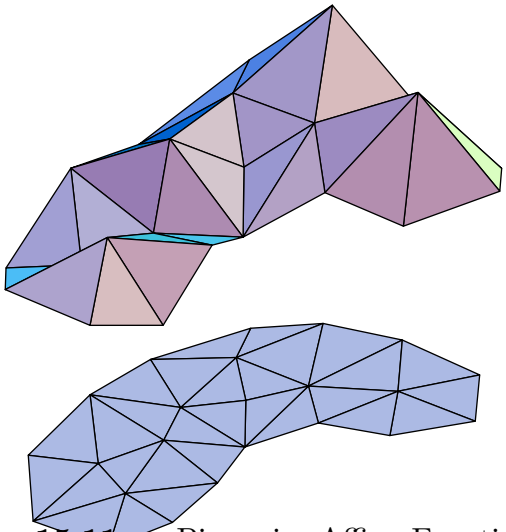


Figure 15.11. Piecewise Affine Function.

forms a connected polyhedral surface whose triangular faces lie above the triangles in the domain; see Figure 15.10 for an illustration.

The next step is to choose a basis of the subspace of piecewise affine functions for the given triangulation. As in the one-dimensional version, the most convenient basis consists of *pyramid functions* $\varphi_k(x, y)$ which assume the value 1 at a single node $\mathbf{x}_k$, and are zero at all the other nodes; thus

$$\varphi_k(x_i, y_i) = \begin{cases} 1, & i = k, \\ 0, & i \neq k. \end{cases} \tag{15.106}$$

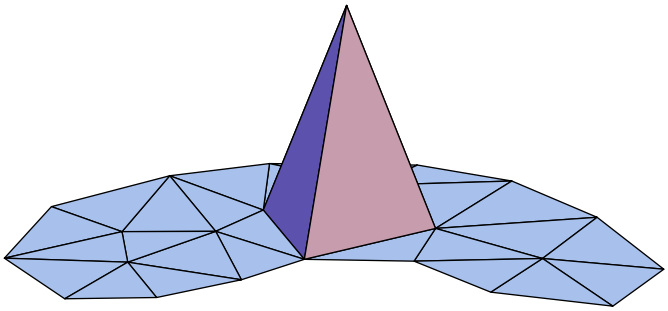


Figure 15.12. Finite Element Pyramid Function.

Note that φ_k will be nonzero only on those triangles which have the node $\mathbf{x}_k$ as one of their vertices, and hence the graph of φ_k looks like a pyramid of unit height sitting on a flat plane, as illustrated in Figure 15.12.

The pyramid functions $\varphi_k(x,y)$ corresponding to the *interior nodes* $\mathbf{x}_k$ automatically satisfy the homogeneous Dirichlet boundary conditions on the boundary of the domain — or, more correctly, on the polygonal boundary of the triangulated domain, which is supposed to be a good approximation to the curved boundary of the original domain Ω. Thus, the finite-dimensional finite element subspace W is the span of the interior node pyramid functions, and so general element $w \in W$ is a linear combination thereof:

$$w(x,y) = \sum_{k=1}^{n} c_k \, \varphi_k(x,y), \tag{15.107}$$

where the sum ranges over the n interior nodes of the triangulation. Owing to the original specification (15.106) of the pyramid functions, the coefficients

$$c_k = w(x_k, y_k) \approx u(x_k, y_k), \qquad k = 1, \dots, n, \tag{15.108}$$

are the *same* as the values of the finite element approximation $w(x,y)$ at the interior nodes. This immediately implies linear independence of the pyramid functions, since the only linear combination that vanishes at all nodes is the trivial one $c_1 = \cdots = c_n = 0$. Thus, the interior node pyramid functions $\varphi_1, \dots \varphi_n$ form a basis for finite element subspace W, which therefore has dimension equal to n, the number of interior nodes.

Determining the explicit formulae for the finite element basis functions is not difficult. On one of the triangles T_ν that has $\mathbf{x}_k$ as a vertex, $\varphi_k(x,y)$ will be the unique affine function (15.105) that takes the value 1 at the vertex $\mathbf{x}_k$ and 0 at its other two vertices $\mathbf{x}_l$ and $\mathbf{x}_m$. Thus, we are in need of a formula for an affine function or *element*

$$\omega_k^\nu(x,y) = \alpha_k^\nu + \beta_k^\nu \, x + \gamma_k^\nu \, y, \qquad (x,y) \in T_\nu, \tag{15.109}$$

that takes the prescribed values

$$\omega_k^\nu(x_i, y_i) = \omega_k^\nu(x_j, y_j) = 0, \qquad \omega_k^\nu(x_k, y_k) = 1,$$

at three distinct points. These three conditions lead to the linear system

$$\begin{aligned} \omega_k^\nu(x_i, y_i) &= \alpha_k^\nu + \beta_k^\nu x_i + \gamma_k^\nu y_i &&= 0, \\ \omega_k^\nu(x_j, y_j) &= \alpha_k^\nu + \beta_k^\nu x_j + \gamma_k^\nu y_j &&= 0, \\ \omega_k^\nu(x_k, y_k) &= \alpha_k^\nu + \beta_k^\nu x_k + \gamma_k^\nu y_k &&= 1. \end{aligned} \tag{15.110}$$

The solution† produces the explicit formulae

$$\alpha_k^\nu = \frac{x_i y_j - x_j y_i}{\Delta_\nu}, \qquad \beta_k^\nu = \frac{y_i - y_j}{\Delta_\nu}, \qquad \gamma_k^\nu = \frac{x_j - x_i}{\Delta_\nu}, \tag{15.111}$$

for the coefficients; the denominator

$$\Delta_\nu = \det \begin{pmatrix} 1 & x_i & y_i \\ 1 & x_j & y_j \\ 1 & x_k & y_k \end{pmatrix} = \pm 2 \operatorname{area} T_\nu \tag{15.112}$$

is, up to sign, twice the area of the triangle T_ν; see Exercise ■.

Example 15.17. Consider an isoceles right triangle T with vertices

$$\mathbf{x}_1 = (0, 0), \qquad \mathbf{x}_2 = (1, 0), \qquad \mathbf{x}_3 = (0, 1).$$

Using (15.111–112) (or solving the linear systems (15.110) directly), we immediately produce the three affine elements

$$\omega_1(x, y) = 1 - x - y, \qquad \omega_2(x, y) = x, \qquad \omega_3(x, y) = y. \tag{15.113}$$

As required, each ω_k equals 1 at the vertex $\mathbf{x}_k$ and is zero at the other two vertices.

The finite element pyramid function is then obtained by piecing together the individual affine elements, whence

$$\varphi_k(x, y) = \begin{cases} \omega_k^\nu(x, y), & \text{if } (x, y) \in T_\nu \text{ which has } \mathbf{x}_k \text{ as a vertex}, \\ 0, & \text{otherwise.} \end{cases} \tag{15.114}$$

Continuity of $\varphi_k(x, y)$ is assured since the constituent affine elements have the same values at common vertices. The support of the pyramid function (15.114) is the polygon

$$\operatorname{supp} \varphi_k = P_k = \bigcup_\nu T_\nu \tag{15.115}$$

consisting of all the triangles T_ν that have the node $\mathbf{x}_k$ as a vertex. In other words, $\varphi_k(x, y) = 0$ whenever $(x, y) \notin P_k$. We will call P_k the k^{th} *vertex polygon*. The node $\mathbf{x}_k$ lies on the interior of its vertex polygon P_k, while the vertices of P_k are all those that are

† Cramer's Rule (1.88) comes in handy here

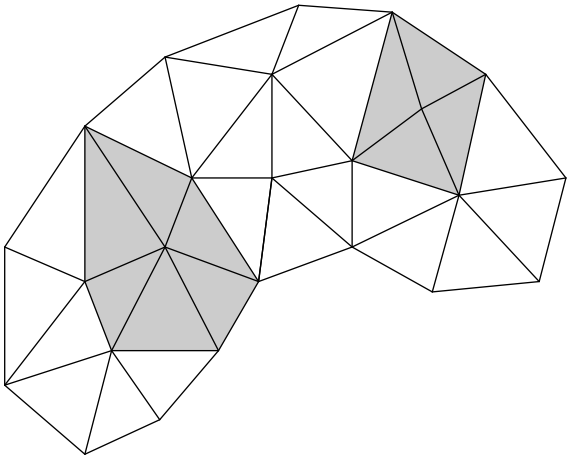


Figure 15.13. Vertex Polygons.

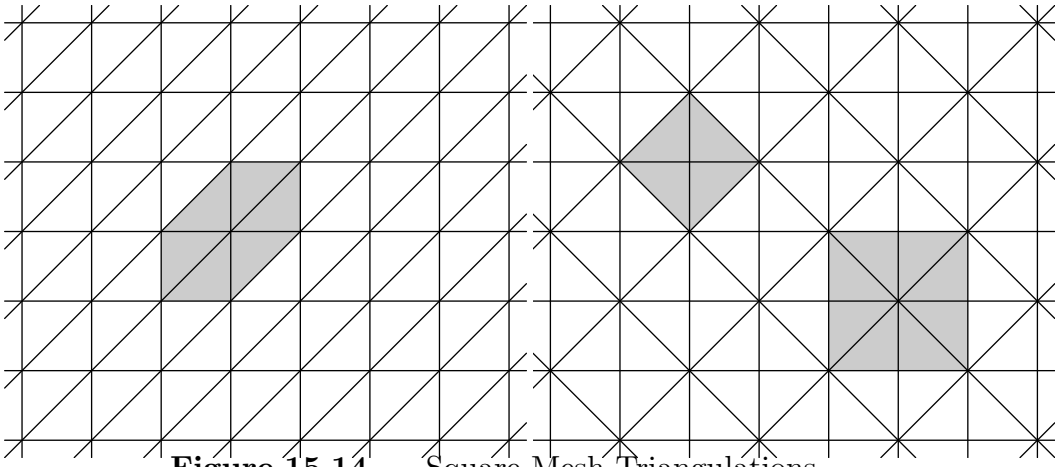


Figure 15.14. Square Mesh Triangulations.

connected to $\mathbf{x}_k$ by a single edge of the triangulation. In Figure 15.13 the shaded regions indicate two of the vertex polygons for the triangulation in Figure 15.10.

Example 15.18. The simplest, and most common triangulations are based on regular meshes. Suppose that the nodes lie on a square grid, and so are of the form $\mathbf{x}_{i,j} = (i\,h + a, j\,h + b)$ where $h > 0$ is the inter-node spacing, and (a, b) represents an overall offset. If we choose the triangles to all have the same orientation, as in the first picture in Figure 15.14, then the vertex polygons all have the same shape, consisting of 6 triangles of total area $3h^2$ — the shaded region. On the other hand, if we choose an alternating, perhaps more æsthetically pleasing triangulation as in the second picture, then there are two types of vertex polygons. The first, consisting of four triangles, has area $2h^2$, while the second, containing 8 triangles, has twice the area, $4h^2$. In practice, there are good reasons to prefer the former triangulation.

In general, in order to ensure convergence of the finite element solution to the true minimizer, one should choose a triangulation with the following properties:

(*a*) The triangles are not too long and skinny. In other words, their sides should have comparable lengths. In particular, obtuse triangles should be avoided.

(*b*) The areas of nearby triangles T_ν should not vary too much.

(*c*) The areas of nearby vertex polygons P_k should also not vary too much.

For adaptive or variable meshes, one might very well have wide variations in area over the entire grid, with small triangles in regions of rapid change in the solution, and large ones in less interesting regions. But, overall, the sizes of the triangles and vertex polygons should not dramatically vary as one moves across the domain.

The Finite Element Equations

We now seek to approximate the solution to the homogeneous Dirichlet boundary value problem by restricting the Dirichlet functional to the selected finite element subspace W. Substituting the formula (15.107) for a general element of W into the quadratic Dirichlet functional (15.100) and expanding, we find

$$\begin{aligned}\mathcal{P}[w] = \mathcal{P}\left[\sum_{i=1}^{n} c_i\,\varphi_i\right] &= \iint_\Omega \left[\left(\sum_{i=1}^{n} c_i\,\nabla\varphi_i\right)^2 - f(x,y)\,\left(\sum_{i=1}^{n} c_i\,\varphi_i\right)\right] dx\,dy\\ &= \frac{1}{2}\sum_{i,j=1}^{n} k_{ij}\,c_i\,c_j - \sum_{i=1}^{n} b_i\,c_i = \tfrac{1}{2}\,\mathbf{c}^T K\,\mathbf{c} - \mathbf{b}^T\mathbf{c}.\end{aligned}$$

Here, $K = (k_{ij})$ is the symmetric $n \times n$ matrix, while $\mathbf{b} = (\,b_1, b_2, \ldots, b_n\,)^T$ is the vector that have the respective entries

$$\begin{aligned} k_{ij} &= \langle\,\nabla\varphi_i\,,\nabla\varphi_j\,\rangle = \iint_\Omega \nabla\varphi_i\cdot\nabla\varphi_j\,dx\,dy,\\ b_i &= \langle\,f\,,\varphi_i\,\rangle = \iint_\Omega f\,\varphi_i\,dx\,dy.\end{aligned} \tag{15.116}$$

Thus, to determine the finite element approximation, we need to minimize the quadratic function

$$P(\mathbf{c}) = \tfrac{1}{2}\,\mathbf{c}^T K\,\mathbf{c} - \mathbf{b}^T\mathbf{c} \tag{15.117}$$

over all possible choices of coefficients $\mathbf{c} = (\,c_1, c_2, \ldots, c_n\,)^T \in \mathbb{R}^n$, i.e., over all possible function values at the interior nodes. Restricting to the finite element subspace has reduced us to a standard finite-dimensional quadratic minimization problem. First, the coefficient matrix $K > 0$ is positive definite due to the positive definiteness of the original functional; the proof in Section 11.6 is easily adapted to the present situation. Theorem 4.1 tells us that the minimizer is obtained by solving the associated linear system

$$K\,\mathbf{c} = \mathbf{b}. \tag{15.118}$$

The solution to (15.118) can be effected by either Gaussian elimination or an iterative technique.

To find explicit formulae for the matrix coefficients k_{ij} in (15.116), we begin by noting that the gradient of the affine element (15.109) is equal to

$$\nabla \omega_k^\nu(x,y) = \mathbf{a}_k^\nu = \begin{pmatrix} \beta_k^\nu \\ \gamma_k^\nu \end{pmatrix} = \frac{1}{\Delta_\nu} \begin{pmatrix} y_i - y_j \\ x_j - x_i \end{pmatrix}, \qquad (x,y) \in T_\nu, \tag{15.119}$$

which is a constant vector inside the triangle T_ν, while outside $\nabla \omega_k^\nu = \mathbf{0}$. Therefore,

$$\nabla \varphi_k(x,y) = \begin{cases} \nabla \omega_k^\nu = \mathbf{a}_k^\nu, & \text{if } (x,y) \in T_\nu \text{ which has } \mathbf{x}_k \text{ as a vertex,} \\ \mathbf{0}, & \text{otherwise,} \end{cases} \tag{15.120}$$

reduces to a piecewise constant function on the triangulation. Actually, (15.120) is not quite correct since if (x,y) lies on the boundary of a triangle T_ν, then the gradient does not exist. However, this technicality will not cause any difficulty in evaluating the ensuing integral.

We will approximate integrals over the domain Ω by integrals over the triangles, which relies on our assumption that the polygonal boundary of the triangulation is a reasonably close approximation to the true boundary $\partial\Omega$. In particular,

$$k_{ij} \approx \sum_\nu \iint_{T_\nu} \nabla \varphi_i \cdot \nabla \varphi_j \, dx \, dy \equiv \sum_\nu k_{ij}^\nu. \tag{15.121}$$

Now, according to (15.120), one or the other gradient in the integrand will vanish on the entire triangle T_ν *unless* both $\mathbf{x}_i$ and $\mathbf{x}_j$ are vertices. Therefore, the only terms contributing to the sum are those triangles T_ν that have both $\mathbf{x}_i$ and $\mathbf{x}_j$ as vertices. If $i \neq j$ there are only two such triangles, while if $i = j$ every triangle in the i^{th} vertex polygon P_i contributes. The individual summands are easily evaluated, since the gradients are constant on the triangles, and so, by (15.120),

$$k_{ij}^\nu = \iint_{T_\nu} \mathbf{a}_i^\nu \cdot \mathbf{a}_j^\nu \, dx \, dy = \mathbf{a}_i^\nu \cdot \mathbf{a}_j^\nu \text{ area } T_\nu = \tfrac{1}{2} \, \mathbf{a}_i^\nu \cdot \mathbf{a}_j^\nu \, | \, \Delta_\nu \, | \, .$$

Let T_ν have vertices $\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k$. Then, by (15.119, 120, 112),

$$\begin{aligned} k_{ij}^\nu &= \frac{1}{2} \, \frac{(y_j - y_k)(y_k - y_i) + (x_k - x_j)(x_i - x_k)}{(\Delta_\nu)^2} \, | \, \Delta_\nu \, | = - \, \frac{(\mathbf{x}_i - \mathbf{x}_k) \cdot (\mathbf{x}_j - \mathbf{x}_k)}{2 \, | \, \Delta_\nu \, |} \, , \quad i \neq j, \\ k_{ii}^\nu &= \frac{1}{2} \, \frac{(y_j - y_k)^2 + (x_k - x_j)^2}{(\Delta_\nu)^2} \, | \, \Delta_\nu \, | = \frac{\| \, \mathbf{x}_j - \mathbf{x}_k \, \|^2}{2 \, | \, \Delta_\nu \, |} \, . \end{aligned} \tag{15.122}$$

In this manner, each triangle T_ν specifies a collection of 6 different coefficients, $k_{ij}^\nu = k_{ji}^\nu$, indexed by its vertices, and known as the *elemental stiffnesses* of T_ν. Interestingly, the elemental stiffnesses depend only on the *angles* of the triangle and not on its size. Thus, similar triangles have the *same* elemental stiffnesses. Indeed, if we denote the angle in T_ν at the vertex $\mathbf{x}_k$ by θ_k^ν, then, according to Exercise ■,

$$k_{ii}^\nu = \tfrac{1}{2} \big(\cot \theta_k^\nu + \cot \theta_j^\nu \big), \qquad \text{while} \qquad k_{ij}^\nu = k_{ji}^\nu = -\tfrac{1}{2} \cot \theta_k^\nu, \quad i \neq j, \tag{15.123}$$

depend only upon the cotangents of the angles.

Figure 15.15. Right and Equilateral Triangles.

Example 15.19. The right triangle with vertices $\mathbf{x}_1 = (0,0)$, $\mathbf{x}_2 = (1,0)$, $\mathbf{x}_3 = (0,1)$ has elemental stiffnesses

$$k_{11} = 1, \qquad k_{22} = k_{33} = \tfrac{1}{2}, \qquad k_{12} = k_{21} = k_{13} = k_{31} = -\tfrac{1}{2}, \qquad k_{23} = k_{32} = 0. \tag{15.124}$$

The same holds for any other isoceles right triangle, as long as we chose the first vertex to be at the right angle. Similarly, an equilateral triangle has all 60° angles, and so its elemental stiffnesses are

$$\begin{aligned} &k_{11} = k_{22} = k_{33} = \tfrac{1}{\sqrt{3}} \approx .577350, \\ &k_{12} = k_{21} = k_{13} = k_{31} = k_{23} = k_{32} = -\tfrac{1}{2\sqrt{3}} \approx -.288675. \end{aligned} \tag{15.125}$$

Assembling the Elements

The elemental stiffnesses of each triangle will contribute, through the summation (15.121), to the finite element coefficient matrix K. We begin by constructing a larger matrix K^*, which we call the *full finite element matrix*, of size $m \times m$ where m is the total number of nodes in our triangulation, including both interior and boundary nodes. The rows and columns of K^* are labeled by the nodes $\mathbf{x}_i$. Let $K_\nu = (k_{ij}^\nu)$ be the corresponding $m \times m$ matrix containing the elemental stiffnesses k_{ij}^ν of T_ν in the rows and columns indexed by its vertices, and all other entries equal to 0. Thus, K_ν will have (at most) 9 nonzero entries. The resulting $m \times m$ matrices are all summed together over all the triangles,

$$K^* = \sum_{\nu=1}^{N} K_\nu, \tag{15.126}$$

to produce the full finite element matrix, in accordance with (15.121).

The full finite element matrix K^* is too large, since its rows and columns include all the nodes, whereas the finite element matrix K appearing in (15.118) only refers to the n interior nodes. The *reduced $n \times n$ finite element matrix* K is simply obtained from K^* by deleting all rows and columns indexed by boundary nodes, retaining only the elements k_{ij} when both $\mathbf{x}_i$ and $\mathbf{x}_j$ are interior nodes. (This may remind the reader of our construction of the reduced incidence matrix for a structure in Chapter 6.) For the homogeneous boundary value problem, this is all we require. As we shall see, inhomogeneous boundary conditions are most easily handled by retaining (part of) the full matrix K^*.

Figure 15.16. The Oval Plate.

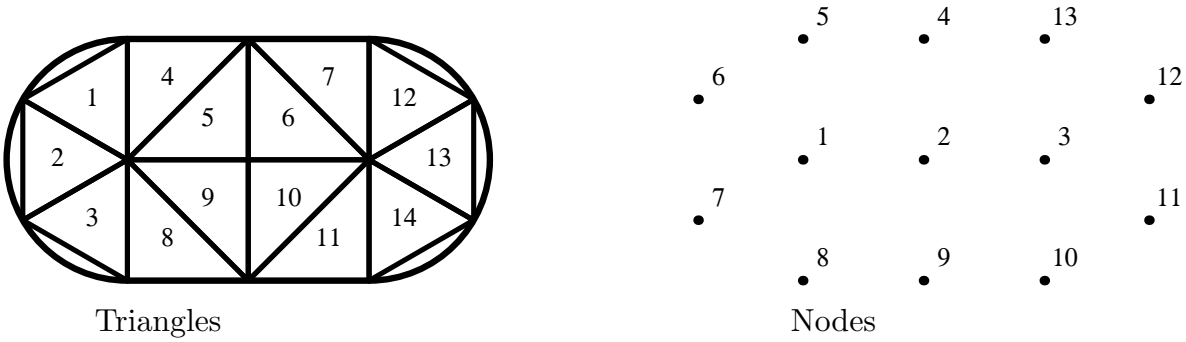


Figure 15.17. A Coarse Triangulation of the Oval Plate.

The easiest way to digest the construction is by working through a particular example.

Example 15.20. A metal plate has the shape of an oval running track, consisting of a rectangle, with side lengths 1 m by 2 m, and two semicircular disks glued onto its shorter ends, as sketched in Figure 15.16. The plate is subject to a heat source while its edges are held at a fixed temperature. The problem is to find the equilibrium temperature distribution within the plate. Mathematically, we must solve the Poisson equation with Dirichlet boundary conditions, for the equilibrium temperature $u(x, y)$.

Let us describe how to set up the finite element approximation to such a boundary value problem. We begin with a very coarse triangulation of the plate, which will not give particularly accurate results, but does serve to illustrate how to go about assembling the finite element matrix. We divide the rectangular part of the plate into 8 right triangles, while each semicircular end will be approximated by three equilateral triangles. The triangles are numbered from 1 to 14 as indicated in Figure 15.17. There are 13 nodes in all, numbered as in the second figure. Only nodes $1, 2, 3$ are interior, while the boundary nodes are labeled 4 through 13, going counterclockwise around the boundary starting at the top. The full finite element matrix K^* will have size 13×13, its rows and columns labeled by all the nodes, while the reduced matrix K appearing in the finite element equations (15.118) consists of the upper left 3×3 submatrix of K^* corresponding to the three interior nodes.

Each triangle T_ν will contribute the summand K_ν whose values are its elemental stiffnesses, as indexed by its vertices. For example, the first triangle T_1 is equilateral, and so has elemental stiffnesses (15.125). Its vertices are labeled 1, 5, and 6, and therefore we place the stiffnesses (15.125) in the rows and columns numbered $1, 5, 6$ to form the summand

$$K_1 = \begin{pmatrix} .577350 & 0 & 0 & 0 & -.288675 & -.288675 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ -.288675 & 0 & 0 & 0 & .577350 & -.288675 & 0 & 0 & \dots \\ -.288675 & 0 & 0 & 0 & -.288675 & .577350 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where all the undisplayed entries in the full 13×13 matrix are 0. The next triangle T_2 has the same equilateral elemental stiffness matrix (15.125), but now its vertices are $1, 6, 7$, and so it will contribute

$$K_2 = \begin{pmatrix} .577350 & 0 & 0 & 0 & 0 & -.288675 & -.288675 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ -.288675 & 0 & 0 & 0 & 0 & .577350 & -.2886750 & 0 & \dots \\ -.288675 & 0 & 0 & 0 & 0 & -.288675 & .5773500 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Similarly for K_3, with vertices $1, 7, 8$. On the other hand, triangle T_4 is an isoceles right triangle, and so has elemental stiffnesses (15.124). Its vertices are labeled 1, 4, and 5, with vertex 5 at the right angle. Therefore, its contribution is

$$K_4 = \begin{pmatrix} .5 & 0 & 0 & 0 & -.5 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & .5 & -.5 & 0 & 0 & 0 & \dots \\ -.5 & 0 & 0 & -.5 & 1.0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

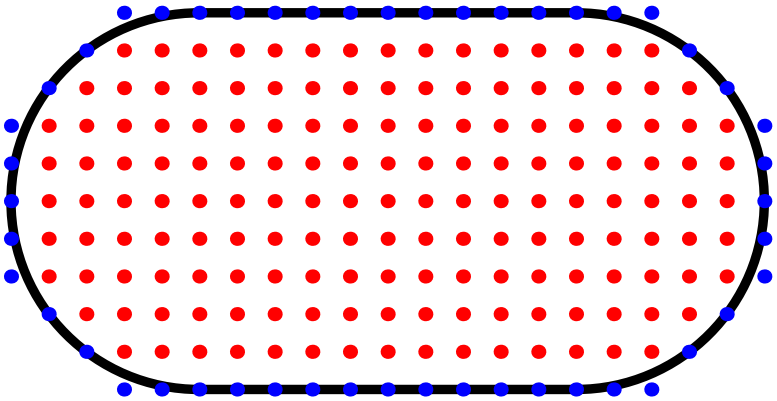


Figure 15.18. A Square Mesh for the Oval Plate.

Continuing in this manner, we assemble 14 contributions $K_1, \dots, K_{14}$, each with (at most) 9 nonzero entries. The full finite element matrix is the sum

$$
\begin{aligned}
K^* &= K_1 + K_2 + \cdots + K_{14} \\
&= \begin{pmatrix}
3.732 & -1 & 0 & 0 & -.7887 & -.5774 & -.5774 & & & & & \\
-1 & 4 & -1 & -1 & 0 & 0 & 0 & & & & & \\
0 & -1 & 3.732 & 0 & 0 & 0 & 0 & & & & & \\
0 & -1 & 0 & 2 & -.5 & 0 & 0 & & & & & \\
-.7887 & 0 & 0 & -.5 & 1.577 & -.2887 & 0 & & & & & \\
-.5774 & 0 & 0 & 0 & -.2887 & 1.155 & -.2887 & & & & & \\
-.5774 & 0 & 0 & 0 & 0 & -.2887 & 1.155 & & & & & \\
-.7887 & 0 & 0 & 0 & 0 & 0 & -.2887 & & & & & \\
0 & -1 & 0 & 0 & 0 & 0 & 0 & & & & & \\
0 & 0 & -.7887 & 0 & 0 & 0 & 0 & & & & & \\
0 & 0 & -.5774 & 0 & 0 & 0 & 0 & & & & & \\
0 & 0 & -.5774 & 0 & 0 & 0 & 0 & & & & & \\
0 & 0 & -.7887 & -.5 & 0 & 0 & 0 & & & & &
\end{pmatrix} \\
&\qquad \begin{pmatrix}
-.7887 & 0 & 0 & 0 & 0 & 0 \\
0 & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & -.7887 & -.5774 & -.5774 & -.7887 \\
0 & 0 & 0 & 0 & 0 & -.5 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
-.2887 & 0 & 0 & 0 & 0 & 0 \\
1.577 & -.5 & 0 & 0 & 0 & 0 \\
-.5 & 2 & -.5 & 0 & 0 & 0 \\
0 & -.5 & 1.577 & -.2887 & 0 & 0 \\
0 & 0 & -.2887 & 1.155 & -.2887 & 0 \\
0 & 0 & 0 & -.2887 & 1.155 & -.2887 \\
0 & 0 & 0 & 0 & -.2887 & 1.577
\end{pmatrix}.
\end{aligned}
\tag{15.127}
$$

Since only nodes $1, 2, 3$ are interior nodes, the reduced finite element matrix only uses the

upper left 3×3 block of K^*, so

$$K = \begin{pmatrix} 3.732 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 3.732 \end{pmatrix}. \tag{15.128}$$

It is not difficult to directly construct K, bypassing K^* entirely.

For a finer triangulation, the construction is similar, but the matrices become much larger. The procedure can, of course, be automated. Fortunately, if we choose a very regular triangulation, then we do not need to be nearly as meticulous in assembling the stiffness matrices, since many of the entries are the same. The simplest case is when we use a uniform square mesh, and so triangulate the domain into isoceles right triangles. This is accomplished by laying out a relatively dense square grid over the domain $\Omega \subset \mathbb{R}^2$. The interior nodes are the grid points that fall inside the oval domain, while the boundary nodes are all those grid points lying adjacent to one or more of the interior nodes, and are near but not necessarily precisely on the boundary $\partial\Omega$. Figure 15.18 shows the nodes in a square grid with intermesh spacing $h = .2$. While a bit crude in its approximation of the boundary of the domain, this procedure does have the advantage of making the construction of the associated finite element matrix relatively painless.

For such a mesh, all the triangles are isoceles right triangles, with elemental stiffnesses (15.124). Summing the corresponding matrices K_ν over all the triangles, as in (15.126), the rows and columns of K^* corresponding to the interior nodes are seen to all have the same form. Namely, if i labels an interior node, then the corresponding diagonal entry is $k_{ii} = 4$, while the off-diagonal entries $k_{ij} = k_{ji}$, $i \neq j$, are equal to either -1 when node i is adjacent to node j on the grid, and is equal to 0 in all other cases. Node j is allowed to be a boundary node. (Interestingly, the result does not depend on how one orients the pair of triangles making up each square of the grid, which only plays a role in the computation of the right hand side of the finite element equation.) Observe that the same computation applies even to our coarse triangulation. The interior node 2 belongs to all right isoceles triangles, and the corresponding entries in (15.127) are $k_{22} = 4$, and $k_{2j} = -1$ for the four adjacent nodes $j = 1, 3, 4, 9$.

Remark: Interestingly, the coefficient matrix arising from the finite element method on a square (or even rectangular) grid is the same as the coefficient matrix arising from a finite difference solution to the Laplace or Poisson equation, as described in Exercise ■. The finite element approach has the advantage of applying to much more general triangulations.

In general, while the finite element matrix K for a two-dimensional boundary value problem is not as nice as the tridiagonal matrices we obtained in our one-dimensional problems, it is still very sparse and, on regular grids, highly structured. This makes solution of the resulting linear system particularly amenable to an iterative matrix solver such as Gauss–Seidel, Jacobi, or, for even faster convergence, successive over-relaxation (SOR).

The Coefficient Vector and the Boundary Conditions

So far, we have been concentrating on assembling the finite element coefficient matrix K. We also need to compute the forcing vector $\mathbf{b} = (\, b_1, b_2, \ldots, b_n \,)^T$ appearing on the right

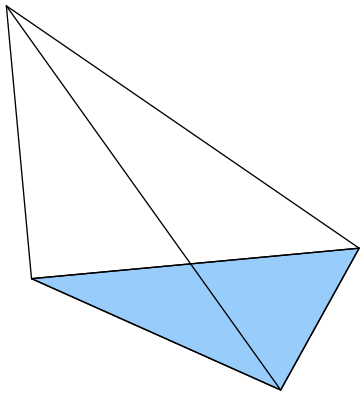


Figure 15.19. Finite Element Tetrahedron.

hand side of the fundamental linear equation (15.118). According to (15.116), the entries b_i are found by integrating the product of the forcing function and the finite element basis function. As before, we will approximate the integral over the domain Ω by an integral over the triangles, and so

$$b_i = \iint_\Omega f\,\varphi_i\,dx\,dy \approx \sum_\nu \iint_{T_\nu} f\,\omega_i^\nu\,dx\,dy \equiv \sum_\nu b_i^\nu. \tag{15.129}$$

Typically, the exact computation of the various triangular integrals is not convenient, and so we resort to a numerical approximation. Since we are assuming that the individual triangles are small, we can adopt a very crude numerical integration scheme. If the function $f(x,y)$ does not vary much over the triangle T_ν — which will certainly be the case if T_ν is sufficiently small — we may approximate $f(x,y) \approx c_i^\nu$ for $(x,y) \in T_\nu$ by a constant. The integral (15.129) is then approximated by

$$b_i^\nu = \iint_{T_\nu} f\,\omega_i^\nu\,dx\,dy \approx c_i^\nu \iint_{T_\nu} \omega_i^\nu(x,y)\,dx\,dy = \tfrac{1}{3}\,c_i^\nu \text{ area } T_\nu = \tfrac{1}{6}\,c_i^\nu\,|\,\Delta_\nu\,|. \tag{15.130}$$

The formula for the integral of the affine element $\omega_i^\nu(x,y)$ follows from solid geometry. Indeed, it equals the volume under its graph, a tetrahedron of height 1 and base T_ν, as illustrated in Figure 15.19.

How to choose the constant c_i^ν? In practice, the simplest choice is to let $c_i^\nu = f(x_i, y_i)$ be the value of the function at the i^{th} vertex. With this choice, the sum in (15.129) becomes

$$b_i \approx \sum_\nu \tfrac{1}{3}\,f(x_i,y_i) \text{ area } T_\nu = \tfrac{1}{3}\,f(x_i,y_i) \text{ area } P_i, \tag{15.131}$$

where P_i is the vertex polygon (15.115) corresponding to the node $\mathbf{x}_i$. In particular, for the square mesh with the uniform choice of triangles, as in Example 15.18,

$$\text{area } P_i = 3\,h^2 \qquad \text{for all } i, \text{ and so} \qquad b_i \approx f(x_i,y_i)\,h^2 \tag{15.132}$$

is well approximated by just h^2 times the value of the forcing function at the node. This is the underlying reason to choose the uniform triangulation for the square mesh; the

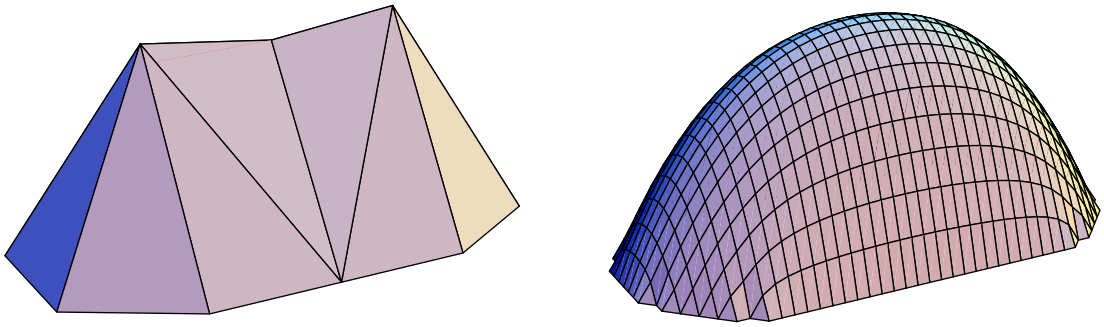


Figure 15.20. Finite Element Solutions to Poisson's Equation for an Oval Plate.

alternating version would give unequal values for the b_i over adjacent nodes, and this would introduce unnecessary errors into the final approximation.

Example 15.21. For the coarsely triangulated oval plate, the reduced stiffness matrix is (15.128). The Poisson equation

$$-\Delta u = 4$$

models a constant external heat source of magnitude 4° over the entire plate. If we keep the edges of the plate fixed at 0°, then we need to solve the finite element equation $K\mathbf{c} = \mathbf{b}$, where K is the coefficient matrix (15.128), while

$$\mathbf{b} = \tfrac{4}{3}\left(2 + \tfrac{3\sqrt{3}}{4},\ 2,\ 2 + \tfrac{3\sqrt{3}}{4}\right)^T = (\,4.39872, 2.66667, 4.39872\,)^T .$$

The entries of $\mathbf{b}$ are, by (15.131), equal to $4 = f(x_i, y_i)$ times one third the area of the corresponding vertex polygon, which for node 2 is the square consisting of 4 right triangles, each of area $\frac{1}{2}$, whereas for nodes 1 and 3 it consists of 4 right triangles of area $\frac{1}{2}$ plus three equilateral triangles, each of area $\frac{\sqrt{3}}{4}$; see Figure 15.17.

The solution to the final linear system is easily found:

$$\mathbf{c} = (\,1.56724, 1.45028, 1.56724\,)^T .$$

Its entries are the values of the finite element approximation at the three interior nodes. The finite element solution is plotted in the first illustration in Figure 15.20. A more accurate solution, based on a square grid triangulation of size $h = .1$ is plotted in the second figure.

Inhomogeneous Boundary Conditions

So far, we have restricted our attention to problems with homogeneous Dirichlet boundary conditions. According to Theorem 15.15, the solution to the inhomogeneous Dirichlet problem

$$-\Delta u = f \quad \text{in} \quad \Omega, \qquad\qquad u = h \quad \text{on} \quad \partial\Omega,$$

is also obtained by minimizing the Dirichlet functional (15.100). However, now the minimization takes place over the affine subspace consisting of all functions that satisfy the inhomogeneous boundary conditions. It is not difficult to fit this problem into the finite element scheme.

The elements corresponding to the interior nodes of our triangulation remain as before, but now we need to include additional elements to ensure that our approximation satisfies the boundary conditions. Note that if $\mathbf{x}_k$ is a boundary node, then the corresponding *boundary element* $\varphi_k(x,y)$ satisfies the interpolation condition (15.106), and so has the same piecewise affine form (15.114). The corresponding finite element approximation

$$w(x,y) = \sum_{i=1}^{m} c_i\, \varphi_i(x,y), \tag{15.133}$$

has the same form as before, (15.107), but now the sum is over all nodes, both interior and boundary. As before, the coefficients $c_i = w(x_i, y_i) \approx u(x_i, y_i)$ are the values of the finite element approximation at the nodes. Therefore, in order to satisfy the boundary conditions, we require

$$c_j = h(x_j, y_j) \qquad \text{whenever} \qquad \mathbf{x}_j = (x_j, y_j) \qquad \text{is a boundary node.} \tag{15.134}$$

Remark: If the boundary node $\mathbf{x}_j$ does not lie precisely on the boundary $\partial\Omega$, we need to approximate the value $h(x_j, y_j)$ appropriately, e.g., by using the value of $h(x,y)$ at the nearest boundary point $(x,y) \in \partial\Omega$.

The derivation of the finite element equations proceeds as before, but now there are additional terms arising from the nonzero boundary values. Leaving the intervening details to the reader, the final outcome can be written as follows. Let K^* denote the full $m \times m$ finite element matrix constructed as above. The reduced coefficient matrix K is obtained by retaining the rows and columns corresponding to only interior nodes, and so will have size $n \times n$, where n is the number of interior nodes. The *boundary coefficient matrix* $\widetilde{K}$ is the $n \times (m-n)$ matrix consisting of the entries of the interior rows that do not appear in K, i.e., those lying in the columns indexed by the boundary nodes. For instance, in the the coarse triangulation of the oval plate, the full finite element matrix is given in (15.127), and the upper 3×3 subblock is the reduced matrix (15.128). The remaining entries of the first three rows form the boundary coefficient matrix

$$\widetilde{K} = \begin{pmatrix} 0 & -.7887 & -.5774 & -.5774 & -.7887 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -.7887 & -.5774 & -.5774 & -.7887 \end{pmatrix}. \tag{15.135}$$

We similarly split the coefficients c_i of the finite element function (15.133) into two groups. We let $\mathbf{c} \in \mathbb{R}^n$ denote the as yet unknown coefficients c_i corresponding to the values of the approximation at the interior nodes $\mathbf{x}_i$, while $\mathbf{h} \in \mathbb{R}^{m-n}$ will be the vector of boundary values (15.134). The solution to the finite element approximation (15.133) is obtained by solving the associated linear system

$$K\,\mathbf{c} + \widetilde{K}\,\mathbf{h} = \mathbf{b}, \qquad \text{or} \qquad K\,\mathbf{c} = \mathbf{f} = \mathbf{b} - \widetilde{K}\,\mathbf{h}. \tag{15.136}$$

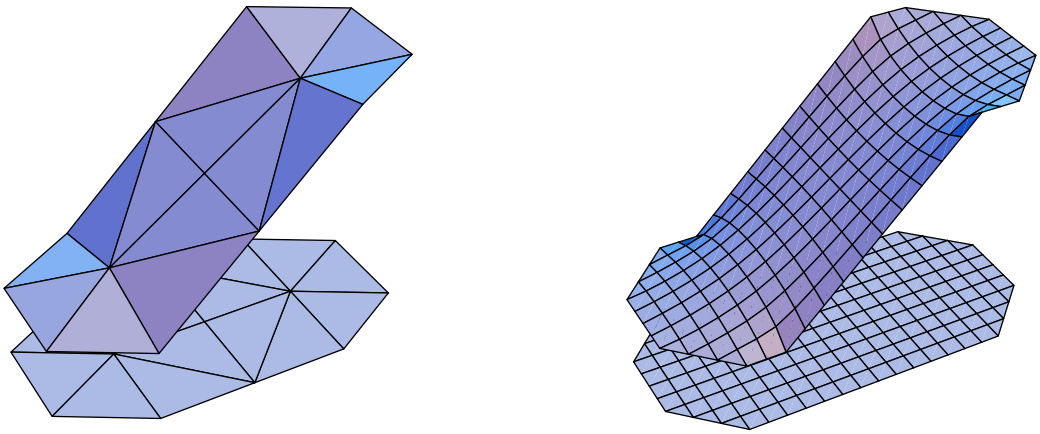


Figure 15.21. Solution to the Dirichlet Problem for the Oval Plate.

Example 15.22. For the oval plate discussed in Example 15.20, suppose the right hand semicircular edge is held at 10°, the left hand semicircular edge at -10°, while the two straight edges have a linearly varying temperature distribution ranging from -10° at the left to 10° at the right, as illustrated in Figure 15.21. Our task is to compute its equilibrium temperature, assuming no internal heat source. Thus, for the coarse triangulation we have the boundary nodes values

$$\mathbf{h} = (\, h_4, \dots, h_{13} \,)^T = (\, 0, -1, -1, -1, -1, 0, 1, 1, 1, 1, 0 \,)^T .$$

Using the previously computed formulae (15.128, 135) for the interior coefficient matrix K and boundary coefficient matrix $\widetilde{K}$, we approximate the solution to the Laplace equation by solving (15.136). We are assuming that there is no external forcing function, $f(x, y) \equiv 0$, and so the right hand side is $\mathbf{b} = \mathbf{0}$, and so we must solve $K\mathbf{c} = \mathbf{f} = -\widetilde{K}\,\mathbf{h} = (\, 2.18564, 3.6, 7.64974 \,)^T$. The finite element function corresponding to the solution $\mathbf{c} = (\, 1.06795, 1.8, 2.53205 \,)^T$ is plotted in the first illustration in Figure 15.21. Even on such a coarse mesh, the approximation is not too bad, as evidenced by the second illustration, which plots the finite element solution for a square mesh with spacing $h = .2$ between nodes.

Second Order Elliptic Boundary Value Problems

While the Laplace and Poisson equations are by far the most important elliptic partial differential equations, they only model homogeneous media, e.g., membranes made out of a uniform material, or heated plates with uniform (constant) heat capacity. Inhomogeneous media lead to more general self-adjoint differential operators, leading to variable coefficient second order elliptic boundary value problems. Even more generally, *elastic shells*, meaning bendable plates, lead to fourth order two-dimensional elliptic boundary value problems similar to the one-dimensional beam equation (11.115). And, these are in turn only linear approximations to the fully nonlinear elliptic boundary value problems occurring in elasticity theory, [**85**]. The latter are beyond the scope of this text, although some of the required mathematical tools appear in Chapter 21.

The most important class of linear, self-adjoint, second order, elliptic partial differential equations in two space variables take the form

$$-\frac{\partial}{\partial x}\left(p(x,y)\,\frac{\partial u}{\partial x}\right)-\frac{\partial}{\partial y}\left(q(x,y)\,\frac{\partial u}{\partial x}\right)+r(x,y)\,u=f(x,y),\qquad (x,y)\in\Omega,\quad (15.137)$$

where $p(x,y), q(x,y) > 0$ are strictly positive functions, while $r(x,y) \geq 0$ is non-negative. For simplicity, we also impose homogeneous Dirichlet boundary conditions $u = 0$ on $\partial\Omega$. Note that the positivity conditions ensure that the partial differential equation is *elliptic* in accordance with the classification of Definition 15.1.

The reader may notice that (15.137) is a two-dimensional version of the Sturm–Liouville ordinary differential equation (11.143). The self-adjoint formulation (11.155) of a Sturm–Liouville boundary value problem serves to inspire the self-adjoint form

$$L^*\circ L[u]=f,\qquad \text{by setting}\qquad L[u]=\begin{pmatrix}u_x\\ u_y\\ u\end{pmatrix},\qquad (15.138)$$

of the boundary value problem for (15.137). Note that the linear operator $L\colon U \to V$ maps the vector space U consisting of all smooth functions $u(x,y)$ satisfying the homogeneous Dirichlet boundary conditions to the vector space V consisting of all vector-valued functions $\mathbf{v} = (\,v_1(x,y), v_2(x,y), v_3(x,y)\,)^T$. We adopt the usual L^2 inner product (15.83) on U, but introduce a weighted inner product[†]

$$\langle\!\langle\, \mathbf{v}\,,\widetilde{\mathbf{v}}\,\rangle\!\rangle=\iint_\Omega\,\big(\,p\,v_1\,\widetilde{v}_1+q\,v_2\,\widetilde{v}_2+r\,v_3\,\widetilde{v}_3\,\big)\,dx\,dy$$

on the vector space V. A straightforward computation based on Green's formula (15.87) produces the "weighted adjoint"

$$L^*[\mathbf{v}]=-\frac{\partial}{\partial x}\,\big[\,p(x,y)\,v_1(x,y)\,\big]-\frac{\partial}{\partial y}\,\big[\,q(x,y)\,v_2(x,y)\,\big]+r(x,y)\,v_3(x,y)\qquad (15.139)$$

of the operator L. Therefore, the formula for the self-adjoint product

$$L^*\circ L[u]=L^*\begin{pmatrix}u_x\\ u_y\\ u\end{pmatrix}=-\frac{\partial}{\partial x}\left(p(x,y)\,\frac{\partial u}{\partial x}\right)-\frac{\partial}{\partial y}\left(q(x,y)\,\frac{\partial u}{\partial y}\right)+r(x,y)\,u(x,y)$$

proves the identification of (15.138) and (15.137). Positive definiteness follows from the observation that $\ker L = \{0\}$. The minimization principle associated with the operator L is, as usual,

$$\mathcal{P}[u]=\tfrac12\,\|\,L[u]\,\|^2-\langle\,f\,,u\,\rangle=\iint_\Omega\,\big[\,\tfrac12\,p\,u_x^2+\tfrac12\,q\,u_y^2+\tfrac12\,r\,u^2-f\,u\,\big]\,dx\,dy.\qquad (15.140)$$

As always, the solution to our boundary value problem is the unique minimizing function for $\mathcal{P}[u]$ among all $u \in U$ satisfying the homogeneous boundary conditions.

[†] Technically, we should require that $r(x,y) \not\equiv 0$ not vanish on any open subdomain in order that this define a nondegenerate inner product.

Remark: Interestingly, in contrast to the Poisson equation, if $r > 0$ the boundary value problem for (15.137) is positive definite with minimization principle (15.140) even in the case of pure Neumann boundary conditions. This is because the operator L always has trivial kernel.

The finite element approximation is constructed as before — by restricting the minimization principle to the finite-dimensional subspace spanned by the finite element basis functions (11.161). This requires the solution of a linear system of the same form (15.136), in which

$$k_{ij} = \langle\, L[\varphi_i]\,, L[\varphi_j]\,\rangle = \iint_\Omega \left(p\, \frac{\partial \varphi_i}{\partial x}\, \frac{\partial \varphi_j}{\partial x} + q\, \frac{\partial \varphi_i}{\partial y}\, \frac{\partial \varphi_j}{\partial y} + r\, \varphi_i\, \varphi_j \right) dx\, dy,$$
$$b_i = \langle\, f\,, \varphi_i\,\rangle = \iint_\Omega f\, \varphi_i\, dx\, dy. \tag{15.141}$$

As before, the double integrals are approximated by a sum of integrals over the triangles T_ν. The only triangles that contribute to the final result for k_{ij} are the ones that have both $\mathbf{x}_i$ and $\mathbf{x}_j$ as vertices. When the triangles are small, the integrals can be approximated by fairly crude numerical integration formulae. This completes our brief outline of the method; full details are left to the reader.

Example 15.23. The *Helmholtz equation*

$$\Delta u + \lambda\, u = 0, \tag{15.142}$$

governs the eigenvalues of the Laplacian, and as such forms the fundamental modes of vibration of a wide variety of mechanical system, including the vibration of plates, scattering of acoustic and electromagnetic waves, and many others.

If $\lambda < 0$, then the Helmholtz equation fits into the positive definite framework (15.137), with $p = q = 1$ and $r = -\lambda$. To solve the problem by finite elements, we restrict the minimization principle

$$\mathcal{P}[u] = \iint_\Omega \left(\tfrac{1}{2} \,\|\, \nabla u \,\|^2 - \tfrac{1}{2}\, \lambda\, u^2 - f\, u \right) dx\, dy. \tag{15.143}$$

to the finite-dimensional finite element subspace determined by a triangulation of the underlying domain. The resulting coefficient matrix has the form

$$\begin{aligned} k_{ij} &= \iint_\Omega \left(\nabla \varphi_i \cdot \nabla \varphi_j - \tfrac{1}{2}\, \lambda\, \varphi_i \varphi_j \right) dx\, dy \\ &\approx \sum_\nu \iint_{T_\nu} \left(\nabla \omega_i^\nu \cdot \nabla \omega_j^\nu - \lambda\, \omega_i^\nu\, \omega_j^\nu \right) dx\, dy \equiv \sum_\nu k_{ij}^\nu. \end{aligned} \tag{15.144}$$

Determination of the explicit formulae for the k_{ij}^ν is left as an exercise for the reader. On the other hand, the forcing vector $\mathbf{b}$ has exactly the same form (15.116) as in the Poisson example.

Unfortunately, the most interesting cases are when $\lambda > 0$ and the boundary value problem is not positive definite; nevertheless, the finite element approach can will still ive quite respectable approximations to the actual solution, but is more challenging to justify here.

Chapter 16

Complex Analysis

The term "complex analysis" refers to the calculus of complex-valued functions $f(z)$ depending on a complex variable z. On the surface, it may seem that this subject should merely be a simple reworking of standard real variable theory that you learned in first year calculus. However, this naïve first impression could not be further from the truth! Complex analysis is the culmination of a deep and far-ranging study of the fundamental notions of complex differentiation and complex integration, and has an elegance and beauty not found in the more familiar real arena. For instance, complex functions are always *analytic*, meaning that they can be represented as convergent power series. As an immediate consequence, a complex function automatically has an *infinite* number of derivatives, and difficulties with degree of smoothness, strange discontinuities, delta functions, and other forms of pathological behavior of real functions never arise in the complex realm.

The driving force behind many applications of complex analysis is the remarkable, profound connection between harmonic functions (solutions of the Laplace equation) of two variables and complex functions. Namely, the real and imaginary parts of a complex analytic function are automatically harmonic. In this manner, complex functions provide a rich lode of new solutions to the two-dimensional Laplace equation to help solve boundary value problems. One of the most useful practical consequences arises from the elementary observation that the composition of two complex functions is also a complex function. We interpret this operation as a complex changes of variables, also known as a *conformal mapping* since it preserves angles. Conformal mappings can be effectively used for constructing solutions to the Laplace equation on complicated planar domains, and play a particularly important role in the solution of physical problems.

Complex integration also enjoys many remarkable properties not found in its real sibling. Integrals of complex functions are similar to the line integrals of planar multivariable calculus. The remarkable theorem due to Cauchy implies that complex integrals are generally path-independent — provided one pays proper attention to the complex singularities of the integrand. In particular, an integral of a complex function around a closed curve can be directly evaluated through the "calculus of residues", which effectively bypasses the Fundamental Theorem of Calculus. Surprisingly, the method of residues can even be applied to evaluate certain types of definite real integrals.

In this chapter, we shall introduce the basic techniques and theorems in complex analysis, paying particular attention to those aspects which are required to solve boundary value problems associated with the planar Laplace and Poisson equations. Complex analysis is an essential tool in a surprisingly broad range of applications, including fluid flow, elasticity, thermostatics, electrostatics, and, in mathematics, geometry, and even number

theory. Indeed, the most famous unsolved problem in all of mathematics, the Riemann hypothesis, is a conjecture about a specific complex function that has profound consequences for the distribution of prime numbers[†].

16.1. Complex Variables.

In this section we shall develop the basics of complex analysis — the calculus of complex functions $f(z)$. Here $z = x + \mathrm{i}\, y$ is a single complex variable and $f\colon \Omega \to \mathbb{C}$ is a complex-valued function defined on a domain $z \in \Omega \subset \mathbb{C}$ in the complex plane. Before diving into this material, you should first make sure you are familiar with the basics of complex numbers, as discussed in Section 3.6.

Any complex function can be written as a complex combination

$$f(z) = f(x + \mathrm{i}\, y) = u(x,y) + \mathrm{i}\, v(x,y), \tag{16.1}$$

of two real functions u, v depending on two real variables x, y, called, respectively, its *real* and *imaginary parts*, and written

$$u(x,y) = \operatorname{Re} f(z), \qquad \text{and} \qquad v(x,y) = \operatorname{Im} f(z). \tag{16.2}$$

For example, the monomial function $f(z) = z^3$ is written as

$$z^3 = (x + \mathrm{i}\, y)^3 = (x^3 - 3\,x\,y^2) + \mathrm{i}\,(3\,x^2 y - y^3),$$

and so

$$\operatorname{Re} z^3 = x^3 - 3\,x\,y^2, \qquad\qquad \operatorname{Im} z^3 = 3\,x^2 y - y^3.$$

We can identify $\mathbb{C}$ with the real, two-dimensional plane $\mathbb{R}^2$, so that the complex number $z = x + \mathrm{i}\, y \in \mathbb{C}$ is identified with the real vector $(\,x, y\,)^T \in \mathbb{R}^2$. Based on this identification, we shall employ the standard terminology of planar vector calculus, e.g., domain, curve, etc., without alteration; see Appendix A for details. In this manner, we may regard a complex function as particular type of real vector field that maps

$$\begin{pmatrix} x \\ y \end{pmatrix} \in \Omega \subset \mathbb{R}^2 \qquad \text{to the vector} \qquad \mathbf{v}(x,y) = \begin{pmatrix} u(x,y) \\ v(x,y) \end{pmatrix} \in \mathbb{R}^2. \tag{16.3}$$

Not every real vector field qualifies as a complex function; the components $u(x,y), v(x,y)$ must satisfy certain fairly stringent requirements, which can be found in Theorem 16.3 below.

Many of the well-known functions appearing in real-variable calculus — polynomials, rational functions, exponentials, trigonometric functions, logarithms, and many others — have natural complex extensions. For example, complex polynomials

$$p(z) = a_n\, z^n + a_{n-1}\, z^{n-1} + \cdots + a_1\, z + a_0 \tag{16.4}$$

[†] Not to mention that a solution will net you a cool \$1,000,000.00. For details on how to claim your prize, check out the web site `http://www.claymath.org`.

are complex linear combinations (meaning that the coefficients a_k are allowed to be complex numbers) of the basic monomial functions $z^k = (x + \mathrm{i}\, y)^k$. Similarly, we have already made sporadic use of complex exponentials such as

$$e^z = e^{x+\mathrm{i}\, y} = e^x \cos y + \mathrm{i}\, e^x \sin y$$

for solving differential equations. Other examples will appear shortly.

There are several ways to motivate[†] the link between harmonic functions $u(x, y)$, meaning solutions of the two-dimensional Laplace equation

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, \tag{16.5}$$

and complex functions. One natural starting point is to return to the d'Alembert solution (14.121) of the one-dimensional wave equation, which was based on the factorization

$$\Box = \partial_t^2 - c^2\, \partial_x^2 = (\partial_t - c\, \partial_x)\, (\partial_t + c\, \partial_x)$$

of the linear wave operator (14.110). The two-dimensional Laplace operator $\Delta = \partial_x^2 + \partial_y^2$ has essentially the same form, except for a "minor" change in sign[‡]. We cannot produce a real factorization of the Laplace operator, but there is a complex factorization,

$$\Delta = \partial_x^2 + \partial_y^2 = (\partial_x - \mathrm{i}\, \partial_y)\, (\partial_x + \mathrm{i}\, \partial_y),$$

into a product of two complex first order differential operators, having complex "wave speed" $c = \mathrm{i}$. Mimicking the solution formula (14.118) for the wave equation, we expect that the solutions to the Laplace equation (16.5) should be expressed in the form

$$u(x, y) = f(x + \mathrm{i}\, y) + g(x - \mathrm{i}\, y), \tag{16.6}$$

i.e., a linear combination of functions of the complex variable $z = x + \mathrm{i}\, y$ and its complex conjugate $\overline{z} = x - \mathrm{i}\, y$. The functions $f(x + \mathrm{i}\, y)$ and $g(x - \mathrm{i}\, y)$ satisfy the first order complex partial differential equations

$$\frac{\partial f}{\partial x} = -\,\mathrm{i}\, \frac{\partial f}{\partial y}\,, \qquad\qquad \frac{\partial g}{\partial x} = \mathrm{i}\, \frac{\partial g}{\partial y}\,, \tag{16.7}$$

and hence (16.6) does indeed define a complex-valued solution to the Laplace equation.

In most applications, we are searching for a real solution to the Laplace equation, and so our d'Alembert-type formula (16.6) is not entirely satisfactory. As we know, a complex number $z = x + \mathrm{i}\, y$ is real if and only if it equals its own conjugate, $z = \overline{z}$. Thus, the solution (16.6) will be real if and only if

$$f(x + \mathrm{i}\, y) + g(x - \mathrm{i}\, y) = u(x, y) = \overline{u(x, y)} = \overline{f(x + \mathrm{i}\, y)} + \overline{g(x - \mathrm{i}\, y)}.$$

[†] A reader uninterested in the motivation can skip ahead to Proposition 16.1 at this point.

[‡] However, the change in sign has serious ramifications for the analytical properties of solutions to the two equations. As noted in Section 15.1, there is a profound difference between the elliptic Laplace equation and the hyperbolic wave equation.

Now, the complex conjugation operation switches $x + \mathrm{i}\, y$ and $x - \mathrm{i}\, y$, and so we expect the first term $\overline{f(x + \mathrm{i}\, y)}$ to be a function of $x - \mathrm{i}\, y$, while the second term $\overline{g(x - \mathrm{i}\, y)}$ will be a function of $x + \mathrm{i}\, y$. Therefore[§], to equate the two sides of this equation, we should require

$$g(x - \mathrm{i}\, y) = \overline{f(x + \mathrm{i}\, y)},$$

and so

$$u(x, y) = f(x + \mathrm{i}\, y) + \overline{f(x + \mathrm{i}\, y)} = 2 \operatorname{Re} f(x + \mathrm{i}\, y).$$

Dropping the inessential factor of 2, we conclude that a real solution to the two-dimensional Laplace equation can be written as the *real part* of a complex function. A direct proof of the following key result will appear below.

Proposition 16.1. *If $f(z)$ is a complex function, then its real part*

$$u(x, y) = \operatorname{Re} f(x + \mathrm{i}\, y) \tag{16.8}$$

is a harmonic function.

The *imaginary part* of a complex function is also harmonic. This is because

$$\operatorname{Im} f(z) = \operatorname{Re} (-\mathrm{i}\, f(z))$$

is the real part of the complex function

$$-\mathrm{i}\, f(z) = -\mathrm{i}\, [\, u(x, y) + \mathrm{i}\, v(x, y)\,] = v(x, y) - \mathrm{i}\, u(x, y).$$

Therefore, if $f(z)$ is any complex function, we can write it as a complex combination

$$f(z) = f(x + \mathrm{i}\, y) = u(x, y) + \mathrm{i}\, v(x, y),$$

of two real harmonic functions: $u(x, y) = \operatorname{Re} f(z)$ and $v(x, y) = \operatorname{Im} f(z)$.

Before delving into the many remarkable properties of complex functions, let us look at some of the most basic examples. In each case, the reader can directly check that the harmonic functions given as the real and imaginary parts of the complex function are indeed solutions to the Laplace equation.

Examples of Complex Functions

(a) *Harmonic Polynomials*: The simplest examples of complex functions are polynomials. Any polynomial is a complex linear combinations, as in (16.4), of the basic complex monomials

$$z^n = (x + \mathrm{i}\, y)^n = u_n(x, y) + \mathrm{i}\, v_n(x, y). \tag{16.9}$$

The real and imaginary parts of a complex polynomial are known as *harmonic polynomials*, and we list the first few below. The general formula for the basic harmonic polynomials

[§] We are ignoring the fact that f and g are not quite uniquely determined since one can add and subtract a constant from them. This does not affect the argument in any significant way.

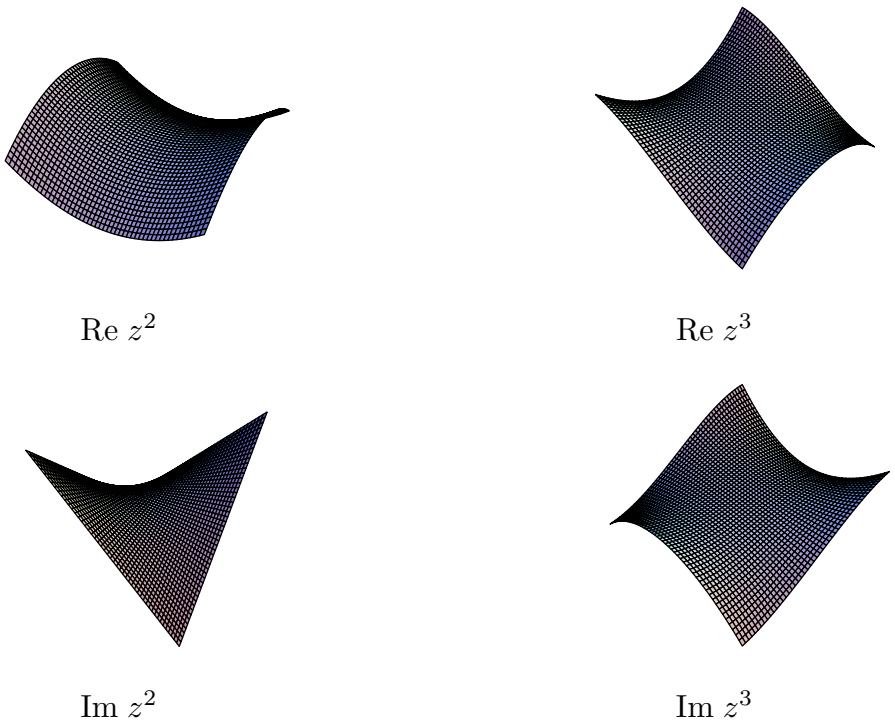


Figure 16.1. Real and Imaginary Parts of z^2 and z^3.

$u_n(x,y)$ and $v_n(x,y)$ is easily found by applying the binomial theorem to expand (16.9), as in Exercise ■.

Harmonic Polynomials

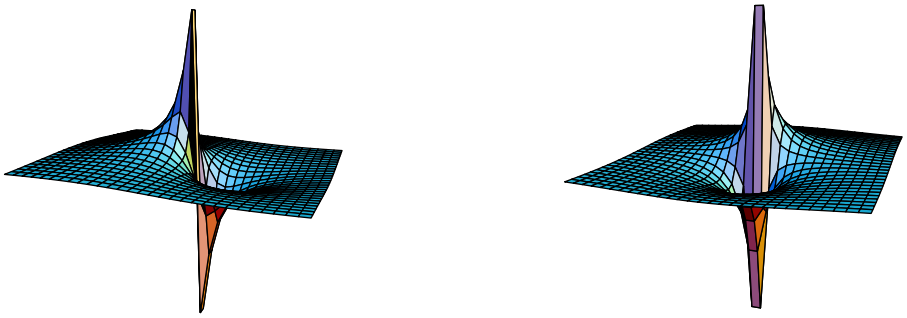
n	z^n	$u_n(x,y)$	$v_n(x,y)$
0	1	1	0
1	$x + \mathrm{i}\, y$	x	y
2	$(x^2 - y^2) + 2\,\mathrm{i}\, x\, y$	$x^2 - y^2$	$2\,x\, y$
3	$(x^3 - 3\,x\, y^2) + \mathrm{i}\,(3\,x^2 y - y^3)$	$x^3 - 3\,x\, y^2$	$3\,x^2 y - y^3$
4	$(x^4 - 6\,x^2 y^2 + y^4) + \mathrm{i}\,(4\,x^3 y - 4\,x\, y^3)$	$x^4 - 6\,x^2 y^2 + y^4$	$4\,x^3 y - 4\,x\, y^3$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

We have, in fact, already encountered these polynomial solutions to the Laplace equation. If we write

$$z = r\, e^{\,\mathrm{i}\,\theta}, \tag{16.10}$$

where

$$r = |\, z\,| = \sqrt{x^2 + y^2}, \qquad \theta = \mathrm{ph}\, z = \tan^{-1}\frac{y}{x}\,,$$



Re $\dfrac{1}{z}$ Im $\dfrac{1}{z}$

Figure 16.2. Real and Imaginary Parts of $f(z) = \dfrac{1}{z}$.

are the usual polar coordinates (modulus and phase) of $z = x + \mathrm{i}\, y$, then Euler's formula (3.84) yields

$$z^n = r^n\, e^{\mathrm{i}\, n\, \theta} = r^n \cos n\,\theta + \mathrm{i}\, r^n \sin n\,\theta,$$

and so

$$u_n = r^n \cos n\,\theta, \qquad\qquad v_n = r^n \sin n\,\theta.$$

Therefore, the harmonic polynomials are just the polar coordinate separablesolutions (15.38) to the Laplace equation. In Figure 16.1 we plot[†] the real and imaginary parts of the monomials z^2 and z^3.

(*b*) *Rational Functions*: Ratios

$$f(z) = \frac{p(z)}{q(z)} \tag{16.11}$$

of complex polynomials provide a large variety of harmonic functions. The simplest case is

$$\frac{1}{z} = \frac{\overline{z}}{z\,\overline{z}} = \frac{\overline{z}}{|\,z\,|^2} = \frac{x}{x^2+y^2} - \mathrm{i}\,\frac{y}{x^2+y^2}\,. \tag{16.12}$$

Its real and imaginary parts are graphed in Figure 16.2. Note that these functions have an interesting singularity at the origin $x = y = 0$, but are harmonic everywhere else.

A slightly more complicated example is the function

$$f(z) = \frac{z-1}{z+1}\,. \tag{16.13}$$

[†] Graphing a complex function $f\colon \mathbb{C} \to \mathbb{C}$ is problematic. The identification (16.3) of f with a real vector-valued function $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ implies that four real dimensions are needed to display its complete graph.

To write out (16.13) in real form, we multiply and divide by the complex conjugate of the denominator, leading to

$$f(z) = \frac{z-1}{z+1} = \frac{(z-1)(\overline{z}+1)}{(z+1)(\overline{z}+1)} = \frac{|\,z\,|^2 + z - \overline{z} - 1}{|\,z+1\,|^2} = \frac{x^2+y^2-1}{(x+1)^2+y^2} \; + \mathrm{i}\, \frac{2\,y}{(x+1)^2+y^2}\,. \tag{16.14}$$

This manipulation can always be used to find the real and imaginary parts of general rational functions.

If we assume that the rational function (16.11) is written in lowest terms, with p and q having no common factors, then $f(z)$ will have a singularity, known as a *pole*, wherever the denominator vanishes: $q(z_0) = 0$. The order[†] of the root z_0 of $q(z)$ tells us the *order* of the pole of $f(z)$. For example, the rational function

$$f(z) = \frac{z+2}{z^5+z^3} = \frac{z+2}{(z+\mathrm{i}\,)(z-\mathrm{i}\,)z^3}$$

has three poles: a simple (of order 1) pole at $z = +\mathrm{i}\,$, another simple pole at $z = -\mathrm{i}\,$ and a triple (order 3) pole at $z = 0$.

(*c*) *Complex Exponentials*: Euler's formula

$$e^z = e^x \cos y + \mathrm{i}\, e^x \sin y \tag{16.15}$$

for the complex exponential, cf. (3.84), yields two important harmonic functions: $e^x \cos y$ and $e^x \sin y$, which are graphed in Figure 3.8. More generally, writing out $e^{c\,z}$ for a complex constant $c = a + \mathrm{i}\,b$ produces the complex exponential function

$$e^{c\,z} = e^{a\,x - b\,y} \cos(b\,x + a\,y) + \mathrm{i}\, e^{a\,x - b\,y} \sin(b\,x + a\,y). \tag{16.16}$$

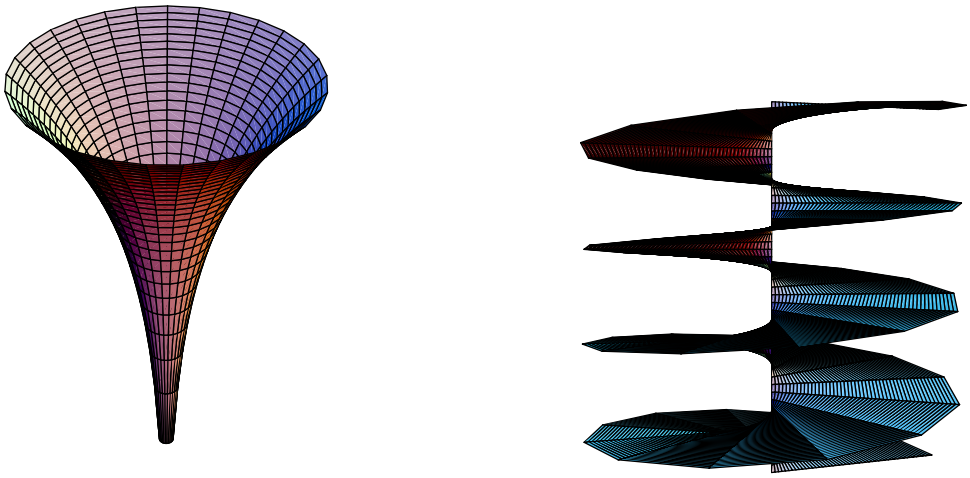
Its real and imaginary parts are harmonic functions for arbitrary $a, b \in \mathbb{R}$. Some of these were found by applying the separation of variables method in Cartesian coordinates; see the table in Section 15.2.

(*d*) *Complex Trigonometric Functions*: The complex trigonometric functions are defined in terms of the complex exponential by adapting our earlier formulae (3.86):

$$\begin{aligned} \cos z &= \frac{e^{\mathrm{i}\,z} + e^{-\mathrm{i}\,z}}{2} = \cos x \cosh y - \mathrm{i}\, \sin x \sinh y, \\ \sin z &= \frac{e^{\mathrm{i}\,z} - e^{-\mathrm{i}\,z}}{2\,\mathrm{i}} = \sin x \cosh y + \mathrm{i}\, \cos x \sinh y. \end{aligned} \tag{16.17}$$

The resulting harmonic functions are products of trigonometric and hyperbolic functions. They can all be written as linear combinations of the harmonic functions (16.16) derived from the complex exponential. Note that when $z = x$ is real, so $y = 0$, these functions reduce to the usual real trigonometric functions $\cos x$ and $\sin x$.

[†] Recall that the *order* of a root z_0 of a polynomial $q(z)$ is the number of times $z - z_0$ occurs as a factor of $q(z)$.



$\operatorname{Re}(\log z) = \log|z|$ $\qquad\qquad$ $\operatorname{Im}(\log z) = \operatorname{ph} z$

Figure 16.3. Real and Imaginary Parts of $\log z$.

(e) *Complex Logarithm*: In a similar fashion, the complex logarithm $\log z$ is a complex extension of the usual real natural (i.e., base e) logarithm. In terms of polar coordinates (16.10), the complex logarithm has the form

$$\log z = \log(r\, e^{\,\mathrm{i}\,\theta}) = \log r + \log e^{\,\mathrm{i}\,\theta} = \log r + \,\mathrm{i}\,\theta, \tag{16.18}$$

Thus, the logarithm of a complex number has real part

$$\operatorname{Re}(\log z) = \log r = \log|z| = \tfrac{1}{2}\,\log(x^2 + y^2),$$

which is a well-defined harmonic function on all of $\mathbb{R}^2$ except for a logarithmic singularity at the origin $x = y = 0$. It is, in fact, the logarithmic potential corresponding to a delta function forcing concentrated at the origin that played a key role in our construction of the Green's function for the Poisson equation in Section 15.3.

The imaginary part

$$\operatorname{Im}(\log z) = \theta = \operatorname{ph} z$$

of the complex logarithm is the *phase* (argument) or polar angle of z. The phase is also not defined at the origin $x = y = 0$. Moreover, it is a multi-valued harmonic function elsewhere, since it is only specified up to integer multiples of 2π. Thus, a given nonzero complex number $z \neq 0$ has an infinite number of possible values for its phase, and hence an infinite number of possible complex logarithms $\log z$, each differing by an integer multiple of $2\pi\,\mathrm{i}$, reflecting the fact that $e^{2\pi\,\mathrm{i}} = 1$. In particular, if $z = x > 0$ is real and positive, then $\log z = \log x$ agrees with the real logarithm, *provided* we choose the angle $\operatorname{ph} z = 0$. Alternative choices for the phase include an integer multiple of $2\pi\,\mathrm{i}$, and so ordinary real, positive numbers $x > 0$ also have complex logarithms! On the other hand, if $z = x < 0$ is real and negative, then $\log z = \log|x| + (2k+1)\,\pi\,\mathrm{i}$ is complex no matter which value of $\operatorname{ph} z$ is chosen. (This explains why we didn't attempt to define the logarithm of a negative number in first year calculus!) As the point z circles around the origin in a counter-clockwise direction, $\operatorname{Im}\log z = \operatorname{ph} z = \theta$ increases by 2π. Thus, its graph can be likened to

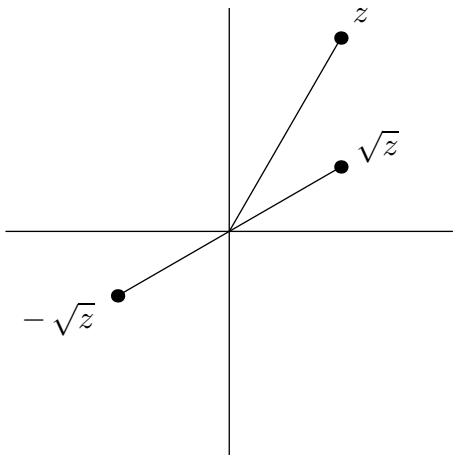


Figure 16.4. Square Roots of a Complex Number.

a parking ramp with infinitely many levels, spiraling ever upwards as one circumabulates the origin; Figure 16.3 attempts to sketch it. For the complex logarithm, the origin is a type of singularity known as a *logarithmic branch point*, indicating that there are an infinite number of possible "branches" meaning values that can be assigned to $\log z$ at any nonzero point.

(f) *Roots and Fractional Powers*: A similar branching phenomenon occurs with the fractional powers and roots of complex numbers. The simplest case is the square root function $\sqrt{z}$. Every nonzero complex number $z \neq 0$ has two different possible square roots: $\sqrt{z}$ and $-\sqrt{z}$. As illustrated in Figure 16.4, the two square roots lie on opposite sides of the origin, and are obtained by multiplying by -1. Writing $z = r\, e^{\,\mathrm{i}\,\theta}$ in polar coordinates, we see that

$$\sqrt{z} = \sqrt{r\, e^{\,\mathrm{i}\,\theta}} = \sqrt{r}\, e^{\,\mathrm{i}\,\theta/2} = \sqrt{r}\, \left(\cos \frac{\theta}{2} + \mathrm{i} \sin \frac{\theta}{2} \right), \tag{16.19}$$

i.e., we take the square root of the modulus and halve the phase:

$$\big|\, \sqrt{z}\, \big| = \sqrt{|\, z\,|} = \sqrt{r}\,, \qquad\qquad \mathrm{ph}\, \sqrt{z} = \tfrac{1}{2}\, \mathrm{ph}\, z = \tfrac{1}{2}\, \theta.$$

Since $\theta = \mathrm{ph}\, z$ is only defined up to an integer multiple of $2\,\pi$, the angle $\frac{1}{2}\,\theta$ is only defined up to an integer multiple of π. The even and odd multiples yield different values for (16.19), which accounts for the two possible values of the square root. For instance, since $\mathrm{ph}\, 4\,\mathrm{i} = \frac{1}{2}\,\pi$ or $\frac{5}{2}\,\pi$, we find

$$\sqrt{4\,\mathrm{i}} = 2\,\sqrt{\mathrm{i}} = \pm\, 2\, e^{\pi\,\mathrm{i}\,/4} = \pm\, 2 \left(\cos \frac{\pi\,\mathrm{i}}{4} + \mathrm{i} \sin \frac{\pi\,\mathrm{i}}{4} \right) = \pm \left(\sqrt{2} + \mathrm{i}\, \sqrt{2} \right).$$

If we start at some $z \neq 0$ and circle once around the origin, we increase $\mathrm{ph}\, z$ by $2\,\pi$, but $\mathrm{ph}\, \sqrt{z}$ only increases by π. Thus, at the end of our circuit, we arrive at the other square root $-\sqrt{z}$. Circling the origin again increases $\mathrm{ph}\, z$ by a further $2\,\pi$, and hence brings us back to the original square root $\sqrt{z}$. Therefore, the graph of the multiply-valued square

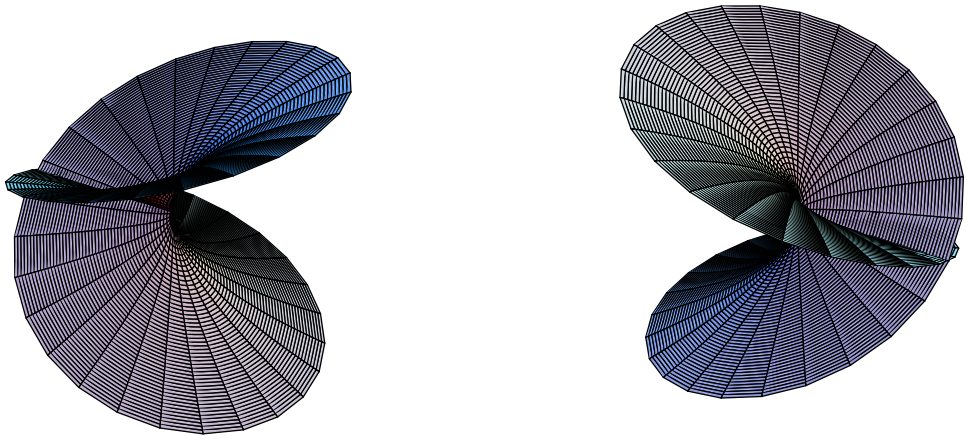


Figure 16.5. Real and Imaginary Parts of $\sqrt{z}$.

root function will look like a weirdly interconnected parking ramp with only two levels, as shown in† Figure 16.5.

Similar remarks apply to the n^{th} root

$$\sqrt[n]{z} = \sqrt[n]{r}\, e^{\,\mathrm{i}\,\theta/n} = \sqrt[n]{r}\, \left(\cos \frac{\theta}{n} + \mathrm{i} \sin \frac{\theta}{n} \right), \tag{16.20}$$

which, except for $z = 0$, has n possible values, depending upon which multiple of 2π is used in the assignment of $\operatorname{ph} z = \theta$. The n different n^{th} roots are obtained by multiplying any one of them by the different n^{th} roots of unity, $\zeta_n^k = e^{2\,k\,\pi\,\mathrm{i}\,/n}$ for $k = 0, \dots, n-1$, as defined in (13.11). In this case, the origin $z = 0$ is called a *branch point* of *order* n since there are n different branches for the function $\sqrt[n]{z}$. Circling around the origin a total of n times leads to the n branches in succession, returning in the end to the original.

The preceding list of elementary examples is far from exhausting the range and variety of complex functions. Lack of space will preclude us from studying the remarkable properties of complex versions of the gamma function, Airy functions, Bessel functions, and Legendre functions that appear in Appendix C, as well as elliptic functions, the Riemann zeta function, modular functions, and many, many other important and fascinating functions arising in complex analysis and its manifold applications; see [**173**, **184**].

16.2. Complex Differentiation.

The bedrock of complex function theory is the notion of the complex derivative. Complex differentiation is defined in the same manner as the usual calculus limit definition of

† These graphs are best appreciated in an interactive three-dimensional graphics viewer.

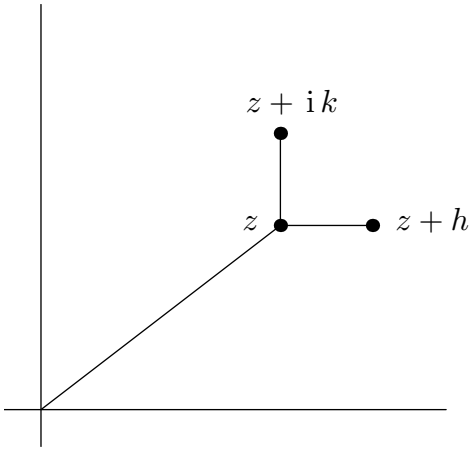


Figure 16.6. Complex Derivative Directions.

the derivative of a real function. Yet, despite a superficial similarity, complex differentiation is profoundly different, and displays an elegance and depth not shared by its real progenitor.

Definition 16.2. A complex function $f(z)$ is *differentiable* at a point $z \in \mathbb{C}$ if and only if the limiting difference quotient exists:

$$f'(z) = \lim_{w \to z} \frac{f(w) - f(z)}{w - z}\,. \tag{16.21}$$

The key feature of this definition is that the limiting value $f'(z)$ of the difference quotient must be *independent* of how w converges to z. On the real line, there are only two directions to approach a limiting point — either from the left or from the right. These lead to the concepts of left and right handed derivatives and their equality is required for the existence of the usual derivative of a real function. In the complex plane, there are an infinite variety of directions to approach the point z, and the definition requires that all of these "directional derivatives" must agree. This is the reason for the more severe restrictions on complex derivatives, and, in consequence, the source of their remarkable properties.

Let us first see what happens when we approach z along the two simplest directions — horizontal and vertical. If we set

$$w = z + h = (x + h) + \mathrm{i}\, y, \qquad \text{where } h \text{ is real},$$

then $w \to z$ along a horizontal line as $h \to 0$, as sketched in Figure 16.6. If we write out

$$f(z) = u(x, y) + \mathrm{i}\, v(x, y)$$

in terms of its real and imaginary parts, then we must have

$$\begin{aligned} f'(z) &= \lim_{h \to 0} \frac{f(z + h) - f(z)}{h} = \lim_{h \to 0} \frac{f(x + h + \mathrm{i}\, y) - f(x + \mathrm{i}\, y)}{h} \\ &= \lim_{h \to 0} \left[\frac{u(x + h, y) - u(x, y)}{h} + \mathrm{i}\, \frac{v(x + h, y) - v(x, y)}{h} \right] = \frac{\partial u}{\partial x} + \mathrm{i}\, \frac{\partial v}{\partial x} = \frac{\partial f}{\partial x}\,, \end{aligned}$$

which follows from the usual definition of the (real) partial derivative. On the other hand, if we set

$$w = z + \mathrm{i}\, k = x + \mathrm{i}\,(y + k), \qquad \text{where } k \text{ is real},$$

then $w \to z$ along a vertical line as $k \to 0$. Therefore, we must also have

$$f'(z) = \lim_{k \to 0} \frac{f(z + \mathrm{i}\, k) - f(z)}{\mathrm{i}\, k} = \lim_{k \to 0} \left[-\mathrm{i}\, \frac{f(x + \mathrm{i}\,(y + k)) - f(x + \mathrm{i}\, y)}{k} \right]$$
$$= \lim_{h \to 0} \left[\frac{v(x, y + k) - v(x, y)}{k} - \mathrm{i}\, \frac{u(x, y + k) - u(x, y)}{k} \right] = \frac{\partial v}{\partial y} - \mathrm{i}\, \frac{\partial u}{\partial y} = -\mathrm{i}\, \frac{\partial f}{\partial y}\,.$$

When we equate the real and imaginary parts of these two distinct formulae for the complex derivative $f'(z)$, we discover that the real and imaginary components of $f(z)$ must satisfy a certain homogeneous linear system of partial differential equations, named after Augustin–Louis Cauchy and Bernhard Riemann†, two of the principal founders of modern complex analysis.

Theorem 16.3. *A function $f(z) = u(x, y) + \mathrm{i}\, v(x, y)$, where $z = x + \mathrm{i}\, y$, has a complex derivative $f'(z)$ if and only if its real and imaginary parts are continuously differentiable and satisfy the* Cauchy–Riemann equations

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}\,, \qquad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}\,. \tag{16.22}$$

In this case, the complex derivative of $f(z)$ is equal to any of the following expressions:

$$f'(z) = \frac{\partial f}{\partial x} = \frac{\partial u}{\partial x} + \mathrm{i}\, \frac{\partial v}{\partial x} = -\mathrm{i}\, \frac{\partial f}{\partial y} = \frac{\partial v}{\partial y} - \mathrm{i}\, \frac{\partial u}{\partial y}\,. \tag{16.23}$$

The proof of the converse — that any function whose real and imaginary components satisfy the Cauchy–Riemann equations is differentiable — will be omitted, but can be found in any basic text on complex analysis, e.g., [**4**, **154**].

Remark: It is worth pointing out that equation (16.23) tells us that f satisfies $\partial f/\partial x = -\mathrm{i}\, \partial f/\partial y$, which, reassuringly, agrees with the first equation in (16.7).

Example 16.4. Consider the elementary function

$$z^3 = (x^3 - 3\, x\, y^2) + \mathrm{i}\,(3\, x^2 y - y^3).$$

Its real part $u = x^3 - 3\, x\, y^2$ and imaginary part $v = 3\, x^2 y - y^3$ satisfy the Cauchy–Riemann equations (16.22), since

$$\frac{\partial u}{\partial x} = 3\, x^2 - 3\, y^2 = \frac{\partial v}{\partial y}\,, \qquad \frac{\partial u}{\partial y} = -6\, x\, y = -\frac{\partial v}{\partial x}\,.$$

† In addition to his contributions to complex analysis, partial differential equations and number theory, Bernhard Riemann also was the inventor of Riemannian geometry, which turned out to be absolutely essential for Einstein's theory of general relativity some 70 years later!

Theorem 16.3 implies that $f(z) = z^3$ is complex differentiable. Not surprisingly, its derivative turns out to be

$$f'(z) = 3\,z^2 = (3\,x^2 - 3\,y^2) + \mathrm{i}\,(6\,x\,y) = \frac{\partial u}{\partial x} + \mathrm{i}\,\frac{\partial v}{\partial x} = \frac{\partial v}{\partial y} - \mathrm{i}\,\frac{\partial u}{\partial y}\,.$$

Fortunately, the complex derivative obeys all of the usual rules that you learned in real-variable calculus. For example,

$$\frac{d}{dz}\,z^n = n\,z^{n-1}, \qquad \frac{d}{dz}\,e^{c\,z} = c\,e^{c\,z}, \qquad \frac{d}{dz}\log z = \frac{1}{z}\,, \tag{16.24}$$

and so on. The power n can even be non-integral or, in view of the identity $z^n = e^{n \log z}$, complex, while c is any complex constant. The exponential formulae (16.17) for the complex trigonometric functions implies that they also satisfy the standard rules

$$\frac{d}{dz}\cos z = -\sin z, \qquad \frac{d}{dz}\sin z = \cos z. \tag{16.25}$$

The formulae for differentiating sums, products, ratios, inverses, and compositions of complex functions are all identical to their real counterparts. Thus, thankfully, you don't need to learn any new rules for performing complex differentiation!

Remark: There are many examples of seemingly reasonable functions which do *not* have a complex derivative. The simplest is the complex conjugate function

$$f(z) = \overline{z} = x - \mathrm{i}\,y.$$

Its real and imaginary parts do *not* satisfy the Cauchy–Riemann equations, and hence $\overline{z}$ does *not* have a complex derivative. More generally, any function $f(x,y) = h(z,\overline{z})$ that explicitly depends on the complex conjugate variable $\overline{z}$ is *not* complex-differentiable.

Power Series and Analyticity

The most remarkable feature of complex analysis, which distinguishes it from real function theory, is that the existence of one complex derivative automatically implies the existence of infinitely many! All complex functions $f(z)$ are infinitely differentiable and, in fact, analytic where defined. The reason for this surprising and profound fact will, however, not become evident until we learn the basics of complex integration in Section 16.5. In this section, we shall take analyticity as a given, and investigate some of its principal consequences.

Definition 16.5. A complex function $f(z)$ is called *analytic* at a point z_0 if it has a power series expansion

$$f(z) = a_0 + a_1(z - z_0) + a_2(z - z_0)^2 + a_3(z - z_0)^3 + \,\cdots\, = \sum_{n=0}^{\infty} a_n\,(z - z_0)^n\,, \tag{16.26}$$

which converges for all z sufficiently close to z_0.

Typically, the standard ratio or root tests for convergence of (real) series that you learned in ordinary calculus, [**9**, **165**], can be applied to determine where a given (complex) power series converges. We note that if $f(z)$ and $g(z)$ are analytic at a point z_0, so is their sum $f(z) + g(z)$, product $f(z)\, g(z)$ and, provided $g(z_0) \neq 0$, ratio $f(z)/g(z)$.

Example 16.6. All of the real power series found in elementary calculus carry over to the complex versions of the functions. For example,

$$e^z = 1 + z + \tfrac{1}{2}\, z^2 + \tfrac{1}{6}\, z^3 + \cdots = \sum_{n=0}^{\infty} \frac{z^n}{n!} \tag{16.27}$$

is the Taylor series for the exponential function based at $z_0 = 0$. A simple application of the ratio test proves that the series converges for all z. On the other hand, the power series

$$\frac{1}{z^2+1} = 1 - z^2 + z^4 - z^6 + \cdots = \sum_{k=0}^{\infty} (-1)^k\, z^{2\,k}\,, \tag{16.28}$$

converges inside the unit disk, where $|\, z \,| < 1$, and diverges outside, where $|\, z \,| > 1$. Again, convergence is established through the ratio test. The ratio test is inconclusive when $|\, z \,| = 1$, and we shall leave the more delicate question of precisely where on the unit disk this complex series converges to a more advanced treatment, e.g., [**4**].

In general, there are three possible options for the domain of convergence of a complex power series (16.26):

(*a*) The series converges for all z.

(*b*) The series converges inside a disk $|\, z - z_0 \,| < \rho$ of radius $\rho > 0$ centered at z_0 and diverges for all $|\, z - z_0 \,| > \rho$ outside the disk. The series may converge at some (but not all) of the points on the boundary of the disk where $|\, z - z_0 \,| = \rho$.

(*c*) The series only converges, trivially, at $z = z_0$.

The number ρ is known as the *radius of convergence* of the series. In case (*a*), we say $\rho = \infty$, while in case (*c*), $\rho = 0$, and the series does *not* represent an analytic function. An example with $\rho = 0$ is the power series $\sum n!\, z^n$. In the intermediate case (*b*), determining precisely where on the boundary of the convergence disk the power series converges is quite delicate, and will not be pursued here. The proof of this result can be found in Exercise ■; see also [**4**, **93**] for further details.

Remarkably, the radius of convergence for the power series of a known analytic function $f(z)$ can be determined by inspection, without recourse to any fancy convergence tests! Namely, ρ is equal to the distance from z_0 to the nearest *singularity* of $f(z)$, meaning a point where the function fails to be analytic. This explains why the Taylor series of e^z converges everywhere, while that of $(z^2+1)^{-1}$ only converges inside the unit disk. Indeed e^z is analytic for all z and has no singularities; therefore the radius of convergence of its power series — centered at any point z_0 — is equal to $\rho = \infty$. On the other hand, the function

$$f(z) = \frac{1}{z^2+1} = \frac{1}{(z + \mathrm{i}\,)(z - \mathrm{i}\,)}$$

has singularities (poles) at $z = \pm \mathrm{i}$, and so the series (16.28) has radius of convergence $\rho = 1$, which is the distance from $z_0 = 0$ to the singularities. Thus, the extension of the theory of power series to the complex plane serves to explain the apparent mystery of why, as a real function, $(1 + x^2)^{-1}$ is well-defined and analytic for all real x, but its power series only converges on the interval $(-1, 1)$. It is the *complex* singularities that prevent its convergence when $|\, x \,| > 1$. If we expand $(z^2 + 1)^{-1}$ in a power series at some other point, say $z_0 = 1 + 2\,\mathrm{i}$, then we need to determine which singularity is closest. We compute $|\, \mathrm{i} - z_0 \,| = |\, -1 - \mathrm{i} \,| = \sqrt{2}$, while $|\, -\mathrm{i} - z_0 \,| = |\, -1 - 3\,\mathrm{i} \,| = \sqrt{10}$, and so the radius of convergence $\rho = \sqrt{2}$ is the smaller. Thus we can determine the radius of convergence without any explicit formula for its (rather complicated) Taylor expansion at $z_0 = 1 + 2\,\mathrm{i}$.

There are, in fact, only three possible types of singularities of a complex function $f(z)$:

(*i*) *Pole*. A singular point $z = z_0$ is called a *pole* of *order* $n > 0$ if and only if the function

$$h(z) = (z - z_0)^n \, f(z) \tag{16.29}$$

is analytic and nonzero, $h(z_0) \neq 0$, at $z = z_0$. The simplest example of such a function is $f(z) = a\,(z - z_0)^{-n}$ for $a \neq 0$ a complex constant. and

(*ii*) *Branch point*. We have already encountered the two basic types: *algebraic branch points*, such as the function $\sqrt[n]{z}$ at $z_0 = 0$, and *logarithmic branch points* such as $\log z$ at $z_0 = 0$. The *degree* of the branch point is n in the first case and ∞ in the second.

(*iii*) *Essential singularity*. By definition, a singularity is *essential* if it is not a pole or a branch point. The simplest example is the essential singularity at $z_0 = 0$ of the function $e^{1/z}$. Details are left as an Exercise ∎.

Example 16.7. For example, the function

$$f(z) = \frac{e^z}{z^3 - z^2 - 5\,z - 3}$$

has a simple (order 1) pole at $z = 3$ and a double (order 2) pole at $z = -1$. Indeed, factorizing the denominator $z^3 - z^2 - 5\,z - 3 = (z + 1)^2 (z - 3)$, we see that the functions

$$h_1(z) = (z - 3)\, f(z) = \frac{e^z}{(z + 1)^2}\,, \qquad\qquad h_2(z) = (z + 1)^2\, f(z) = \frac{e^z}{z - 3}\,,$$

are analytic and non-zero at, respectively, $z = 3$ and $z = -1$.

A complicated complex function can have a variety of singularities. For example, the function

$$f(z) = \frac{\sqrt[3]{z + 2}\; e^{-1/z}}{z^2 + 1} \tag{16.30}$$

has simple poles at $z = \pm \mathrm{i}$, a branch point of degree 3 at $z = -2$, and an essential singularity at $z = 0$.

As in the real case, and unlike Fourier series, convergent power series can always be repeatedly term-wise differentiated. Therefore, given the convergent series (16.26), we have

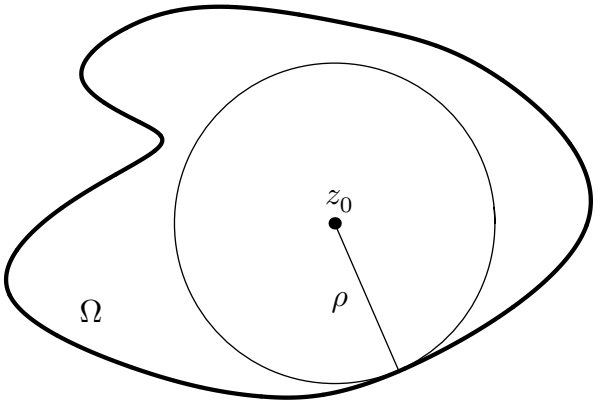


Figure 16.7. Radius of Convergence.

the corresponding series

$$f'(z) = a_1 + 2\,a_2(z - z_0) + 3\,a_3(z - z_0)^2 + 4\,a_4(z - z_0)^3 + \cdots = \sum_{n=0}^{\infty} (n+1)\,a_{n+1}(z - z_0)^n,$$

$$f''(z) = 2\,a_2 + 6\,a_3(z - z_0) + 12\,a_4(z - z_0)^2 + 20\,a_5(z - z_0)^3 + \cdots = \sum_{n=0}^{\infty} (n+1)(n+2)\,a_{n+2}(z - z_0)^n, \qquad (16.31)$$

and so on, for its derivatives. The proof that the differentiated series have the same radius of convergence can be found in [**4**, **154**]. As a consequence, we deduce the following important result.

Theorem 16.8. *Any analytic function is infinitely differentiable.*

In particular, when we substitute $z = z_0$ into the successively differentiated series, we discover that

$$a_0 = f(z_0), \qquad a_1 = f'(z_0), \qquad a_2 = \tfrac{1}{2}\,f''(z_0),$$

and, in general,

$$a_n = \frac{f^{(n)}(z)}{n!}\,. \qquad (16.32)$$

Therefore, a convergent power series (16.26) is, inevitably, the usual *Taylor series*

$$f(z) = \sum_{n=0}^{\infty} \frac{f^{(n)}(z_0)}{n!}\,(z - z_0)^n\,, \qquad (16.33)$$

for the function $f(z)$ at the point z_0.

Let us conclude this section by summarizing the fundamental theorem that characterizes complex functions. A complete, rigorous proof relies on complex integration theory, which is the topic of Section 16.5.

Theorem 16.9. *Let $\Omega \subset \mathbb{C}$ be an open set. The following properties are equivalent:*

(a) *The function $f(z)$ has a continuous complex derivative $f'(z)$ for all $z \in \Omega$.*

(b) *The real and imaginary parts of $f(z)$ have continuous partial derivatives and satisfy the Cauchy–Riemann equations* (16.22) *in Ω.*

(c) *The function $f(z)$ is analytic for all $z \in \Omega$, and so is infinitely differentiable and has a convergent power series expansion at each point $z_0 \in \Omega$. The radius of convergence ρ is at least as large as the distance from z_0 to the boundary $\partial\Omega$; see Figure 16.7.*

From now on, we reserve the term *complex function* to signifiy one that satisfies the conditions of Theorem 16.9. Sometimes one of the equivalent adjectives "analytic" or "holomorphic", is added for emphasis. From now on, all complex functions are assumed to be analytic everywhere on their domain of definition, except, possibly, at certain isolated singularities.

Remark: Since an analytic function is uniquely prescribed by its power series at a point, two analytic function which are the same on any common open set (e.g., a small disk) in their domain of definition must agree everywhere in the connected component of the intersection of their domains. However, they may not agree elsewhere. ■ Analytic continuation.

16.3. Harmonic Functions.

We began this section by motivating the analysis of complex functions through applications to the solution of the two-dimensional Laplace equation. Let us now formalize the precise relationship between the two subjects.

Theorem 16.10. *If $f(z) = u(x,y) + \mathrm{i}\, v(x,y)$ is any complex analytic function, then its real and imaginary parts, $u(x,y), v(x,y)$, are both harmonic functions.*

Proof: Differentiating† the Cauchy–Riemann equations (16.22), and invoking the equality of mixed partial derivatives, we find that

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}\right) = \frac{\partial}{\partial x}\left(\frac{\partial v}{\partial y}\right) = \frac{\partial^2 v}{\partial x\, \partial y} = \frac{\partial}{\partial y}\left(\frac{\partial v}{\partial x}\right) = \frac{\partial}{\partial y}\left(-\frac{\partial u}{\partial y}\right) = -\frac{\partial^2 u}{\partial y^2}\,.$$

Therefore, u is a solution to the Laplace equation $u_{xx} + u_{yy} = 0$. The proof for v is similar. *Q.E.D.*

Thus, every complex function gives rise to two harmonic functions. It is, of course, of interest to know whether we can invert this procedure. Given a harmonic function $u(x,y)$, does there exist a harmonic function $v(x,y)$ such that $f = u + \mathrm{i}\, v$ is a complex analytic function? If so, the harmonic function $v(x,y)$ is known as a *harmonic conjugate* to u. The harmonic conjugate is found by solving the Cauchy–Riemann equations

$$\frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y}\,, \qquad \frac{\partial v}{\partial y} = \frac{\partial u}{\partial x}\,, \tag{16.34}$$

† Theorem 16.9 allows us to differentiate u and v as often as desired.

which, for a prescribed function $u(x,y)$, constitutes an inhomogeneous linear system of partial differential equations for $v(x,y)$. As such, it is usually not hard to solve, as the following example illustrates.

Example 16.11. As the reader can verify, the harmonic polynomial

$$u(x,y) = x^3 - 3x^2y - 3xy^2 + y^3$$

satisfies the Laplace equation everywhere. To find a harmonic conjugate, we solve the Cauchy–Riemann equations (16.34). First of all,

$$\frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y} = 3x^2 + 6xy - 3y^2,$$

and hence, by direct integration with respect to x,

$$v(x,y) = x^3 + 3x^2y - 3xy^2 + h(y),$$

where $h(y)$ — the "constant of integration" — is a function of y alone. To determine h we substitute our formula into the second Cauchy–Riemann equation:

$$3x^2 - 6xy + h'(y) = \frac{\partial v}{\partial y} = \frac{\partial u}{\partial x} = 3x^2 - 6xy - 3y^2.$$

Therefore, $h'(y) = -3y^2$, and so $h(y) = -y^3 + c$, where c is a real constant. We conclude that every harmonic conjugate to $u(x,y)$ has the form

$$v(x,y) = x^3 + 3x^2y - 3xy^2 - y^3 + c.$$

Note that the corresponding complex function

$$\begin{aligned} u(x,y) + \mathrm{i}\, v(x,y) &= (x^3 - 3x^2y - 3xy^2 + y^3) + \mathrm{i}\,(x^3 + 3x^2y - 3xy^2 - y^3 + c) \\ &= (1 - \mathrm{i}\,)z^3 + c \end{aligned}$$

turns out to be a complex cubic polynomial.

Remark: On a connected domain, all harmonic conjugates to a given function $u(x,y)$ only differ by a constant: $\widetilde{v}(x,y) = v(x,y) + c$; see Exercise ■.

Although most harmonic functions have harmonic conjugates, unfortunately this is not always the case. Interestingly, the existence or non-existence of a harmonic conjugate can depend on the underlying geometry of the domain of definition of the function. If the domain is simply-connected, and so contains no holes, then one can *always* find a harmonic conjugate. Otherwise, if the domain of definition Ω of our harmonic function $u(x,y)$ is not simply-connected, then there may not exist a single-valued harmonic conjugate $v(x,y)$ to serve as the imaginary part of a complex function $f(z)$.

Example 16.12. The simplest example where the latter possibility occurs is the logarithmic potential

$$u(x,y) = \log r = \tfrac{1}{2}\log(x^2 + y^2).$$

This function is harmonic on the non-simply-connected domain $\Omega = \mathbb{C} \setminus \{0\}$, but it is not the real part of any single-valued complex function. Indeed, according to (16.18), the logarithmic potential is the real part of the multiply-valued complex logarithm $\log z$, and so its harmonic conjugate[†] is $\operatorname{ph} z = \theta$, which cannot be consistently and continuously defined on all of Ω. On the other hand, restricting z to a simply connected subdomain $\widetilde{\Omega} \not\ni 0$ allows us to select a continuous, single-valued branch of the angle $\theta = \operatorname{ph} z$, and so $\log r$ does have a genuine harmonic conjugate on $\widetilde{\Omega}$.

The harmonic function

$$u(x,y) = \frac{x}{x^2+y^2}$$

is also defined on the same non-simply-connected domain $\Omega = \mathbb{C} \setminus \{0\}$ with a singularity at $x = y = 0$. In this case, there is a single valued harmonic conjugate, namely

$$v(x,y) = -\frac{y}{x^2+y^2}\,,$$

which is defined on all of Ω. Indeed, according to (16.12), these functions define the real and imaginary parts of the complex function $u + \mathrm{i}\, v = 1/z$. Alternatively, one can directly check that they satisfy the Cauchy–Riemann equations (16.22).

Remark: On the "punctured" plane $\Omega = \mathbb{C} \setminus \{0\}$, the logarithmic potential is, in a sense, the only counterexample that prevents a harmonic conjugate from being constructed. It can be shown, [**XC**], that if $u(x,y)$ is a harmonic function defined on a punctured disk $\Omega_R = \big\{\, 0 < |\, z \,| < R \,\big\}$, where $0 < R \leq \infty$, then there exists a constant c such that $\widetilde{u}(x,y) = u(x,y) - c \log r$ is also harmonic and possess a single-valued harmonic conjugate $\widetilde{v}(x,y)$. As a result, the function $\widetilde{f} = \widetilde{u} + \mathrm{i}\,\widetilde{v}$ is analytic on all of Ω_R, and so our original function $u(x,y)$ is the real part of the multiply-valued analytic function $f(z) = \widetilde{f}(z) + c \log z$. We shall use this fact in our later analysis of airfoils.

Theorem 16.13. *Every harmonic function $u(x,y)$ defined on a simply-connected domain Ω is the real part of a complex valued function $f(z) = u(x,y) + \mathrm{i}\, v(x,y)$ which is defined for all $z = x + \mathrm{i}\, y \in \Omega$.*

Proof: We first rewrite the Cauchy–Riemann equations (16.34) in vectorial form as an equation for the gradient of v:

$$\nabla v = \nabla^{\perp} u, \qquad \text{where} \qquad \nabla^{\perp} u = \begin{pmatrix} -\,u_y \\ u_x \end{pmatrix} \tag{16.35}$$

is the vector field that is everywhere orthogonal to the gradient of u and of the same length:

$$\nabla u \cdot \nabla^{\perp} u = 0, \qquad \| \nabla^{\perp} u \| = \| \nabla u \|.$$

[†] We can, by a previous remark, add in any constant to the harmonic conjugate, but this does not affect the subsequent argument.

These properties along with the right hand rule serve to uniquely characterize $\nabla u^\perp$. Thus, the gradient of a harmonic function and that of its harmonic conjugate are mutually orthogonal vector fields having the same Euclidean lengths:

$$\nabla u \cdot \nabla v \equiv 0, \qquad \| \nabla u \| \equiv \| \nabla v \|. \tag{16.36}$$

Now, according to Theorem A.8, provided we work on a simply-connected domain, the gradient equation

$$\nabla v = \mathbf{f} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}$$

has a solution if and only if the vector field $\mathbf{f}$ satisfies the curl-free constraint

$$\nabla \wedge \mathbf{f} = \frac{\partial f_2}{\partial x} - \frac{\partial f_1}{\partial y} \equiv 0.$$

In our specific case, the curl of the perpendicular vector field $\nabla u^\perp$ coincides with the divergence of ∇u, which, in turn, coincides with the Laplacian:

$$\nabla \wedge \nabla u^\perp = \nabla \cdot \nabla u = \Delta u = 0, \quad \text{i.e.,} \quad \frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}\right) - \frac{\partial}{\partial y}\left(-\frac{\partial u}{\partial y}\right) = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0.$$

The result is zero because we are assuming that u is harmonic. Equation (A.41) permits us to reconstruct the harmonic conjugate $v(x, y)$ from its gradient ∇v through line integration

$$v(x, y) = \int_C \nabla v \cdot d\mathbf{x} = \int_C \nabla u^\perp \cdot d\mathbf{x} = \int_C \nabla u \cdot \mathbf{n}\, ds, \tag{16.37}$$

where C is any curve connecting a fixed point (x_0, y_0) to (x, y). Therefore, the harmonic conjugate to a given potential function u can be obtained by evaluating its (path-independent) flux integral (16.37). *Q.E.D.*

Remark: As a consequence of (16.23) and the Cauchy–Riemann equations (16.34),

$$f'(z) = \frac{\partial u}{\partial x} - \mathrm{i}\, \frac{\partial u}{\partial y} = \frac{\partial v}{\partial y} + \mathrm{i}\, \frac{\partial v}{\partial x}\,. \tag{16.38}$$

Thus, the individual components of the gradients ∇u and ∇v appear as the real and imaginary parts of the complex derivative $f'(z)$.

The orthogonality (16.35) of the gradient of a function and of its harmonic conjugate has the following important geometric consequence. Recall, Theorem A.14, that the gradient ∇u of a function $u(x, y)$ points in the normal direction to its *level curves*, that is, the sets $\{\, u(x, y) = c \,\}$ where it assumes a fixed constant value. Since ∇v is orthogonal to ∇u, this must mean that ∇v is *tangent* to the level curves of u. Vice versa, ∇v is normal to its level curves, and so ∇u is tangent to the level curves of its harmonic conjugate v. Since their tangent directions ∇u and ∇v are orthogonal, the level curves of the real and imaginary parts of a complex function form a mutually orthogonal system of plane curves — but with one key exception. If we are at a *critical point*, where $\nabla u = \mathbf{0}$, then $\nabla v = \nabla u^\perp = \mathbf{0}$, and the vectors do not define tangent directions. Therefore, the orthogonality of the level

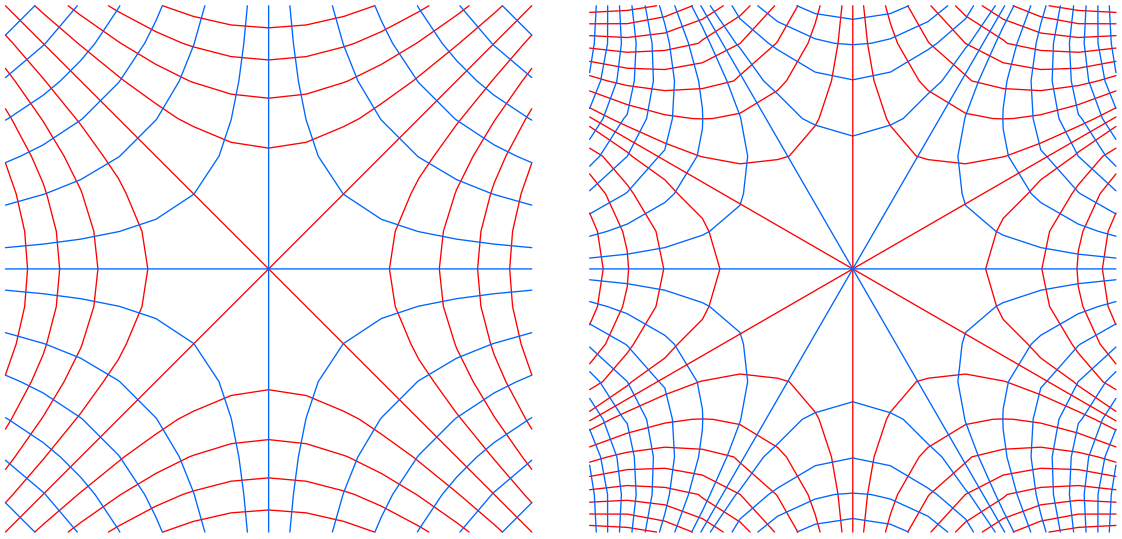


Figure 16.8. Level Curves of the Real and Imaginary Parts of z^2 and z^3.

curves does not necessarily hold at critical points. It is worth pointing out that, in view of (16.38), the critical points of u are the same as those of v and also the same as the critical points of the corresponding complex function $f(z)$, i.e., those points where its complex derivative vanishes: $f'(z) = 0$.

In Figure 16.8, we illustrate the preceding discussion by plotting the level curves of the real and imaginary parts of the monomials z^2 and z^3. Note that, except at the origin, where the derivative vanishes, the level curves intersect everywhere at right angles.

Applications to Fluid Mechanics

Consider a planar[†] steady state fluid flow, with velocity vector field

$$\mathbf{v}(\mathbf{x}) = \begin{pmatrix} u(x,y) \\ v(x,y) \end{pmatrix} \qquad \text{at the point} \quad \mathbf{x} = (x,y) \in \Omega.$$

Here $\Omega \subset \mathbb{R}^2$ is the domain occupied by the fluid, while the vector $\mathbf{v}(\mathbf{x})$ represents the instantaneous velocity of the fluid at the point $\mathbf{x}$. Recall that the flow is *incompressible* if and only if it has vanishing divergence:

$$\nabla \cdot \mathbf{v} = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0. \tag{16.39}$$

Incompressibility means that the fluid volume does not change as it flows. Most liquids, including water, are, for all practical purposes, incompressible. On the other hand, the flow is *irrotational* if and only if it has vanishing curl:

$$\nabla \wedge \mathbf{v} = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} = 0. \tag{16.40}$$

[†] See the remarks in Appendix A on the interpretation of a planar fluid flow as the cross-section of a fully three-dimensional fluid motion that does not depend upon the vertical coordinate.

Irrotational flows has no vorticity or circulation and model fluids in non-turbulent conditions. In many physical situations, the flow of liquids (and, although less often, gases) is both incompressible and irrotational, which for short, is designated an *ideal fluid flow*.

The two constraints (16.39–40) are almost identical to the Cauchy–Riemann equations (16.22)! The only difference is the change in sign in front of the derivatives of v, but this can be easily remedied by replacing v by its negative $-v$. As a result, we deduce a profound connection between ideal planar fluid flows and complex functions.

Theorem 16.14. *The vector field* $\mathbf{v} = (\, u(x,y), v(x,y)\,)^T$ *is the velocity vector of an ideal fluid flow if and only if*

$$f(z) = u(x,y) - \mathrm{i}\, v(x,y) \tag{16.41}$$

is a complex analytic function of $z = x + \mathrm{i}\, y$.

Thus, the components $u(x,y)$ and $-v(x,y)$ of the velocity vector field for an ideal fluid are harmonic conjugates. The complex function (16.41) is known as the *complex velocity* of the fluid flow. When applying this result, *do not forget* the minus sign that appears in front of the imaginary part of $f(z)$.

As discussed in Example A.7, the fluid particles will follow the trajectories $z(t) = x(t) + \mathrm{i}\, y(t)$ obtained by integrating the differential equations

$$\frac{dx}{dt} = u(x,y), \qquad \frac{dy}{dt} = v(x,y). \tag{16.42}$$

In view of the representation (16.41), we can rewrite the system in complex form

$$\frac{dz}{dt} = \overline{f(z)}\,. \tag{16.43}$$

In fluid mechanics, the curves[†] parametrized by $z(t)$ are known as the *streamlines*. Each fluid particle's motion $z(t)$ is uniquely prescribed by its position $z(t_0) = z_0 = x_0 + \mathrm{i}\, y_0$ at an initial time t_0. In particular, if the complex velocity vanishes, $f(z_0) = 0$, then the solution $z(t) \equiv z_0$ to (16.43) is constant, and hence z_0 is a *stagnation point* of the flow. Our steady state assumption, which is reflected in the fact that the ordinary differential equations (16.42) are autonomous, i.e., there is no explicit t dependence, means that, although the fluid is in motion, the stream lines and stagnation point do not change over time. This is a consequence of the standard existence and uniqueness theorems for solutions to ordinary differential equations, to be discussed in detail in Chapter 20.

Example 16.15. The simplest example is when the velocity is constant, corresponding to a uniform, steady flow. Consider first the case

$$f(z) = 1,$$

[†] See below for more details on complex curves.

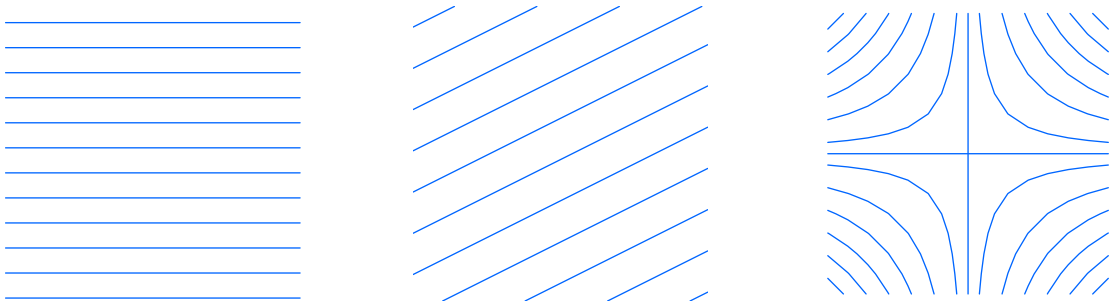


Figure 16.9. Complex Fluid Flows.

which corresponds to the horizontal velocity vector field $\mathbf{v} = (\,1, 0\,)^T$. The actual fluid flow is found by integrating the system

$$\dot{z} = 1, \qquad \text{or} \qquad \dot{x} = 1, \qquad \dot{y} = 0.$$

Thus, the solution $z(t) = t + z_0$ represents a uniform horizontal fluid motion whose streamlines are straight lines parallel to the real axis; see Figure 16.9.

Consider next a more general constant velocity

$$f(z) = c = a + \mathrm{i}\, b.$$

The fluid particles will solve the ordinary differential equation

$$\dot{z} = \overline{c} = a - \mathrm{i}\, b, \qquad \text{so that} \qquad z(t) = \overline{c}\, t + z_0.$$

The streamlines remain parallel straight lines, but now at an angle $\theta = \mathrm{ph}\, \overline{c} = -\mathrm{ph}\, c$ with the horizontal. The fluid particles move along the streamlines at constant speed $|\,\overline{c}\,| = |\,c\,|$.

The next simplest complex velocity function is

$$f(z) = z = x + \mathrm{i}\, y. \tag{16.44}$$

The corresponding fluid flow is found by integrating the system

$$\dot{z} = \overline{z}, \qquad \text{or, in real form,} \qquad \dot{x} = x, \qquad \dot{y} = -\,y.$$

The origin $x = y = 0$ is a stagnation point. The trajectories of the nonstationary solutions

$$z(t) = x_0\, e^t + \mathrm{i}\, y_0\, e^{-t} \tag{16.45}$$

are the hyperbolas $x\,y = c$, and the positive and negative coordinate semi-axes, as illustrated in Figure 16.9.

On the other hand, if we choose

$$f(z) = -\,\mathrm{i}\, z = y - \mathrm{i}\, x,$$

then the flow is the solution to

$$\dot{z} = \mathrm{i}\, \overline{z}, \qquad \text{or, in real form,} \qquad \dot{x} = y, \qquad \dot{y} = x.$$

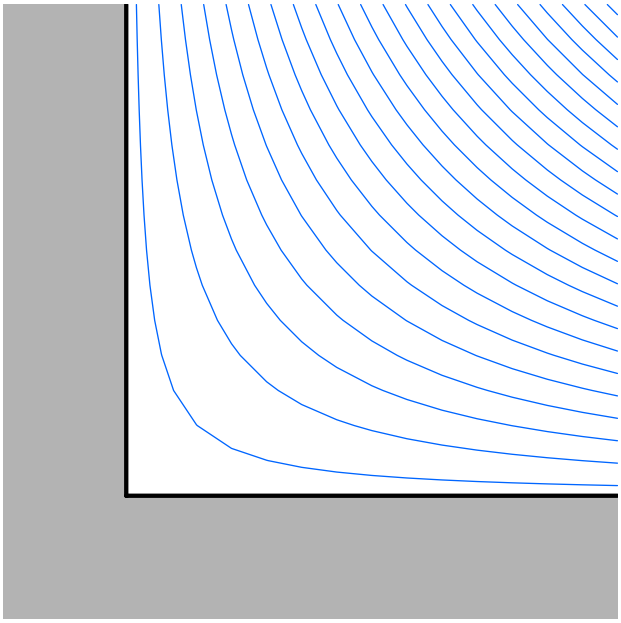


Figure 16.10. Flow Inside a Corner.

The solutions

$$z(t) = (x_0 \cosh t + y_0 \sinh t) + \mathrm{i}\,(x_0 \sinh t + y_0 \cosh t),$$

move along the hyperbolas (and rays) $x^2 - y^2 = c^2$. Observe that this flow can be obtained by rotating the preceding example by 45°.

In general, a solid object in a fluid flow is characterized by the no-flux condition that the fluid velocity $\mathbf{v}$ is everywhere tangent to the boundary, and hence no fluid flows into or out of the object. As a result, the boundary will consist of streamlines and stagnation points of the idealized fluid flow. For example, the boundary of the upper right quadrant $Q = \{x > 0, y > 0\} \subset \mathbb{C}$ consists of the positive x and y axes (along with the origin). Since these are streamlines of the flow with complex velocity (16.44), its restriction to Q represents the flow past a 90° interior corner, which appears in Figure 16.10. The fluid particles move along hyperbolas as they flow past the corner.

Remark: We could also restrict this flow to the domain $\Omega = \mathbb{C} \setminus \{x < 0, y < 0\}$ consisting of three quadrants, corresponding to a 90° exterior corner. However, this flow is not as physically relevant since it has an unrealistic asymptotic behavior at large distances. See Exercise ■ for the "correct" physical flow around an exterior corner.

Now, suppose that the complex velocity $f(z)$ admits a complex anti-derivative, i.e., a complex analytic function

$$\chi(z) = \varphi(x,y) + \mathrm{i}\,\psi(x,y) \qquad \text{that satisfies} \qquad \frac{d\chi}{dz} = f(z). \tag{16.46}$$

Using the formula (16.23) for the complex derivative,

$$\frac{d\chi}{dz} = \frac{\partial \varphi}{\partial x} - \mathrm{i}\,\frac{\partial \varphi}{\partial y} = u - \mathrm{i}\,v, \qquad \text{so} \qquad \frac{\partial \varphi}{\partial x} = u, \qquad \frac{\partial \varphi}{\partial y} = v.$$

Thus, $\nabla\varphi = \mathbf{v}$, and hence the real part $\varphi(x,y)$ of the complex function $\chi(z)$ defines a *velocity potential* for the fluid flow. For this reason, the anti-derivative $\chi(z)$ is known as a *complex potential function* for the given fluid velocity field.

Since the complex potential is analytic, its real part — the potential function — is harmonic, and therefore satisfies the Laplace equation $\Delta\varphi = 0$. Conversely, any harmonic function can be viewed as the potential function for some fluid flow. The real fluid velocity is its gradient $\mathbf{v} = \nabla\varphi$. The harmonic conjugate $\psi(x,y)$ to the velocity potential also plays an important role, and, in fluid mechanics, is known as the *stream function*. It also satisfies the Laplace equation $\Delta\psi = 0$, and the potential and stream function are related by the Cauchy–Riemann equations (16.22). Thus, the potential and stream function satisfy

$$\frac{\partial\varphi}{\partial x} = u = \frac{\partial\psi}{\partial y}\,, \qquad \frac{\partial\varphi}{\partial y} = v = -\,\frac{\partial\psi}{\partial x}\,. \tag{16.47}$$

The level sets of the velocity potential, $\{\,\varphi(x,y) = c\,\}$ where $c \in \mathbb{R}$ is fixed, are known as *equipotential curves*. The velocity vector $\mathbf{v} = \nabla\varphi$ points in the normal direction to the equipotentials. On the other hand, as we noted above, $\mathbf{v} = \nabla\varphi$ is tangent to the level curves $\{\,\psi(x,y) = d\,\}$ of its harmonic conjugate stream function. But $\mathbf{v}$ is the velocity field, and so tangent to the streamlines followed by the fluid particles. Thus, these two systems of curves must coincide, and we infer that *the level curves of the stream function are the streamlines of the flow*, whence its name! Summarizing, for an ideal fluid flow, the equipotentials $\{\,\varphi = c\,\}$ and streamlines $\{\,\psi = d\,\}$ form mutually orthogonal systems of plane curves. The fluid velocity $\mathbf{v} = \nabla\varphi$ is tangent to the stream lines and normal to the equipotentials, whereas the gradient of the stream function $\nabla\psi$ is tangent to the equipotentials and normal to the streamlines.

The discussion in the preceding paragraph implicitly relied on the fact that the velocity is nonzero, $\mathbf{v} = \nabla\varphi \neq 0$, which means we are not at a *stagnation point*, where the fluid is not moving. While streamlines and equipotentials might begin or end at a stagnation point, there is no guarantee, and, indeed, in general it is not the case that they meet at mutually orthogonal directions there.

Example 16.16. The simplest example of a complex potential function is

$$\chi(z) = z = x + \mathrm{i}\,y.$$

Thus, the velocity potential is $\varphi(x,y) = x$, while its harmonic conjugate stream function is $\psi(x,y) = y$. The complex derivative of the potential is the complex velocity,

$$f(z) = \frac{d\chi}{dz} = 1,$$

which corresponds to the uniform horizontal fluid motion considered first in Example 16.15. Note that the horizontal stream lines coincide with the level sets $\{\,y = d\,\}$ of the stream function, whereas the equipotentials $\{\,x = c\,\}$ are the orthogonal system of vertical lines; see Figure 16.11.

Next, consider the complex potential function

$$\chi(z) = \tfrac{1}{2}\,z^2 = \tfrac{1}{2}\,(x^2 - y^2) + \mathrm{i}\,x\,y.$$

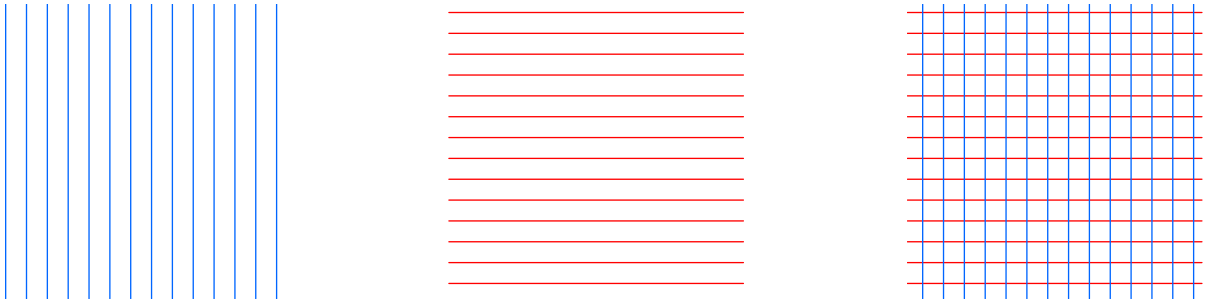


Figure 16.11. Equipotentials and Streamlines for $\chi(z) = z$.

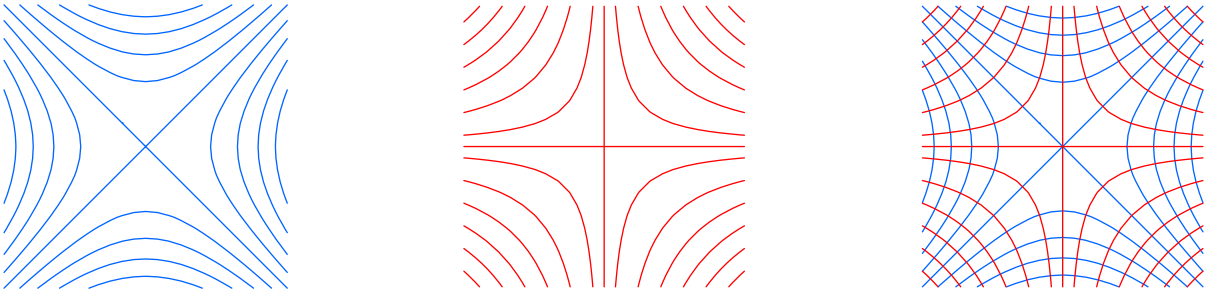


Figure 16.12. Equipotentials and Streamlines for $\chi(z) = \frac{1}{2}\, z^2$.

The associated complex velocity

$$f(z) = \chi'(z) = z = x + \mathrm{i}\, y$$

leads to the hyperbolic flow (16.45). The hyperbolic streamlines $x\, y = d$ are the level curves of the stream function $\psi(x, y) = x\, y$. The equipotential lines $\frac{1}{2}\,(x^2 - y^2) = c$ form a system of orthogonal hyperbolas. Figure 16.12 shows (some of) the equipotentials in the first plot, the stream lines in the second, and combines them together in the third picture.

Example 16.17. *Flow Around a Disk.* Consider the complex potential function

$$\chi(z) = z + \frac{1}{z} = \left(x + \frac{x}{x^2 + y^2} \right) + \mathrm{i} \left(y - \frac{y}{x^2 + y^2} \right). \tag{16.48}$$

The corresponding complex fluid velocity is

$$f(z) = \frac{d\chi}{dz} = 1 - \frac{1}{z^2} = 1 - \frac{x^2 - y^2}{(x^2 + y^2)^2} \;+\; \mathrm{i}\, \frac{2\, x\, y}{(x^2 + y^2)^2}\,. \tag{16.49}$$

The equipotential curves and streamlines are plotted in Figure 16.13. The points $z = \pm 1$ are stagnation points of the flow, while $z = 0$ is a singularity. In particular, fluid particles that move along the positive x axis approach the leading stagnation point $z = 1$ as $t \to \infty$. Note that the streamlines

$$\psi(x, y) = y - \frac{y}{x^2 + y^2} = d$$

are asymptotically horizontal at large distances, and hence, far away from the origin, the flow is indistinguishable from uniform horizontal motion with complex velocity $f(z) \equiv 1$.

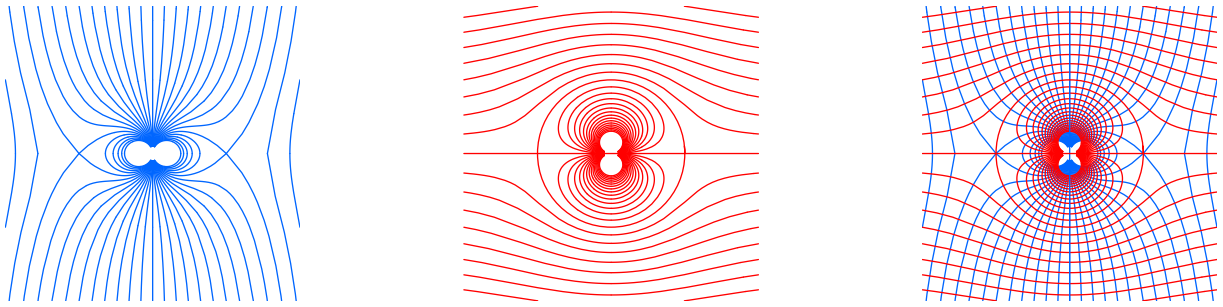


Figure 16.13. Equipotentials and Streamlines for $z + \dfrac{1}{z}$.

The level curve for the particular value $d = 0$ consists of the unit circle $|\, z \,| = 1$ and the real axis $y = 0$. In particular, the unit circle $|\, z \,| = 1$ consists of semicircular two stream lines and the two stagnation points. The flow velocity vector field $\mathbf{v} = \nabla\varphi$ is everywhere tangent to the unit circle, and hence satisfies the no flux condition along the boundary of the unit disk. Thus, we can interpret (16.49), when restricted to the domain $\Omega = \big\{\, |\, z \,| > 1 \,\big\}$, as the complex velocity of a uniformly moving fluid around the outside of a solid circular disk of radius 1. In three dimensions, this would correspond to the steady flow of a fluid around a solid cylinder; see Figure fcyl■.

In this section, we have focused on the fluid mechanical roles of a harmonic function and its conjugate. An analogous interpretation applies when $\varphi(x, y)$ represents an electromagnetic potential function; the level curves of its harmonic conjugate $\psi(x, y)$ are the paths followed by charged particles under the electromotive force field $\mathbf{v} = \nabla\varphi$. Similarly, if $\varphi(x, y)$ represents the equilibrium temperature distribution in a planar domain, its level lines represent the isotherms or curves of constant temperature, while the level lines of its harmonic conjugate are the curves of heat flow, whose mutual orthogonality was already noted in Appendix A. Finally, if $\varphi(x, y)$ represents the height of a deformed membrane, then its level curves are the contour lines of elevation. The level curves of its harmonic conjugate are the curves of steepest descent along the membrane, i.e., the routes followed by, say, water flowing down the membrane.

16.4. Conformal Mapping.

As we now know, complex functions provide an almost inexhaustible supply of harmonic functions, i.e., solutions to the Laplace equation. Thus, to solve a boundary value problem for Laplace's equation we "merely" need to find the complex function whose real part matches the prescribed boundary conditions. Unfortunately, even for relatively simple domains, this remains a daunting task.

The one case where we do have an explicit solution is that of a circular disk, where the Poisson integral formula (15.48) provides a complete solution to the Dirichlet boundary value problem. (See Exercise ■ for the Neumann problem.) Thus, one evident strategy for solving the corresponding boundary value problem on a more complicated domain is to convert it into the solved case by an inspired change of variables.

The intimate connections between complex analysis and solutions to the Laplace equation inspires us to look at changes of variables defined by complex functions. Thus, we

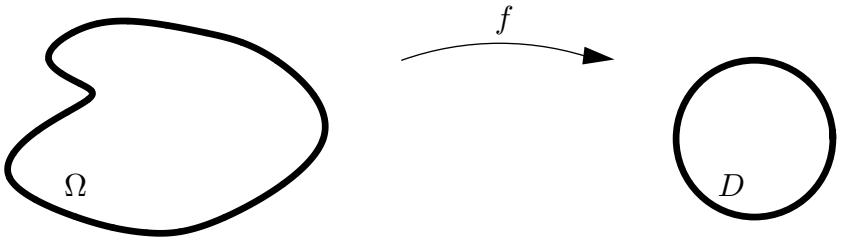


Figure 16.14. Mapping to the Unit Disk.

will now interpret a complex analytic function

$$\zeta = g(z) \qquad \text{or} \qquad \xi + \mathrm{i}\,\eta = p(x,y) + \mathrm{i}\,q(x,y) \tag{16.50}$$

as a *mapping* that takes a point $z = x + \mathrm{i}\,y$ belonging to a prescribed domain $\Omega \subset \mathbb{C}$ to a point $\zeta = \xi + \mathrm{i}\,\eta$ belonging to the image domain $D = g(\Omega) \subset \mathbb{C}$, as in Figure 16.14. In many cases, D is the unit disk, but may be something else in more general examples. In order unambigouously relate functions on Ω to functions on D, we require that the analytic mapping (16.50) be one-to-one so that each point $\zeta \in D$ comes from a unique point $z \in \Omega$. As a result, the inverse function $z = g^{-1}(\zeta)$ is a well-defined map from D back to Ω, which we assume is also analytic on all of D. The calculus formula for the derivative of the inverse function

$$\frac{d}{d\zeta}\, g^{-1}(\zeta) = \frac{1}{g'(z)} \qquad \text{at} \qquad \zeta = g(z), \tag{16.51}$$

which remains valid for complex functions, implies that the derivative of $g(z)$ must be nonzero everywhere in order that $g^{-1}(\zeta)$ be differentiable. This condition,

$$g'(z) \neq 0 \qquad \text{at every point} \qquad z \in \Omega, \tag{16.52}$$

will play a crucial role in the development of the method. Finally, ijn order to match the boundary conditions, we will assume that the mapping extends continuously to the boundary $\partial\Omega$ and maps it to the boundary ∂D of the image domain.

Before trying to apply this idea to solve boundary value problems for the Laplace equation, we introduce some of the most important examples of analytic mappings.

Example 16.18. The simplest nontrivial analytic maps are the *translations*

$$\zeta = z + c = (x + a) + \mathrm{i}\,(y + b), \tag{16.53}$$

where $c = a + \mathrm{i}\,b$ is a fixed cokmplex constant. The effect of (16.53) is to translate the entire complex plane in the direction given by the vector $(\,a, b\,)^T$, a special case of (7.31). In particular, it maps a disk $\Omega = \{\,|\,z + c\,| < 1\,\}$ of radius 1 and center at $-c$ to the unit disk $D = \{\,|\,\zeta\,| < 1\,\}$.

There are two types of linear analytic transformations. First are the scaling maps

$$\zeta = \rho\, z = \rho\, x + \mathrm{i}\,\rho\, y, \tag{16.54}$$

where $\rho \neq 0$ is a fixed nonzero real number. These map the disk $|\, z \,| < 1/|\, \rho \,|$ to the unit disk $|\, \zeta \,| < 1$. Second are the rotations

$$\zeta = e^{\,\mathrm{i}\,\varphi}\, z = (x \cos \varphi - y \sin \varphi) + \mathrm{i}\,(x \sin \varphi + y \cos \varphi) \tag{16.55}$$

around the origin by a fixed (real) angle φ. These map the unit disk to itself.

Any non-constant *affine transformation*

$$\zeta = \alpha\, z + \beta, \qquad\qquad \alpha \neq 0, \tag{16.56}$$

defines an invertible analytic map on all of $\mathbb{C}$, whose inverse $z = \alpha^{-1}(\zeta - \beta)$ is also affine. Writing $\alpha = \rho\, e^{\,\mathrm{i}\,\varphi}$ in polar coordinates, we see that the affine map (16.56) can be built up from a rotation followed by a scaling followed by a translation. As such, it takes the disk $|\, \alpha\, z + \beta \,| < 1$ of radius $1/|\, \alpha \,| = 1/|\, \rho \,|$ and center $-\beta/\alpha$ to the unit disk $|\, \zeta \,| < 1$.

Example 16.19. A more interesting example is the complex function

$$\zeta = g(z) = \frac{1}{z}\,, \qquad \text{or} \qquad \xi = \frac{x}{x^2 + y^2}\,, \qquad \eta = -\,\frac{y}{x^2 + y^2}\,, \tag{16.57}$$

which defines an *inversion*[†] of the complex plane. The inversion is a one-to-one analytic map everywhere except at the origin $z = 0$; indeed $g(z)$ is its own inverse: $g^{-1}(\zeta) = 1/\zeta$. Note that $g'(z) = -1/z^2$ is never zero, and so the derivative condition (16.52) is satisfied everywhere. Note that $|\, \zeta \,| = 1/|\, z \,|$, while $\mathrm{ph}\, \zeta = -\,\mathrm{ph}\, z$. Thus, if $\Omega = \big\{\, |\, z \,| > \rho \,\big\}$ denotes the exterior of the circle of radius ρ, then the image points $\zeta = 1/z$ satisfy $|\, \zeta \,| = 1/|\, z \,|$, and hence the image domain is the *punctured disk* $D = \big\{\, 0 < |\, \zeta \,| < 1/\rho \,\big\}$. In particular, the inversion maps the outside of the unit disk to its inside, but with the origin removed, and vice versa. The reader may enjoy seeing what the inversion does to other domains, e.g., the unit square $0 < x, y < 1$.

Example 16.20. The complex exponential

$$\zeta = g(z) = e^z\,, \qquad\qquad \text{or} \qquad\qquad \xi = e^x \cos y, \qquad \eta = e^x \sin y, \tag{16.58}$$

satisfies the condition

$$g'(z) = e^z \neq 0$$

everywhere. Nevertheless, it is *not* one-to-one because $e^{z + 2\,\pi\,\mathrm{i}} = e^z$, and so all points differing by an integer multiple of $2\,\pi\,\mathrm{i}$ are mapped to the same point.

Under the exponential map, the horizontal line $\mathrm{Im}\; z = b$ is mapped to the curve $\zeta = e^{x + \mathrm{i}\, b} = e^x(\cos b + \mathrm{i} \sin b)$, which, as x varies from $-\infty$ to ∞, traces out the ray emanating from the origin that makes an angle $\mathrm{ph}\, \zeta = b$ with the real axis. Therefore, the exponential map will map a horizontal strip

$$S_{a,b} = \big\{\, a < \mathrm{Im}\; z < b \,\big\}$$

[†] This is slightly different than the real inversion (15.75); see Exercise ■.

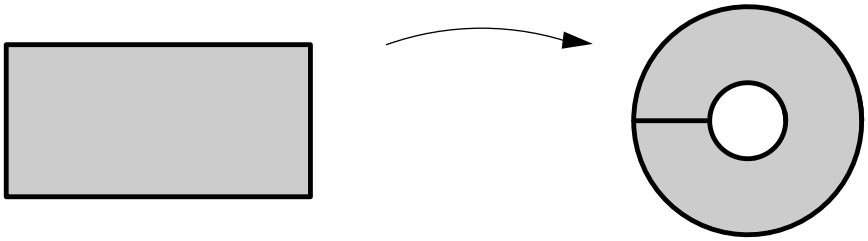


Figure 16.15. The mapping $\zeta = e^z$.

to a wedge-shaped domain

$$\Omega_{a,b} = \big\{\, a < \operatorname{ph} \zeta < b \,\big\},$$

and is one-to-one provided $|\, b - a \,| < 2\pi$. In particular, the horizontal strip

$$S_{-\pi/2,\pi/2} = \big\{\, -\tfrac{1}{2}\pi < \operatorname{Im} z < \tfrac{1}{2}\pi \,\big\}$$

of width π centered around the real axis is mapped, in a one-to-one manner, to the right half plane

$$R = \Omega_{-\pi/2,\pi/2} = \big\{\, -\tfrac{1}{2}\pi < \operatorname{ph} \zeta < \tfrac{1}{2}\pi \,\big\} = \big\{\, \operatorname{Im} \zeta > 0 \,\big\},$$

while the horizontal strip $S_{-\pi,\pi} = \big\{\, -\pi < \operatorname{Im} z < \pi \,\big\}$ of width 2π is mapped onto the domain

$$\Omega_* = \Omega_{-\pi,\pi} = \big\{\, -\pi < \operatorname{ph} \zeta < \pi \,\big\} = \mathbb{C} \setminus \{\operatorname{Im} z = 0,\ \operatorname{Re} z \leq 0\}$$

obtained by cutting the complex plane along the negative real axis.

On the other hand, vertical lines $\operatorname{Re} z = a$ are mapped to circles $|\, \zeta \,| = e^a$. Thus, a vertical strip $a < \operatorname{Re} z < b$ is mapped to an annulus $e^a < |\, \zeta \,| < e^b$, albeit many-to-one, since the strip is effectively wrapped around and around the annulus. The rectangle $R = \big\{\, a < x < b, -\pi < y < \pi \,\big\}$ of height 2π is mapped in a one-to-one fashion on an annulus that has been cut along the negative real axis. See Figure 16.15.

Example 16.21. The squaring map

$$\zeta = g(z) = z^2, \qquad \text{or} \qquad \xi = x^2 - y^2, \quad \eta = 2xy, \tag{16.59}$$

is analytic on all of $\mathbb{C}$, but is not one-to-one. Its inverse is the square root function $z = \sqrt{\zeta}$, which, as we noted in Section 16.1, is doubly-valued, except at the origin $z = 0$. Furthermore, the derivative $g'(z) = 2z$ vanishes at $z = 0$, violating the invertibility condition (16.52). However, once we restrict to a simply connected subdomain Ω that does not contain 0, the function $g(z) = z^2$ does define a one-to-one mapping, whose inverse $z = g^{-1}(\zeta) = \sqrt{\zeta}$ is a well-defined, analytic and single-valued branch of the square root function.

The effect of the squaring map on a point z is to square its modulus, $|\, \zeta \,| = |\, z \,|^2$, while doubling its angle, $\operatorname{ph} \zeta = \operatorname{ph} z^2 = 2 \operatorname{ph} z$. Thus, for example, the upper right quadrant

$$Q = \big\{\, x > 0, y > 0 \,\big\} = \big\{\, 0 < \operatorname{ph} z < \tfrac{1}{2}\pi \,\big\}$$

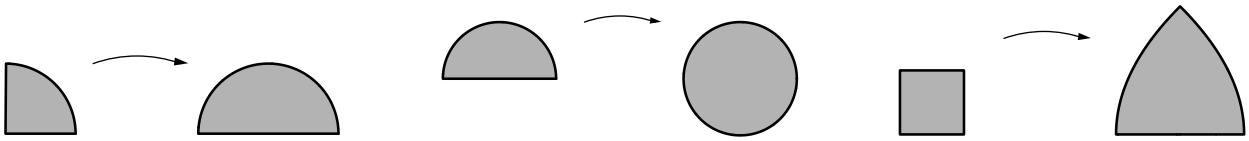


Figure 16.16. The Effect of $\zeta = z^2$ on Various Domains.

is mapped onto the upper half plane

$$U = g(Q) = \big\{\, \eta = \operatorname{Im} \zeta > 0 \,\big\} = \big\{\, 0 < \operatorname{ph} \zeta < \pi \,\big\}.$$

The inverse function maps a point $\zeta \in U$ back to its unique square root $z = \sqrt{\zeta}$ that lies in the quadrant Q. Similarly, a quarter disk

$$Q_\rho = \big\{\, 0 < |\, z \,| < \rho,\ 0 < \operatorname{ph} z < \tfrac{1}{2}\,\pi \,\big\}$$

of radius ρ is mapped to a half disk

$$U_{\rho^2} = g(\Omega) = \big\{\, 0 < |\, \zeta \,| < \rho^2,\ \operatorname{Im} \zeta > 0 \,\big\}$$

of radius ρ^2. On the other hand, the unit square $\Omega = \big\{\, 0 < x < 1, 0 < y < 1 \,\big\}$ is mapped to a curvilinear triangular domain, as indicated in Figure 16.16; the edges of the square on the real and imaginary axes map to the two halves of the straight base of the triangle, while the other two edges become its curved sides.

Example 16.22. A particularly important example is the analytic map

$$\zeta = \frac{z-1}{z+1} = \frac{x^2+y^2-1}{(x+1)^2+y^2} \;+\; \mathrm{i}\, \frac{2\,y}{(x+1)^2+y^2}\,, \tag{16.60}$$

where we used (16.14) to derive the formulae for its real and imaginary parts. The map is one-to-one with analytic inverse

$$z = \frac{1+\zeta}{1-\zeta} = \frac{1-\xi^2-\eta^2}{(1-\xi)^2+\eta^2} \;+\; \mathrm{i}\, \frac{2\,\eta}{(1-\xi)^2+\eta^2}\,, \tag{16.61}$$

provided $z \neq -1$ and $\zeta \neq 1$. This particular analytic map has the important property of mapping the right half plane $R = \big\{\, x = \operatorname{Re} z > 0 \,\big\}$ to the unit disk $D = \big\{\, |\, \zeta \,|^2 < 1 \,\big\}$. Indeed, by (16.61)

$$|\, \zeta \,|^2 = \xi^2 + \eta^2 < 1 \qquad \text{if and only if} \qquad x = \frac{1-\xi^2-\eta^2}{(1-\xi)^2+\eta^2} > 0.$$

Note that the denominator does not vanish on the interior of the disk.

The complex functions (16.56, 57, 60) are particular examples of *linear fractional transformations*

$$\zeta = \frac{\alpha\, z + \beta}{\gamma\, z + \delta}\,, \tag{16.62}$$

which form one of the most important classes of analytic maps. Here $\alpha, \beta, \gamma, \delta$ are arbitrary complex constants, subject to the restriction

$$\alpha\,\delta - \beta\,\gamma \neq 0,$$

since otherwise (16.62) reduces to a trivial constant (and non-invertible) map. (Why?)

Example 16.23. The linear fractional transformation

$$\zeta = \frac{z - \alpha}{\overline{\alpha}\, z - 1} \qquad \text{where} \qquad |\,\alpha\,| < 1, \tag{16.63}$$

maps the unit disk to itself, moving the origin $z = 0$ to the point $\zeta = \alpha$. To prove this, we note that

$$\begin{aligned} |\,z - \alpha\,|^2 &= (z - \alpha)(\overline{z} - \overline{\alpha}) = |\,z\,|^2 - \alpha\,\overline{z} - \overline{\alpha}\, z + |\,\alpha\,|^2, \\ |\,\overline{\alpha}\, z - 1\,|^2 &= (\overline{\alpha}\, z - 1)(\alpha\,\overline{z} - 1) = |\,\alpha\,|^2\,|\,z\,|^2 - \alpha\,\overline{z} - \overline{\alpha}\, z + 1. \end{aligned}$$

Subtracting these two formulae,

$$|\,z - \alpha\,|^2 - |\,\overline{\alpha}\, z - 1\,|^2 = \big(1 - |\,\alpha\,|^2\big)\,\big(|\,z\,|^2 - 1\big) < 0, \quad \text{whenever} \quad |\,z\,| < 1, \quad |\,\alpha\,| < 1.$$

Thus, $|\,z - \alpha\,| < |\,\overline{\alpha}\, z - 1\,|$, which implies that

$$|\,\zeta\,| = \frac{|\,z - \alpha\,|}{|\,\overline{\alpha}\, z - 1\,|} < 1 \qquad \text{provided} \qquad |\,z\,| < 1, \qquad |\,\alpha\,| < 1,$$

and hence ζ lies within the unit disk.

The rotations (16.55) also map the unit disk to itself, while preserving the origin. It can be proved, [**4**], that the only invertible analytic mappings that take the unit disk to itself are obtained by composing such a linear fractional transformation with a rotation.

Proposition 16.24. *If $\zeta = g(z)$ is a one-to-one analytic map that takes the unit disk to itself, then*

$$g(z) = e^{\,\mathrm{i}\,\varphi}\, \frac{z - \alpha}{\overline{\alpha}\, z - 1} \qquad \text{for some} \qquad |\,\alpha\,| < 1, \qquad 0 \leq \varphi < 2\,\pi. \tag{16.64}$$

Additional specific properties of linear fractional transformations are outlined in the exercises. The most important is that they map circles to circles; see Exercise ■ for details.

Conformality

A remarkable geometrical characterization of complex analytic functions is that, at non-critical points, they preserve angles. The mathematical term for this property is *conformal mapping*. Conformality makes sense for any inner product space, although in practice one usually deals with Euclidean space equipped with the standard dot product.

Definition 16.25. A function $g\colon \mathbb{R}^n \to \mathbb{R}^n$ is called *conformal* if it preserves angles.

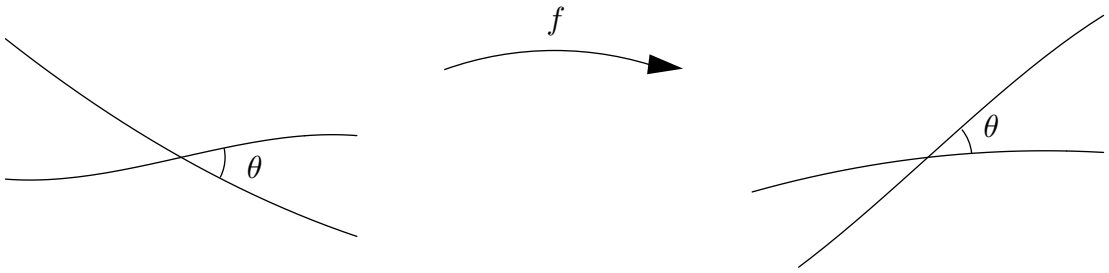


Figure 16.17. A Conformal Map.

But what does it mean to "preserve angles"? In the Euclidean norm, the angle between two vectors is defined by their dot product, as in (3.20). However, most analytic maps are nonlinear, and so will not map vectors to vectors since they will typically map straight lines to curves. However, if we interpret "angle" to mean the angle between two curves, as illustrated in Figure 16.17, then we can make sense of the conformality requirement. Consequently, in order to realize complex functions as conformal maps, we first need to understand their effect on curves.

In general, a *curve* $C \in \mathbb{C}$ in the complex plane is parametrized by a complex-valued function

$$z(t) = x(t) + \mathrm{i}\, y(t), \qquad a < t < b, \tag{16.65}$$

that depends on a real parameter t. Note that there is no essential difference between a complex plane curve (16.65) and a real plane curve, as in (A.1); we have merely switched from vector notation $\mathbf{x}(t) = (\, x(t), y(t)\,)^T$ to complex notation $z(t) = x(t) + \mathrm{i}\, y(t)$. All the usual vectorial curve terminology (closed, simple, piecewise smooth, etc.), as summarized in Appendix A, is used without any modification here. In particular, the *tangent vector* to the curve can be identified as the complex number $\dot z(t) = \dot x(t) + \mathrm{i}\, \dot y(t)$. Smoothness of the curve is guaranteed by the requirement that $\dot z(t) \neq 0$.

Example 16.26. (*a*) The curve

$$z(t) = e^{\,\mathrm{i}\, t} = \cos t + \mathrm{i}\, \sin t, \qquad \text{for} \qquad 0 \leq t \leq 2\,\pi,$$

parametrizes the unit circle $|\, z\,| = 1$ in the complex plane, which is a simple closed curve. Its complex tangent $\dot z(t) = \mathrm{i}\, e^{\,\mathrm{i}\, t} = \mathrm{i}\, z(t)$ is obtained by rotating z through 90°.

(*b*) The complex curve

$$z(t) = \cosh t + \mathrm{i}\, \sinh t = \frac{1 + \mathrm{i}}{2}\, e^t + \frac{1 - \mathrm{i}}{2}\, e^{-\,t}, \qquad -\infty < t < \infty,$$

parametrizes the right hand branch of the hyperbola

$$\mathrm{Re}\, z^2 = x^2 - y^2 = 1.$$

The complex tangent vector is $\dot z(t) = \sinh t + \mathrm{i}\, \cosh t = \mathrm{i}\, \overline{z}(t)$.

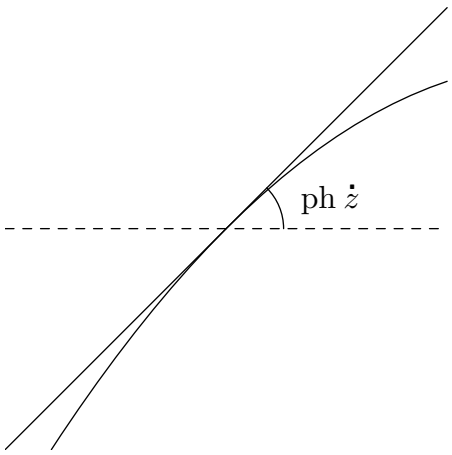


Figure 16.18. Complex Curve and Tangent.

In order to better understand curve geometry, it will help to rewrite the tangent $\dot z$ in polar coordinates. We interpret the curve as the motion of a particle in the complex plane, so that $z(t)$ is the position of the particle at time t, and the tangent $\dot z(t)$ its instantaneous velocity. The modulus of the tangent, $|\,\dot z\,| = \sqrt{\dot x^2 + \dot y^2}$, indicates the particle's speed, while its phase ph $\dot z$ measures the direction of motion, as measured by the angle that the curve makes with the horizontal; see Figure 16.18.

The (signed) angle between two curves is defined as the angle between their tangents at the point of intersection. If the curve C_1 makes an angle $\theta_1 = \operatorname{ph} \dot z_1(t_1)$ while the curve C_2 has angle $\theta_2 = \operatorname{ph} \dot z_2(t_2)$ at the common point $z = z_1(t_1) = z_2(t_2)$, then the angle θ between C_1 and C_2 at z is their difference

$$\theta = \theta_2 - \theta_1 = \operatorname{ph} \dot z_2 - \operatorname{ph} \dot z_1 = \operatorname{ph} \left(\frac{\dot z_2}{\dot z_1} \right) . \tag{16.66}$$

Now, suppose we are given an analytic map $\zeta = g(z)$. A curve C parametrized by $z(t)$ will be mapped to a new curve $\Gamma = g(C)$ parametrized by the composition $\zeta(t) = g(z(t))$. The tangent to the image curve is related to that of the original curve by the chain rule:

$$\frac{d\zeta}{dt} = \frac{dg}{dz}\,\frac{dz}{dt}\,, \qquad \text{or} \qquad \dot\zeta(t) = g'(z(t))\,\dot z(t). \tag{16.67}$$

Therefore, the effect of the analytic map on the tangent vector $\dot z$ is to multiply it by the complex number $g'(z)$. If the analytic map satisfies our key assumption $g'(z) \neq 0$, then $\dot\zeta \neq 0$, and so the image curve is guaranteed to be smooth.

According to equation (16.67),

$$|\,\dot\zeta\,| = |\,g'(z)\,\dot z\,| = |\,g'(z)\,|\,|\,\dot z\,|. \tag{16.68}$$

Thus, the speed of motion along the new curve $\zeta(t)$ is multiplied by a factor $\rho = |\,g'(z)\,| > 0$. The magnification factor ρ depends only upon the point z and not how the curve passes through it. All curves passing through the point z are speeded up (or slowed down if $\rho < 1$)

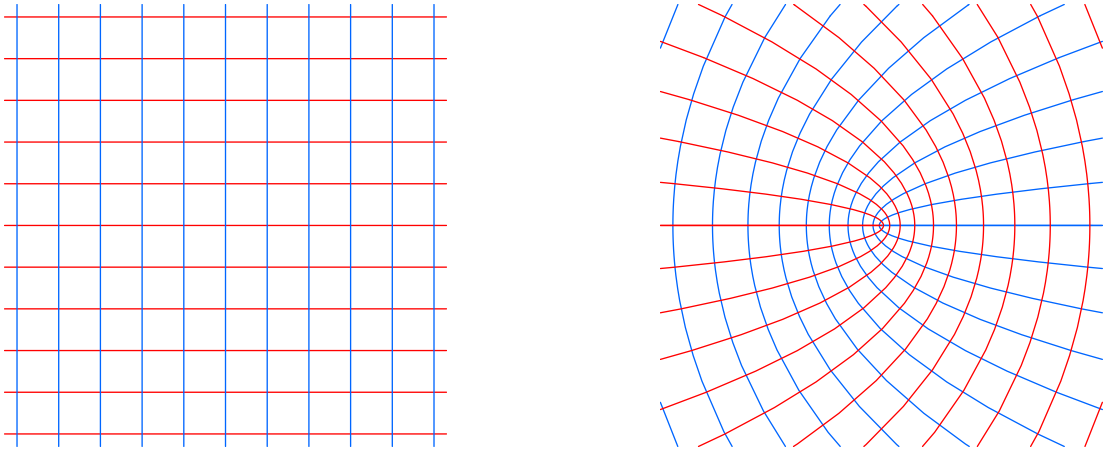


Figure 16.19. Conformality of z^2.

by the same factor! Similarly, the angle that the new curve makes with the horizontal is given by

$$\operatorname{ph} \dot{\zeta} = \operatorname{ph}\big(g'(z)\, \dot{z} \big) = \operatorname{ph} g'(z) + \operatorname{ph} \dot{z}, \tag{16.69}$$

since the phase of the product of two complex numbers is the sum of their individual phases, (3.82). Therefore, the tangent angle of the curve is increased by an amount $\phi = \operatorname{ph} g'(z)$, which means that the tangent is been rotated through an angle ϕ. Again, the increase in tangent angle only depends on the point z, and all curves passing through z are rotated by the same amount ϕ. As a result, the angle between any two curves is preserved. More precisely, if C_1 is at angle θ_1 and C_2 at angle θ_2 at a point of intersection, then their images $\Gamma_1 = g(C_1)$ and $\Gamma_2 = g(C_2)$ are at angles $\psi_1 = \theta_1 + \phi$ and $\psi_2 = \theta_2 + \phi$. The angle between the two image curves is the difference

$$\psi_2 - \psi_1 = (\theta_2 + \phi) - (\theta_1 + \phi) = \theta_2 - \theta_1,$$

which is *the same* as the angle between the original curves. This establishes the conformality or angle-preservation property of analytic maps.

Theorem 16.27. *If $\zeta = g(z)$ is an analytic function and $g'(z) \neq 0$, then g defines a conformal map.*

Remark: The converse is also valid: Every planar conformal map comes from a complex analytic function with nonvanishing derivative. A proof is outlined in Exercise ∎.

The conformality of analytic functions is all the more surprising when one revisits elementary examples. In Example 16.21, we discovered that the function $w = z^2$ maps a quarter plane to a half plane, and therefore *doubles* the angle between the coordinate axes at the origin! Thus $g(z) = z^2$ is most definitely not conformal at $z = 0$. The explanation is, of course, that $z = 0$ is a critical point, $g'(0) = 0$, and Theorem 16.27 only guarantees conformality when the derivative is nonzero. Amazingly, the map preserves angles everywhere else! Somehow, the angle at the origin is doubled, while angles at all nearby points are preserved. Figure 16.19 illustrates this remarkable and counter-intuitive feat. The left hand figure shows the coordinate grid, while on the right are the images of

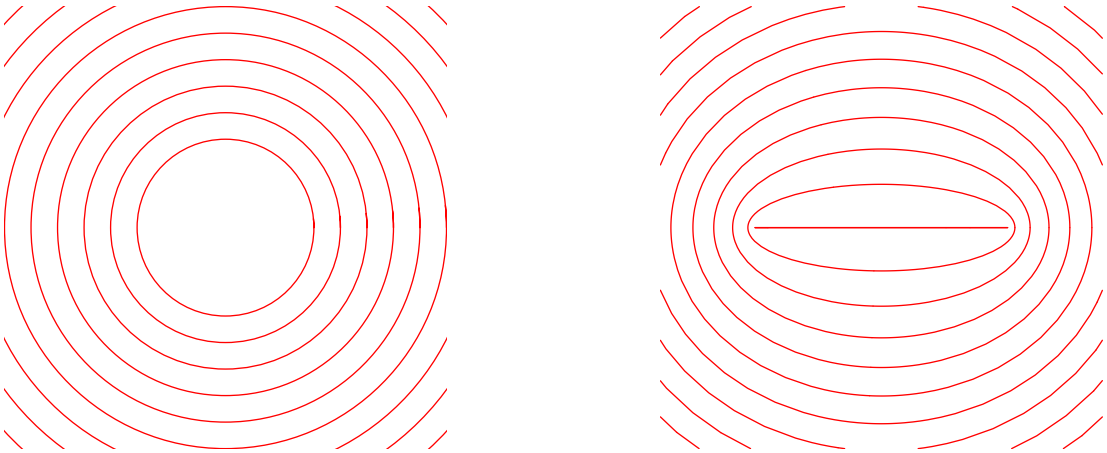


Figure 16.20. The Joukowski Map.

the horizontal and vertical lines under the map z^2. Note that, except at the origin, the image curves continue to meet at 90° angles, in accordance with conformality.

Example 16.28. A particularly interesting conformal transformation is given by the *Joukowski map*

$$\zeta = \frac{1}{2}\left(z + \frac{1}{z} \right). \tag{16.70}$$

It is used in the study of flows around airplane wings, and named after the pioneering Russian aero- and hydro-dynamics researcher Nikolai Zhukovskii (Joukowski). Since

$$\frac{d\zeta}{dz} = \frac{1}{2}\left(1 - \frac{1}{z^2} \right) = 0 \qquad \text{if and only if} \qquad z = \pm 1,$$

the Joukowski map is conformal except at the critical points $z = \pm 1$, as well as at the singularity $z = 0$ where it is not defined.

If $z = e^{\,\mathrm{i}\,\theta}$ lies on the unit circle, then

$$\zeta = \tfrac{1}{2}\big(e^{\,\mathrm{i}\,\theta} + e^{-\,\mathrm{i}\,\theta} \big) = \cos\theta,$$

lies on the real axis, with $-1 \le \zeta \le 1$. Thus, the Joukowski map squashes the unit circle down to the real line segment $[-1, 1]$. The images of points outside the unit circle fill the rest of the ζ plane, as do the images of the (nonzero) points inside the unit circle. Indeed, if we solve (16.70) for

$$z = \zeta \pm \sqrt{\zeta^2 - 1}\,, \tag{16.71}$$

we see that every ζ except ± 1 comes from two different points z; for ζ not on the critical line segment $[-1, 1]$, one point lies inside and and one lies outside the unit circle, whereas if $-1 < \zeta < 1$, the points are situated directly above and below it on the circle. Therefore, (16.70) defines a one-to-one conformal map from the exterior of the unit circle $\big\{\, |\, z\,| > 1 \,\big\}$ onto the exterior of the unit line segment $\mathbb{C} \setminus [-1, 1]$.

Under the Joukowski map, the concentric circles $|\, z\,| = r \neq 1$ are mapped to ellipses with foci at ± 1 in the ζ plane; see Figure 16.20. The effect on circles not centered at

the origin is quite interesting. The image curves take on a wide variety of shapes; several examples are plotted in Figure airfoil. If the circle passes through the singular point $z = 1$, then its image is no longer smooth, but has a cusp at $\zeta = 1$. Some of the image curves have the shape of the cross-section through an airplane wing or *airfoil*. Later we will see how to construct the physical fluid flow around such an airfoil, which proved to be a critical step in early airplane design.

Composition and The Riemann Mapping Theorem

One of the strengths of the method of conformal mapping is that one can build up lots of complicated examples by simply composing elementary mappings. The method rests on the simple fact that the composition of two complex analytic functions is also complex analytic. This is the complex counterpart of the result, learned in first year calculus, that the composition of two differentiable functions is itself differentiable.

Proposition 16.29. *If $w = f(z)$ is an analytic function of the complex variable $z = x + \mathrm{i}\, y$, and $\zeta = g(w)$ is an analytic function of the complex variable $w = u + \mathrm{i}\, v$, then the composition*[†] *$\zeta = h(z) \equiv g \circ f(z) = g(f(z))$ is an analytic function of z.*

Proof: The proof that the composition of two differentiable functions is differentiable is identical to the real variable version, [**9**, **165**], and need not be reproduced here. The derivative of the composition is explicitly given by the usual chain rule:

$$\frac{d}{dz}\, g \circ f(z) = g'(f(z))\, f'(z), \quad \text{or, in Leibnizian notation,} \quad \frac{d\zeta}{dz} = \frac{d\zeta}{dw}\, \frac{dw}{dz}\,. \tag{16.72}$$

If both f and g are one-to-one, so is the composition $h = g \circ f$. Moreover, the composition of two conformal maps is also conformal, a fact that is immediate from the definition, or by using the chain rule (16.72) to show that

$$h'(z) = g'(f(z))\, f'(z) \neq 0 \qquad \text{provided} \qquad g'(f(z)) \neq 0 \qquad \text{and} \qquad f'(z) \neq 0.$$

Thus, if f and g satisfy the conformality condition (16.52), so does $h = g \circ f$. *Q.E.D.*

Example 16.30. As we learned in Example 16.20, the exponential function

$$w = e^z$$

maps the horizontal strip $S = \{ -\frac{1}{2}\pi < \operatorname{Im} z < \frac{1}{2}\pi \}$ conformally onto the right half plane $R = \{ \operatorname{Re} w > 0 \}$. On the other hand, Example 16.22 tells us that the linear fractional transformation

$$\zeta = \frac{w - 1}{w + 1}$$

maps the right half plane R conformally to the unit disk $D = \{ |\, \zeta \,| < 1 \}$. Therefore, the composition

$$\zeta = \frac{e^z - 1}{e^z + 1} \tag{16.73}$$

[†] Of course, to properly define the composition, we need to ensure that the range of the function $w = f(z)$ is contained in the domain of the function $\zeta = g(w)$.

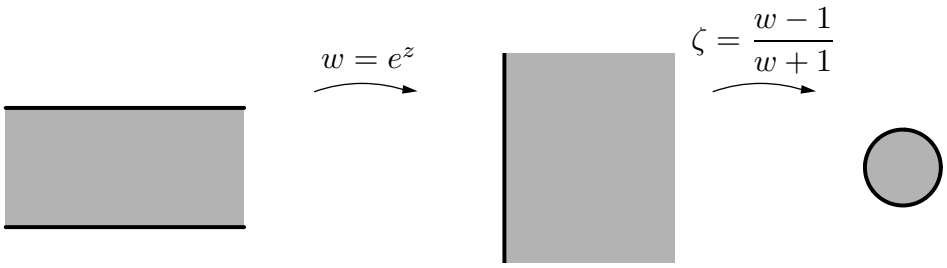


Figure 16.21. Composition of Conformal Maps.

is a one-to-one conformal map from the horizontal strip S to the unit disk D, as illustrated in Figure 16.21.

Recall that our motivating goal is to use analytic functions/conformal maps to solve boundary value problems for the Laplace equation on a complicated domain Ω by transforming them to boundary value problems on the unit disk. Of course, the key question the student should be asking at this point is: Is there, in fact, a conformal map $\zeta = g(z)$ from a given domain Ω to the unit disk $D = g(\Omega)$? The theoretical answer is the celebrated *Riemann Mapping Theorem*.

Theorem 16.31. *If $\Omega \subsetneq \mathbb{C}$ is any simply connected open subset,* not *equal to the entire complex plane, then there exists a one-to-one complex analytic map $\zeta = g(z)$, satisfying the conformality condition $g'(z) \neq 0$ for all $z \in \Omega$, that maps Ω to the unit disk $D = \{\, |\, \zeta\, | < 1 \,\}$.*

Thus, *any* simply connected domain — with one exception, the entire complex plane — can be conformally mapped the unit disk. (Exercise ■ provides a reason for this exception.) Note that Ω need *not* be bounded for this to hold. Indeed, the conformal map (16.60) takes the unbounded right half plane $R = \{\, \mathrm{Re}\, z > 0 \,\}$ to the unit disk. The proof of this important theorem relies some more advanced results in complex analysis, and can be found, for instance, in [**4**].

The Riemann Mapping Theorem guarantees the existence of a conformal map from any simply connected domain to the unit disk, but its proof is not constructive, and so provides no clue as to how to actually construct the desired mapping. And, in general, this is not an easy task. In practice, one assembles a repertoire of useful conformal maps that apply to particular domains of interest. An extensive catalog can be found in [**Cmap**]. More complicated maps can then be built up by composition of the basic examples. Ultimately, though, the determination of a suitable conformal map is more an art than a systematic science. Numerical methods for constructing conformal maps can be found in [**TH**].

Let us consider a few additional examples beyond those already encountered:

Example 16.32. Suppose we are asked to conformally map the upper half plane $U = \big\{\, \mathrm{Im}\, z > 0 \,\big\}$ to the unit disk $D = \big\{\, |\, \zeta\, | < 1 \,\big\}$. We already know that the linear fractional transformation

$$\zeta = g(w) = \frac{w-1}{w+1}$$

maps the right half plane $R = \big\{\, \mathrm{Re}\, z > 0 \,\big\}$ to $D = g(R)$. On the other hand, multiplication by $\mathrm{i} = e^{\,\mathrm{i}\,\pi/2}$, with $z = h(w) = \mathrm{i}\, w$, rotates the complex plane by 90° and so maps the

right half plane R to the upper half plane $U = h(R)$. Its inverse $h^{-1}(z) = -\mathrm{i}\, z$ will therefore map U to $R = h^{-1}(U)$. Therefore, to map the upper half plane to the unit disk, we compose these two maps, leading to the conformal map

$$\zeta = g \circ h^{-1}(z) = \frac{-\mathrm{i}\, z - 1}{-\mathrm{i}\, z + 1} = \frac{\mathrm{i}\, z + 1}{\mathrm{i}\, z - 1} \tag{16.74}$$

from U to D.

In a similar vein, we already know that the squaring map $w = z^2$ maps the upper right quadrant $Q = \left\{ 0 < \mathrm{ph}\, z < \frac{1}{2}\pi \right\}$ to the upper half plane U. Composing this with our previously constructed map — which requires using w instead of z in the previous formula (16.74) — leads to the conformal map

$$\zeta = \frac{\mathrm{i}\, z^2 + 1}{\mathrm{i}\, z^2 - 1} \tag{16.75}$$

that maps the quadrant Q to the unit disk D.

Example 16.33. The goal of this example is to construct an conformal map that takes a half disk

$$D_+ = \left\{ |\, z\,| < 1,\ y = \mathrm{Im}\, z > 0 \right\} \tag{16.76}$$

to the full unit disk $D = \left\{ \|\, \zeta\, \| < 1 \right\}$. The answer is *not* $\zeta = z^2$ because the image of D_+ omits the positive real axis, resulting in a disk with a slit cut out of it: $\left\{ \|\, \zeta\, \| < 1,\ 0 < \mathrm{ph}\, \zeta < 2\pi \right\}$. To obtain the entire disk as the image of the conformal map, we must think a little harder. The first observation is that the map $z = (w-1)/(w+1)$ that we analyzed in Example 16.22 takes the right half plane $R = \left\{ \mathrm{Re}\, w > 0 \right\}$ to the unit disk. Moreover, it maps the upper right quadrant $Q = \left\{ 0 < \mathrm{ph}\, w < \frac{1}{2}\pi \right\}$ to the half disk (16.76). Its inverse,

$$w = \frac{z+1}{z-1} \tag{16.77}$$

will therefore map the half disk, $z \in D_+$, to the upper right quadrant $w \in Q$.

On the other hand, we just constructed a conformal map (16.75) that takes the upper right quadrant Q to the unit disk D. Therefore, if compose the two maps (replacing z by w in (16.75) and then using (16.77)), we obtain the desired conformal map:

$$\zeta = \frac{\mathrm{i}\, w^2 + 1}{\mathrm{i}\, w^2 - 1} = \frac{\mathrm{i}\left(\dfrac{z+1}{z-1}\right)^2 + 1}{\mathrm{i}\left(\dfrac{z+1}{z-1}\right)^2 - 1} = \frac{(\mathrm{i}+1)(z^2+1) + 2(\mathrm{i}-1)z}{(\mathrm{i}-1)(z^2+1) + 2(\mathrm{i}+1)z}\,.$$

The formula can be further simplified by multiplying numerator and denominator by $\mathrm{i}+1$, and so

$$\zeta = -\mathrm{i}\, \frac{z^2 + 2\,\mathrm{i}\, z + 1}{z^2 - 2\,\mathrm{i}\, z + 1}\,.$$

The leading factor $-\mathrm{i}$ is unimportant and can be omitted, since it merely rotates the disk by -90°, and so

$$\zeta = \frac{z^2 + 2\,\mathrm{i}\,z + 1}{z^2 - 2\,\mathrm{i}\,z + 1} \tag{16.78}$$

is an equally valid solution to our problem.

Finally, as noted in the preceding example, the conformal map guaranteed by the Riemann Mapping Theorem is *not* unique. Since the linear fractional transformations (16.63) map the unit disk to itself, we can compose them with any conformal Riemann mapping to produce additional conformal maps from a simply-connected domain to the unit disk. For example, composing (16.63) with (16.73) produces a family of mappings

$$\zeta = \frac{1 + e^z - \alpha(1 - e^z)}{\overline{\alpha}\,(1 + e^z) - 1 + e^z}\,, \tag{16.79}$$

which, for any $|\,\alpha\,| < 1$, maps the strip $S = \left\{ -\frac{1}{2}\,\pi < \mathrm{Im}\; z < \frac{1}{2}\,\pi \right\}$ onto the unit disk. With a little more work, it can be shown that this is the only ambiguity, and so, for instance, (16.79) forms a complete list of one-to-ne conformal maps from S to D.

Annular Domains

The Riemann Mapping Theorem does not apply to non-simply connected domains. For purely topological reasons, a hole cannot be made to disappear under a one-to-one continuous mapping — much less a conformal map — and so a non-simply connected domain cannot be mapped in a on-to-one manner onto the unit disk. So we must look elsewhere for a simple model domain.

The simplest non-simply connected domain is an *annulus* consisting of the points between two concentric circles

$$A_{r,R} = \left\{ r < |\,\zeta\,| < R \right\}, \tag{16.80}$$

which, for simplicity, is centered around the origin; see Figure 16.22. The case $r = 0$ corresponds to a punctured disk, while $R = \infty$ gives the exterior of a disk or radius r. It can be proved, [**Cmap**], that any other domain with a single hole can be mapped to an annulus. The annular radii r, R are not uniquely specified; indeed the linear map $\zeta = \alpha\,z$ maps the annulus (16.80) to a rescaled annulus $A_{\rho\,r,\rho\,R}$ whose inner and outer radii have both been scaled by the factor $\rho = |\,\alpha\,|$. But the ratio[†] r/R of the inner to outer radius of the annulus is uniquely specified; annuli with different ratios *cannot* be mapped to each other by a conformal map.

Example 16.34. Let $c > 0$. Consider the domain

$$\Omega = \left\{ |\,z\,| < 1 \quad \text{and} \quad |\,z - c\,| > c \right\}$$

[†] If $r = 0$ or $R = \infty$, but not both, then $r/R = 0$ by convention. The punctured plane, where $r = 0$ and $R = \infty$ remains a separate case.

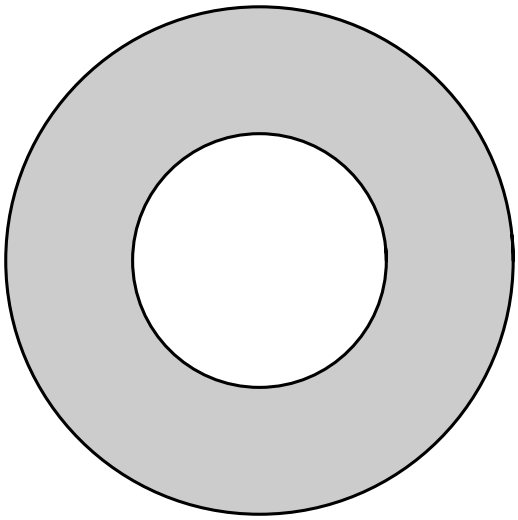


Figure 16.22. An Annulus.

contained between two nonconcentric circles. To keep the computations simple, we take the outer circle to have radius 1 (which can always be arranged by scaling, anyway) while the inner circle has center at the point $z = c$ on the real axis and radius c, which means that it passes through the origin. We must restrict $c < \frac{1}{2}$ in order that the inner circle not overlap with the outer circle. Our goal is to conformally map this non-concentric annular domain to a concentric annulus of the form

$$A_{r,1} = \{\, r < |\, \zeta \,| < 1 \,\}$$

by a conformal map $\zeta = g(z)$.

Now, according, to Example 16.23, a linear fractional transformation of the form

$$\zeta = g(z) = \frac{z - \alpha}{\overline{\alpha}\, z - 1} \qquad \text{with} \qquad |\, \alpha \,| < 1 \tag{16.81}$$

maps the unit disk to itself. Moreover, as remarked earlier, and demonstrated in Exercise ■, linear fractional transformations always map circles to circles. Therefore, we seek a particular value of α that maps the inner circle $|\, z - c \,| = c$ to a circle of the form $|\, \zeta \,| = r$ centered at the origin. We choose α real and try to map the points 0 and $2c$ on the inner circle to the points r and $-r$ on the circle $|\, \zeta \,| = r$. This requires

$$g(0) = \alpha = r, \qquad g(2c) = \frac{2c - \alpha}{2c\alpha - 1} = -r. \tag{16.82}$$

Substituting the first into the second leads to the quadratic equation

$$c\alpha^2 - \alpha + c = 0.$$

There are two real solutions:

$$\alpha = \frac{1 - \sqrt{1 - 4c^2}}{2c} \qquad \text{and} \qquad \alpha = \frac{1 + \sqrt{1 - 4c^2}}{2c}\,. \tag{16.83}$$

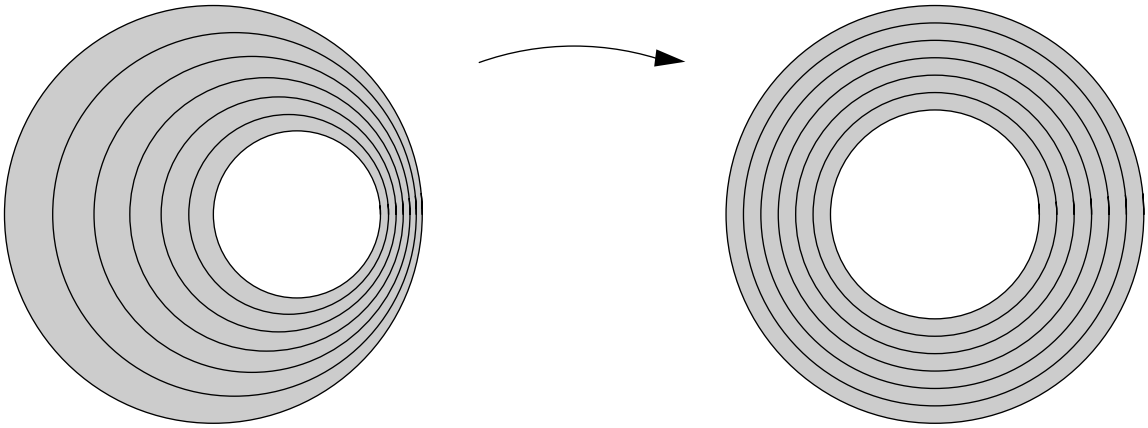


Figure 16.23. Conformal Map for Non-concentric Annulus.

Since $0 < c < \frac{1}{2}$, the second solution has $\alpha > 1$, and hence is inadmissible. Therefore, the first solution yields the required conformal map

$$\zeta = \frac{z - 1 + \sqrt{1 - 4\,c^2}}{(1 - \sqrt{1 - 4\,c^2}\,)\,z - 2\,c}\,.$$

Note in particular that the radius $r = \alpha$ of the inner circle in $A_{r,1}$ is *not the same* as the radius c of the inner circle in Ω.

For example, taking $c = \frac{2}{5}$, equation (16.83) implies $\alpha = \frac{1}{2}$, and hence the linear fractional transformation $\zeta = \dfrac{2\,z - 1}{z - 2}$ maps the annular domain $\Omega = \left\{\, |\, z\,| < 1,\ \left|\, z - \tfrac{2}{5}\,\right| > \tfrac{2}{5}\,\right\}$ to the concentric annulus $A = A_{1/2,1} = \left\{\, \frac{1}{2} < |\, \zeta\,| < 1\,\right\}$. In Figure 16.23, we plot the non-concentric circles in Ω that map to concentric circles in the annulus A. In Exercise ∎ the reader is asked to adapt this construction to a general non-concentric annular domain.

Applications to Harmonic Functions and Laplace's Equation

Let us now apply our newly developed expertise in conformal mapping to the study of harmonic functions and boundary value problems for the Laplace equation. Our goal was to change a boundary value problem on a domain Ω into a boundary value problem on the unit disk D that we know how to solve. To this end, suppose we know a conformal map $\zeta = g(z)$ that takes $z \in \Omega$ to $\zeta \in D$. As we know, the real and imaginary parts of an analytic function $F(\zeta)$ defined on D define harmonic functions. Moreover, according to Proposition 16.29, the composition $f(z) = F(g(z))$ defines an analytic function whose real and imaginary parts are harmonic functions on Ω. Thus, the conformal mapping can be regarded as a change of variables between their harmonic real and imaginary parts. In fact, this property does not even require the harmonic function to be the real part of an analytic function, i.e., we are not required to assume the existence of a harmonic conjugate.

Proposition 16.35. *If $U(\xi, \eta)$ is a harmonic function of ξ, η, and*

$$\zeta = \xi + \mathrm{i}\,\eta = p(x, y) + \mathrm{i}\,q(x, y) = g(z) \tag{16.84}$$

is any analytic function, then the composition

$$u(x,y) = U(p(x,y), q(x,y)) \tag{16.85}$$

is a harmonic function of x, y.

Proof: This is a straightforward application of the chain rule:

$$\frac{\partial u}{\partial x} = \frac{\partial U}{\partial \xi}\,\frac{\partial \xi}{\partial x} + \frac{\partial U}{\partial \eta}\,\frac{\partial \eta}{\partial x}, \qquad \frac{\partial u}{\partial y} = \frac{\partial U}{\partial \xi}\,\frac{\partial \xi}{\partial y} + \frac{\partial U}{\partial \eta}\,\frac{\partial \eta}{\partial y},$$

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 U}{\partial \xi^2}\left(\frac{\partial \xi}{\partial x}\right)^2 + 2\,\frac{\partial^2 U}{\partial \xi\,\partial \eta}\,\frac{\partial \xi}{\partial x}\,\frac{\partial \eta}{\partial x} + \frac{\partial^2 U}{\partial \eta^2}\left(\frac{\partial \eta}{\partial x}\right)^2 + \frac{\partial U}{\partial \xi}\,\frac{\partial^2 \xi}{\partial x^2} + \frac{\partial U}{\partial \eta}\,\frac{\partial^2 \eta}{\partial x^2},$$

$$\frac{\partial^2 u}{\partial y^2} = \frac{\partial^2 U}{\partial \xi^2}\left(\frac{\partial \xi}{\partial y}\right)^2 + 2\,\frac{\partial^2 U}{\partial \xi\,\partial \eta}\,\frac{\partial \xi}{\partial y}\,\frac{\partial \eta}{\partial y} + \frac{\partial^2 U}{\partial \eta^2}\left(\frac{\partial \eta}{\partial y}\right)^2 + \frac{\partial U}{\partial \xi}\,\frac{\partial^2 \xi}{\partial y^2} + \frac{\partial U}{\partial \eta}\,\frac{\partial^2 \eta}{\partial y^2}.$$

Using the Cauchy–Riemann equations

$$\frac{\partial \xi}{\partial x} = -\,\frac{\partial \eta}{\partial y}, \qquad \frac{\partial \xi}{\partial y} = \frac{\partial \eta}{\partial x},$$

for the analytic function $\zeta = \xi + \mathrm{i}\,\eta$, we find, after some algebra,

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \left[\left(\frac{\partial \xi}{\partial x}\right)^2 + \left(\frac{\partial \eta}{\partial x}\right)^2\right]\left[\frac{\partial^2 U}{\partial \xi^2} + \frac{\partial^2 U}{\partial \eta^2}\right] = |\,g'(z)\,|^2\,\Delta U,$$

where $|\,g'(z)\,|^2 = (\,\partial \xi/\partial x\,)^2 + (\,\partial \eta/\partial x\,)^2$. We conclude that whenever $U(\xi, \eta)$ is any harmonic function, and so solves the Laplace equation $\Delta U = 0$ (in the ξ, η variables), then $u(x, y)$ is a solution to the Laplace equation $\Delta u = 0$ in the x, y variables, and is thus also harmonic. *Q.E.D.*

This observation has profound consequences for boundary value problems arising in physical applications. Suppose we wish to solve the Dirichlet problem

$$\Delta u = 0 \qquad \text{in} \qquad \Omega, \qquad\qquad u = h \qquad \text{on} \qquad \partial\Omega, \tag{16.86}$$

on a simply connected domain[†] $\Omega \subsetneq \mathbb{C}$. Let $\zeta = g(z) = p(x, y) + \mathrm{i}\,q(x, y)$ be a one-to-one conformal mapping from the domain Ω to the unit disk D, whose existence is guaranteed by the Riemann Mapping Theorem 16.31. (Although its explicit construction may be much more problematic.) Then the change of variables formula (16.85) will map the harmonic function $u(x, y)$ on Ω to a harmonic function $U(\xi, \eta)$ on D. Moreover, the boundary values of $U = H$ on the unit circle ∂D correspond to those of $u = h$ on $\partial\Omega$ by the same change of variables formula:

$$h(x,y) = H(p(x,y), q(x,y)), \qquad \text{for} \qquad (x,y) \in \partial\Omega. \tag{16.87}$$

[†] The Riemann Mapping Theorem 16.31 tells us to exclude the case $\Omega = \mathbb{C}$. Indeed, this case is devoid of boundary conditions, and so the problem does not admit a unique solution.

We conclude that $U(\xi, \eta)$ solves the Dirichlet problem

$$\Delta U = 0 \qquad \text{in} \qquad D, \qquad\qquad U = H \qquad \text{on} \qquad \partial D.$$

But we already know how to solve the Dirichlet problem on the unit disk by the Poisson integral formula (15.48)! Then the solution to the original boundary value problem is given by the composition formula $u(x,y) = U\big(p(x,y), q(x,y)\big)$. Thus, the solution to the Dirichlet problem on a unit disk can be used to solve the Dirichlet problem on more complicated planar domains — provided we know the appropriate conformal map.

Example 16.36. According to Example 16.22, the analytic function

$$\xi + \mathrm{i}\,\eta = \zeta = \frac{z-1}{z+1} = \frac{x^2+y^2-1}{(x+1)^2+y^2} + \mathrm{i}\,\frac{2y}{(x+1)^2+y^2} \tag{16.88}$$

maps the right half plane $R = \big\{\, x = \operatorname{Re} z > 0 \,\big\}$ to the unit disk $D = \big\{\, |\,\zeta\,| < 1 \,\big\}$. Proposition 16.35 implies that if $U(\xi, \eta)$ is a harmonic function in the unit disk, then

$$u(x,y) = U\left(\frac{x^2+y^2-1}{(x+1)^2+y^2}\,,\ \frac{2y}{(x+1)^2+y^2} \right) \tag{16.89}$$

is a harmonic function on the right half plane. (This can, of courtse, be checked directly by a rather unpleasant chain rule computation.)

To solve the Dirichlet boundary value problem

$$\Delta u = 0, \qquad x > 0, \qquad\qquad u(0,y) = h(y), \tag{16.90}$$

on the right half plane, we adopt the change of variables (16.88) and use the Poisson integral formula to construct the solution to the transformed Dirichlet problem

$$\Delta U = 0, \qquad \xi^2 + \eta^2 < 1, \qquad\qquad U(\cos\varphi, \sin\varphi) = H(\varphi), \tag{16.91}$$

on the unit disk. The relevant boundary conditions are found as follows. Using the explicit form

$$x + \mathrm{i}\,y = z = \frac{1+\zeta}{1-\zeta} = \frac{(1+\zeta)(1-\overline{\zeta})}{|\,1-\zeta\,|^2} = \frac{1+\zeta-\overline{\zeta}-|\,\zeta\,|^2}{|\,1-\zeta\,|^2} = \frac{1-\xi^2-\eta^2+2\,\mathrm{i}\,\eta}{(\xi-1)^2+\eta^2}$$

for the inverse map, we see that the boundary point $\zeta = \xi + \mathrm{i}\,\eta = e^{\,\mathrm{i}\,\varphi}$ on the unit circle ∂D will correspond to the boundary point

$$\mathrm{i}\,y = \frac{2\,\eta}{(\xi-1)^2+\eta^2} = \frac{2\,\mathrm{i}\,\sin\varphi}{(\cos\varphi-1)^2+\sin^2\varphi} = \mathrm{i}\,\cot\frac{\varphi}{2} \tag{16.92}$$

on the imaginary axis $\partial R = \big\{\, \operatorname{Re} z = 0 \,\big\}$. Thus, the boundary data $h(y)$ on ∂R corresponds to the boundary data

$$H(\varphi) = h\left(\cot\tfrac{1}{2}\,\varphi\right)$$

on the unit circle. The Poisson integral formula (15.48) can then be applied to solve (16.91), from which we are able to reconstruct the solution (16.89) to the boundary value problem (16.89) on the half plane.

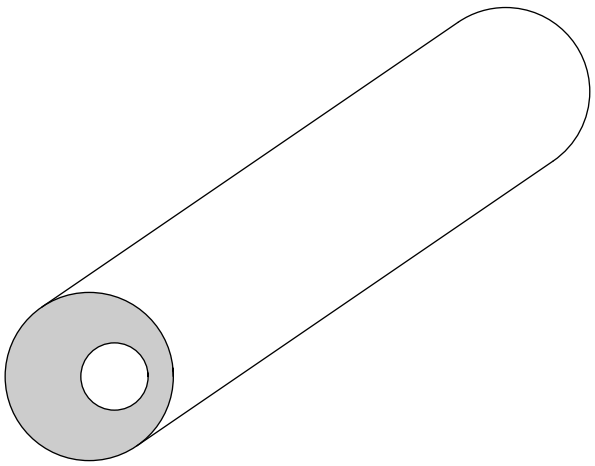


Figure 16.24. Non–Coaxial Cable.

Let's look at an explicit example. If the boundary data on the imaginary axis is provided by the step function

$$u(0,y) = h(y) \equiv \begin{cases} 1, & y > 0, \\ 0, & y < 0, \end{cases}$$

then the corresponding boundary data on the unit disk is a (periodic) step function

$$H(\varphi) = \begin{cases} 1, & 0 < \varphi < \pi, \\ 0, & \pi < \varphi < 2\pi, \end{cases}$$

that has values $+1$ on the upper semicircle, -1 on the lower semicircle, and jump discontinuities at $\zeta = \pm 1$. According to the Poisson formula (15.48), the solution to the latter boundary value problem is given by

$$\begin{aligned} U(\xi,\eta) &= \frac{1}{2\pi} \int_0^{\pi} \frac{1-\rho^2}{1+\rho^2-2\rho\cos(\varphi-\phi)}\, d\phi \\ &= \frac{1}{\pi}\left[\tan^{-1}\left(\frac{1+\rho}{1-\rho}\, \cot\frac{\varphi}{2} \right) + \tan^{-1}\left(\frac{1+\rho}{1-\rho}\, \tan\frac{\varphi}{2} \right) \right] \end{aligned} \qquad \text{where} \quad \begin{aligned} \xi &= \rho\, \cos\varphi, \\ \eta &= \rho\, \sin\varphi. \end{aligned}$$

Finally, we use (16.89) to construct the solution on the upper half plane, although we shall spare the reader the messy details of the final formula. The result is depicted in Figure zpm1h■.

Remark: The solution to the preceding Dirichlet boundary value problem is not, in fact, unique, owing to the unboundedness of the domain. The solution that we pick out by using the conformal map to the unit disk is the one that remains bounded at ∞. There are other solutions, but they are unbounded as $|\, z\,| \to \infty$ and would correspond to solutions on the unit disk that have some form of delta function singularity in their boundary data at the point -1; see Exercise ■.

Example 16.37. *A non-coaxial cable.* The goal of this example is to determine the electrostatic potential inside a non-coaxial cylindrical cable with prescribed constant potential values on the two bounding cylinders, as illustrated in Figure 16.24. Assume

for definiteness that the larger cylinder has radius 1, and centered at the origin, while the smaller cylinder has radius $\frac{2}{5}$, and is centered at $z = \frac{2}{5}$. The resulting electrostatic potential will be independent of the longitudinal coordinate, and so can be viewed as a planar potential in the annular domain contained between two circles representing the cross-sections of our cylinders. The desired potential must satisfy the Dirichlet boundary value problem

$$\begin{aligned} \Delta u = 0 \qquad &\text{when} \qquad |\, z \,| < 1 \quad \text{and} \quad \left| \, z - \tfrac{2}{5} \, \right| > \tfrac{2}{5}, \\ u = a, \quad &\text{when} \quad |\, z \,| = 1, \qquad \text{and} \qquad u = b \quad \text{when} \quad \left| \, z - \tfrac{2}{5} \, \right| = \tfrac{2}{5}. \end{aligned}$$

According to Example 16.34, the linear fractional transformation

$$\zeta = \frac{2\,z - 1}{z - 2} \tag{16.93}$$

will map this non-concentric annular domain to the annulus $A_{1/2,1} = \left\{ \, \frac{1}{2} < |\, \zeta \,| < 1 \, \right\}$, which is the cross-section of a coaxial cable. The corresponding transformed potential $U(\xi, \eta)$ has the constant Dirichlet boundary conditions

$$U = a, \quad \text{when} \quad |\, \zeta \,| = \tfrac{1}{2}, \qquad \text{and} \qquad U = b \quad \text{when} \quad |\, \zeta \,| = 1. \tag{16.94}$$

Clearly the coaxial potential U must be a radially symmetric solution to the Laplace equation, and hence, according to (15.64), of the form

$$U(\xi, \eta) = \alpha \, \log |\, \zeta \,| + \beta,$$

for constants α, β. A short computation shows that the particular potential function

$$U(\xi, \eta) = \frac{b - a}{\log 2} \, \log |\, \zeta \,| + b = \frac{b - a}{2 \log 2} \, \log(\xi^2 + \eta^2) + b$$

satisfies the prescribed boundary conditions (16.94). Therefore, the desired non-coaxial electrostatic potential

$$u(x, y) = \frac{b - a}{\log 2} \, \log \left| \, \frac{2\,z - 1}{z - 2} \, \right| + b = \frac{b - a}{2 \log 2} \, \log \left(\, \frac{(2\,x - 1)^2 + y^2}{(x - 2)^2 + y^2} \, \right) + b \tag{16.95}$$

is obtained by composition with the conformal map (16.93). The particular case $a = 0, b = 1$ is plotted in Figure ncoaxep■.

Remark: The same harmonic function solves the problem of determining the equilibrium temperature in an annular plate whose inner boundary is kept at a temperature $u = a$ while the outer boundary is kept at temperature $u = b$. One could also interpret this solution as the equilibrium temperature of a three-dimensional cylindrical body contained between two non-coaxial cylinders that are held at fixed temperatures. The body's temperature (16.95) will only depend upon the transverse coordinates x, y and not upon the longitudinal coordinate z.

Applications to Fluid Flow

Conformal mappings are particularly useful in the analysis of planar ideal fluid flow. Recall that if $\Theta(\zeta) = \Phi(\xi, \eta) + \mathrm{i}\,\Psi(\xi, \eta)$ is an analytic function that represents the complex potential function for a steady state fluid flow in a planar domain $\zeta \in D$, then we can interpret its real part $\Phi(\xi, \eta)$ as the velocity potential, while the imaginary part $\Psi(\xi, \eta)$ is the harmonic conjugate stream function. The level curves of Φ are the equipotential lines, and, except at stagnation points where $\Theta'(\zeta) = 0$, are orthogonal to the level curves of Ψ, which are the streamlines followed by the individual fluid particles.

Composing the complex potential with a conformal map $\zeta = g(z)$ leads to a transformed complex potential $\Theta(g(z)) = \chi(z) = \varphi(x, y) + \mathrm{i}\,\psi(x, y)$ on the corresponding domain $z \in \Omega$. A key fact is that the conformal map will take isopotential lines of φ to isopotential lines of Φ and streamlines of ψ to streamlines of Ψ. Conformality implies that the orthogonality relations among isopotentials and streamlines away from stagnation points is maintained.

Let us concentrate on the case of flow past a solid object. In three dimensions, the object is assumed to have a uniform shape in the axial direction, and so we can restrict our attention to a planar fluid flow around a closed, bounded subset $D \subset \mathbb{R}^2 \simeq \mathbb{C}$ representing the cross-section of our cylindrical object, as in Figure crss. The (complex) velocity and potential are defined on the complementary domain $\Omega = \mathbb{C} \setminus D$ occupied by the fluid. The ideal flow assumptions of incompressibility and irrotationality are reasonably accurate if the flow is laminar, meaning far away from turbulent. The velocity potential $\varphi(x, y)$ will satisfy the Laplace equation $\Delta\varphi = 0$ in the exterior domain Ω. For a solid object, we should impose the homogeneous Neumann boundary conditions

$$\frac{\partial \varphi}{\partial \mathbf{n}} = 0 \qquad \text{on the boundary} \qquad \partial\Omega = \partial D, \tag{16.96}$$

indicating that there no fluid flux into the object. We note that, according to Exercise , a conformal map will automatically preserve the Neumann boundary conditions.

In addition, since the flow is taking place on an unbounded domain, we need to specify the fluid motion at large distances. We shall assume our object is placed in a uniform horizontal flow, as in Figure hflow. Thus, far away, the object will not affect the flow, and so the velocity should approximate the uniform velocity field $\mathbf{v} = (\,1, 0\,)^T$, where, for simplicity, we choose our physical units so that the asymptotic speed of the fluid is equal to 1. Equivalently, the velocity potential should satisfy

$$\varphi(x, y) \approx x, \qquad \text{so} \qquad \nabla\varphi \approx \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad \text{when} \qquad x^2 + y^2 \gg 0.$$

Remark: An alternative physical interpretation is that the fluid is at rest, while the object moves through the fluid at unit speed 1 in a horizontal direction. For example, think of an airplane flying through the air at constant speed. If we adopt a moving coordinate system by sitting inside the airplane, then the effect is as if the object is sitting still while the air is moving towards us at unit speed.

Example 16.38. The simplest example is a flat plate moving through the fluid in a horizontal direction. The plate's cross-section is a horizontal line segment, and, for simplicity, we take it to be the segment $D = [-1, 1]$ lying on the real axis. If the plate is very thin, it will have absolutely no effect on the horizontal flow of the fluid, and, indeed, the velocity potential is given by

$$\varphi(x, y) = x, \qquad \text{for} \qquad x + \mathrm{i}\, y \in \Omega = \mathbb{C} \setminus [-1, 1].$$

Note that $\nabla\varphi = (1, 0)^T$, and hence this flow satisfies the Neumann boundary conditions (16.96) on the horizontal segment $D = \partial\Omega$. The corresponding complex potential is $\chi(z) = z$, with complex velocity $f(z) = \chi'(z) = 1$.

Example 16.39. Recall that the Joukowski conformal map defined by the analytic function

$$\zeta = g(z) = \frac{1}{2}\left(z + \frac{1}{z}\right) \tag{16.97}$$

squashes the unit circle $|z| = 1$ down to the real line segment $[-1, 1]$ in the ζ plane. Therefore, it will map the fluid flow outside the unit disk (which is the cross-section of a circular cylinder) to the fluid flow past the line segment, which, according to the previous example, has complex potential $\Theta(\zeta) = \zeta$. As a result, the complex potential for the flow past a disk is the Joukowski function

$$\chi(z) = \Theta \circ g(z) = g(z) = \frac{1}{2}\left(z + \frac{1}{z}\right). \tag{16.98}$$

Except for a factor of $\frac{1}{2}$, this agrees with the flow potential we derived in Example 16.17. The difference is that, at large distances, the potential

$$\chi(z) \approx \tfrac{1}{2} z \qquad \text{for} \qquad |z| \gg 1.$$

corresponds to uniform horizontal flow whose velocity $\left(\frac{1}{2}, 0\right)^T$ is half as fast. The discrepancy between the two flows can easily be rectified by multiplying (16.98) by 2, whose only effect is to speed up the flow.

Example 16.40. Let us next consider the case of a tilted plate in a uniformly horizontal fluid flow. Thus, the cross-section is the line segment

$$z(t) = t\, e^{\mathrm{i}\,\theta}, \qquad -1 \le t \le 1,$$

obtained by rotating the horizontal line segment $[-1, 1]$ through an angle θ, as in Figure tilt■. The goal is to construct a fluid flow past the tilted segment that is asymptotically horizontal at large distance.

The critical observation is that, while the effect of rotating a plate in a fluid flow is not so evident, rotating a circularly symmetric disk has no effect on in the flow around it. Thus, the rotation $w = e^{-\mathrm{i}\,\theta} z$ maps the Joukowski potential (16.98) to the complex potential

$$\Upsilon(w) = \chi(e^{\mathrm{i}\,\theta} w) = \frac{1}{2}\left(e^{\mathrm{i}\,\theta} w + \frac{e^{-\mathrm{i}\,\theta}}{w}\right).$$

The streamlines of the induced flow are no longer asymptotically horizontal, but rather at an angle $-\theta$. If we now apply the original Joukowski map (16.97) to the rotated flow, the circle is again squashed down to the horizontal line segment, but the flow lines continue to be at angle $-\theta$ at large distances. Thus, if we then rotate the resulting flow through an angle θ, the net effect will be to tilt the segment to the desired angle θ while rotating the streamlines to be asymptotically horizontal. Putting the pieces together, we have the final complex potential in the form

$$\chi(z) = e^{\,\mathrm{i}\,\theta}\left(z\,\cos\theta - \,\mathrm{i}\,\sin\theta\,\sqrt{z^2 - e^{-2\,\mathrm{i}\,\theta}}\right). \tag{16.99}$$

Sample streamlines for the flow at several attack angles are plotted in Figure tilt■.

Example 16.41. As we discovered in Example 16.28, applying the Joukowski map to off-center disks will, in favorable configurations, produce airfoil-shaped objects. The fluid motion around such airfoils can thus be obtained by applying the Joukowski map to the flow past such an off-center circle.

First, an affine map

$$w = \alpha\, z + \beta$$

has the effect of moving the unit disk $|\,z\,| \le 1$ to the disk $|\,w - \beta\,| \le |\,\alpha\,|$ with center β and radius $|\,\alpha\,|$. In particular, the boundary circle will continue to pass through the point $w = 1$ provided $|\,\alpha\,| = |\,1 - \beta\,|$. Moreover, as noted in Example 16.18, the angular component of α has the effect of a rotation, and so the streamlines around the new disk will, asymptotically, be at an angle $\varphi = \mathrm{ph}\,\alpha$ with the horizontal. We then apply the Joukowski transformation

$$\zeta = \frac{1}{2}\left(w + \frac{1}{w}\right) = \frac{1}{2}\left(\alpha\, z + \beta + \frac{1}{\alpha\, z + \beta}\right) \tag{16.100}$$

to map the disk to the airfoil shape. The resulting complex potential for the flow past the airfoil is obtained by substituting the inverse map

$$z = \frac{w - \beta}{\alpha} = \frac{\zeta - \beta + \sqrt{\zeta^2 - 1}}{\alpha}\,,$$

into the original potential (16.98), whereby

$$\Theta(\zeta) = \frac{1}{2}\left(\frac{\zeta - \beta + \sqrt{\zeta^2 - 1}}{\alpha} + \frac{\alpha\big(\,\zeta - \beta - \sqrt{\zeta^2 - 1}\,\big)}{\beta^2 + 1 - 2\,\beta\,\zeta}\right).$$

Since the streamlines have been rotated through an angle $\varphi = \mathrm{ph}\,\alpha$, we then rotate the final result back by multiplying by $e^{\,\mathrm{i}\,\varphi}$ in order to see the effect of the airfoil tiled at an angle $-\varphi$ in a horizontal flow. Sample streamlines are graphed in Figure airfoilnolift■.

We can interpret all these examples as planar cross-sections of three-dimensional fluid flows past an airplane wing oriented in the longitudinal z direction. The wing is assumed to have a uniform cross-section shape, and the flow not dependent upon the axial z coordinate. For sufficiently long wings flying in laminar (non-turbulent) flows, this model will be valid away from the wing tips. Underestanding the dynamics of more complicated airfoils with

varying cross-section and/or faster motion requires a fully three-dimensional fluid model. For such problems, complex analysis is no longer applicable, and, for the most part, one must rely on large scale numerical integration. Only in recent years have computers become sufficiently powerful to compute realistic three-dimensional fluid motions — and then only in reasonably "mild" scenarios†. The two-dimensional versions that have been analyzed here still provide important clues to the behavior of a three-dimensional flow, as well as useful initial approximations to the three-dimensional airplane wing design problem.

Unfortunately, there is a major flaw with the airfoils that we have just designed. As we will see, potential flows do not produce any lift, and hence such airplanes would not fly. Fortunately for us, the physical flow is not of this nature! In order to understand how lift enters into the picture, we need to study complex integration, and so we will return to this example later. In Example 16.56, we shall construct an alternative flow past an airfoil that continues to have the correct asymptotic behavior at large distances, while inducing a nonzero lift on the wing. This is the secret to flight.

Poisson's Equation and the Green's Function

Although designed for solving the homogeneous Laplace equation, the method of conformal mapping can also be used to solve its inhomogeneous counterpart — the Poisson equation. As we learned in Chapter 15, to solve an inhomogeneous boundary value problem it suffices to solve the problem when the right hand side is a delta function concentrated at a point in the domain:

$$-\Delta u = \delta_\zeta(x,y) = \delta(x-\xi)\,\delta(y-\eta), \qquad \zeta = \xi + \mathrm{i}\,\eta \in \Omega,$$

sucject to homogeneous boundary conditions (Dirichlet or mixed) on $\partial\Omega$. (As usual, we exclude pure Neumann boundary conditions due to lack of existence/uniqueness.) The solution

$$u(x,y) = G_\zeta(x,y) = G(x,y;\xi,\eta)$$

is the *Green's function* for the given boundary value problem. With the Green's function in hand, the solution to the homogeneous boundary value problem under a general external forcing,

$$-\Delta u = f(x,y),$$

is then provided by the superposition principle

$$u(x,y) = \iint_\Omega G(x,y;\xi,\eta)\, f(\xi,\eta)\, d\xi\, d\eta. \tag{16.101}$$

For the planar Poisson equation, the starting point is the logarithmic potential function

$$u(x,y) = \operatorname{Re} \frac{1}{2\pi} \log z = \frac{1}{2\pi} \log |\, z\,| = \frac{1}{4\pi} \log(x^2+y^2), \tag{16.102}$$

† The definition of mild relies on the magnitude of the Reynolds number, [**14**], an overall measure of the flow's complexity.

which solves the Dirichlet problem

$$-\Delta u = \delta_0(x,y), \qquad (x,y) \in D, \qquad\qquad u = 0 \qquad \text{on} \qquad \partial D,$$

on the unit disk D for an impulse concentrated at the origin; see Section 15.3 for details. How do we obtain the corresponding solution when the unit impulse is concentrated at another point $\zeta = \xi + \mathrm{i}\,\eta \in D$ instead of the origin? According to Example 16.23, the linear fractional transformation

$$w = g(z) = \frac{z - \zeta}{\overline{\zeta}\, z - 1}, \qquad \text{where} \qquad |\,\zeta\,| < 1, \tag{16.103}$$

maps the unit disk to itself, moving the point $z = \zeta$ to the origin $w = g(\zeta) = 0$. The logarithmic potential $U = \dfrac{1}{2\,\pi}\, \log |\,w\,|$ will thus be mapped to the Green's function

$$G(x,y;\xi,\eta) = \frac{1}{2\,\pi}\, \log \left| \frac{z - \zeta}{\overline{\zeta}\, z - 1} \right| \tag{16.104}$$

at the point $\zeta = \xi + \mathrm{i}\,\eta$. Indeed, by the properties of conformal mapping, since U is harmonic except at the singularity $w = 0$, the function (16.104) will also be harmonic except at the image point $z = \zeta$. The fact that the mapping does not affect the delta function singularity is not hard to check; the details are relegated to Exercise ■. Moreover, since the conformal map does not alter the boundary $|\,z\,| = 1$, the function (16.104) continues to satisfy the homogeneous Dirichlet boundary conditions.

Formula (16.104) reproduces the Poisson formula (15.78) for the Green's function that we previously derived by the method of images. This identification can be verified by substituting $z = r\,e^{\mathrm{i}\,\theta}$, $\zeta = \rho\, e^{\mathrm{i}\,\varphi}$, or, more simply, by noting that the numerator in the logarithmic fraction gives the potential due to a unit impulse at $z = \zeta$, while the denominator represents the image potential at $z = 1/\,\overline{\zeta}$ required to cancel out the effect of the interior potential on the boundary of the unit disk.

Now that we know the Green's function on the unit disk, we can use the methods of conformal mapping to produce the Green's function for any other simply connected domain $\Omega \subsetneq \mathbb{C}$.

Proposition 16.42. *Let $w = g(z)$ denote the conformal map that takes the domain $z \in \Omega$ to the unit disk $w \in D$, guaranteed by the Riemann Mapping Theorem 16.31. Then the Green's function associated with homogeneous Dirichlet boundary conditions on Ω is explicitly given by*

$$G(z;\zeta) = \frac{1}{2\,\pi}\, \log \left| \frac{g(z) - g(\zeta)}{\overline{g(\zeta)}\, g(z) - 1} \right|. \tag{16.105}$$

Example 16.43. According to Example 16.22, the analytic function

$$w = \frac{z - 1}{z + 1}$$

maps the right half plane $x = \operatorname{Re} z > 0$ to the unit disk $|\,\zeta\,| < 1$. Therefore, by (16.105), the Green's function for the right half plane has the form

$$G(z;\zeta) = \frac{1}{2\,\pi}\,\log\left|\,\frac{\dfrac{z-1}{z+1} - \dfrac{\zeta-1}{\zeta+1}}{\dfrac{z-1}{z+1}\,\dfrac{\overline{\zeta}-1}{\overline{\zeta}+1} - 1}\,\right| = \frac{1}{2\,\pi}\,\log\left|\,\frac{(\overline{\zeta}+1)(z-\zeta)}{(z+1)(z-\overline{\zeta})}\,\right|. \tag{16.106}$$

One can then write the solution to the Poisson equation in a superposition as in (16.101).

16.5. Complex Integration.

The magic and power of calculus ultimately rests on the amazing fact that differentiation and integration are mutually inverse operations. And, just as complex functions have many remarkable differentiability properties not enjoyed by their real siblings, so the sublime beauty and innate structure complex integration goes far beyond its more mundane real counterpart. In the remaining two sections of this chapter, we shall develop the basics of complex integration theory and present a few of its important applications.

The first step is to motivate the definition of a complex integral. As you know, the (definite) integral of a real function, $\displaystyle\int_a^b f(t)\,dt$, is evaluated on an interval $[\,a,b\,] \subset \mathbb{R}$. In complex function theory, integrals are taken along curves in the complex plane, and thus have the flavor of the line integrals appearing in real vector calculus. Indeed, the identification of a complex number $z = x + \mathrm{i}\,y$ with a planar vector $\mathbf{x} = (\,x,y\,)^T$ will serve to connect the two theories.

Consider a curve C in the complex plane, parametrized, as in (16.65), by $z(t) = x(t) + \mathrm{i}\,y(t)$ for $a \le t \le b$. We define the *integral* of the complex function $f(z)$ along the curve C to be the complex number

$$\int_C f(z)\,dz = \int_a^b f(z(t))\,\frac{dz}{dt}\,dt. \tag{16.107}$$

We shall always assume that the integrand $f(z)$ is a well-defined complex function at each point on the curve. Let us write out the integrand

$$f(z) = u(x,y) + \mathrm{i}\,v(x,y)$$

in terms of its real and imaginary parts. Also,

$$dz = \frac{dz}{dt}\,dt = \left(\,\frac{dx}{dt} + \mathrm{i}\,\frac{dy}{dt}\,\right)dt = dx + \mathrm{i}\,dy.$$

As a result, the complex integral (16.107) splits up into a pair of real line integrals:

$$\int_C f(z)\,dz = \int_C (u + \mathrm{i}\,v)(dx + \mathrm{i}\,dy) = \int_C (u\,dx - v\,dy) + \mathrm{i}\int_C (v\,dx + u\,dy). \tag{16.108}$$

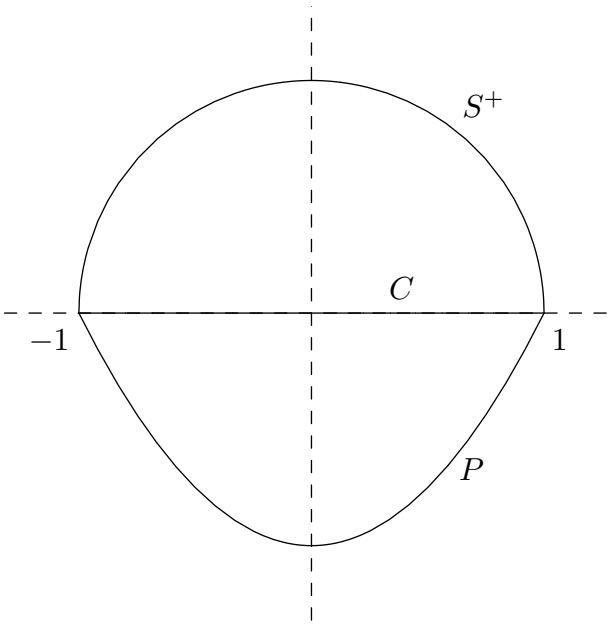


Figure 16.25. Curves for Complex Integration.

Example 16.44. Let us compute complex integrals

$$\int_C z^n \, dz, \tag{16.109}$$

of the monomial function $f(z) = z^n$, where n is an integer, along several different curves. We begin with a straight line segment along the real axis connecting the points -1 to 1, which we parametrize by $z(t) = t$ for $-1 \le t \le 1$. The defining formula (16.107) implies that the complex integral (16.109) reduces to a real integral:

$$\int_C z^n \, dz = \int_{-1}^{1} t^n \, dt = \begin{cases} 0, & n = 2\,k + 1 > 0 \text{ is odd} \\ \dfrac{2}{n+1}, & n = 2\,k \ge 0 \text{ is even.} \end{cases},$$

If $n \le -1$ is negative, then the singularity of the integrand at at the origin implies that the integral diverges, and so the complex integral is not defined.

Let us evaluate the same complex integral, but now along a parabolic arc P parametrized by

$$z(t) = t + \mathrm{i}\,(t^2 - 1), \qquad -1 \le t \le 1.$$

Note that, as graphed in Figure 16.25, the parabola connects the same two points. We again refer back to the basic definition (16.107) to evaluate the integral, so

$$\int_P z^n \, dz = \int_{-1}^{1} \big[\, t + \mathrm{i}\,(t^2 - 1) \,\big]^n (1 + 2\,\mathrm{i}\,t) \, dt.$$

We could, at this point, expand the resulting complex polynomial integrand, and then integrate term by term. A more elegant approach is to recognize that the integrand is an

exact derivative; namely, by the chain rule

$$\frac{d}{dt}\,\frac{\big[\,t+\mathrm{i}\,(t^2-1)\,\big]^{n+1}}{n+1} = \big[\,t+\mathrm{i}\,(t^2-1)\,\big]^{n}(1+2\,\mathrm{i}\,t),$$

as long as $n \neq -1$. Therefore, we can use the Fundamental Theorem of Calculus (which works equally well for real integrals of complex-valued functions), to evaluate

$$\int_P z^n\,dz = \left.\frac{\big[\,t+\mathrm{i}\,(t^2-1)\,\big]^{n+1}}{n+1}\,\right|_{t=-1}^{1} = \begin{cases} 0, & -1 \neq n = 2\,k+1 \text{ odd}, \\ \dfrac{2}{n+1}, & n = 2\,k \text{ even}. \end{cases}$$

Thus, when $n \geq 0$ is a positive integer, we obtain the same result as before. Interestingly, in this case the complex integral is well-defined even when n is a negative integer because, unlike the real line segment, the parabolic path does *not* go through the singularity of z^n at $z = 0$. The case $n = -1$ needs to be done slightly differently, and integration of $1/z$ along the parabolic path is left as an exercise for the reader — one that requires some care. We recommend trying the exercise now, and then verifying your answer once we have become a little more familiar with basic complex integration techniques.

Finally, let us try integrating around a semi-circular arc, again with the same endpoints -1 and 1. If we parametrize the semi-circle S^+ by $z(t) = e^{\,\mathrm{i}\,t}$, $0 \leq t \leq \pi$, we find

$$\begin{aligned}\int_{S^+} z^n\,dz &= \int_0^{\pi} z^n\,\frac{dz}{dt}\,dt = \int_0^{\pi} e^{\,\mathrm{i}\,n t}\,\mathrm{i}\,e^{\,\mathrm{i}\,t}\,dt = \int_0^{\pi} \mathrm{i}\,e^{\,\mathrm{i}\,(n+1)t}\,dt \\ &= \left.\frac{e^{\,\mathrm{i}\,(n+1)t}}{n+1}\,\right|_{t=0}^{\pi} = \frac{1-e^{\,\mathrm{i}\,(n+1)\pi}}{n+1} = \begin{cases} 0, & -1 \neq n = 2\,k+1 \text{ odd}, \\ -\dfrac{2}{n+1}, & n = 2\,k \text{ even}. \end{cases}\end{aligned}$$

This value is the negative of the previous cases — but this can be explained by the fact that the circular arc is oriented to go *from* 1 *to* -1 whereas the line segment and parabola both go *from* -1 *to* 1. Just as with line integrals, the direction of the curve determines the sign of the complex integral; if we reverse direction, replacing t by $-t$, we end up with the same value as the preceding two complex integrals. Moreover — again provided $n \neq -1$ — it does not matter whether we use the upper semicircle or lower semicircle to go from -1 to 1 — the result is exactly the same. However, the case $n = -1$ is an exception to this "rule". Integrating along the upper semicircle S^+ from 1 to -1 yields

$$\int_{S^+} \frac{dz}{z} = \int_0^{\pi} \mathrm{i}\,dt = \pi\,\mathrm{i}\,, \tag{16.110}$$

whereas integrating along the lower semicircle S^- from 1 to -1 yields the negative

$$\int_{S^-} \frac{dz}{z} = \int_0^{-\pi} \mathrm{i}\,dt = -\pi\,\mathrm{i}\,. \tag{16.111}$$

Hence, when integrating the function $1/z$, it makes a difference which direction we go around the origin.

Integrating z^n for any integer $n \neq -1$ around an entire circle gives zero — irrespective of the radius. This can be seen as follows. We parametrize a circle of radius r by $z(t) = re^{\,\mathrm{i}\,t}$ for $0 \leq t \leq 2\pi$. Then, by the same computation,

$$\oint_C z^n\,dz = \int_0^{2\pi} (r^n e^{\,\mathrm{i}\,nt})(r\,\mathrm{i}\,e^{\,\mathrm{i}\,t})\,dt = \int_0^{2\pi} \mathrm{i}\,r^{n+1} e^{\,\mathrm{i}\,(n+1)t}\,dt = \frac{r^{n+1}}{n+1}\,e^{\,\mathrm{i}\,(n+1)t}\,\Big|_{t=0}^{2\pi} = 0, \tag{16.112}$$

provided $n \neq -1$. Here, as in Appendix A, the circle on the integral sign serves to remind us that we are integrating around a closed curve. The case $n = -1$ remains special. Integrating once around the circle in the counter-clockwise direction yields a nonzero result

$$\oint_C \frac{dz}{z} = \int_0^{2\pi} \mathrm{i}\;dt = 2\pi\,\mathrm{i}\,. \tag{16.113}$$

Let us note that a complex integral does not depend on the particular parametrization of the curve C. It does, however, depend upon the orientation of the curve: if we traverse the curve in the reverse direction, then the complex integral changes its sign:

$$\int_{-C} f(z)\,dz = -\int_C f(z)\,dz. \tag{16.114}$$

Moreover, if we chop up the curve into two non-overlapping pieces, $C = C_1 \cup C_2$, with a common orientation, then the complex integral can be decomposed into a sum over the pieces:

$$\int_{C_1 \cup C_2} f(z) = \int_{C_1} f(z)\,dz \;+\; \int_{C_2} f(z)\,dz. \tag{16.115}$$

For instance, the integral (16.113) of $1/z$ around the circle is the difference of the individual semicircular integrals (16.110–111); the lower semicircular integral acquires a negative sign to flip its orientation so as to agree with that of the entire circle. All these facts are immediate consequences of the basic properties of line integrals, or can be easily proved directly from the defining formula (16.107).

Note: In complex integration theory, a simple closed curve is often referred to as a *contour*, and so complex integration is sometimes referred to as *contour integration*. Unless explicitly stated, we always go around contours in the *counter-clockwise* direction.

Further experiments lead us to suspect that complex integrals are usually path-independent, and hence evaluate to zero around closed contours. One must be careful, though, as the integral (16.113) makes clear. Path independence, in fact, follows from the complex version of the Fundamental Theorem of Calculus.

Theorem 16.45. *Let $f(z) = F'(z)$ be the derivative of a single-valued complex function $F(z)$ defined on a domain $\Omega \subset \mathbb{C}$. Let $C \subset \Omega$ be any curve with initial point α and final point β. Then*

$$\int_C f(z)\,dz = \int_C F'(z)\,dz = F(\beta) - F(\alpha). \tag{16.116}$$

Proof: This follows immediately from the definition (16.107) and the chain rule:

$$\int_C F'(z)\,dz = \int_a^b F'(z(t))\,\frac{dz}{dt}\,dt = \int_a^b \frac{d}{dt}\,F(z(t))\,dt = F(z(b)) - F(z(a)) = F(\beta) - F(\alpha),$$

where $\alpha = z(a)$ and $\beta = z(b)$ are the endpoints of the curve. *Q.E.D.*

For example, when $n \neq -1$, the function $f(z) = z^n$ is the derivative of the single-valued function $F(z) = \dfrac{1}{n+1}\,z^{n+1}$. Hence

$$\int_C z^n\,dz \;=\; \frac{\beta^{n+1}}{n+1} \;-\; \frac{\alpha^{n+1}}{n+1}$$

whenever C is a curve connecting α to β. When $n < 0$, the curve is not allowed to pass through the origin, $z = 0$, as it is a singularity for z^n.

In contrast, the function $f(z) = 1/z$ is the derivative of the complex logarithm

$$\log z = \log|\,z\,| + \mathrm{i}\ \mathrm{ph}\,z,$$

which is *not* single-valued on all of $\mathbb{C} \setminus \{0\}$, and so Theorem 16.45 cannot be applied directly. However, if our curve is contained within a simply connected subdomain that does not include the origin, $0 \notin \Omega \subset \mathbb{C}$, then we can use *any* single-valued branch of the complex logarithm to evaluate the integral

$$\int_C \frac{dz}{z} = \log\beta - \log\alpha,$$

where α, β are the endpoints of the curve. Since the common multiples of $2\pi\,\mathrm{i}$ cancel, the answer does not depend upon which particular branch of the complex logarithm is chosen as long as we are consistent in our choice. For example, on the upper semicircle S^+ of radius 1 going from 1 to -1,

$$\int_{S^+} \frac{dz}{z} = \log(-1) - \log 1 = \pi\,\mathrm{i}\,,$$

where we use the branch of $\log z = \log|\,z\,| + \mathrm{i}\,\mathrm{ph}\,z$ with $0 \leq \mathrm{ph}\,z \leq \pi$. On the other hand, if we integrate on the lower semi-circle S^- going from 1 to -1, we need to adopt a different branch, say that with $-\pi \leq \mathrm{ph}\,z \leq 0$. With this choice, the integral becomes

$$\int_{S^-} \frac{dz}{z} = \log(-1) - \log 1 = -\pi\,\mathrm{i}\,,$$

thus reproducing (16.110, 111). Pay particular attention to the different values of $\log(-1)$ in the two cases!

The most important consequence of Theorem 16.45 is that, as long as the integrand $f(z)$ has a single-valued anti-derivative, its complex integral is independent of the path connecting two points — the value only depends on the endpoints of the curve and not how one gets from point α to point β.

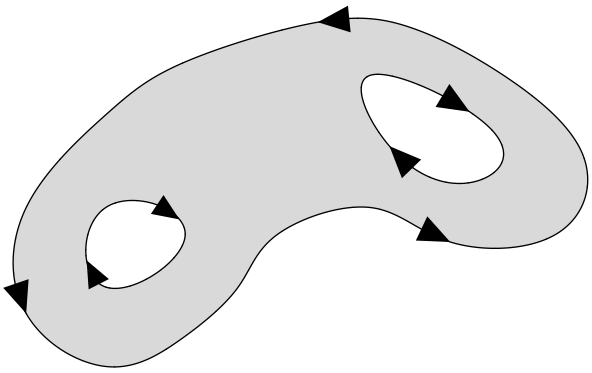


Figure 16.26. Orientation of Domain Boundary.

Theorem 16.46. *Let $f(z) = F'(z)$, where $F(z)$ is a single-valued complex function for $z \in \Omega$. If $C \subset \Omega$ is any closed curve, then*

$$\oint_C f(z)\, dz = 0. \tag{16.117}$$

Conversely, if (16.117) *holds for all closed curves $C \subset \Omega$ contained in the domain of definition of $f(z)$, then f admits a single-valued complex anti-derivative with $F'(z) = f(z)$.*

Proof: We have already demonstrated the first statement. As for the second, we define

$$F(z) = \int_{z_0}^{z} f(z)\, dz,$$

where $z_0 \in \Omega$ is any fixed point, and we choose any convenient curve $C \subset \Omega$ connecting[†] z_0 to z. (16.117) assures us that the value does not depend on the chosen path. The proof that this formula does define an anti-derivative of f is left as an exercise, which can be solved in the same fashion as the case of a real line integral, cf. (22.41). *Q.E.D.*

The preceding considerations suggest the following fundamental theorem, due in its general form to Cauchy. Before stating it, we introduce the convention that a complex function $f(z)$ is to be called *analytic on a domain* $\Omega \subset \mathbb{C}$ provided it is analytic at every point inside Ω and, in addition, remains (at least) continuous on the boundary $\partial\Omega$. When Ω is bounded, its boundary $\partial\Omega$ consists of one or more simple closed curves. In general, as in Green's Theorem A.26, we orient $\partial\Omega$ so that the domain is always on our left hand side. This means that the outermost boundary curve is traversed in the counter-clockwise direction, but any interior holes are take on a clockwise orientation. Our convention is depicted in Figure 16.26.

Theorem 16.47. *If $f(z)$ is analytic on a bounded domain $\Omega \subset \mathbb{C}$, then*

$$\oint_{\partial\Omega} f(z)\, dz = 0. \tag{16.118}$$

[†] This assumes Ω is a connected domain; otherwise, apply the result to its individual connected components.

Proof: If we apply Green's Theorem to the two real line integrals in (16.108), we find

$$\oint_{\partial\Omega} u\,dx - v\,dy = \iint_{\Omega} \left(-\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \right) = 0, \quad \oint_{\partial\Omega} v\,dx + u\,dy = \iint_{\Omega} \left(\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y} \right) = 0,$$

both of which vanish by virtue of the Cauchy–Riemann equations (16.22). *Q.E.D.*

If the domain of definition of our complex function $f(z)$ is simply connected, then, by definition, the interior of any closed curve $C \subset \Omega$ is contained in Ω, and hence Cauchy's Theorem 16.47 implies path independence of the complex integral within Ω.

Corollary 16.48. *If $f(z)$ is analytic on a simply connected domain $\Omega \subset \mathbb{C}$, then its complex integral $\displaystyle\int_C f(z)\,dz$ for $C \subset \Omega$ is independent of path. In particular,*

$$\oint_C f(z)\,dz = 0 \tag{16.119}$$

for any closed curve $C \subset \Omega$.

Remark: Simple connectivity of the domain is an essential hypothesis — our evaluation (16.113) of the integral of $1/z$ around thye unit circle provides a simple counterexample to (16.119) in the non-simply connected domain $\Omega = \mathbb{C} \setminus \{0\}$. Interestingly, this result also admits a converse: a continuous function $f(z)$ that satisfies (16.119) for *all* closed curves is necessarily analytic; see [**4**] for a proof.

We will also require a slight generalization of this result.

Proposition 16.49. *If $f(z)$ is analytic in a domain that contains two simple closed curves S and C, and the entire region lying between them, then, assuming they are oriented in the same direction,*

$$\oint_C f(z)\,dz = \oint_S f(z)\,dz. \tag{16.120}$$

Proof: If C and S do not cross each other, we let Ω denote the domain contained between them, so that $\partial\Omega = C \cup S$; see the first plot in Figure 16.27. According to Cauchy's Theorem 16.47, $\displaystyle\oint_{\partial\Omega} f(z) = 0$. Now, our orientation convention for $\partial\Omega$ means that the outer curve, say C, is traversed in the counter-clockwise direction, while the inner curve S has the opposite, clockwise orientation. Therefore, if we assign both curves the same counter-clockwise orientation,

$$0 = \oint_{\partial\Omega} f(z) = \oint_C f(z)\,dz - \oint_S f(z)\,dz,$$

proving (16.120).

If the two curves cross, we can construct a nearby curve $K \subset \Omega$ that neither crosses, as in the second sketch in Figure 16.27. By the preceding paragraph, each integral is equal to to that over the third curve,

$$\oint_C f(z)\,dz = \oint_K f(z)\,dz = \oint_S f(z)\,dz,$$

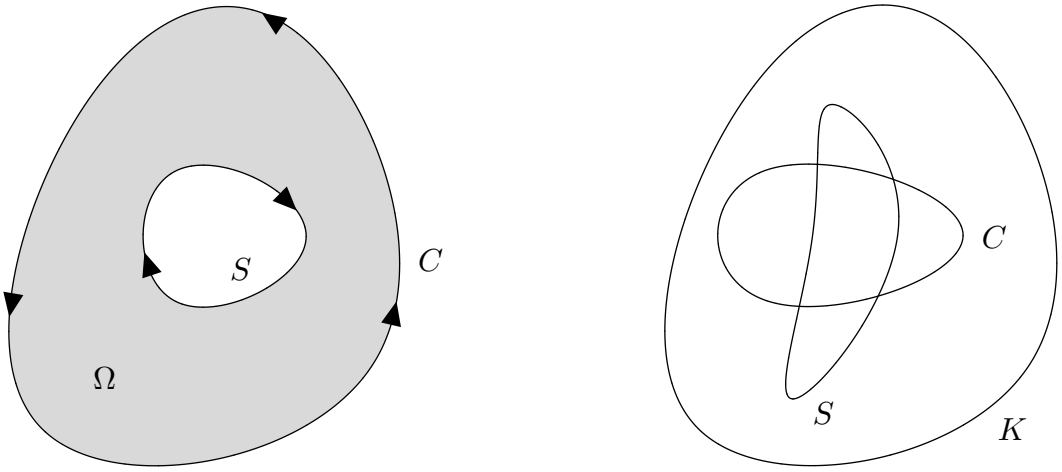


Figure 16.27. Integration Around Two Closed Curves.

and formula (16.120) remains valid. *Q.E.D.*

Example 16.50. Consider the function $f(z) = z^n$ where n is an integer[†]. In (16.112), we already computed

$$\oint_C z^n \, dz = \begin{cases} 0, & n \neq -1, \\ 2\pi \, \mathrm{i}\,, & n = -1, \end{cases} \tag{16.121}$$

when C is a circle centered at $z = 0$. When $n \geq 0$, Theorem 16.45 immediately implies that the integral of z^n is 0 over *any* closed curve in the plane. The same applies in the cases $n \leq -2$ provided the curve does not pass through the singular point $z = 0$. In particular, the integral is zero around closed curves encircling the origin, even though z^n for $n \leq -2$ has a singularity inside the curve and so Cauchy's Theorem 16.47 does not apply as stated.

The case $n = -1$ has particular significance. Here, Proposition 16.49 implies that the integral is the same as the integral around a circle — provided the curve C also goes once around the origin in a counter-clockwise direction. Thus (16.113) holds for any closed curve that goes counter-clockwise once around the origin. More generally, if the curve goes several times around the origin[‡], then

$$\oint_C \frac{dz}{z} = 2k\pi \, \mathrm{i} \tag{16.122}$$

is an integer multiple of $2\pi \, \mathrm{i}$. The integer k is called the *winding number* of the curve C, and measures the total number of times C goes around the origin. For instance, if C winds three times around 0 in a counter-clockwise fashion, then $k = 3$, while $k = -5$ indicates

[†] When n is fractional or irrational, the integrals are not well-defined owing to the branch point at the origin.

[‡] Such a curve is undoubtedly not simple and must necessarily cross over itself.

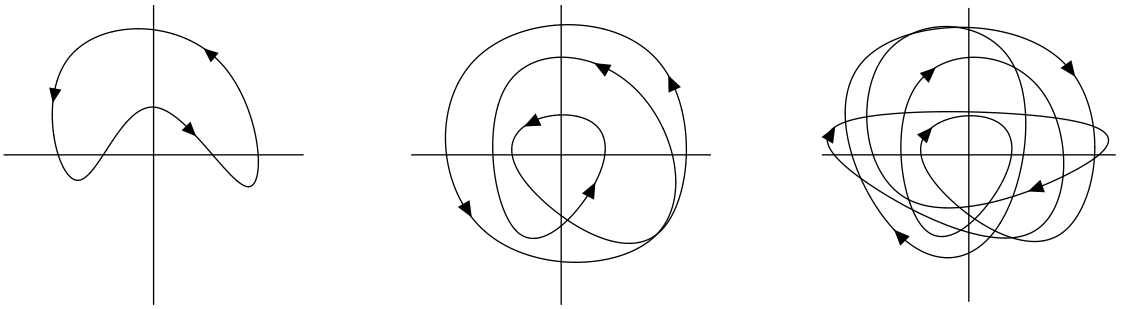


Figure 16.28. Winding Numbers.

that the curve winds 5 times around 0 in a clockwise direction, as in Figure 16.28. In particular, a winding number $k = 0$ indicates that C is not wrapped around the origin. If C represents a loop of string wrapped around a pole (the *pole* of $1/z$ at 0) then a winding number $k = 0$ would indicate that the string can be disentangled from the pole without cutting; nonzero winding numbers would indicate that the string is truly entangled[§].

Lemma 16.51. *If C is any simple closed curve, and a is any point not lying on C, then*

$$\oint_C \frac{dz}{z-a} = \begin{cases} 2\pi\,\mathrm{i}\,, & a \quad \text{inside } C \\ 0 & a \quad \text{outside } C. \end{cases} \tag{16.123}$$

If $a \in C$, then the integral does not converge.

Proof: Note that the integrand $f(z) = 1/(z-a)$ is analytic everywhere except at $z = a$, where it has a simple pole. If a is outside C, then Cauchy's Theorem 16.47 applies, and the integral is zero. On the other hand, if a is inside C, then Proposition 16.49 implies that the integral is equal to the integral around a circle centered at $z = a$. The latter integral can be computed directly by using the parametrization $z(t) = a + r\,e^{\,\mathrm{i}\,t}$ for $0 \le t \le 2\pi$, as in (16.113). *Q.E.D.*

Example 16.52. Let $D \subset \mathbb{C}$ be a closed and *connected* domain. Let $a, b \in D$ be two points in D. Then

$$\oint_C \left(\frac{1}{z-a} - \frac{1}{z-b} \right) dz = \oint_C \frac{dz}{z-a} - \oint_C \frac{dz}{z-b} = 0$$

for any closed curve $C \subset \Omega = \mathbb{C} \setminus D$ lying outside the domain D. This is because, by connectivity of D, either C contains both points in its interior, in which case both integrals equal $2\pi\,\mathrm{i}\,$, or C contains neither point, in which case both integrals are 0. According to Theorem 16.46, the integrand admits a single-valued anti-derivative on the domain Ω, even

[§] Actually, there are more subtle three-dimensional considerations that come into play, and even strings with zero winding number cannot be removed from the pole without cutting if they are linked in some nontrivial manner, cf. [**113**]. Can you think of an example?

though each individual term is the derivative of a multiply-valued complex logarithm. The conclusion is that, while the individual logarithms are multiply-valued, their difference

$$F(z) = \log(z-a) - \log(z-b) \tag{16.124}$$

is a consistent, single-valued complex function on all of $\Omega = \mathbb{C} \setminus D$. The difference (16.124), in fact, an infinite number of possible values, differing by integer multiples of $2\pi\,\mathrm{i}$; the ambiguity can be removed by choosing one of its value at a single point in Ω. These conclusions rest on the fact that D is connected, and are *not* valid, say, for the twice-punctured plane $\mathbb{C} \setminus \{a, b\}$.

We are sometimes interested in estimating the size of a complex integral. The basic inequality bounds it in terms of an arc length integral.

Proposition 16.53. *The modulus of the integral of the complex function f along a curve C is bounded by the integral of its modulus with respect to arc length:*

$$\left| \int_C f(z)\,dz \right| \;\le\; \int_C |\, f(z)\,|\,ds. \tag{16.125}$$

Proof: We begin with a simple lemma on real integrals of complex functions.

Lemma 16.54. *Let $f(t)$ be a complex-valued function depending on the real variable $a \le t \le b$. Then*

$$\left| \int_a^b f(t)\,dt \right| \;\le\; \int_a^b |\, f(t)\,|\,dt. \tag{16.126}$$

Proof: If $\displaystyle\int_a^b f(t)\,dt = 0$, the inequality is trivial. Otherwise, let $\theta = \operatorname{ph} \displaystyle\int_a^b f(t)\,dt$. Then, using Exercise 3.6.9,

$$\left| \int_a^b f(t)\,dt \right| = \operatorname{Re}\, \left[e^{-\mathrm{i}\,\theta} \int_a^b f(t)\,dt \right] = \int_a^b \operatorname{Re}\, \left[e^{-\mathrm{i}\,\theta}\, f(t) \right] dt \le \int_a^b |\, f(t)\,|\,dt,$$

which proves the lemma. *Q.E.D.*

To prove the proposition, we write out the complex integral, and use (16.126) as follows:

$$\left| \int_C f(z)\,dz \right| = \left| \int_a^b f(z(t))\,\frac{dz}{dt}\,dt \right| \;\le\; \int_a^b |\, f(z(t))\,|\, \left| \frac{dz}{dt} \right| dt = \int_C |\, f(z)\,|\,ds,$$

since $|\,dz\,| = |\,\dot z\,|\,dt = \sqrt{\dot x^2 + \dot y^2}\,dt = ds$ is the arc length integral element (A.30). *Q.E.D.*

Corollary 16.55. *If the curve C has length $L = \mathcal{L}(C)$, and $f(z)$ is an analytic function such that $|\, f(z)\,| \le M$ for all points $z \in C$, then*

$$\left| \int_C f(z)\,dz \right| \;\le\; M\,L. \tag{16.127}$$

Lift and Circulation

In fluid mechanical applications, the complex integral can be assigned an important physical interpretation. As above, we consider the steady state flow of an incompressible, irrotational fluid. Let $f(z) = u(x,y) - \mathrm{i}\, v(x,y)$ denote the complex velocity corresponding to the real velocity vector $\mathbf{v} = (\, u(x,y), v(x,y)\,)^T$ at the point (x,y).

As we noted in (16.108), the integral of the complex velocity $f(z)$ along a curve C can be written as a pair of real line integrals. In the present situation,

$$\int_C f(z)\,dz = \int_C (u - \mathrm{i}\, v)(dx + \mathrm{i}\, dy) = \int_C (u\,dx + v\,dy) - \mathrm{i} \int_C (v\,dx - u\,dy). \tag{16.128}$$

According to (A.37, 42), the real part is the circulation integral

$$\int_C \mathbf{v} \cdot d\mathbf{x} = \int_C u\,dx + v\,dy, \tag{16.129}$$

while the imaginary part is minus the flux integral

$$\int_C \mathbf{v} \cdot \mathbf{n}\, ds = \int_C \mathbf{v} \wedge d\mathbf{x} = \int_C v\,dx - u\,dy, \tag{16.130}$$

along the curve C under the associated steady state fluid flow.

If the complex velocity admits a single-valued complex potential

$$\chi(z) = \varphi(z) - \mathrm{i}\,\psi(z), \qquad \text{where} \qquad \chi'(z) = f(z)$$

— which is always the case if its domain of definition is simply connected — then the complex integral is independent of path, and one can use the Fundamental Theorem 16.45 to evaluate it:

$$\int_C f(z)\,dz = \chi(\beta) - \chi(\alpha) \tag{16.131}$$

for any curve C connecting α to β. Path independence of the complex integral reconfirms the path independence of the flux and circulation integrals for irrotational, incompressible fluid dynamics. The real part of formula (16.131) evaluates the circulation integral

$$\int_C \mathbf{v} \cdot d\mathbf{x} = \int_C \nabla\varphi \cdot d\mathbf{x} = \varphi(\beta) - \varphi(\alpha), \tag{16.132}$$

as the difference in the values of the (real) potential at the endpoints α, β of the curve C. On the other hand, the imaginary part of formula (16.131) computes the flux integral

$$\int_C \mathbf{v} \wedge d\mathbf{x} = \int_C \nabla\psi \cdot d\mathbf{x} = \psi(\beta) - \psi(\alpha), \tag{16.133}$$

as the difference in the values of the stream function at the endpoints of the curve. Thus, the stream function acts as a "flux potential" for the flow. Thus, for ideal fluid flows, flux is independent of path, and depends only upon the endpoints of the curve. In particular, if C is a closed contour,

$$\oint_C \mathbf{v} \cdot d\mathbf{x} \;=\; 0 \;=\; \oint_C \mathbf{v} \wedge d\mathbf{x}, \tag{16.134}$$

and so there is no net circulation or flux along any closed curve in this situation.

In aerodynamics, lift is the result of the circulation of the fluid (air) around the body, [**14**, **164**]. More precisely, let $D \subset \mathbb{C}$ be a closed, bounded subset representing the cross-section of a cylindrical body, e.g., an airplane wing. The velocity vector field $\mathbf{v}$ of a steady state flow around the exterior of the body is defined on the domain $\Omega = \mathbb{C} \setminus D$. According to Blasius' Theorem, the body will experience a net lift if and only if it has nonvanishing circulation integral $\displaystyle\oint_C \mathbf{v} \cdot d\mathbf{x} \neq 0$, where C is any simple closed contour encircling the body. However, if the complex velocity admits a single-valued complex potential in Ω, then (16.134) tells us that the circulation is automatically zero, and so the body cannot experience any lift!

Example 16.56. Let us investigate the role of lift in flow around an airfoil. Consider first the flow around a disk, as discussed in Examples 16.17 and 16.39. The Joukowski potential $\chi(z) = z + z^{-1}$ is a single-valued analytic function everywhere except at the origin $z = 0$. Therefore, the circulation integral (16.132) around any contour encircling the disk will vanish, and hence the disk experiences no net lift. This is more or less evident from the Figure 16.13 graphing the streamlines of the flow; they are symmetric above and below the disk, and hence there cannot be any net force in the vertical direction.

Any conformal map will maintain single-valuedness of the complex potentials, and hence preserve the zero-circulation property. In particular, all the flows past airfoils constructed in Example 16.41 also admit single-valued potentials, and so also have zero circulation integral. Such an airplane will not fly, because its wings experience no lift! Of course, physical airplanes do fly, and so there must be some physical assumption we are neglecting in our treatment of flow past a body. Abandoning incompressibility or irrotationality would draw us outside the manicured gardens of complex variable theory, and into the jungles of fully nonlinear partial differential equations of fluid mechanics. Moreover, although air is slightly compressible, water is, for all practical purposes, incompressible, and hydrofoils do experience lift when traveling through water.

The only way to introduce lift into the picture is through a (single-valued) complex velocity with a non-zero circulation integral, and this requires that its complex potential be multiply-valued. The one function that we know that has such a property is the complex logarithm

$$\lambda(z) = \log(a\, z + b), \qquad \text{whose derivative} \qquad \lambda'(z) = \frac{1}{a\, z + b}$$

is single-valued away from the singularity at $z = -b/a$. Thus, we are naturally led to introduce the family of complex potentials[†]

$$\chi_k(z) = \frac{1}{2}\left(z + \frac{1}{z} \right) - \mathrm{i}\, k\, \log z. \tag{16.135}$$

[†] We center the logarithmic singularity at the origin in order to maintain the no flux boundary conditions on the unit circle. Moreover, Example 16.52 tells us that more than one logarithm in the potential is redundant, since the difference of any two logarithms is effectively a single-valued function, and hence contributes nothing to the circulation integral.

According to Exercise ■, the coefficient k must be real in order to maintain the no flux boundary conditions on the unit circle. By (16.128), the circulation is equal to the real part of the integral of the complex velocity

$$f_k(z) = \frac{d\chi_k}{dz} = \frac{1}{2} - \frac{1}{2\,z^2} - \frac{\mathrm{i}\,k}{z}\,. \tag{16.136}$$

By Cauchy's Theorem 16.47 coupled with formula (16.123), if C is a curve going once around the disk in a counter-clockwise direction, then

$$\oint_C f_k(z)\,dz = \oint_C \left(\frac{1}{2} - \frac{1}{2\,z^2} - \frac{\mathrm{i}\,k}{z}\right) dz = 2\,\pi\,k.$$

Therefore, when $\operatorname{Re} k \neq 0$, the circulation integral is non-zero, and the cylinder experiences a net lift. In Figure liftc■, the streamlines for the flow corresponding to a few representative values of k are plotted. Note the asymmetry of the streamlines that accounts for the lift experienced by the disk.

When we compose the modified lift potentials (16.135) with the Joukowski transformation (16.100), we obtain a complex potential

$$\Theta_k(\zeta) = \chi_k(z) \qquad \text{when} \qquad \zeta = \frac{1}{2}\left(w + \frac{1}{w}\right) = \frac{1}{2}\left(a\,z + \beta + \frac{1}{a\,z+\beta}\right)$$

for flow around the corresponding airfoil — the image of the unit disk. The conformal mapping does not affect the value of the complex integrals, and hence, for any $k \neq 0$, there is a nonzero circulation around the airfoil under the modified fluid flow. This circulation is the cause of a net lift on the airfoil, and at last our airplane will fly!

However, there is now a slight embarrassment of riches, since we have designed flows around the airfoil with an *arbitrary* value $2\,\pi\,k$ for the circulation integral, and hence having an *arbitrary* amount of lift! Which of these possible flows most closely realizes the true physical version with the correct amount of lift? In his 1902 thesis, Martin Kutta hypothesized that Nature chooses the constant k so as to keep the velocity of the flow at the trailing edge of the airfoil, $\zeta = 1$, to be finite. With some additional analysis, it turns out that this condition serves to uniquely specify k, and yields a reasonably good physical approximation to the actual lift experienced by such an airfoil in flight, provided the tilt or attack angle of the airfoil in the flow is not too large. Further details, can be found in several references, including [**14**, **124**, **114**, **164**].

16.6. Cauchy's Integral Formulae and the Calculus of Residues.

Cauchy's Integral Theorem 16.47 and its consequences underlie almost all applications of complex integration. The fact that we can move the contours of complex integrals around freely — as long as we do not cross over singularities of the integrand — grants us great flexibility in their evaluation. A key consequence of Cauchy's Theorem is that the value of a complex integral around a closed contour depends only upon the nature of the singularities of the integrand that happen to lie inside the contour. This observation inspires us to develop a direct method, known as the "calculus of residues", for evaluating

such integrals. The residue method effectively bypasses the Fundamental Theorem of Calculus — no antiderivatives are required! Remarkably, the method of residues can even be applied to evaluate certain types of real definite integrals, as the final examples in this section shall demonstrate.

Cauchy's Integral Formula

The first important consequence of Cauchy's Theorem is the justly famous *Cauchy integral formulae*. It enables us to compute the value of an analytic function at a point by evaluating a contour integral around a closed curve encircling the point.

Theorem 16.57. *Let $\Omega \subset \mathbb{C}$ be a bounded domain with boundary $\partial\Omega$, and let $a \in \Omega$. If $f(z)$ is analytic on Ω, then*

$$f(a) = \frac{1}{2\,\pi\,\mathrm{i}} \oint_{\partial\Omega} \frac{f(z)}{z-a}\,dz. \tag{16.137}$$

Remark: As always, we traverse the boundary curve $\partial\Omega$ so that the domain Ω lies on our left. In most applications, Ω is simply connected, and so $\partial\Omega$ is a simple closed curve oriented in the counter-clockwise direction.

It is worth emphasizing that Cauchy's formula (16.137) is *not* a form of the Fundamental Theorem of Calculus, since we are reconstructing the function by *integration* — not its anti-derivative! Cauchy's formula is a cornerstone of complex analysis, and has no real counterpart, once again underscoring the profound difference between complex and real analysis.

Proof: We first prove that the difference quotient

$$g(z) = \frac{f(z) - f(a)}{z-a}$$

is an analytic function on all of Ω. The only problematic point is at $z = a$ where the denominator vanishes. First, by the definition of complex derivative,

$$g(a) = \lim_{z \to a} \frac{f(z) - f(a)}{z-a} = f'(a)$$

exists and therefore $g(z)$ is well-defined and, in fact, continuous at $z = a$. Secondly, we can compute its derivative at $z = a$ directly from the definition:

$$g'(a) = \lim_{z \to a} \frac{g(z) - g(a)}{z-a} = \lim_{z \to a} \frac{f(z) - f(a) - f'(a)\,(z-a)}{(z-a)^2} = \tfrac{1}{2}\,f''(a),$$

where we use Taylor's Theorem C.1 (or l'Hôpital's rule) to evaluate the final limit. Knowing that g is differentiable at $z = a$ suffices to establish that it is analytic on all of Ω. Thus, we may appeal to Cauchy's Theorem 16.47, and conclude that

$$\begin{aligned} 0 = \oint_{\partial\Omega} g(z)\,dz &= \oint_{\partial\Omega} \frac{f(z) - f(a)}{z-a}\,dz = \oint_{\partial\Omega} \frac{f(z)}{z-a}\,dz - f(a) \oint_{\partial\Omega} \frac{dz}{z-a} \\ &= \oint_{\partial\Omega} \frac{f(z)}{z-a}\,dz - 2\,\pi\,\mathrm{i}\,f(a). \end{aligned}$$

The second integral was evaluated using (16.123). Rearranging terms completes the proof of the Cauchy formula. *Q.E.D.*

Remark: The proof shows that if, in contrast, $a \notin \overline{\Omega}$, then the Cauchy integral vanishes:

$$\frac{1}{2\,\pi\,\mathrm{i}} \oint_{\partial\Omega} \frac{f(z)}{z-a}\,dz = 0.$$

If $a \in \partial\Omega$, then the integral does not converge.

Let us see how we can apply this result to evaluate seemingly intractable complex integrals.

Example 16.58. Suppose that you are asked to compute the contour integral

$$\oint_C \frac{e^z\,dz}{z^2-2\,z-3}$$

where C is a circle of radius 2 centered at the origin. A direct evaluation is not possible, since the integrand does not have an elementary anti-derivative[†]. However, we note that

$$\frac{e^z}{z^2-2\,z-3} = \frac{e^z}{(z+1)(z-3)} = \frac{f(z)}{z+1} \qquad \text{where} \qquad f(z) = \frac{e^z}{z-3}$$

is analytic in the disk $|\,z\,| \le 2$ since its only singularity, at $z = 3$, lies outside the contour C. Therefore, by Cauchy's formula (16.137), we immediately obtain the integral

$$\oint_C \frac{e^z\,dz}{z^2-2\,z-3} = \oint_C \frac{f(z)}{z+1}\,dz = 2\,\pi\,\mathrm{i}\,f(-1) = -\,\frac{\pi\,\mathrm{i}}{2\,e}\,.$$

Note: Path independence implies that the integral has the same value on any other simple closed contour, provided it is oriented in the usual counter-clockwise direction and encircles the point $z = 1$ but not the point $z = 3$.

In this example, if the contour encloses both singularities, at $z = 1$ and 3, then we cannot apply Cauchy's formula directly. However, as we will see, Theorem 16.57 can be adapted in a direct manner to such situations. This more general result will lead us directly to the calculus of residues, to be discussed shortly.

Derivatives by Integration

The fact that we can recover values of complex functions by integration is noteworthy. Even more amazing[‡] is the fact that we can compute *derivatives* of complex functions by *integration* — turning the Fundamental Theorem on its head! Let us differentiate both sides of Cauchy's formula (16.137) with respect to a. The integrand in the Cauchy

[†] At least not one listed in any integration tables, e.g., [**81**]. A more profound analysis, [**Int**], confirms that its anti-derivative *cannot* be expressed in closed form using elementary functions.

[‡] Readers who have successfully tackled Exercise ■ may be less shocked by this fact.

formula is sufficiently nice so as to allow us to bring the derivative inside the integral sign. Moreover, the derivative of the Cauchy integrand with respect to a is easily found:

$$\frac{\partial}{\partial a}\left(\frac{f(z)}{z-a}\right) = \frac{f(z)}{(z-a)^2}\,.$$

In this manner, we deduce an integral formulae for the *derivative* of an analytic function:

$$f'(a) = \frac{1}{2\pi\,\mathrm{i}}\oint_C \frac{f(z)}{(z-a)^2}\,dz, \tag{16.138}$$

where, as before, C is any closed curve that goes once around the point $z = a$ in a counterclockwise direction†. Further differentiation yields the general integral formulae

$$f^{(n)}(a) = \frac{n!}{2\pi\,\mathrm{i}}\oint_C \frac{f(z)}{(z-a)^{n+1}}\,dz \tag{16.139}$$

that expresses the n^{th} order derivative of a complex function in terms of a contour integral.

These remarkable formulae, which again have no counterpart in real function theory, can be used to prove our earlier claim that an analytic function is infinitely differentiable, and thereby complete the proof of Theorem 16.9.

Example 16.59. Let us compute the integral

$$\oint_C \frac{e^z\,dz}{z^3 - z^2 - 5\,z - 3} = \oint_C \frac{e^z\,dz}{(z+1)^2(z-3)}\,,$$

around the circle of radius 2 centered at the origin. We use (16.138) with

$$f(z) = \frac{e^z}{z-3}\,, \qquad \text{whereby} \qquad f'(z) = \frac{(z-4)\,e^z}{(z-3)^2}\,.$$

Since $f(z)$ is analytic inside C, the integral formula (16.138) that

$$\oint_C \frac{e^z\,dz}{z^3 - z^2 - 5\,z - 3} = \oint_C \frac{f(z)}{(z+1)^2}\,dz = 2\pi\,\mathrm{i}\,f'(-1) = -\frac{5\pi\,\mathrm{i}}{8\,e}\,.$$

One application is the following remarkable result due to Liouville, whom we already met in Section 11.5. It says that the only bounded complex functions are the constants!

Theorem 16.60. *If $f(z)$ is analytic at all $z \in \mathbb{C}$, and satisfies $|\,f(z)\,| \le M$ for some fixed positive number $M > 0$, then $f(z) \equiv c$ is constant.*

Proof: According to Cauchy's formula (16.137), for any point $a \in \mathbb{C}$,

$$f'(a) = \frac{1}{2\pi\,\mathrm{i}}\oint_{C_R} \frac{f(z)}{(z-a)^2}\,dz,$$

† Or, more generally, has winding number $+1$ around the point $z = a$.

where we take $C_R = \{ z \,|\, |z-a| = R \}$ to be a circle of radius R centered at $z = a$. We then estimate the complex integral using (16.125), whence

$$|f'(a)| = \frac{1}{2\pi} \left| \oint_{C_R} \frac{f(z)}{(z-a)^2}\, dz \right| \le \frac{1}{2\pi} \oint_{C_R} \frac{|f(z)|}{|z-a|^2}\, ds \le \frac{1}{2\pi} \oint_{C_R} \frac{M}{R^2}\, ds = \frac{M}{R},$$

since the length of C_R is $2\pi R$. Since $f(z)$ is analytic everywhere, we can let $R \to \infty$ and conclude that $f'(a) = 0$. Since this occurs for all possible points a, we conclude that $f'(z) \equiv 0$ is everywhere zero, which suffices to prove constancy of $f(z)$. *Q.E.D.*

One outstanding application of Liouville's Theorem 16.60 is a proof of the Fundamental Theorem of Algebra, first proved by Gauss in 1799; see [**69**] for an extensive discussion. Although it is, in essence, a purely algebraic result, this proof relies in an essential way on complex analysis and complex integration.

Theorem 16.61. *Every nonconstant (complex or real) polynomial $f(z)$ has at least one complex root $z_0 \in \mathbb{C}$.*

Proof: Suppose

$$f(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0 \neq 0 \qquad \text{for all} \qquad z \in \mathbb{C}. \tag{16.140}$$

Then we claim that its reciprocal

$$g(z) = \frac{1}{f(z)} = \frac{1}{a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0}$$

satisfies the hypotheses of Theorem 16.60, and hence must be constant, in contradiction to our hypothesis. Therefore, $f(z)$ cannot be non-zero for all z, and this establishes the result.

To prove the claim, note first that our nonvanishing assumption (16.140) implies that $g(z)$ is analytic for all $z \in \mathbb{C}$. Moreover,

$$\begin{aligned} |f(z)| &= |z|^n \left| a_n + \frac{a_{n-1}}{z} + \cdots + \frac{a_1}{z^{n-1}} + \frac{a_0}{z^n} \right| \\ &\le |z|^n \big(|a_n| + |a_{n-1}| + \cdots + |a_1| + |a_0| \big), \end{aligned} \qquad \text{whenever} \qquad |z| \ge 1,$$

which implies that $|f(z)| \to \infty$ as $|z| \to \infty$, and so

$$|g(z)| = \frac{1}{|f(z)|} \longrightarrow 0 \qquad \text{as} \qquad |z| \to \infty.$$

This suffices to prove that $|g(z)| \le M$ is bounded for $z \in \mathbb{C}$. *Q.E.D.*

Corollary 16.62. *Every complex polynomial of degree n can be factored,*

$$f(z) = a_n (z - z_1)(z - z_2) \cdots (z - z_n)$$

where $z_1, \dots, z_n$ are the roots of $f(z)$, listed in accordance with their multiplicity.

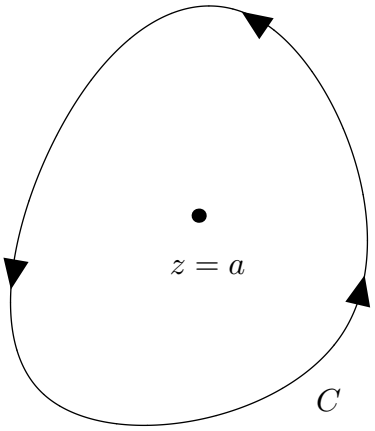


Figure 16.29. Computing a Residue.

Proof: Theorem 16.61 guarantees that there is at least one point $z_1 \in \mathbb{C}$ where $f(z_1) = 0$. Therefore, by the rules of polynomial factorization, we can write

$$f(z) = (z - z_1)\, g(z)$$

where $g(z)$ is a polynomial of degree $n - 1$. A straightforward induction on the degree of the polynomial completes the proof. *Q.E.D.*

The Calculus of Residues

Cauchy's Theorem and Integral Formulae provide us with some amazingly versatile tools for evaluating complicated complex integrals. The upshot is that one only needs to understand the singularities of the integrand within the domain of integration — no indefinite integration is required! With a little more work, we are led to a general method for efficiently computing contour integrals, known as the *Calculus of Residues*. While the residue method has no counterpart in real integration theory, it can, remarkably, be used to evaluate a large variety of interesting definite real integrals, including many without an explicitly known anti-derivative.

Definition 16.63. Let $f(z)$ be an analytic function for all z near, but not equal to a. The *residue* of $f(z)$ at the point $z = a$ is defined by the contour integral

$$\operatorname*{Res}_{z=a} f(z) = \frac{1}{2\pi\,\mathrm{i}} \oint_C f(z)\, dz, \tag{16.141}$$

where C is any simple, closed curve that contains a in its interior, oriented, as always, in a counter-clockwise direction, and such that $f(z)$ is analytic everywhere inside C except at the point $z = a$; see Figure 16.29. For example, C could be a small circle centered at a. The residue is a complex number, and tells us important information about the singularity of $f(z)$ at $z = a$.

The simplest example is the monomial function $f(z) = cz^n$, where c is a complex constant and the exponent n is assumed to be an integer. (Residues are not defined at branch points.) According to (16.112),

$$\operatorname*{Res}_{z=0} cz^n = \frac{1}{2\pi\,\mathrm{i}} \oint_C cz^n\,dz = \begin{cases} 0, & n \neq -1, \\ c, & n = -1. \end{cases} \tag{16.142}$$

Thus, only the case $n = -1$ gives a nonzero residue. The residue singles out the function $1/z$, which, not coincidentally, is the only one with a logarithmic, and multiply-valued, antiderivative.

Cauchy's Theorem 16.47, when applied to the integral in (16.141), implies that if $f(z)$ is analytic at $z = a$, then it has zero residue at a. Therefore, all the monomials, including $1/z$, have zero residue at any nonzero point:

$$\operatorname*{Res}_{z=a} cz^n = 0 \qquad \text{for} \qquad a \neq 0. \tag{16.143}$$

Since integration is a linear operation, the residue is a linear operator, mapping complex functions to complex numbers:

$$\operatorname*{Res}_{z=a} \big[\, f(z) + g(z) \,\big] = \operatorname*{Res}_{z=a} f(z) + \operatorname*{Res}_{z=a} g(z), \qquad \operatorname*{Res}_{z=a} \big[\, c\,f(z) \,\big] = c \operatorname*{Res}_{z=a} f(z), \tag{16.144}$$

for any complex constant c. Thus, by linearity, the residue of any finite linear combination of monomials,

$$f(z) = \frac{c_{-m}}{z^m} + \frac{c_{-m+1}}{z^{m-1}} + \,\cdots\, + \frac{c_{-1}}{z} + c_0 + c_1\,z + \,\cdots\, + c_n\,z^n = \sum_{k=-m}^{n} c_k\,z^k,$$

is equal to

$$\operatorname*{Res}_{z=0} f(z) = c_{-1}.$$

Thus, the residue effectively picks out the coefficient of the term $1/z$ in such an expansion.

The easiest nontrivial residues to compute are at the poles of a function. According to (16.29), the function $f(z)$ has a *simple pole* at $z = a$ if

$$h(z) = (z - a)\,f(z) \tag{16.145}$$

is analytic at $z = a$ and $h(a) \neq 0$. The next result allows us to bypass the contour integral when evaluating such a residue.

Lemma 16.64. *If $f(z) = \dfrac{h(z)}{z - a}$ has a simple pole at $z = a$, then $\operatorname*{Res}_{z=a} f(z) = h(a)$.*

Proof: We substitute the formula for $f(z)$ into the definition (16.141), and so

$$\operatorname*{Res}_{z=a} f(z) = \frac{1}{2\pi\,\mathrm{i}} \oint_C f(z)\,dz = \frac{1}{2\pi\,\mathrm{i}} \oint_C \frac{h(z)\,dz}{z - a} = h(a)$$

by Cauchy's formula (16.137). *Q.E.D.*

Example 16.65. Consider the function

$$f(z) = \frac{e^z}{z^2 - 2\,z - 3} = \frac{e^z}{(z+1)(z-3)}\,.$$

From the factorization of the denominator, we see that $f(z)$ has simple pole singularities at $z = -1$ and $z = 3$. The residues are given, respectively, by

$$\operatorname*{Res}_{z=-1} \frac{e^z}{z^2 - 2\,z - 3} = \left.\frac{e^z}{z-3}\right|_{z=-1} = -\,\frac{1}{4\,e}\,, \qquad \operatorname*{Res}_{z=3} \frac{e^z}{z^2 - 2\,z - 3} = \left.\frac{e^z}{z+1}\right|_{z=3} = \frac{e^3}{4}\,.$$

Since $f(z)$ is analytic everywhere else, its residue at any other point is automatically 0.

Recall that a function $g(z)$ is said to have *simple zero* at $z = a$ provided

$$g(z) = (z-a)\,k(z)$$

where $k(z)$ is analytic at $z = a$ and $k(a) = g'(a) \neq 0$. If $f(z)$ is analytic at $z = a$, then the quotient

$$\frac{f(z)}{g(z)} = \frac{f(z)}{(z-a)\,k(z)}$$

has a simple pole at $z = a$, with residue

$$\operatorname*{Res}_{z=a} \frac{f(z)}{g(z)} = \operatorname*{Res}_{z=a} \frac{f(z)}{(z-a)\,k(z)} = \frac{f(a)}{k(a)} = \frac{f(a)}{g'(a)} \tag{16.146}$$

by Lemma 16.64. More generally, if $z = a$ is a zero or order m of

$$g(z) = (z-a)^m\,k(z), \qquad \text{so that} \qquad k(a) = \frac{g^{(m)}(a)}{m!} \neq 0,$$

then

$$\operatorname*{Res}_{z=a} \frac{f(z)}{g(z)} = \frac{1}{(m-1)!}\,\frac{d^{m-1}}{dz^{m-1}}\left(\frac{f(z)}{k(z)}\right)\,\Bigg|_{z=a}. \tag{16.147}$$

The proof of the latter formula is left as an exercise.

Example 16.66. As an illustration, let us compute the residue of $\sec z = 1/\cos z$ at the point $z = \frac{1}{2}\,\pi$. Note that $\cos z$ has a simple zero at $z = \frac{1}{2}\,\pi$ since its derivative, namely $-\sin z$, is nonzero there. Thus, according to (16.146) with $f(z) \equiv 1$,

$$\operatorname*{Res}_{z=\pi/2} \sec z = \operatorname*{Res}_{z=\pi/2} \frac{1}{\cos z} = \frac{-1}{\sin \frac{1}{2}\,\pi} = -1.$$

The direct computation of the residue using the defining contour integral (16.141) is much harder.

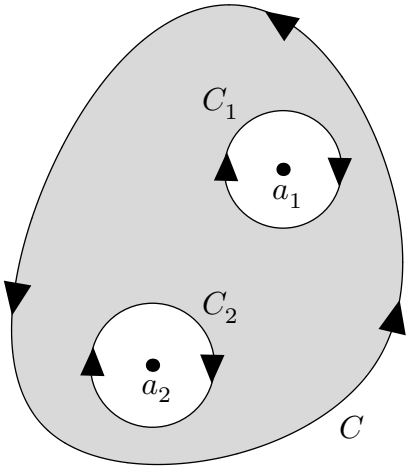


Figure 16.30. Residues Inside a Contour.

The Residue Theorem

Residues are the building blocks of a general method for computing contour integrals of analytic functions. The *Residue Theorem* says that the value of the integral of a complex function around a closed curve depends only on its residues at the enclosed singularities. Since the residues can be computed directly from the function, the resulting formula provides an effective mechanism for painless evaluation of complex integrals, that completely avoids the anti-derivative. Indeed, the residue method copntinues to be effective even when the integrand does not have an anti-derivative that can be expressed in terms of elementary functions.

Theorem 16.67. *Let C be a simple, closed curve, oriented in the counter-clockwise direction. Suppose $f(z)$ is analytic everywhere inside C except at a finite number of singularities, $a_1, \dots, a_n$. Then*

$$\frac{1}{2\pi\,\mathrm{i}} \oint_C f(z)\,dz \;=\; \operatorname*{Res}_{z=a_1} f(z) + \;\cdots\; + \operatorname*{Res}_{z=a_n} f(z). \tag{16.148}$$

Keep in mind that only the singularities that lie *inside* the contour C contribute to the residue formula (16.148).

Proof: We draw a small circle C_i around each singularity a_i. We assume the circles all lie inside the contour C and do not cross each other, so that a_i is the only singularity contained within C_i; see Figure 16.30. Definition 16.63 implies that

$$\operatorname*{Res}_{z=a_i} f(z) = \frac{1}{2\pi\,\mathrm{i}} \oint_{C_i} f(z)\,dz, \tag{16.149}$$

where the line integral is taken in the counter-clockwise direction around C_i.

Consider the domain Ω consisting of all points z which lie inside the given curve C, but outside all the small circles $C_1, \dots, C_n$; this is the shaded region in Figure 16.30. By our

construction, the function $f(z)$ is analytic on Ω, and hence by Cauchy's Theorem 16.47, the integral of $f(z)$ around the boundary $\partial\Omega$ is zero. The boundary $\partial\Omega$ must be oriented consistently, so that the domain is always lying on one's left hand side. This means that the outside contour C should be traversed in a counter-clockwise direction, whereas the inside circles C_i are taken in a *clockwise* direction. Therefore, the integral around the boundary of the domain Ω can be broken up into a difference

$$0 = \frac{1}{2\pi\,\mathrm{i}} \oint_{\partial\Omega} f(z)\,dz = \frac{1}{2\pi\,\mathrm{i}} \oint_{C} f(z) \;-\; \sum_{i=1}^{n} \frac{1}{2\pi\,\mathrm{i}}\; dz \oint_{C_i} f(z)\,dz$$
$$= \frac{1}{2\pi\,\mathrm{i}} \oint_{C} f(z) \;-\; \sum_{i=1}^{n} \operatorname*{Res}_{z=a_i} f(z)\,dz.$$

The minus sign converts the circular integrals to the counterclockwise orientation used in the definition (16.149) of the residues. Rearranging the final identity leads to the residue formula (16.148). *Q.E.D.*

Example 16.68. Let us use residues to evaluate the contour integral

$$\oint_{C_r} \frac{e^z}{z^2 - 2\,z - 3}\,dz$$

where C_r denotes the circle of radius r centered at the origin. According to Example 16.65, the integrand has two singularities at -1 and 3, with respective residues $-1/(4\,e)$ and $e^3/4$. If the radius of the circle is $r > 3$, then it goes around both singularities, and hence by the residue formula (16.148)

$$\oint_{C} \frac{e^z\,dz}{z^2 - 2\,z - 3} = 2\pi\,\mathrm{i}\,\left(-\frac{1}{4e} + \frac{e^3}{4}\right) = \frac{(e^4 - 1)\,\pi\,\mathrm{i}}{2\,e}, \qquad r > 3.$$

If the circle has radius $1 < r < 3$, then it only encircles the singularity at -1, and hence

$$\oint_{C} \frac{e^z}{z^2 - 2\,z - 3}\,dz = -\frac{\pi\,\mathrm{i}}{2\,e}, \qquad 1 < r < 3.$$

If $0 < r < 1$, the function has no singularities inside the circle and hence, by Cauchy's Theorem 16.47, the integral is 0. Finally, when $r = 1$ or $r = 3$, the contour passes through a singularity, and the integral does not converge.

Evaluation of Real Integrals

One important — and unexpected — application of the Residue Theorem 16.67 is to aid in the evaluation of certain definite real integrals. Of particular note is that it even applies to cases in which one is unable to evaluate the corresponding indefinite integral in closed form. Nevertheless, converting the definite real integral into (part of a) complex contour integral leads to a direct evaluation via the calculus of residues that sidesteps the difficulties in finding the antiderivative.

The method treats two basic types of real integral, although numerous variations appear in more extensive treatments of the subject. The first category are real trigonometric integrals of the form

$$I = \int_0^{2\pi} F(\cos\theta, \sin\theta)\, d\theta. \tag{16.150}$$

Such integrals can often be evaluated by converting them into complex integrals around the unit circle $C = \big\{\, |\, z\,| = 1 \,\big\}$. If we set

$$z = e^{\,\mathrm{i}\,\theta} \qquad \text{so} \qquad \frac{1}{z} = e^{-\,\mathrm{i}\,\theta},$$

then

$$\cos\theta = \frac{e^{\,\mathrm{i}\,\theta} + e^{-\,\mathrm{i}\,\theta}}{2} = \frac{1}{2}\left(z + \frac{1}{z}\right), \qquad \sin\theta = \frac{e^{\,\mathrm{i}\,\theta} - e^{-\,\mathrm{i}\,\theta}}{2\,\mathrm{i}} = \frac{1}{2\,\mathrm{i}}\left(z - \frac{1}{z}\right). \tag{16.151}$$

Moreover,

$$dz = de^{\,\mathrm{i}\,\theta} = \mathrm{i}\, e^{\,\mathrm{i}\,\theta}\, d\theta = \mathrm{i}\, z\, d\theta, \qquad \text{and so} \qquad d\theta = \frac{dz}{\mathrm{i}\, z}\,. \tag{16.152}$$

Therefore, the integral (16.150) can be written in the complex form

$$I = \oint_C F\left(\frac{z + z^{-1}}{2}, \frac{z + z^{-1}}{2\,\mathrm{i}}\right) \frac{dz}{\mathrm{i}\, z}\,. \tag{16.153}$$

If we know that the resulting complex integrand is well-defined and single-valued, except, possibly, for a finite number of singularities inside the unit circle, then the residue formula (16.148) tells us that the integral can be directly evaluated by adding together its residues and multiplying by $2\pi\,\mathrm{i}$.

Example 16.69. We compute the relatively simple example

$$\int_0^{2\pi} \frac{d\theta}{2 + \cos\theta}\,.$$

Applying the substitution (16.153), we find

$$\int_0^{2\pi} \frac{d\theta}{2 + \cos\theta} = \oint_C \frac{dz}{\mathrm{i}\, z\left[\, 2 + \frac{1}{2}(z + z^{-1})\,\right]} = -\,\mathrm{i} \oint_C \frac{2\, dz}{z^2 + 4\, z + 1}\,.$$

The complex integrand has singularities where its denominator vanishes:

$$z^2 + 4\, z + 1 = 0, \qquad \text{so that} \qquad z = -2 \pm \sqrt{3}\,.$$

Only one of these singularities, namely $-2 + \sqrt{3}$ lies inside the unit circle. Therefore, applying (16.146), we find

$$-\,\mathrm{i} \oint_C \frac{2\, dz}{z^2 + 4\, z + 1} = 2\pi \operatorname*{Res}_{z = -2+\sqrt{3}} \frac{2}{z^2 + 4\, z + 1} = \frac{4\pi}{2\, z + 4}\,\bigg|_{z = -2+\sqrt{3}} = \frac{2\pi}{\sqrt{3}}\,.$$

As you may recall from first year calculus, this particular integral can, in fact, be computed directly via a trigonometric substitution. However, the integration is not particularly pleasant, and, with a little practice, the residue method is seen to be an easier method. Moreover, it straightforwardly applies to situations where no elementary anti-derivative exists.

Example 16.70. The goal is to evaluate the definite integral

$$\int_0^\pi \frac{\cos 2\theta}{3-\cos\theta}\, d\theta.$$

The first thing to note is that the integral omnly runs from 0 to π and so is not explicitly of the form (16.150). However, note that the integrand is even, and so

$$\int_0^\pi \frac{\cos 2\theta}{3-\cos\theta}\, d\theta = \frac{1}{2}\int_{-\pi}^\pi \frac{\cos 2\theta}{3-\cos\theta}\, d\theta,$$

which will turn into a contour integral around the entire unit circle under the substitution (16.151). Also note that

$$\cos 2\,\theta = \frac{e^{2\,\mathrm{i}\,\theta}+e^{-2\,\mathrm{i}\,\theta}}{2} = \frac{1}{2}\left(z^2+\frac{1}{z^2}\right),$$

and so

$$\int_{-\pi}^\pi \frac{\cos 2\theta}{3-\cos\theta}\, d\theta = \oint_C \frac{\frac{1}{2}(z^2+z^{-2})}{3-\frac{1}{2}(z+z^{-1})}\,\frac{dz}{\mathrm{i}\,z} = \mathrm{i}\,\oint_C \frac{z^4+1}{z^2(z^2-6\,z+1)}\,dz.$$

The denominator has 4 roots — at 0, $3-2\sqrt{2}$, and $3+2\sqrt{2}$ — but the last one does not lie inside the unit circle and so cn be ignored. We use (16.146) with $f(z)=(z^4+1)/z^2$ and $g(z)=z^2-6\,z+1$ to compute

$$\operatorname*{Res}_{z=3-2\sqrt{2}} \frac{z^4+1}{z^2(z^2-6\,z+1)} = \left.\frac{z^4+1}{z^2}\right|_{z=-2+\sqrt{3}} \frac{1}{2\,z-6} = \tfrac{17}{4}\,\sqrt{2}\,,$$

whereas (16.147) is used to compute

$$\operatorname*{Res}_{z=0} \frac{z^4+1}{z^2(z^2-6\,z+1)} = \frac{d}{dz}\left(\frac{z^4+1}{z^2-6\,z+1}\right)\Bigg|_{z=0} = \left.\frac{2(z^5-9\,z^4+2\,z^3-z+3)}{(z^2-6\,z+1)^2}\right|_{z=0} = 6\,.$$

Therefore,

$$\int_0^\pi \frac{\cos 2\theta}{3-\cos\theta}\, d\theta = \pi\left[\operatorname*{Res}_{z=0} \frac{z^4+1}{z^2(z^2-6\,z+1)} + \operatorname*{Res}_{z=3-2\sqrt{2}} \frac{z^4+1}{z^2(z^2-6\,z+1)}\right] = -6\,\pi + \tfrac{17}{4}\,\sqrt{2}\,\pi.$$

A second type of real integral that can often be evaluated by complex residues are integrals over the entire real line, from $-\infty$ to ∞. Here the technique is a little more subtle, and we sneak up on the integral by using larger and larger closed contours that include more and more of the real axis. The basic idea is contained in the following example.

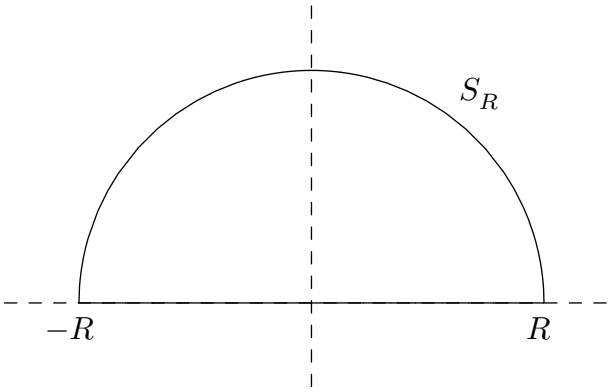


Figure 16.31. The Semicircular Contour.

Example 16.71. The problem is to evaluate the real integral

$$I = \int_0^\infty \frac{\cos x}{1+x^2}\, dx. \tag{16.154}$$

The corresponding indefinite integral cannot be evaluated in elementary terms, and so we are forced to rely on the calculus of residues. We begin by noting that the integrand is even, and hence the integral $I = \frac{1}{2}\, J$ is one half the integral

$$J = \int_{-\infty}^\infty \frac{\cos x}{1+x^2}\, dx$$

over the entire real line. Moreover, for x real, we can write

$$\frac{\cos x}{1+x^2} = \operatorname{Re} \frac{e^{\,\mathrm{i}\, x}}{1+x^2}\,, \qquad \text{and hence} \qquad J = \operatorname{Re} \int_{-\infty}^\infty \frac{e^{\,\mathrm{i}\, x}}{1+x^2}\, dx. \tag{16.155}$$

Let C_R be the closed contour consisting of a large semicircle of radius $R \gg 0$, which we denote by S_R, connected at its ends by the real interval $-R \le x \le R$, as plotted in Figure 16.31 Figure 16.31, and having the usual counterclockwise orientation. The corresponding contour integral

$$\oint_{C_R} \frac{e^{\,\mathrm{i}\, z}\, dz}{1+z^2} \;=\; \int_{-R}^R \frac{e^{\,\mathrm{i}\, x}\, dx}{1+x^2} \;+\; \int_{S_R} \frac{e^{\,\mathrm{i}\, z}\, dz}{1+z^2} \tag{16.156}$$

breaks up into two pieces: the first over the real interval, and the second over the semicircle. As the radius $R \to \infty$, the semicircular contour C_R includes more and more of the real axis, and so the first integral gets closer and closer to our desired integral (16.155). If we can prove that the second, semicircular integral goes to zero, then we will be able to evaluate the integral over the real axis by contour integration, and hence by the method of residues. Our hope that the semicircular integral is small seems reasonable, since the integrand $(1+z^2)^{-1} e^{\,\mathrm{i}\, z}$ gets smaller and smaller as $|\, z\,| \to \infty$ *provided* $\operatorname{Im} z \ge 0$. (Why?) A rigorous verification of this fact will appear at the end of the example.

According to the Residue Theorem 16.67, the integral (16.156) is equal to the sum of all the residues at the singularities of $f(z)$ lying inside the contour C_R. Now e^z is analytic

everywhere, and so the singularities occur where the denominator vanishes, i.e., $z^2 = -1$, and so are at $z = \pm \mathrm{i}$. Since the semicircle lies in the upper half plane $\operatorname{Im} z > 0$, only the singularity $z = +\mathrm{i}$ lies inside — and then only when $R > 1$. To compute the residue, we use (16.146) to evaluate

$$\operatorname*{Res}_{z=\mathrm{i}} \frac{e^{\mathrm{i}\,z}}{1+z^2} = \left.\frac{e^{\mathrm{i}\,z}}{2\,z}\right|_{z=\mathrm{i}} = \frac{e^{-1}}{2\,\mathrm{i}} = \frac{1}{2\,\mathrm{i}\,e}\,.$$

Therefore, by (16.148),

$$\frac{1}{2\,\pi\,\mathrm{i}} \oint_{C_R} \frac{e^{\mathrm{i}\,z}\,dz}{1+z^2} = \frac{1}{2\,\mathrm{i}\,e}\,, \qquad \text{provided} \qquad R > 1.$$

Thus, assuming the semicircular part of the integral does indeed become vanishingly small as $R \to \infty$, we conclude that

$$\int_{-\infty}^{\infty} \frac{e^{\mathrm{i}\,x}\,dx}{1+x^2} = \lim_{R\to\infty} \oint_{C_R} \frac{e^{\mathrm{i}\,z}\,dz}{1+z^2} = 2\,\pi\,\mathrm{i}\;\frac{1}{2\,\mathrm{i}\,e} = \frac{\pi}{e}\,.$$

Incidentally, the integral is real because its imaginary part,

$$\int_{-\infty}^{\infty} \frac{\sin x\,dx}{1+x^2} = 0,$$

is the integral of an odd function which is automatically zero. Consequently,

$$I = \int_0^{\infty} \frac{\cos x\,dx}{1+x^2} = \frac{1}{2}\,\operatorname{Re} \int_{-\infty}^{\infty} \frac{e^{\mathrm{i}\,x}\,dx}{1+x^2} = \frac{\pi}{2\,e}\,, \tag{16.157}$$

which is the desired result.

To complete the argument, let us estimate the size of the semicircular integral. The integrand is bounded by

$$\left|\frac{e^{\mathrm{i}\,z}}{1+z^2}\right| \;\le\; \frac{1}{1+|\,z\,|^2} = \frac{1}{1+R^2} \qquad \text{whenever} \qquad |\,z\,| = R, \qquad \operatorname{Im} z \ge 0,$$

where we are using the fact that

$$\left|\,e^{\mathrm{i}\,z}\,\right| = e^{-y} \le 1 \qquad \text{whenever} \qquad z = x + \mathrm{i}\,y \qquad \text{with} \qquad y \ge 0.$$

According to Corollary 16.55, the size of the integral of a complex function is bounded by its maximum modulus along the curve times the length of the semicircle, namely πR. Thus, in our case,

$$\left|\int_{S_R} \frac{e^{\mathrm{i}\,z}\,dz}{1+z^2}\right| \;\le\; \frac{\pi\,R}{1+R^2} \;\le\; \frac{\pi}{R} \quad\longrightarrow\quad 0 \qquad \text{as} \qquad R \longrightarrow \infty,$$

as required.

Example 16.72. Here we will use residues to evaluate the real integral

$$\int_{-\infty}^{\infty} \frac{dx}{1+x^4}\,. \tag{16.158}$$

The indefinite integral can, in fact, be found by the method of partial fractions, but, as you may know, this is not a particularly pleasant task. Let us try the method of residues. Let C_R denote the same semicircular contour as in Figure 16.31. The integrand has pole singularities where the denominator vanishes, i.e., $z^4 = -1$, and so at the four fourth roots of -1. These are

$$e^{\pi\,\mathrm{i}\,/4} = \frac{1+\mathrm{i}}{\sqrt{2}}\,, \qquad e^{3\,\pi\,\mathrm{i}\,/4} = \frac{-1+\mathrm{i}}{\sqrt{2}}\,, \qquad e^{5\,\pi\,\mathrm{i}\,/4} = \frac{1-\mathrm{i}}{\sqrt{2}}\,, \qquad e^{7\,\pi\,\mathrm{i}\,/4} = \frac{-1-\mathrm{i}}{\sqrt{2}}\,.$$

Only the first two roots lie inside C_R when $R > 1$. Their residues can be computed using (16.146):

$$\operatorname*{Res}_{z\,=\,e^{\pi\,\mathrm{i}\,/4}} \frac{1}{1+z^4} = \frac{1}{4\,z^3}\,\bigg|_{z\,=\,e^{\pi\,\mathrm{i}\,/4}} = \frac{e^{-3\,\pi\,\mathrm{i}\,/4}}{4} = \frac{-1-\mathrm{i}}{4\sqrt{2}}\,,$$

$$\operatorname*{Res}_{z\,=\,e^{3\,\pi\,\mathrm{i}\,/4}} \frac{1}{1+z^4} = \frac{1}{4\,z^3}\,\bigg|_{z\,=\,e^{3\,\pi\,\mathrm{i}\,/4}} = \frac{e^{-9\,\pi\,\mathrm{i}\,/4}}{4} = \frac{1-\mathrm{i}}{4\sqrt{2}}\,.$$

Therefore, by the residue formula (16.148),

$$\oint_{C_R} \frac{dz}{1+z^4} = 2\,\pi\,\mathrm{i}\,\left(\frac{-1-\mathrm{i}}{4\sqrt{2}} + \frac{1-\mathrm{i}}{4\sqrt{2}}\right) = \frac{\pi}{\sqrt{2}}\,. \tag{16.159}$$

On the other hand, we can break up the contour integral into an integral along the real axis and an integral around the semicircle:

$$\oint_{C_R} \frac{dz}{1+z^4} = \int_{-\,R}^{R} \frac{dx}{1+x^4} \;+\; \int_{S_R} \frac{dz}{1+z^4}.$$

The first integral goes to the desired real integral as the radius $R \to \infty$. On the other hand, on a large semicircle $|\,z\,| = R$, the integrand is small:

$$\left|\,\frac{1}{1+z^4}\,\right| \;\le\; \frac{1}{1+|\,z\,|^4} = \frac{1}{1+R^4} \qquad \text{when} \qquad |\,z\,| = R.$$

Thus, using Corollary 16.55, the integral around the semicircle can be bounded by

$$\left|\,\int_{S_R} \frac{dz}{1+z^4}\,\right| \;\le\; \frac{1}{1+R^4}\,\pi R \;\le\; \frac{\pi}{R^3} \quad \longrightarrow \quad 0 \qquad \text{as} \qquad R \;\longrightarrow\; \infty.$$

Thus, as $R \to \infty$, the complex integral (16.159) goes to the desired real integral (16.158), and so

$$\int_{-\infty}^{\infty} \frac{dx}{1+x^4} = \frac{\pi}{\sqrt{2}}\,.$$

Note that the result is real and positive, as it must be.

Chapter 17

Dynamics of Planar Media

In this chapter, we press on in our ascent of the dimensional ladder for linear systems. In Chapter 6, we embarked on our journey with equilibrium configurations of discrete systems — mass–spring chains, circuits, and structures — which are governed by certain linear algebraic systems. In Chapter 9, the dynamical behavior of such discrete systems was modeled by systems of linear ordinary differential equations. Chapter 11 began our treatment of continuous media with the boundary value problems that describe the equilibria of one-dimensional bars, strings and beams. Their dynamical motions formed the topic of Chapter 14, in the simplest case leading to two fundamental partial differential equations: the heat equation describing thermal diffusion, and the wave equation modeling vibrations. In Chapters 15 and 16, we focussed our attention on the boundary value problems describing equilibria of planar bodies — plates and membranes — with primary emphasis on solving the ubiquitous Laplace equation, both analytically or numerically. We now turn to the analysis of their dynamics, as governed by the two-dimensional[†] forms of the heat and wave equations. The heat equation describes diffusion of, say, heat energy, or population, or pollutants in a homogeneous two-dimensional domain. The wave equation models small vibrations of two-dimensional membranes such as a drum.

Although the increase in dimension does challenge our analytical prowess, we have, in fact, already mastered the key techniques: separation of variables and fundamental solutions. (Disappointingly, conformal mappings are not particularly helpful in the dynamical universe.) When applied to partial differential equation in higher dimensions, separation of variables often leads to new ordinary differential equations, whose solutions are no longer elementary functions. These so-called *special functions*, which include the Bessel functions appearing in the present chapter, and the Legendre functions, spherical harmonics, and spherical Bessel functions in three-dimensional problems, play a ubiquitous role in more advanced applications in physics, engineering and mathematics. Basic series solution techniques for ordinary differential equations, and key properties of the most important classes of special functions, can be found in Appendix C

In Appendix C, we collect together the required results about the most important classes of special functions, including a short presentation of the series approach for solving non-elementary ordinary differential equations.

[†] Throughout, "dimension" refers to the number of space variables. In Newtonian dynamics, the time "dimension" is accorded a separate status, which distinguishes dynamics from equilibrium. Of course, in the more complicated relativistic universe, time and space must be regarded on an equal footing, and the dimension count modified accordingly.

Numerical methods for solving boundary value and initial value problems are, of course, essential in all but the simplest situations. The two basic methods — finite element and finite difference — have already appeared, and the only new aspect is the (substantial) complication of working in higher dimensions. Thus, in the interests of brevity, we defer the discussion of the numerical aspects of multi-dimensional partial differential equations to more advanced texts, e.g., [**107**], and student projects outlined in the exercises. However, the student should be assured that, without knowledge of the qualitative features based on direct analysis and explicit solutions, the design, implementation, and testing of numerical solution techniques would be severely hampered.

17.1. Diffusion in Planar Media.

As we learned in Chapter 15, the equilibrium temperature $u(x, y)$ of a homogeneous plate is governed by the two-dimensional Laplace equation

$$\Delta u = u_{xx} + u_{yy} = 0.$$

In conformity with our general framework, the dynamical diffusion of such a plate will be modeled by the two-dimensional heat equation

$$u_t = \gamma\, \Delta u = \gamma \left(u_{xx} + u_{yy} \right), \tag{17.1}$$

where the diffusivity coefficient $\gamma > 0$ measures the relative speed of diffusion of heat energy throughout the plate; its positivity is required on physical grounds, and also avoids ill-posedness of the dynamical system. In this simplest model of two-dimensional diffusion, we are assuming that there are no loss of heat or external heat sources in the plate's interior, which can be arranged by covering it with insulation.

The solution $u(t, \mathbf{x}) = u(t, x, y)$ to (17.1) measures the temperature at time t at each point $\mathbf{x} = (x, y) \in \Omega$ in the domain $\Omega \subset \mathbb{R}^2$ occupied by the plate. To uniquely specify $u(t, x, y)$ at each point $(x, y) \in \Omega$ and each positive $t > 0$, we must impose both initial and boundary conditions. The most important are:

(a) *Dirichlet boundary conditions*:

$$u = h \qquad \text{on} \qquad \partial\Omega, \tag{17.2}$$

which fix the temperature on the boundary of the plate.

(b) *Neumann boundary conditions*:

$$\frac{\partial u}{\partial \mathbf{n}} = \frac{\partial u}{\partial \mathbf{n}} = k \qquad \text{on} \qquad \partial\Omega, \tag{17.3}$$

that prescribe the heat flux along the boundary; the case $k = 0$ corresponds to an insulated boundary.

(c) *Mixed boundary conditions*: we impose Dirichlet conditions on part of the boundary $D \subsetneq \partial\Omega$ and Neumann conditions on the remainder $N = \partial\Omega \setminus D$. For instance, the homogeneous mixed boundary conditions

$$u = 0 \qquad \text{on} \qquad D, \qquad\qquad \frac{\partial u}{\partial \mathbf{n}} = 0 \qquad \text{on} \qquad N, \tag{17.4}$$

correspond to freezing part of the boundary and insulating the remainder.
The initial conditions specify the temperature of the plate

$$u(0, x, y) = f(x, y), \qquad (x, y) \in \Omega, \tag{17.5}$$

at an initial time, which for simplicity, we take as $t_0 = 0$. Under reasonable assumptions, e.g., the domain Ω is bounded, its boundary is piecewise smooth, and the boundary data is piecewise continuous, say, a general theorem, [**46**], guarantees the existence of a unique solution $u(t, x, y)$ to any of these initial-boundary value problems for all subsequent times $t > 0$. Our practical goal is to both compute and understand the behavior of the solution in specific situations.

Derivation of the Diffusion Equation

This section is for those interested in understanding how the heat equation arises as a model for heat flow. The physical derivation of the two-dimensional (and three-dimensional) heat equation relies upon the same two basic thermodynamical laws that were used, in Section 14.1, to establish the one-dimensional version. The first principle is that heat energy flows from hot to cold as rapidly as possible. According to the multivariable calculus Theorem 19.37, the negative temperature gradient $-\nabla u$ points in the direction of the steepest decrease in the function u at a point, and so heat energy will flow in that direction. Therefore, the heat flux vector $\mathbf{w}$, which measures the magnitude and direction of the flow of heat energy, should be proportional to the temperature gradient:

$$\mathbf{w}(t, x, y) = -\kappa(x, y)\, \nabla u. \tag{17.6}$$

The scalar quantity $\kappa(x, y) > 0$ measures the *thermal conductivity* of the material, so (17.6) is the multi-dimensional form of *Fourier's Law of Cooling* (14.3). We are assuming that the thermal conductivity depends only on the position $(x, y) \in \Omega$, which means that the material in the plate

(a) is not changing in time;

(b) is *isotropic*, meaning that its conductivity is the same in all directions, and, further,

(c) does not depend upon temperature.

Dropping either (b) or (c) would result in a much more complicated nonlinear diffusion equation.

The second thermodynamical principle is that, in the absence of external heat sources, heat can only enter any subregion $D \subset \Omega$ through its boundary ∂D. (Keep in mind that the plate is insulated above and below.) In other words, the rate of change of the heat energy in D is prescribed by the heat flux across its boundary ∂D. Let $\varepsilon(t, x, y)$ denote the heat energy density at each time and point in the domain, so that $\iint_D \varepsilon(t, x, y)\, dx\, dy$ represents the total heat contained within the region D at time t. The amount of additional heat energy entering D at a boundary point $\mathbf{x} \in \partial D$ is the inward normal component of the heat flux vector: $-\mathbf{w} \cdot \mathbf{n}$, where $\mathbf{n}$ denotes the *outward* unit normal to ∂D. Thus, the total heat flux entering the region D is given by the flux line integral $-\oint_{\partial D} \mathbf{w} \cdot \mathbf{n}\, ds$,

cf. (A.42). Equating the rate of change of heat energy to the heat flux yields

$$\frac{\partial}{\partial t} \iint_D \varepsilon(t,x,y)\, dx\, dy = - \oint_{\partial D} \mathbf{w} \cdot \mathbf{n}\; ds = - \iint_D \nabla \cdot \mathbf{w}\; dx\, dy,$$

where we applied the divergence form of Green's Theorem, (A.58), to convert the flux line integral into a double integral. We bring the time derivative inside the first integral and collect the terms, whence

$$\iint_D \left(\frac{\partial \varepsilon}{\partial t} + \nabla \cdot \mathbf{w} \right) dx\, dy = 0. \tag{17.7}$$

Keep in mind that this integral formula must hold for *any* subdomain $D \subset \Omega$. Now, the only way in which an integral of a continuous function can vanish for all subdomains is if the integrand is identically zero, cf. Exercise ■, and so

$$\frac{\partial \varepsilon}{\partial t} + \nabla \cdot \mathbf{w} = 0. \tag{17.8}$$

In this way, we derive the basic *conservation law* relating heat energy ε and heat flux $\mathbf{w}$.

As in our one-dimensional model, (14.2), the heat energy $\varepsilon(t,x,y)$ at each time and point in the domain is proportional to the temperature, so

$$\varepsilon(t,x,y) = \sigma(x,y)\, u(t,x,y), \qquad \text{where} \qquad \sigma(x,y) = \rho(x,y)\, \chi(x,y) \tag{17.9}$$

is the product of the *density* and the *heat capacity* of the material. Combining this with the Fourier Law (17.6) and the energy balance equation (17.9) leads to the general two-dimensional *diffusion equation*

$$\sigma\, \frac{\partial u}{\partial t} = \nabla \cdot \big(\kappa\, \nabla u \big) \tag{17.10}$$

governing the temperature dynamics of an isotropic medium in the absence of external heat sources. In full detail, this second order partial differential equation is

$$\sigma(x,y)\, \frac{\partial u}{\partial t} = \frac{\partial}{\partial x} \left(\kappa(x,y)\, \frac{\partial u}{\partial x} \right) + \frac{\partial}{\partial y} \left(\kappa(x,y)\, \frac{\partial u}{\partial y} \right). \tag{17.11}$$

In particular, if the body is homogeneous, then both σ and κ are constant, and so general diffusion equation (17.10) reduces to the heat equation (17.1) with thermal diffusivity

$$\gamma = \frac{\kappa}{\sigma} = \frac{\kappa}{\rho\, \chi}. \tag{17.12}$$

The heat and diffusion equations are also used to model movements of populations, e.g., bacteria in a petri dish or wolves in the Canadian Rockies, [**biol**]. The solution $u(t,x,y)$ represents the number of individuals near position (x,y) at time t, and diffuses over the domain due to random motions of the individuals. Similar diffusion processes model the mixing of chemical reagents in solutions, with the diffusion induced by the random Brownian motion from molecular collisions. Convection due to fluid motion and/or changes due to chemical reactions lead to the more general class of *reaction–diffusion and convective–diffusion equations*, [**chem**].

Self-Adjoint Formulation

The general diffusion equation (17.11) can be readily fit into the self-adjoint framework of Section 14.7, taking the form

$$u_t = -K[u] = -\nabla^* \circ \nabla u. \tag{17.13}$$

The gradient operator ∇ maps scalar fields u to vector fields $\mathbf{v} = \nabla u$; its adjoint ∇^*, which goes in the reverse direction, is taken with respect to the weighted inner products

$$\langle\, u\,,\widetilde{u}\,\rangle = \iint_\Omega u(x,y)\,\widetilde{u}(x,y)\,\sigma(x,y)\,dx\,dy, \quad \langle\!\langle\, \mathbf{v}\,,\widetilde{\mathbf{v}}\,\rangle\!\rangle = \iint_\Omega \mathbf{v}(x,y)\cdot\widetilde{\mathbf{v}}(x,y)\;\kappa(x,y)\,dx\,dy, \tag{17.14}$$

between, respectively, scalar and vector fields. A straightforward integration by parts argument, done in detail in Section 15.4, tells us that

$$\nabla^*\mathbf{v} = -\frac{1}{\sigma}\,\nabla\cdot(\kappa\,\mathbf{v}) = -\frac{1}{\sigma}\left[\frac{\partial(\kappa\,v_1)}{\partial x} + \frac{\partial(\kappa\,v_2)}{\partial y}\right], \qquad \text{when} \qquad \mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}. \tag{17.15}$$

Therefore, the right hand side of (17.13) is equal to

$$-K[u] = -\nabla^* \circ \nabla u = \frac{1}{\sigma}\,\nabla\cdot(\kappa\,\nabla u), \tag{17.16}$$

which recovers (17.10). As always, we need to impose suitable homogeneous boundary conditions — Dirichlet, Neumann or mixed — to ensure the validity of the integration by parts argument used to establish the adjoint formula (17.15).

In particular, to obtain the heat equation, we take σ and κ to be constant, and so (17.14) reduce, up to a constant factor, to the usual L^2 inner products between scalar and vector fields. In this case, the adjoint of the gradient is, up to a scale factor, minus the divergence: $\nabla^* = -\gamma\,\nabla\cdot$, where $\gamma = \kappa/\sigma$. In this scenario, (17.13) reduces to the two-dimensional heat equation (17.1).

The self-adjoint operator $K = \nabla^* \circ \nabla$ is always positive semi-definite, and is positive definite if and only if $\ker\nabla = \{0\}$. As we saw in Section 15.4, positive definiteness holds in the Dirichlet and mixed cases, whereas the Neumann boundary conditions lead to only a positive semi-definite operator. Indeed, assuming Ω is connected, the only functions in the kernel of the gradient operator are the constants. Moreover, only the zero constant function satisfies the Dirichlet or mixed boundary conditions, and hence the gradient's kernel is trivial, ensuring positive definiteness.

The heat and diffusion equations are examples of *parabolic* partial differential equations, the terminology being adapted from Definition 15.1 to apply to partial differential equations in more than two variables. As we will see, all of the basic qualitative features we learned when studying the one-dimensional heat equation carry over to solutions to parabolic partial differential equations in higher dimensions.

Separation of Variables

In Section 14.1, we applied the method of separation of variables to express the solution to the one-dimensional heat equation as a series involving the eigenfunctions of an

associated boundary value problem. Section 14.7 argued that the eigenfunction series solution method carries over, essentially unchanged, to the general class of self-adjoint diffusion equations. In particular, the solutions to the two-dimensional heat and diffusion equations are expressed in series form based on the eigenfunctions of an associated two-dimensional boundary value problem.

As we know, the separable solutions to any diffusion equation (17.13) are of exponential form

$$u(t,x,y) = e^{-\lambda t}\, v(x,y). \tag{17.17}$$

Since the linear operator K only involves differentiation with respect to the spatial variables x, y, we find

$$\frac{\partial u}{\partial t} = -\lambda\, e^{-\lambda t}\, v(x,y), \qquad \text{while} \qquad K[u] = e^{-\lambda t}\, K[v].$$

Substituting back into the diffusion equation (17.13) and canceling the exponentials, we conclude that

$$K[v] = \lambda\, v, \tag{17.18}$$

Thus, $v(x,y)$ must be an eigenfunction for the linear operator K, subject to the relevant homogeneous boundary conditions. In the case of the heat equation (17.1),

$$K[u] = -\gamma\, \Delta u,$$

and hence the eigenvalue equation (17.18) takes the form

$$\gamma\, \Delta v + \lambda\, v = 0, \qquad \text{or, in detail,} \qquad \gamma \left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} \right) + \lambda\, v = 0. \tag{17.19}$$

This generalization of the Laplace equation is known as the *Helmholtz equation*, and was briefly discussed in Example 15.23.

According to Theorem 14.18, the eigenvalues of the self-adjoint operator $K = \nabla^* \circ \nabla$ are all real and non-negative: $\lambda \geq 0$. When K is positive definite, they are strictly positive: $\lambda > 0$. Let us index the eigenvalues in increasing order:

$$0 < \lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \ \cdots \ . \tag{17.20}$$

Under reasonable conditions on the underlying linear boundary value problem, it can be proved, [**46**; Chapter V], that there are, in fact, an infinite number of eigenvalues, and, moreover, their size increases without bound, so $\lambda_k \to \infty$ as $k \to \infty$. Each eigenvalue is repeated according to the number (which is necessarily finite) of linearly independent eigenfunctions it admits. The problem has a zero eigenvalue, $\lambda_0 = 0$, corresponding to the constant null eigenfunction $v(x,y) \equiv 1$, if and only if K is not positive definite, i.e., only in the case of pure Neumann boundary conditions.

Each eigenvalue and eigenfunction pair will produce a separable solution

$$u_k(t,x,y) = e^{-\lambda_k t}\, v_k(x,y)$$

to the diffusion equation (17.13). The solutions corresponding to positive eigenvalues are exponentially decaying in time, while a zero eigenvalue, which only occurs in the Neumann

problem, produces a constant solution. The general solution to the homogeneous boundary value problem can then be built up as a linear combination of these basic solutions, in the form of an *eigenfunction series*

$$u(t,x,y) = \sum_{k=1}^{\infty} c_k\, u_k(t,x,y) = \sum_{k=1}^{\infty} c_k\, e^{-\lambda_k t}\, v_k(x,y), \tag{17.21}$$

which is a form of generalized Fourier series. The eigenfunction coefficients c_k are prescribed by the initial conditions, which require

$$\sum_{k=1}^{\infty} c_k\, v_k(x,y) = f(x,y). \tag{17.22}$$

Theorem 14.19 guarantees orthogonality of the eigenfunctions, and hence the coefficients are given by our standard orthogonality formula

$$c_k = \frac{\langle\, f\,, v_k\,\rangle}{\|\, v_k\,\|^2} = \frac{\displaystyle\iint_\Omega f(x,y)\, v_k(x,y)\, \sigma(x,y)\, dx\, dy}{\displaystyle\iint_\Omega v_k(x,y)^2\, \sigma(x,y)\, dx\, dy}\,, \tag{17.23}$$

where the weighting function $\sigma(x,y)$ was defined in (17.9). In the case of the heat equation, σ is constant and so can be canceled from both numerator and denominator, leaving the simpler formula

$$c_k = \frac{\displaystyle\iint_\Omega f(x,y)\, v_k(x,y)\, dx\, dy}{\displaystyle\iint_\Omega v_k(x,y)^2\, dx\, dy}\,. \tag{17.24}$$

Under fairly general hypotheses, it can be shown that the eigenfunctions form a *complete system*, which means that the eigenfunction series (17.22) will converge (at least in norm) to the function $f(x,y)$, provided it is not too bizarre. Moreover, the resulting series (17.21) converges to the solution to the initial-boundary value problem for the diffusion equation. See [**46**; p. 369] for a precise statement and proof of the general theorem.

Qualitative Properties

Before tackling simple examples where we are able to construct explicit formulae for the eigenfunctions and eigenvalues, let us see what the eigenfunction series solution (17.21) can tell us about general diffusion processes. Based on our experience with the case of a one-dimensional bar, the final conclusions will not be especially surprising. Indeed, they also apply, word for word, to diffusion processes in three-dimensional solid bodies. A reader who prefers to see explicit solution formulae may wish to skip ahead to the following section, returning here later.

Keep in mind that we are still dealing with the solution to the homogeneous boundary value problem. The first observation is that all terms in the series solution (17.21), with the possible exception of a null eigenfunction term that appears in the semi-definite case,

are tending to zero exponentially fast. Since most eigenvalues are large, all the higher order terms in the series become almost instantaneously negligible, and hence the solution can be accurately approximated by a finite sum over the first few eigenfunction modes. As time goes on, more and more of the modes can be neglected, and the solution decays to thermal equilibrium at an exponentially fast rate. The rate of convergence to thermal equilibrium is, for most initial data, governed by the smallest positive eigenvalue $\lambda_1 > 0$ for the Helmholtz boundary value problem on the domain.

In the positive definite cases of homogeneous Dirichlet or mixed boundary conditions, thermal equilibrium is $u(t, x, y) \to u_\star(x, y) \equiv 0$. Thus, in these cases, the equilibrium temperature is equal to the boundary temperature — even if this temperature is only fixed on a small part of the boundary. The initial heat is eventually dissipated away through the non-insulated part of the boundary. In the semi-definite Neumann case, corresponding to a completely insulated plate, In this case, the general solution has the form

$$u(t, x, y) = c_0 + \sum_{k=1}^{\infty} c_k \, e^{-\lambda_k t} \, v_k(x, y), \tag{17.25}$$

where the sum is over the positive eigenmodes, $\lambda_k > 0$. Since all the summands are exponentially decaying, the final equilibrium temperature $u_\star = c_0$ is the same as the constant term in the eigenfunction expansion. We evaluate this term using the orthogonality formula (17.23), and so, as $t \to \infty$,

$$u(t, x, y) \;\longrightarrow\; c_0 = \frac{\langle\, f \,, 1 \,\rangle}{\| \, 1 \, \|^2} = \frac{\displaystyle\iint_\Omega f(x, y)\, \sigma(x, y)\, dx\, dy}{\displaystyle\iint_\Omega \sigma(x, y)\, dx\, dy} \,,$$

representing a suitably weighted average of the initial temperature over the domain. In particular, in the case of the heat equation, the weighting function σ is constant, and so the equilibrium temperature

$$u(t, x, y) \quad\longrightarrow\quad c_0 = \frac{1}{\text{area } \Omega} \iint_\Omega f(x, y)\, dx\, dy \tag{17.26}$$

equals the average initial temperature distribution. In this case, the heat energy is not allowed to escape through the boundary, and thus redistributes itself in a uniform manner over the domain.

Diffusion has a smoothing effect on the initial temperature distribution $f(x, y)$. Assume that the eigenfunction coefficients are uniformly bounded, so $| \, c_k \, | \le M$ for some constant M. This will certainly be the case if $f(x, y)$ is piecewise continuous, but even holds for quite rough initial data, including delta functions. Then, at any time $t > 0$ after the initial instant, the coefficients $c_k \, e^{-\lambda_k t}$ in the eigenfunction series solution (17.21) are exponentially small as $k \to \infty$, which is enough to ensure smoothness† of the solution $u(t, x, y)$ for each $t > 0$. Therefore, the diffusion process serves to immediately smooth out

† For a general diffusion equation, this requires that the functions $\sigma(x, y)$ and $\kappa(x, y)$ be smooth.

jumps, corners and other discontinuities in the initial data. As time progresses, the local variations in the solution become less and less pronounced, as it asymptotically reaches a constant equilibrium state.

As a result, diffusion processes can be effectively applied to clean and denoise planar images. The initial data $f(x, y)$ represents the grey-scale value of the image at position (x, y), so that $0 \leq f(x, y) \leq 1$ with 0 representing black, and 1 representing white. As time progresses, the solution $u(t, x, y)$ represents a more and more smoothed version of the image. Although this has the effect of removing unwanted noise from the image, there is also a gradual blurring of the actual features. Thus, the "time" or "multiscale" parameter t needs to be chosen to optimally balance between the two effects — the larger t is the more noise is removed, but the more noticeable the blurring. A representative illustration appears in Figure im2. To further suppress undesirable blurring effects, recent image processing filters are based on anisotropic (and thus *nonlinear*) diffusion equations. See Sapiro, [**156**], for a survey of recent progress in this active field.

Since the forward heat equation effectively blurs the features in an image, we might be tempted to try going backwards in time to sharpen the image. However, the argument presented in Section 14.1 tells us that the backwards heat equation is ill-posed, and hence cannot be used directly for this purpose. Various "regularization" strategies have been devised to circumvent this mathematical barrier, and thereby design effective image enhancement algorithms, [**reg**].

Inhomogeneous Boundary Conditions and Forcing

Finally, let us briefly mention how to incorporate inhomogeneous boundary conditions and external heat sources into the problem. Consider, as a specific example, the forced heat equation

$$u_t = \gamma\, \Delta u + F(x, y), \qquad \text{for} \qquad (x, y) \in \Omega, \tag{17.27}$$

where $F(x, y)$ represents an unvarying external heat source, subject to inhomogeneous Dirichlet boundary conditions

$$u = h \qquad \text{for} \qquad (x, y) \in \partial\Omega, \tag{17.28}$$

that fixes the temperature of the plate on its boundary. When the external forcing is fixed for all t, we expect the solution to eventually settle down to an equilibrium configuration: $u(t, x, y) \to u_\star(x, y)$ as $t \to \infty$, which will be justified below.

The time-independent equilibrium temperature $u = u_\star(x, y)$ satisfies the equation obtained by setting $u_t = 0$ in the differential equation (17.27), namely Poisson equation

$$-\gamma\, \Delta u_\star = F, \qquad \text{for} \qquad (x, y) \in \Omega, \tag{17.29}$$

and subject to the same inhomogeneous Dirichlet boundary conditions (17.28). Positive definiteness of the Dirichlet boundary value problem implies that there is a unique equilibrium solution.

Once we have determined the equilibrium solution — usually through a numerical approximation — we set

$$v(t, x, y) = u(t, x, y) - u_\star(x, y),$$

so that v measures the deviation of the solution u from its eventual equilibrium. By linearity $v(t,x,y)$ satisfies the unforced heat equation subject to homogeneous boundary conditions:

$$v_t = \gamma\,\Delta v, \qquad (x,y) \in \Omega, \qquad\qquad u = 0, \qquad (x,y) \in \partial\Omega. \tag{17.30}$$

Therefore, v can be expanded in an eigenfunction series (17.21), and will decay to zero, $v(t,x,y) \to 0$, at a exponentially fast rate prescribed by the smallest eigenvalue λ_1 of the associated homogeneous Helmholtz boundary value problem. Consequently, the solution to the forced, inhomogeneous problem

$$u(t,x,y) = v(t,x,y) + u_\star(x,y) \;\longrightarrow\; u_\star(x,y)$$

will approach thermal equilibrium, namely $u_\star(x,y)$, at exactly the same exponential rate as its homogeneous counterpart.

17.2. Explicit Solutions for the Heat Equation.

Thus, solving the two-dimensional heat equation in series form requires knowing the eigenfunctions for the associated Helmholtz boundary value problem. Unfortunately, as with any partial differential equation, explicit solution formulae are few and far between. In this section, we discuss two specific cases where the required eigenfunctions can be found in closed form. The calculations rely on separation of variables, which, as we know, works in only a very limited class of domains. Nevertheless, interesting solution features can be gleaned from these particular cases.

Heating of a Rectangle

A homogeneous rectangular plate

$$R = \big\{\, 0 < x < a,\ 0 < y < b \,\big\}$$

is heated to a prescribed initial temperature

$$u(0,x,y) = f(x,y), \qquad \text{for} \qquad (x,y) \in R, \tag{17.31}$$

and then insulated. The sides of the plate are held at zero temperature. Our task is to determine how fast the plate returns to thermal equilibrium.

The temperature $u(t,x,y)$ evolves according to the two-dimensional heat equation

$$u_t = \gamma(u_{xx} + u_{yy}), \qquad \text{for} \qquad (x,y) \in R, \qquad t > 0, \tag{17.32}$$

subject to homogeneous Dirichlet conditions

$$u(0,y) = u(a,y) = 0 = u(x,0) = u(x,b), \qquad 0 < x < a, \quad 0 < y < b, \tag{17.33}$$

along the boundary of the rectangle. As in (17.17), the eigensolutions to the heat equation are obtained from the usual exponential ansatz $u(t,x,y) = e^{-\lambda t}\, v(x,y)$. Substituting this expression into the heat equation, we conclude that the function $v(x,y)$ solves the Helmholtz eigenvalue problem

$$\gamma(v_{xx} + v_{yy}) + \lambda\, v = 0, \qquad (x,y) \in R, \tag{17.34}$$

subject to the same homogeneous Dirichlet boundary conditions

$$v(x,y) = 0, \qquad (x,y) \in \partial R. \tag{17.35}$$

To solve the rectangular Helmholtz eigenvalue problem (17.34–35), we shall, as in (15.13), introduce a further separation of variables, writing

$$v(x,y) = p(x)\, q(y)$$

as the product of functions depending upon the individual Cartesian coordinates. Substituting this ansatz into the Helmholtz equation (17.34), we find

$$\gamma\, p''(x)\, q(y) + \gamma\, p(x)\, q''(y) + \lambda\, p(x)\, q(y) = 0.$$

To effect the variable separation, we collect all terms involving x on one side and all terms involving y on the other side of the equation. This is accomplished by dividing by $v = pq$ and rearranging the terms; the result is

$$\gamma\, \frac{p''(x)}{p(x)} = -\gamma\, \frac{q''(y)}{q(y)} - \lambda \equiv -\mu.$$

The left hand side of this equation only depends on x, whereas the right hand side only depends on y. As argued in Section 15.2, the only way this can occur is if the two sides equal a common *separation constant*, denoted by $-\mu$. (The minus sign is for later convenience.) In this manner, we reduce our partial differential equation to a pair of one-dimensional eigenvalue problems

$$\gamma\, p'' + \mu\, p = 0, \qquad \gamma\, q'' + (\lambda - \mu)\, q = 0,$$

each of which is subject to homogeneous Dirichlet boundary conditions

$$p(0) = p(a) = 0, \qquad q(0) = q(b) = 0,$$

stemming from the boundary conditions (17.35). To obtain a nontrivial solution to the Helmholtz equation, we seek nonzero solutions to these two supplementary eigenvalue problems. The fact that we are dealing with a rectangular domain is critical to the success of this procedure.

We have already solved these particular two boundary value problems many times; see, for instance, (14.17). The eigenfunctions are, respectively,

$$p_m(x) = \sin \frac{m\,\pi\, x}{a}, \qquad m = 1, 2, 3, \ldots, \qquad q_n(y) = \sin \frac{n\,\pi\, y}{b}, \qquad n = 1, 2, 3, \ldots,$$

with

$$\mu = \frac{m^2\,\pi^2\,\gamma}{a^2}, \qquad \lambda - \mu = \frac{n^2\,\pi^2\,\gamma}{b^2}, \qquad \text{so that} \qquad \lambda = \frac{m^2\,\pi^2\,\gamma}{a^2} + \frac{n^2\,\pi^2\,\gamma}{b^2}.$$

Therefore, the separable eigenfunction solutions to the Helmholtz boundary value problem (17.33–34) have the doubly trigonometric form

$$v_{m,n}(x,y) = \sin \frac{m\,\pi\, x}{a}\ \sin \frac{n\,\pi\, y}{b}, \tag{17.36}$$

with associated eigenvalues

$$\lambda_{m,n} = \frac{m^2\,\pi^2\,\gamma}{a^2} + \frac{n^2\,\pi^2\,\gamma}{b^2} = \left(\frac{m^2}{a^2} + \frac{n^2}{b^2}\right)\pi^2\,\gamma\,. \tag{17.37}$$

Each of these corresponds to an exponentially decaying, separable solution

$$u_{m,n}(t,x,y) = e^{-\lambda_{m,n}\,t}\,v_{m,n}(x,y) = \exp\left[-\left(\frac{m^2}{a^2} + \frac{n^2}{b^2}\right)\pi^2\,\gamma\,t\right]\,\sin\frac{m\,\pi\,x}{a}\,\sin\frac{n\,\pi\,y}{b} \tag{17.38}$$

to the original rectangular boundary value problem for the heat equation.

Using the fact that the univariate sine functions form a complete system, it is not hard to prove, [**181**], that the separable eigenfunction solutions (17.38) are complete, which implies that there are no non-separable eigenfunctions. As a consequence, the general solution to the initial-boundary value problem can be expressed as a linear combination

$$u(t,x,y) = \sum_{m,n=1}^{\infty} c_{m,n}\,u_{m,n}(t,x,y) = \sum_{m,n=1}^{\infty} c_{m,n}\,e^{-\lambda_{m,n}\,t}\,v_{m,n}(x,y) \tag{17.39}$$

of our eigenfunction modes. The coefficients $c_{m,n}$ are prescribed by the initial conditions, which take the form of a double Fourier sine series

$$f(x,y) = u(0,x,y) = \sum_{m,n=1}^{\infty} c_{m,n}v_{m,n}(x,y) = \sum_{m,n=1}^{\infty} c_{m,n}\,\sin\frac{m\,\pi\,x}{a}\,\sin\frac{n\,\pi\,y}{b}\,.$$

Self-adjointness of the Laplacian coupled with the boundary conditions implies that the eigenfunctions $v_{m,n}(x,y)$ are orthogonal with respect to the L^2 inner product on the rectangle:

$$\langle\, v_{k,l}\,,v_{m,n}\,\rangle = \int_0^b\int_0^a v_{k,l}(x,y)\,v_{m,n}(x,y)\,dx\,dy = 0 \qquad \text{unless} \qquad k=m \quad \text{and} \quad l=n.$$

(The skeptical reader can verify the orthogonality relations directly from the eigenfunction formulae (17.36).) Thus, we can appeal to our usual orthogonality formula (17.24) to evaluate the coefficients

$$c_{m,n} = \frac{\langle\, f\,,v_{m,n}\,\rangle}{\|\,v_{m,n}\,\|^2} = \frac{4}{a\,b}\int_0^b\int_0^a f(x,y)\,\sin\frac{m\,\pi\,x}{a}\,\sin\frac{n\,\pi\,y}{b}\,dx\,dy, \tag{17.40}$$

where the formula for the norms of the eigenfunctions

$$\|\,v_{m,n}\,\|^2 = \int_0^b\int_0^a v_{m,n}(x,y)^2\,dx\,dy = \int_0^b\int_0^a \sin^2\frac{m\,\pi\,x}{a}\,\sin^2\frac{n\,\pi\,y}{b}\,dx\,dy = \frac{1}{4}\,a\,b. \tag{17.41}$$

follows from a direct evaluation of the double integral. (Unfortunately, while the orthogonality is automatic, the computation of the norm must inevitably be done "by hand".)

The rectangle approaches thermal equilibrium at the rate equal to the smallest eigenvalue:

$$\lambda_{1,1} = \left(\frac{1}{a^2} + \frac{1}{b^2}\right)\pi^2\,\gamma, \tag{17.42}$$

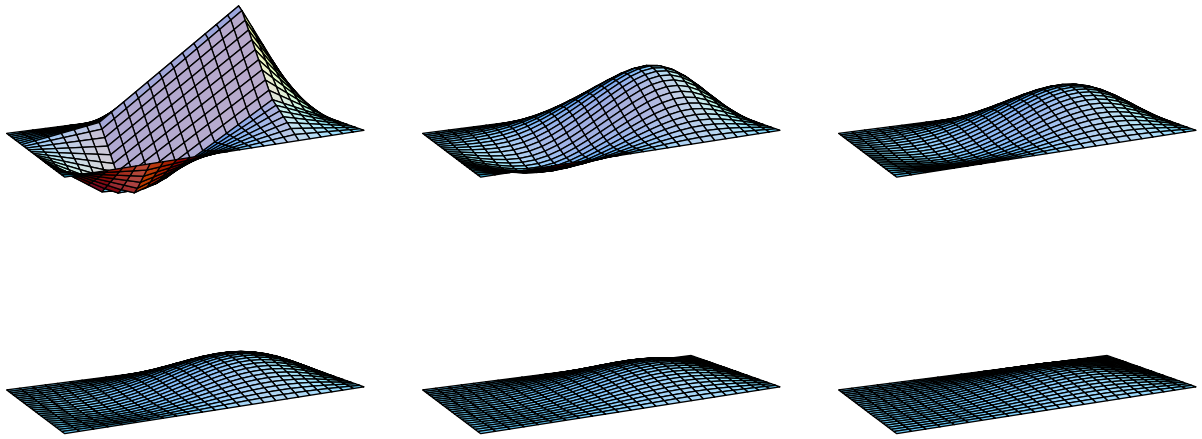


Figure 17.1. Heat Diffusion in a Rectangle.

i.e., the sum of the reciprocals of the squared lengths of its sides multiplied by the diffusion coefficient. The larger the rectangle, or the smaller the diffusion coefficient, the smaller $\lambda_{1,1}$, and hence slower the return to thermal equilibrium. The exponentially fast decay rate of the Fourier series implies that the solution immediately smooths out any initial discontinuites in the initial temperature profile. Indeed, the higher modes, with m and n large, decay to zero almost instantaneously, and so the solution immediately behaves like a finite sum over a few low order modes. Assuming that $c_{1,1} \neq 0$, the slowest decaying mode in the Fourier series (17.39) is

$$c_{1,1}\, u_{1,1}(t,x,y) = c_{1,1} \exp\left[-\left(\frac{1}{a^2} + \frac{1}{b^2} \right) \pi^2\, \gamma\, t \right] \, \sin \frac{\pi\, x}{a} \, \sin \frac{\pi\, y}{b}\,. \tag{17.43}$$

Thus, in the long run, the temperature is of one sign — either positive or negative depending upon the sign of $c_{1,1}$ — throughout the rectangle. This observation is, in fact, indicative of the general phenomenon that the eigenfunction associated with the smallest positive eigenvalue of a self-adjoint elliptic operator is of one sign throughout the domain. A typical solution is plotted in Figure 17.1.

Heating of a Disk

Let us perform a similar analysis of the thermodynamics of a circular disk. For simplicity (or by choice of suitable physical units), we will assume that the disk

$$D = \{\, x^2 + y^2 \leq 1 \,\} \subset \mathbb{R}^2$$

has unit radius and unit diffusion coefficient $\gamma = 1$. We shall solve the heat equation on D subject to homogeneous Dirichlet boundary values of zero temperature at the circular edge

$$\partial D = C = \{\, x^2 + y^2 = 1 \,\}.$$

Thus, the full initial-boundary value problem is

$$\begin{aligned} &\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\,, && x^2 + y^2 < 1, \qquad t > 0, \\ &u(t,x,y) = 0, && x^2 + y^2 = 1, \\ &u(0,x,y) = f(x,y), && x^2 + y^2 \le 1. \end{aligned} \tag{17.44}$$

We remark that a simple rescaling of space and time, as outlined in Exercise ■, can be used to recover the solution for an arbitrary diffusion coefficient and a disk of arbitrary radius from this particular case.

Since we are working in a circular domain, we instinctively pass to polar coordinates (r,θ). In view of the polar coordinate formula (15.29) for the Laplace operator, the heat equation and boundary conditions assume the form

$$\begin{aligned} &\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\,\frac{\partial u}{\partial r} + \frac{1}{r^2}\,\frac{\partial^2 u}{\partial \theta^2}\,, && 0 \le r < 1, \qquad t > 0, \\ &u(t,1,\theta) = 0, && u(0,r,\theta) = f(r,\theta), \qquad r \le 1, \end{aligned} \tag{17.45}$$

where the solution $u(t,r,\theta)$ is defined for all $0 \le r \le 1$ and $t \ge 0$. To ensure that the solution represents a single-valued function on the entire disk, it is required to be a 2π periodic function of the angular variable:

$$u(t,r,\theta + 2\pi) = u(t,r,\theta)$$

To obtain the separable solutions

$$u(t,r,\theta) = e^{-\lambda t}\, v(r,\theta), \tag{17.46}$$

we need to solve the polar coordinate form of the Helmholtz equation

$$\frac{\partial^2 v}{\partial r^2} + \frac{1}{r}\,\frac{\partial v}{\partial r} + \frac{1}{r^2}\,\frac{\partial^2 v}{\partial \theta^2} + \lambda\, v = 0, \qquad \begin{aligned} &0 \le r < 1, \\ &0 \le \theta \le 2\pi, \end{aligned} \tag{17.47}$$

subject to the boundary conditions

$$v(1,\theta) = 0, \qquad v(r,\theta + 2\pi) = v(r,\theta). \tag{17.48}$$

We invoke a further separation of variables by writing

$$v(r,\theta) = p(r)\, q(\theta). \tag{17.49}$$

Substituting this ansatz into the polar Helmholtz equation (17.47), and then collecting together all terms involving r and all terms involving θ, we are led to the pair of ordinary differential equations

$$r^2 p'' + r\, p' + (\lambda r^2 - \mu)\, p = 0, \qquad q'' + \mu\, q = 0, \tag{17.50}$$

where λ is the Helmholtz eigenvalue, and μ the separation constant.

Let us start with the equation for $q(\theta)$. The periodicity condition (17.48) requires that $q(\theta)$ be 2π periodic. Therefore, the required solutions are the elementary trigonometric functions

$$q(\theta) = \cos m\theta \qquad \text{or} \qquad \sin m\theta, \qquad \text{where} \qquad \mu = m^2, \tag{17.51}$$

with $m = 0, 1, 2, \dots$ a non-negative integer.

Substituting the formula, $\mu = m^2$, for the separation constant, the differential equation for $p(r)$ takes the form

$$r^2 \frac{d^2p}{dr^2} + r \frac{dp}{dr} + (\lambda r^2 - m^2)\, p = 0, \qquad 0 \le r \le 1. \tag{17.52}$$

Ordinarily, one imposes two boundary conditions in order to pin down a solution to such a second order ordinary differential equation. But our Dirichlet condition, namely $p(1) = 0$, only specifies its value at one of the endpoints. The other endpoint is a *singular point* for the ordinary differential equation, because the coefficient of the highest order derivative, namely r^2, vanishes at $r = 0$. This situation might remind you of our solution to the Euler differential equation (15.35) in the context of separable solutions to the Laplace equation on the disk. As there, we only require the solution to be bounded at $r = 0$:

$$|\, p(0)\,| < \infty, \qquad p(1) = 0. \tag{17.53}$$

As we now show, this pair of boundary conditions suffices to distinguish the relevant eigenfunction solutions to (17.52).

Although the ordinary differential equation (17.52) appears in a variety of applications, this may be the first time that you have encountered it. It is not an elementary equation; indeed, most solutions cannot be written in terms of the elementary functions you see in first year calculus. Nevertheless, owing to their significance in a wide range of physical applications, the solutions have been extensively studied and tabulated, and so are, in a sense, well-known. After some preparatory manipulations, we shall summarize their relevant properties, relegating details and proofs to Appendix C.

To simplify the analysis, we make a preliminary rescaling of the independent variable, replacing r by

$$z = \sqrt{\lambda}\, r.$$

Note that, by the chain rule,

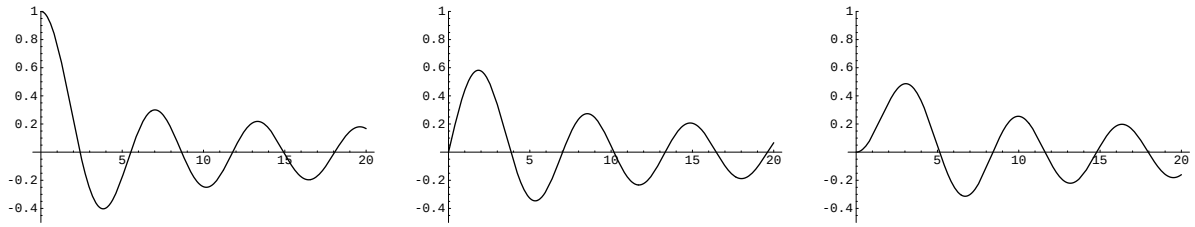
$$\frac{dp}{dr} = \sqrt{\lambda}\, \frac{dp}{dz}, \qquad \frac{d^2p}{dr^2} = \lambda\, \frac{d^2p}{dz^2},$$

and hence

$$r \frac{dp}{dr} = z \frac{dp}{dz}, \qquad r^2 \frac{d^2p}{dr^2} = z^2 \frac{d^2p}{dz^2}.$$

The net effect is to eliminate the eigenvalue parameter λ (or, rather, hide it in the change of variables), so that (17.52) assumes the slightly simpler form

$$z^2 \frac{d^2p}{dz^2} + z \frac{dp}{dz} + (z^2 - m^2)\, p = 0. \tag{17.54}$$



$J_0(z)$ $J_1(z)$ $J_2(z)$

Figure 17.2. Bessel Functions.

The ordinary differential equation (17.54) is known as *Bessel's equation*, named after the early 19th century astronomer Wilhelm Bessel, who first used its solutions to analyze planetary orbits. The solutions to Bessel's equation have become an indispensable tool in applied mathematics, physics, and engineering.

To begin, the one thing we know for sure is that, as with any second order ordinary differential equation, there are two linearly independent solutions. However, it turns out that, up to a constant multiple, only one solution remains bounded as $z \to 0$. This solution is known as the *Bessel function* of *order* m, and is denoted by $J_m(z)$. Applying the general systematic method for finding power series solutions to linear ordinary differential equations presented in Appendix C, it can be shown that the Bessel function of order m has the Taylor expansion

$$J_m(z) = \sum_{k=0}^{\infty} \frac{(-1)^k\, z^{m+2\,k}}{2^{m+2\,k}\, k!\, (m+k)!} \tag{17.55}$$
$$= \frac{z^m}{2^m\, m!}\left[1 - \frac{z^2}{4\,(m+1)} + \frac{z^4}{32\,(m+1)(m+2)} - \frac{z^6}{384\,(m+1)(m+2)(m+3)} + \cdots\right].$$

A simple application of the ratio test tells us that the power series converges for all (complex) values of z, and hence $J_m(z)$ is everywhere analytic. Indeed, the convergence is quite rapid when z is of moderate size, and so summing the series is a reasonably effective method for computing the Bessel function $J_m(z)$ — although in serious applications one adopts more sophisticated numerical techniques based on asymptotic expansions and integral formulae, [**3**, **140**]. Moreover, verification that it indeed solves the Bessel equation (17.54) is a straightforward computation. Figure 17.2 displays graphs of the first three Bessel functions for $z > 0$. Most software packages, both symbolic and numerical, include routines for accurately evaluating and graphing Bessel functions.

Reverting back to our original radial coordinate $r = z/\sqrt{\lambda}$, we conclude that every solution to the radial equation (17.52) which is bounded at $r = 0$ is a constant multiple

$$p(r) = J_m\big(\sqrt{\lambda}\, r\,\big) \tag{17.56}$$

of the rescaled Bessel function of order m. So far, we have only dealt with the boundary condition at the singular point $r = 0$. The Dirichlet condition at the other end requires

$$p(1) = J_m\big(\sqrt{\lambda}\,\big) = 0.$$

Therefore, in order that λ be a legitimate eigenvalue, $\sqrt{\lambda}$ must be a *root* of the m^{th} order Bessel function J_m.

Remark: We already know, thanks to the positive definiteness of the Dirichlet boundary value problem, that the Helmholtz eigenvalues $\lambda > 0$ must be positive, and so there is no problem taking the square root. Indeed, it can be proved, [**180**], that the Bessel functions do not have any negative roots!

The graphs of $J_m(z)$ strongly indicate, and, indeed, it can be rigorously proved, that each Bessel function oscillates between positive and negative values as z increases above 0, with slowly decreasing amplitude. In fact, it can be proved that, asymptotically,

$$J_m(z) \quad \sim \quad \sqrt{\frac{2}{\pi\, z}}\, \cos\big(\, z - \big(\, \tfrac{1}{2}\, m + \tfrac{1}{4}\,\big)\, \pi\,\big) \qquad \text{as} \qquad z \longrightarrow \infty, \tag{17.57}$$

and so the oscillations become essentially the same as a (phase-shifted) cosine whose amplitude decreases, relatively slowly, like $z^{-1/2}$. As a consequence, there exists an infinite sequence of *Bessel roots*, which we number in the order in which they appear:

$$\begin{aligned} &J_m(\zeta_{m,n}) = 0, \qquad \text{where} \\ &\qquad 0 < \zeta_{m,1} < \zeta_{m,2} < \zeta_{m,3} < \, \cdots \quad \text{with} \quad \zeta_{m,n} \longrightarrow \infty \quad \text{as} \quad n \longrightarrow \infty. \end{aligned} \tag{17.58}$$

It is worth noting that the Bessel functions are *not* periodic, and their roots are not initially evenly spaced.

Owing to their physical importance in a wide range of problems, the Bessel roots have been extensively tabulated in the literature, cf. [**3**, **58**]. The accompanying table displays all Bessel roots that are < 12 in magnitude. The columns of the table are indexed by m, the order of the Bessel function, and the rows by n, the root number.

Table of Bessel Roots $\zeta_{m,n}$

n \ m	0	1	2	3	4	5	6	7	...
1	2.4048	3.8317	5.1356	6.3802	7.5883	8.7715	9.9361	11.0860	...
2	5.5201	7.0156	8.4172	9.761	11.0650	⋮	⋮	⋮	
3	8.6537	10.1730	11.6200	⋮	⋮				
4	11.7920	⋮	⋮						
⋮	⋮								

Remark: According to (17.55),

$$J_m(0) = 0 \qquad \text{for} \qquad m > 0, \qquad \text{while} \qquad J_0(0) = 1.$$

However, we do not count 0 as a *bona fide* Bessel root, since it does not lead to a valid eigenfunction for the Helmholtz boundary value problem.

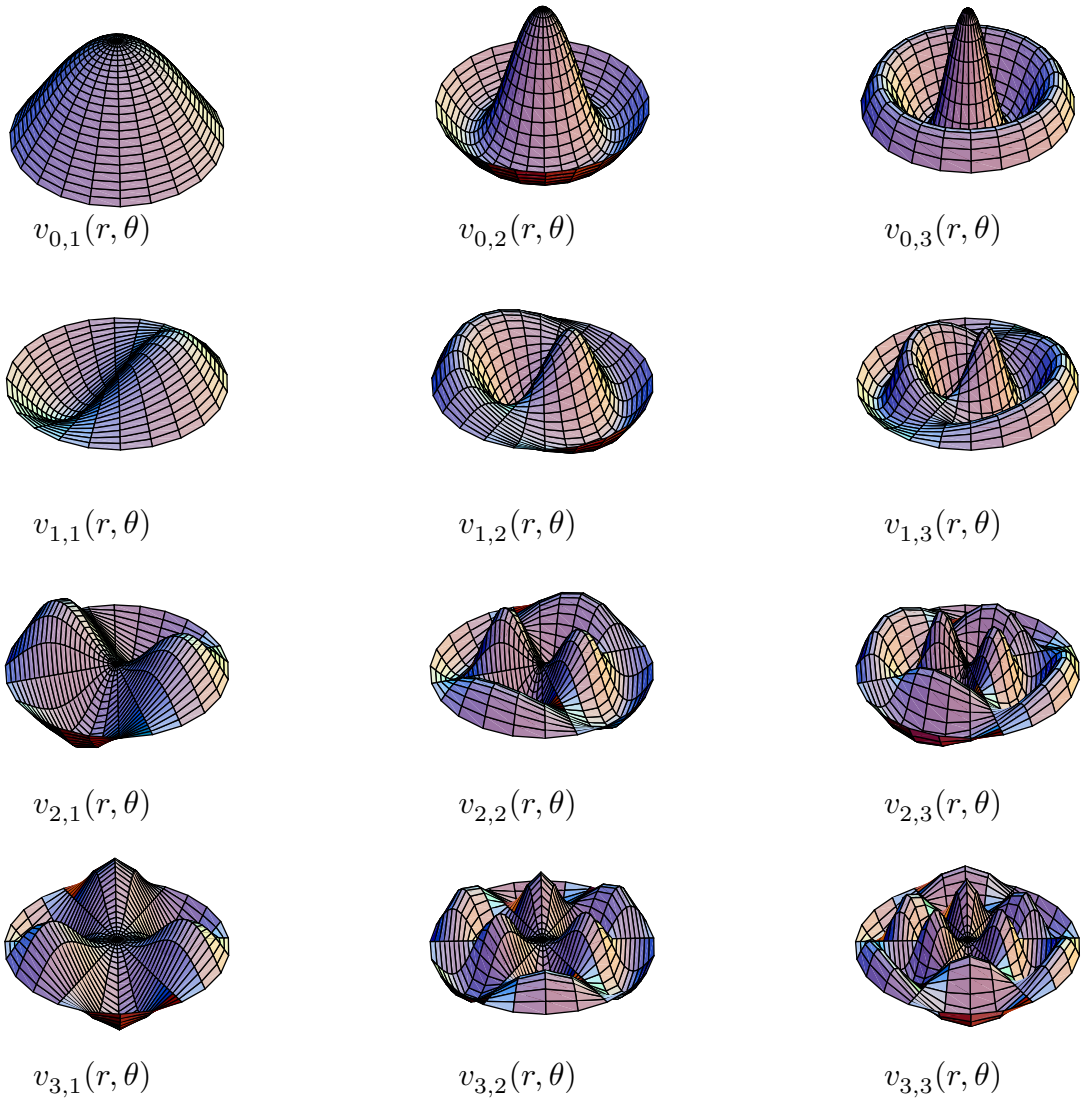


Figure 17.3. Normal Modes for a Disk.

Summarizing our progress, the eigenvalues

$$\lambda_{m,n} = \zeta_{m,n}^2, \qquad n = 1, 2, 3, \ldots, \qquad m = 0, 1, 2, \ldots, \tag{17.59}$$

of the Bessel boundary value problem (17.52–53) are the squares of the roots of the Bessel function of order m. The corresponding eigenfunctions are

$$w_{m,n}(r) = J_m(\zeta_{m,n}\, r)\,, \qquad n = 1, 2, 3, \ldots, \qquad m = 0, 1, 2, \ldots, \tag{17.60}$$

defined for $0 \le r \le 1$. Combining (17.60) with the formula (17.51) for the angular components, we conclude that the separable solutions (17.49) to the polar Helmholtz boundary

value problem (17.47) are

$$\begin{aligned} v_{0,n}(r,\theta) &= J_0(\zeta_{0,n}\, r), \\ v_{m,n}(r,\theta) &= J_m(\zeta_{m,n}\, r)\, \cos m\theta, \\ \widehat{v}_{m,n}(r,\theta) &= J_m(\zeta_{m,n}\, r)\, \sin m\theta, \end{aligned} \qquad \text{where} \qquad \begin{aligned} &n = 1, 2, 3, \dots, \\ &m = 0, 1, 2, \dots. \end{aligned} \tag{17.61}$$

These solutions define the so-called *normal modes* for the unit disk, and Figure 17.3 plots the first few of them. The eigenvalues $\lambda_{0,n}$ are simple, and contribute radially symmetric eigenfunctions, whereas the eigenvalues $\lambda_{m,n}$ for $m > 0$ are double, and produce two linearly independent separable eigenfunctions, with trigonometric dependence on the angular variable.

We have at last produced the basic separable solutions

$$\begin{aligned} u_{0,n}(t,r) &= e^{-\,\zeta_{0,n}^2\, t}\, J_0(\zeta_{0,n}\, r), \\ u_{m,n}(t,r,\theta) &= e^{-\,\zeta_{m,n}^2\, t}\, J_m(\zeta_{m,n}\, r)\, \cos m\theta, \\ \widehat{u}_{m,n}(t,r,\theta) &= e^{-\,\zeta_{m,n}^2\, t}\, J_m(\zeta_{m,n}\, r)\, \sin m\theta, \end{aligned} \qquad \begin{aligned} &n = 1, 2, 3, \dots, \\ &m = 1, 2, \dots. \end{aligned} \tag{17.62}$$

to the homogeneous Dirichlet boundary value problem for the heat equation on the unit disk (17.45). The general solution is a linear superposition, in the form of an infinite series

$$u(t,r,\theta) = \frac{1}{2} \sum_{n=1}^{\infty} a_{0,n}\, u_{0,n}(t,r) + \sum_{m,n=1}^{\infty} \big[\, a_{m,n}\, u_{m,n}(t,r,\theta) + b_{m,n}\, \widehat{u}_{m,n}(t,r,\theta) \,\big], \tag{17.63}$$

where the initial factor of $\frac{1}{2}$ is included, as with ordinary Fourier series, for later convenience. As usual, the coefficients $a_{m,n}, b_{m,n}$ are determined by the initial condition, so

$$u(0,r,\theta) = \frac{1}{2} \sum_{n=1}^{\infty} a_{0,n}\, v_{0,n}(r) + \sum_{m,n=1}^{\infty} \big[\, a_{m,n}\, v_{m,n}(r,\theta) + b_{m,n}\, \widehat{v}_{m,n}(r,\theta) \,\big] = f(r,\theta). \tag{17.64}$$

Thus, we must expand the initial data into a *Fourier–Bessel series* in the eigenfunctions. As in the rectangular case, it is possible to prove, [**46**], that the separable eigenfunctions are *complete* — there are no other eigenfunctions — and, moreover, every (reasonable) function defined on the unit disk can be written as a convergent series in the Bessel eigenfunctions.

Theorem 14.19 gurantees that the eigenfunctions are orthogonal[†] with respect to the standard L^2 inner product

$$\langle\, u\,, v\,\rangle = \iint_D u(x,y)\, v(x,y)\, dx\, dy = \int_0^1 \int_0^{2\,\pi} u(r,\theta)\, v(r,\theta)\; r\, d\theta\, dr$$

[†] For the two eigenfunctions corresponding to one of the double eigenvalues, orthogonality must be verified by hand.

Norms of the Fourier–Bessel Eigenfunctions $\| v_{m,n} \| = \| \widehat{v}_{m,n} \|$

n \ m	0	1	2	3	4	5	6	7
1	.6507	.5048	.4257	.3738	.3363	.3076	.2847	.2658
2	.4265	.3761	.3401	.3126	.2906	.2725	.2572	.2441
3	.3402	.3130	.2913	.2736	.2586	.2458	.2347	.2249
4	.2913	.2737	.2589	.2462	.2352	.2255	.2169	.2092
5	.2589	.2462	.2353	.2257	.2171	.2095	.2025	.1962

on the unit disk. (Note the extra factor of r coming from the polar coordinate form (A.51) of the infinitesimal element of area $dx\,dy = r\,dr\,d\theta$.) The norms of the Fourier–Bessel functions are given by the interesting formula

$$\| v_{m,n} \| = \| \widehat{v}_{m,n} \| = \sqrt{\frac{\pi}{2}}\, \big| J_{m+1}(\zeta_{m,n}) \big| \tag{17.65}$$

that involves the value of the Bessel function of the next higher order at the appropriate Bessel root; numerical values appear in the accompanying table. A proof of this formula can be found in ■–■.

Orthogonality of the eigenfunctions implies that the coefficients in the Fouier–Bessel series (17.64) are given by

$$\begin{aligned} a_{m,n} &= \frac{\langle\, f\,, v_{m,n}\,\rangle}{\| v_{m,n} \|^2} = \frac{2}{\pi\, J_{m+1}(\zeta_{m,n})^2} \int_0^1 \int_0^{2\pi} f(r,\theta)\, J_m(\zeta_{m,n}\, r)\; r \cos m\theta \, d\theta \, dr, \\ b_{m,n} &= \frac{\langle\, f\,, \widehat{v}_{m,n}\,\rangle}{\| \widehat{v}_{m,n} \|^2} = \frac{2}{\pi\, J_{m+1}(\zeta_{m,n})^2} \int_0^1 \int_0^{2\pi} f(r,\theta)\, J_m(\zeta_{m,n}\, r)\; r \sin m\theta \, d\theta \, dr. \end{aligned} \tag{17.66}$$

In accordance with the general theory, each individual separable solution (17.62) to the heat equation decays exponentially fast, at a rate $\lambda_{m,n} = \zeta_{m,n}^2$ prescribed by the square of the corresponding Bessel root. In particular, the dominant mode, meaning the one that persists the longest, is

$$u_{0,1}(t,r,\theta) = e^{-\,\zeta_{0,1}^2\, t}\, J_0(\zeta_{0,1}\, r). \tag{17.67}$$

Its decay rate

$$\zeta_{0,1}^2 \;\approx\; 5.783 \tag{17.68}$$

is the square of the smallest non-zero root of the Bessel function $J_0(z)$. The dominant eigenfunction $v_{0,1}(r,\theta) = J_0(\zeta_{0,1}\, r) > 0$ is strictly positive within the entire disk and radially symmetric. Consequently, for most initial conditions (specifically those for which $a_{0,1} \neq 0$), the disk's temperature distribution eventually becomes entirely of one sign and radially symmetric, while decaying exponentially fast to zero at the rate given by

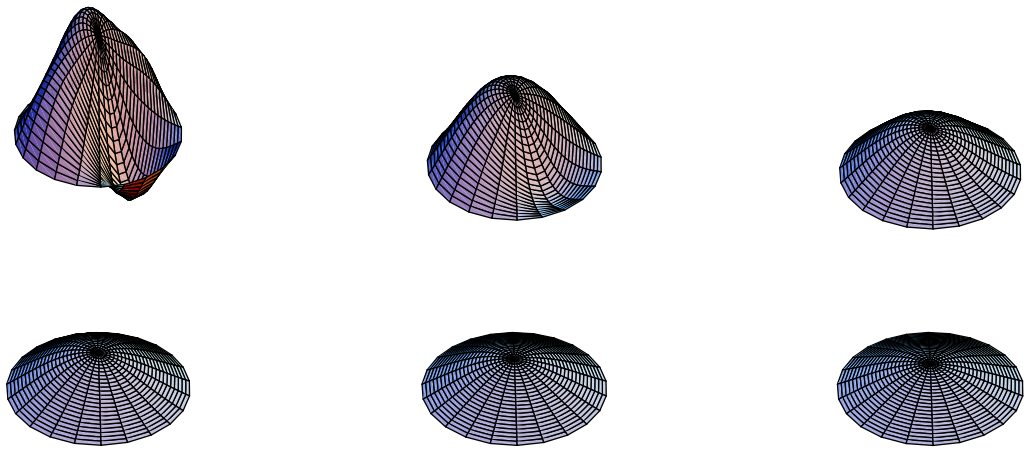


Figure 17.4. Heat Diffusion in a Disk.

(17.68). See Figure 17.4 for a plot of a typical solution, displayed as successive times $t = 0, .04, .08, .12, .16, .2$. Note how, in accordance with the theory, the solution almost immediately acquires a radial symmetry, followed by a fairly rapid decay to thermal equilibrium.

17.3. The Fundamental Solution.

As we learned in Section 14.1, the *fundamental solution* to the heat equation measures the temperature distribution resulting from a concentrated initial heat source, e.g., a hot soldering iron applied instantaneously at one pointe. The physical problem is modeled mathematically by imposing a delta function as the initial condition for the heat equation, along with the chosen homogeneous boundary conditions. Knowledge of the fundamental solution enables one to use linear superposition to recover the solution for any other initial data.

As in our one-dimensional analysis, we shall concentrate on the most tractable case, when the domain is the entire plane: $\Omega = \mathbb{R}^2$. Our first goal will be to solve the initial value problem

$$u_t = \gamma\, \Delta u, \qquad u(0, x, y) = \delta(x - \xi, y - \eta) = \delta(x - \xi)\, \delta(y - \eta), \tag{17.69}$$

for $t > 0$ and $(x, y) \in \mathbb{R}^2$. The initial data is a delta function representing a concentrated unit heat source placed at position (ξ, η). The resulting solution $u = F(t, x, y; \xi, \eta)$ is called the *fundamental solution* for the heat equation on $\mathbb{R}^2$.

The quickest route to the desired solution relies on the following simple lemma that combines solutions of the one-dimensional heat equation to produce solutions of the two-dimensional version.

Lemma 17.1. *If $v(t, x)$ and $w(t, x)$ are any two solutions to the one-dimensional heat equation $u_t = \gamma\, u_{xx}$, then the product*

$$u(t, x, y) = v(t, x)\, w(t, y) \tag{17.70}$$

is a solution to the two-dimensional heat equation $u_t = \gamma\,(u_{xx} + u_{yy})$.

Proof: Our assumptions imply that that $v_t = \gamma\, v_{xx}$, while $w_t = \gamma\, w_{yy}$ when we write $w(t,y)$ as a function of t and y. Therefore, differentiating (17.70), we find

$$\frac{\partial u}{\partial t} = \frac{\partial v}{\partial t}\, w + v\, \frac{\partial w}{\partial t} = \gamma\, \frac{\partial^2 v}{\partial x^2}\, w + \gamma\, v\, \frac{\partial^2 w}{\partial y^2} = \gamma \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right),$$

and hence $u(t,x,y)$ solves the heat equation. *Q.E.D.*

For example, if

$$v(t,x) = e^{-\gamma \omega^2 t} \sin \omega x, \qquad w(t,y) = e^{-\gamma \nu^2 t} \sin \nu y,$$

are separable solutions of the one-dimensional heat equation, then

$$u(t,x,y) = e^{-\gamma (\omega^2 + \nu^2) t} \sin \omega x \sin \nu y$$

are the separable solutions we used to solve the heat equation on a rectangle. A more interesting case is to chose

$$v(t,x) = \frac{1}{2\sqrt{\pi\,\gamma\, t}}\, e^{-(x-\xi)^2/4\gamma t}, \qquad w(t,y) = \frac{1}{2\sqrt{\pi\,\gamma\, t}}\, e^{-(y-\eta)^2/4\gamma t}, \tag{17.71}$$

to be the fundamental solutions (14.58) to the one-dimensional heat equation at respective locations $x = \xi$ and $y = \eta$. Multiplying these two solutions together produces the fundamental solution for the two-dimensional problem.

Proposition 17.2. *The fundamental solution to the heat equation* $u_t = \gamma\, \Delta u$ *corresponding to a unit delta function placed at position* $(\xi, \eta) \in \mathbb{R}^2$ *at the initial time* $t_0 = 0$ *is*

$$F(t,x,y;\xi,\eta) = \frac{1}{4\,\pi\,\gamma\, t}\, e^{-[(x-\xi)^2 + (y-\eta)^2]/4\gamma t}. \tag{17.72}$$

Proof: Since we already know that both function (17.71) are solutions to the one-dimensional heat equation, Lemma 17.1 guarantees that their product $u(t,x,y) = v(t,x)\, w(t,y)$, which equals (17.72), solves the two-dimensional heat equation for $t > 0$. Moreover, at the initial time

$$u(0,x,y) = v(0,x)\, w(0,y) = \delta(x-\xi)\, \delta(y-\eta)$$

is a product of delta functions, and hence the result follows. Indeed, the total heat

$$\iint u(t,x,y)\, dx\, dy = \int_{-\infty}^{\infty} v(t,x)\, dx \int_{-\infty}^{\infty} w(t,y)\, dy = 1, \qquad t \geq 0,$$

remains constant, while

$$\lim_{t \to 0^+} u(t,x,y) = \begin{cases} \infty, & (x,y) = (\xi,\eta), \\ 0, & \text{otherwise.} \end{cases}$$

has the standard delta function limit at the initial time instant. *Q.E.D.*

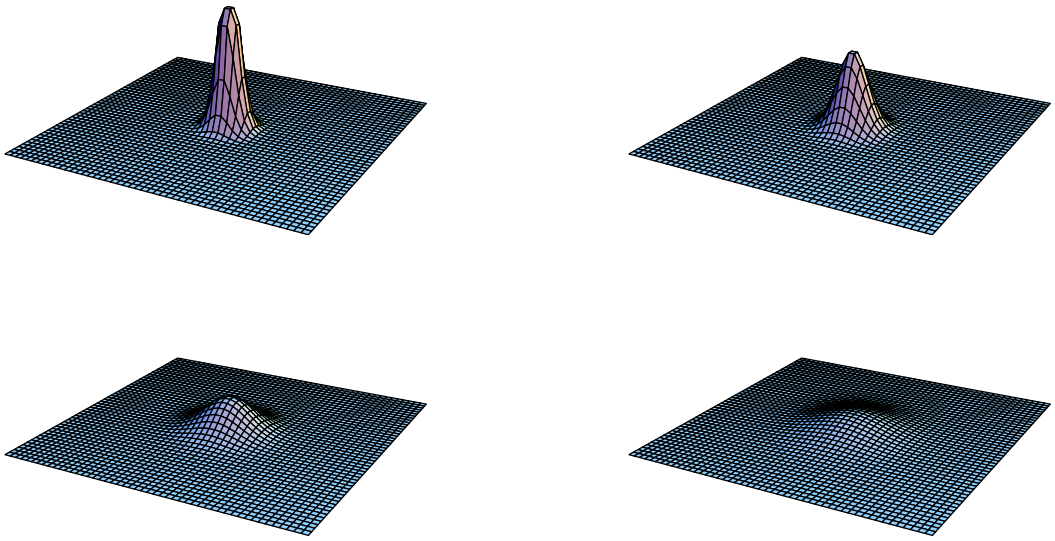


Figure 17.5. Fundamental Solution to the Planar Heat Equation.

Figure 17.5 depicts the evolution of the fundamental solution when $\gamma = 1$ at successive times $t = .01, .02, .05, .1$. Observe that the initially concentrated heat spreads out in a radially symmetric manner. The total amount of heat remains constant. At any individual point $(x, y) \neq (0, 0)$, the initially zero temperature rises slightly at first, but then decays monotonically back to zero at a rate proportional to $1/t$.

Both the one- and two-dimensional fundamental solutions have a bell-shaped profile known as a *Gaussian* function. The most important difference is the initial factor. In a one-dimensional medium, the fundamental solution decays in proportion to $1/\sqrt{t}$, whereas in the plane the decay is more rapid, being proportional to $1/t$. The physical explanation is that the energy is able to spread out in two independent directions, and hence diffuses away from its initial source more rapidly. As we shall see, the decay in three-dimensional space is more rapid still, being proportional to $t^{-3/2}$ for similar reasons; see (18.103).

The principal purpose of the fundamental solution is to solve the general initial value problem. We express the initial temperature distribution as a superposition of delta function sources,

$$u(0, x, y) = f(x, y) = \iint f(\xi, \eta)\, \delta(x - \xi, y - \eta)\, d\xi\, d\eta,$$

where, at the point $(\xi, \eta) \in \mathbb{R}^2$, the source has magnitude $f(\xi, \eta)$. Linearity implies that the solution is then given by the same superposition of fundamental solutions.

Theorem 17.3. *The solution to the initial value problem*

$$u_t = \gamma\, \Delta u, \qquad u(t, x, y) = f(x, y), \qquad (x, y) \in \mathbb{R}^2,$$

for the planar heat equation is given by the linear superposition formula

$$u(t, x, y) = \frac{1}{4\,\pi\,\gamma\, t} \iint f(\xi, \eta)\, e^{-[(x-\xi)^2 + (y-\eta)^2]/4\,\gamma\, t}\, d\xi\, d\eta. \tag{17.73}$$

We can interpret the solution formula (17.73) as a two-dimensional *convolution*

$$u(t,x,y) = F(t,x,y) * f(x,y) \tag{17.74}$$

of the initial data with a one-parameter family of progressively wider and shorter Gaussian filters

$$F(t,x,y) = F(t,x,y;0,0) = \frac{1}{4\pi\gamma t}\, e^{-[x^2+x^2]/4\gamma t}. \tag{17.75}$$

As in (13.126), such a convolution can be interpreted as a weighted averaging of the function, which has the effect of smoothing out the initial signal $f(x,y)$.

Example 17.4. If our initial temperature distribution is constant on a circular region, say

$$u(0,x,y) = \begin{cases} 1 & x^2+y^2 < 1, \\ 0, & \text{otherwise}, \end{cases}$$

Then the solution can be evaluated using (17.73), as follows:

$$u(t,x,y) = \frac{1}{4\pi t} \iint_D e^{-[(x-\xi)^2+(y-\eta)^2]/4t}\, d\xi\, d\eta,$$

where the integral is over the unit disk $D = \{\, \xi^2+\eta^2 \le 1 \,\}$. Unfortunately, the integral cannot be expressed in terms of elementary functions. On the other hand, numerical evaluation of the integral is straightforward. A plot of the resulting radially symmetric solution appears in Figure h2disk■.

For more general configurations, when analytical formulas are no longer available, one turns to numerical approximation methods. The most popular are based on a two-dimensional variant of the Crank–Nicholson scheme (14.156), relying on either finite differences or finite elements to discretize the space coordinates. We will not dwell on the details, but refer the interested reader to [**34**, **148**, **107**].

17.4. The Planar Wave Equation.

The second important class of dynamical equations are those governing vibrational motions. The simplest planar system of this type is the two-dimensional wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \Delta u = c^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right), \tag{17.76}$$

which models the free (unforced) vibrations of a uniform two-dimensional membrane, e.g., a drum. Here $u(t,x,y)$ represents the displacement of the membrane at time t and position $(x,y) \in \Omega$, where the domain $\Omega \subset \mathbb{R}^2$ represents the undeformed shape of the membrane. The constant $c^2 > 0$ encapsulates the membrane's physical properties — density, tension, stiffness, etc.; its square root c is called, as in the one-dimensional case, the *wave speed*, since it is the speed at which localized signals propagate through the membrane.

Remark: In this simplified model, we are only allowing small, transverse (vertical) displacements of the membrane. Large elastic vibrations lead to the nonlinear partial differential equations of elastodynamics, [**8**]. The bending vibrations of a flexible plate, which can be viewed as the two-dimensional version of a beam, are governed by a more complicated fourth order partial differential equation; see Exercise ■.

The solution $u(t, x, y)$ to the wave equation will be uniquely specified once we impose suitable boundary and initial conditions. The Dirichlet conditions

$$u = h \qquad \text{on} \qquad \partial\Omega, \tag{17.77}$$

correspond to gluing our membrane to a fixed boundary — a rim. On the other hand, the homogeneous Neumann conditions

$$\frac{\partial u}{\partial \mathbf{n}} = 0 \qquad \text{on} \qquad \partial\Omega, \tag{17.78}$$

represent a free boundary where the membrane is not attached to any support. Mixed boundary conditions attach part of the boundary and leave the remaining portion free to vibrate:

$$u = h \qquad \text{on} \qquad D, \qquad\qquad \frac{\partial u}{\partial \mathbf{n}} = 0 \qquad \text{on} \qquad N, \tag{17.79}$$

where $\partial\Omega = D \cup N$ with D and N non-overlapping. Since the wave equation is second order in time, we also need to impose two initial conditions:

$$u(0, x, y) = f(x, y), \qquad\qquad \frac{\partial u}{\partial t}\,(0, x, y) = g(x, y), \qquad\qquad (x, y) \in \Omega. \tag{17.80}$$

The first specifies the initial displacement of the membrane, while the second prescribes its initial velocity.

The wave equation is the simplest example of a general second order system of Newtonian form

$$\frac{\partial^2 u}{\partial t^2} = -\,K[\,u\,] = -\,\nabla^* \circ \nabla u, \tag{17.81}$$

as presented in Section 14.7. As in (17.15), adopting general weighted inner products

$$\langle\, u\,, \widetilde{u}\,\rangle = \iint_\Omega u(x, y)\, \widetilde{u}(x, y)\, \rho(x, y)\, dx\, dy, \quad \langle\!\langle\, \mathbf{v}\,, \widetilde{\mathbf{v}}\,\rangle\!\rangle = \iint_\Omega \mathbf{v}(x, y) \cdot \widetilde{\mathbf{v}}(x, y)\, \kappa(x, y)\, dx\, dy, \tag{17.82}$$

on, respectively, the spaces of scalar and vector fields, the adjoint to the gradient is a rescaled version of the divergence operator

$$\nabla^* \mathbf{v} = -\,\frac{1}{\rho}\,\nabla \cdot (\kappa\, \mathbf{v}).$$

Therefore, the system (17.81) assumes the self-adjoint form

$$u_{tt} = -\,K[\,u\,] = \frac{1}{\rho}\,\nabla \cdot (\kappa\, \mathbf{v}),$$

or, in full detail,

$$\rho(x,y)\,\frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial x}\left(\kappa(x,y)\,\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(\kappa(x,y)\,\frac{\partial u}{\partial y}\right). \tag{17.83}$$

The resulting hyperbolic partial differential equation models the small transverse vibrations of a nonuniform membrane, in which $\rho(x,y) > 0$ represents the density of the membrane at the point $(x,y) \in \Omega$, while $\kappa(x,y) > 0$ represents its stiffness — in direct analogy with the one-dimensional version (14.189). In particular, if the material is homogeneous, then both ρ and κ are constant, and (17.83) reduces to the two-dimensional wave equation (17.76) with wave speed

$$c = \sqrt{\frac{\kappa}{\rho}}\,. \tag{17.84}$$

Separation of Variables

According to the general framework establisahed in Section 14.7, the separable solutions to the vibration equation (17.81) have the trigonometric form

$$u_k(t,x,y) = \cos\omega_k t\; v_k(x,y) \qquad \text{and} \qquad \widetilde{u}_k(t,x,y) = \cos\omega_k t\; v_k(x,y), \tag{17.85}$$

in which $v_k(x,y)$ is an eigenfunction of the assocaited boundary value problem

$$K[v] = \omega_k^2\, v = \lambda_k\, v. \tag{17.86}$$

The eigenvalue $\lambda_k = \omega_k^2$ equals the square of the vibrational frequency. As in (17.20), there are an infinite number of such *normal modes*, having progressively faster and faster vibrational frequencies: $\omega_k \to \infty$ as $k \to \infty$. In addition, in the positive semi-definite case — which occurs under homogeneous Neumann boundary conditions — there is a single constant null eigenfunction, leading to the additional separable solutions

$$u_0(t,x,y) = 1 \qquad \text{and} \qquad \widetilde{u}_0(t,x,y) = t. \tag{17.87}$$

The first represents a stationary membrane that has been displaced by a fixed amount in the vertical direction, while the second represents a membrane that is moving off in the vertical direction with constant unit speed. (Think of the membrane moving in outer space unaffected by any external gravitational force.) Specializing to the wave equation (17.76), the required eigenvalue problem (17.86) reduces to the Helmholtz eigenvalue problem

$$c^2\,\Delta v + \lambda\, v = c^2\,(v_{xx} + v_{yy}) + \lambda\, v = 0 \tag{17.88}$$

that we analyzed earlier in this chapter.

In the stable cases (Drichlet or mixed), the general solution to the initial value problem can be built up as a quasi-periodic *eigenfunction series*

$$u(t,x,y) = \sum_{k=1}^{\infty} a_k\, u_k(t,x,y) + b_k\, \widetilde{u}_k(t,x,y) = \sum_{k=1}^{\infty} \left(a_k \cos\omega_k t + b_k \sin\omega_k t\right) v_k(x,y) \tag{17.89}$$

in the fundamental vibrational modes. The coefficients a_k, b_k are prescribed by the initial conditions:

$$\sum_{k=1}^{\infty} a_k\, v_k(x,y) = f(x,y), \qquad \sum_{k=1}^{\infty} \omega_k\, b_k\, v_k(x,y) = g(x,y), \tag{17.90}$$

whence, by orthogonality of the eigenfunctions,

$$a_k = \frac{\langle\, f\,, v_k\,\rangle}{\|\, v_k\,\|^2} = \frac{\displaystyle\iint_\Omega f\, v_k\, \rho\; dx\, dy}{\displaystyle\iint_\Omega v_k^2\, \rho\; dx\, dy}\,, \qquad b_k = \frac{1}{\omega_k}\,\frac{\langle\, g\,, v_k\,\rangle}{\|\, v_k\,\|^2} = \frac{\displaystyle\iint_\Omega g\, v_k\, \rho\; dx\, dy}{\omega_k \displaystyle\iint_\Omega v_k^2\, \rho\; dx\, dy}\,. \tag{17.91}$$

In the case of the wave equation, the density ρ is constant, and hence can be canceled from the numerator and denominator of the orthogonality formulae (17.90).

For the Neumann boundary value problem, the eigenfunction series solution takes an amended form

$$u(t,x,y) = a_0 + b_0\, t + \sum_{k=1}^{\infty} \big(\, a_k \cos \omega_k\, t + b_k \sin \omega_k\, t\,\big)\, v_k(x,y). \tag{17.92}$$

The coefficients a_k, b_k for $k > 0$ are given by the same orthogonality formulae (17.91). The only unstable, non-periodic mode is the linearly growing term $b_0\, t$; its coefficient

$$b_0 = \frac{\langle\, g\,, 1\,\rangle}{\|\, 1\,\|^2} = \frac{\displaystyle\iint_\Omega g\, \rho\; dx\, dy}{\displaystyle\iint_\Omega \rho\; dx\, dy}\,,$$

is a weighted average of the initial velocity $g(x,y) = u_t(0,x,y)$ over the domain. In the case of the wave equation, the density ρ is constant, and hence

$$b_0 = \frac{1}{\text{area }\Omega} \iint_\Omega g(x,y)\, dx\, dy$$

equals the average initial velocity. If the (weighted) average initial velocity $b_0 \neq 0$ is nonzero, then the membrane will move off at an average vertical speed b_0 — while quasiperiodically vibrating in any of the normal modes that have been excited by the initial displacement and/or initial velocity. Again, this is a two-dimensional formulation of our observations of a free, vibrating bar — which in turn was the continuum version of an unsupported mass–spring chain.

Remark: An interesting question is whether two different drums can have identical vibrational frequencies. Or, more descriptively, can one hear the shape of a drum? The answer turns out to be "no", but for quite subtle reasons. See [**drum**] for a discussion.

17.5. Analytical Solutions of the Wave Equation.

The previous section summarized the general, qualitative features exhibited by solutions to two-dimensional vibration and wave equations. Exact analytical formulas are, of course, harder to come by. In this section, we analyze the two most important special cases — rectangular and circular membranes.

Vibration of a Rectangular Drum

Let us first consider the vibrations of a membrane in the shape of a rectangle

$$R = \big\{\, 0 < x < a,\ 0 < y < b \,\big\}$$

with side lengths a and b, whose sides are fixed to the (x, y)–plane. Thus, we seek to solve the wave equation

$$u_{tt} = c^2 \Delta u = c^2 (u_{xx} + u_{yy}), \qquad 0 < x < a, \qquad 0 < y < b, \tag{17.93}$$

subject to the initial and boundary conditions

$$\begin{aligned} u(t,0,y) = v(t,a,y) = 0 = v(t,x,0) = v(t,x,b), &\qquad 0 < x < a, \\ u(0,x,y) = f(x,y), \qquad u_t(0,x,y) = g(x,y), &\qquad 0 < y < b. \end{aligned} \tag{17.94}$$

As we saw in Section 17.2 the eigenvalues and eigenfunctions for the associated Helmholtz equation

$$c^2 (v_{xx} + v_{yy}) + \lambda\, v = 0, \qquad (x,y) \in R, \tag{17.95}$$

on a rectangle, subject to the homogeneous Dirichlet boundary conditions

$$v(0,y) = v(a,y) = 0 = v(x,0) = v(x,b), \qquad 0 < x < a, \qquad 0 < y < b, \tag{17.96}$$

are

$$v_{m,n}(x,y) = \sin \frac{m\pi x}{a}\ \sin \frac{n\pi y}{b}\,, \qquad \text{where} \qquad \lambda_{m,n} = \pi^2 c^2 \left(\tfrac{m^2}{a^2} + \tfrac{n^2}{b^2} \right), \tag{17.97}$$

with $m, n = 1, 2, \ldots$. The fundamental frequencies of vibration are the square roots of the eigenvalues, so

$$\omega_{m,n} = \sqrt{\lambda_{m,n}} = \pi\, c \sqrt{\frac{m^2}{a^2} + \frac{n^2}{b^2}}\ . \tag{17.98}$$

The frequencies will depend upon the underlying geometry — meaning the side lengths — of the rectangle, as well as the wave speed c, which is turn is a function of the membrane's density and stiffness, (17.84). The higher the wave speed, or the smaller the rectangle, the faster the vibrations. In layman's terms, (17.98) quantifies the observation that smaller, stiffer drums made of less dense material vibrate faster.

According to (17.85), the normal modes of vibration of our rectangle are

$$\begin{aligned} u_{m,n}(t,x,y) &= \cos \pi\, c \sqrt{\frac{m^2}{a^2} + \frac{n^2}{b^2}}\ t\ \sin \frac{m\pi x}{a}\ \sin \frac{n\pi y}{b}\,, \\ \widetilde{u}_{m,n}(t,x,y) &= \sin \pi\, c \sqrt{\frac{m^2}{a^2} + \frac{n^2}{b^2}}\ t\ \sin \frac{m\pi x}{a}\ \sin \frac{n\pi y}{b}\,. \end{aligned} \tag{17.99}$$

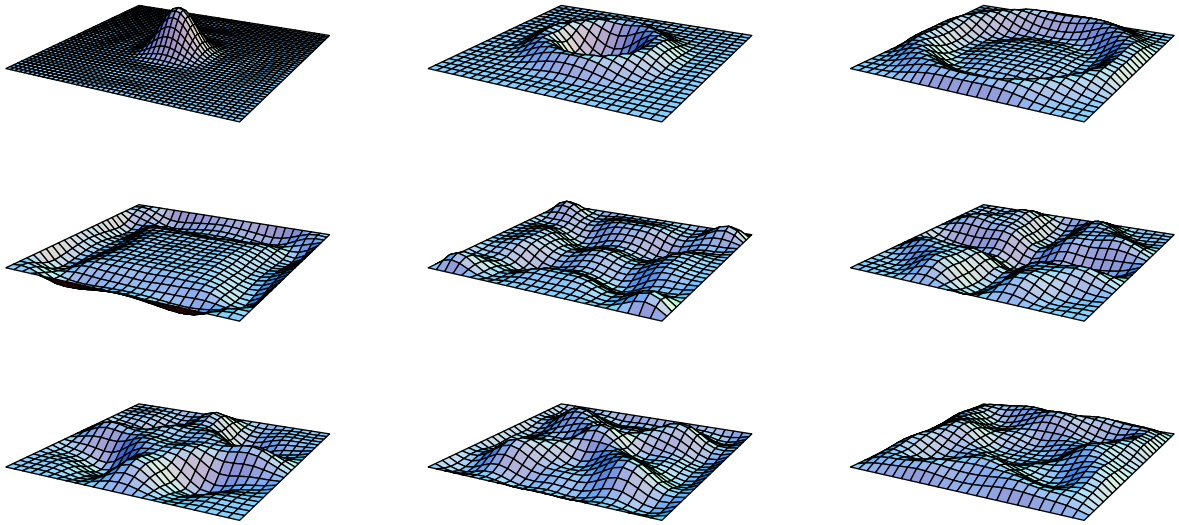


Figure 17.6. Vibrations of a Square.

The general solution can then be written as a double Fourier series

$$u(t,x,y) = \sum_{m,n=1}^{\infty} \big[\, a_{m,n}\, u_{m,n}(t,x,y) + b_{m,n}\, \widetilde{u}_{m,n}(t,x,y) \,\big].$$

in the normal modes. The coefficients $a_{m,n}, b_{m,n}$ are fixed by the initial displacement $u(0,x,y) = f(x,y)$ and the initial velocity $u_t(0,x,y) = g(x,y)$, as in (17.90). The usual orthogonality relations among the eigenfunctions imply

$$a_{m,n} = \frac{\langle\, v_{m,n}\,, f\,\rangle}{\|\, v_{m,n}\,\|^2} = \frac{4}{a\,b}\int_0^b\int_0^a f(x,y)\, \sin\frac{m\,\pi\,x}{a}\, \sin\frac{n\,\pi\,y}{b}\, dx\, dy, \tag{17.100}$$

$$b_{m,n} = \frac{\langle\, v_{m,n}\,, g\,\rangle}{\omega_{m,n}\,\|\, v_{m,n}\,\|^2} = \frac{4}{\pi\, c\,\sqrt{m^2\,b^2 + n^2\,a^2}}\int_0^b\int_0^a g(x,y)\, \sin\frac{m\,\pi\,x}{a}\, \sin\frac{n\,\pi\,y}{b}\, dx\, dy.$$

Since the fundamental frequencies are not rational multiples of each other, the general solution is a genuinely quasi-periodic superposition of the various normal modes.

In Figure 17.6, we plot the solution resulting from the initially concentrated displacement[†]

$$u(0,x,y) = f(x,y) = e^{-100[(x-.5)^2+(y-.5)^2]}$$

[†] The alert reader may object that the initial displacement $f(x,y)$ does not exactly satisfy the Dirichlet boundary conditions on the edges of the rectangle. But this does not prevent the existence of a well-defined solution to the initial value problem, whose initial boundary discontinuities will subsequently propagate inside the rectagle. However, they are so tiny as to be unnoticeable in the solution graphs.

at the center of a unit square, so $a = b = 1$ and the wave speed $c = 1$. The plots are at successive times $0, .02, .04, \dots, 1.6$. Note that, unlike a one-dimensional string where a concentrated displacement remains concentrated at all subsequent times and periodically repeats, the initial displacement spreads out in a radially symmetric manner and propagates to the edges of the rectangle, where it reflects and then interacts with itself. However, owing to the quasi-periodicity of the solution, the displacement of the drum never exactly repeats itself, and the initial concentrated signal never quite reforms in the rectangle's center.

Vibration of a Circular Drum

Let us next analyze the vibrations of a circular membrane. As always, we build up the solution as a quasi-periodic linear combination of the normal modes, which, by (17.85), are specified by the eigenfunctions for the associated Helmholtz boundary value problem.

As we saw in Section 17.2, the eigenfunctions of the Helmholtz equation on a disk of radius 1, say, subject to homogeneous Dirichlet boundary conditions, are products of trigonometric and Bessel functions:

$$\begin{aligned} v_{m,n}(r,\theta) &= J_m(\zeta_{m,n}\, r)\cos m\theta, & & m = 0, 1, 2, \dots, \\ \widetilde{v}_{m,n}(r,\theta) &= J_m(\zeta_{m,n}\, r)\sin m\theta, & & n = 1, 2, 3, \dots. \end{aligned} \tag{17.101}$$

Here r, θ are the usual polar coordinates, while $\zeta_{m,n} > 0$ denotes the n^{th} (positive) root of the m^{th} order Bessel function $J_m(z)$, cf. (17.58). The corresponding eigenvalue is its square, $\lambda_{m,n} = \zeta_{m,n}^2$, and hence the natural frequencies of vibration are equal to the Bessel roots, scaled by the wave speed:

$$\omega_{m,n} = c\,\sqrt{\lambda_{m,n}}\; = c\,\zeta_{m,n}. \tag{17.102}$$

A table of their values (for the case $c = 1$) can be found in the preceding section. The Bessel roots do not follow any easily discernible pattern, and are certainly not rational multiples of each other. Thus, the vibrations of a circular drum are also truly quasi-periodic.

The frequencies $\omega_{0,n} = c\,\zeta_{0,n}$ correspond to simple eigenvalues, with a single radially symmetric eigenfunction $J_0(\zeta_{0,n}\, r)$, while the "angular modes" $\omega_{m,n}$, for $m > 0$, are double, each possessing two linearly independent eigenfunctions (17.101). According to the general formula (17.85), each eigenfunction engenders two independent normal modes of vibration, having the explicit forms

$$\begin{aligned} &\cos c\,\zeta_{m,n}\, t\;\cos m\theta\;J_m(\zeta_{m,n}\, r), \qquad &&\cos c\,\zeta_{m,n}\, t\;\sin m\theta\;J_m(\zeta_{m,n}\, r), \\ &\sin c\,\zeta_{m,n}\, t\;\cos m\theta\;J_m(\zeta_{m,n}\, r), \qquad &&\sin c\,\zeta_{m,n}\, t\;\sin m\theta\;J_m(\zeta_{m,n}\, r). \end{aligned} \tag{17.103}$$

The general solution is written as a Fourier–Bessel series:

$$\begin{aligned} u(t,r,\theta) = \sum_{m,n} \Big[\big(a_{m,n}\cos c\,\zeta_{m,n}\, t + b_{m,n}\sin c\,\zeta_{m,n}\, t \big)\cos m\theta \\ + \big(c_{m,n}\cos c\,\zeta_{m,n}\, t + d_{m,n}\sin c\,\zeta_{m,n}\, t \big)\sin m\theta \Big]\, J_m(\zeta_{m,n}\, r), \end{aligned} \tag{17.104}$$

whose coefficients $a_{m,n}, b_{m,n}, c_{m,n}, d_{m,n}$ are determined, as usual, by the initial displacement and velocity of the membrane. In Figure vdisk■, the vibrations due to an initially

concentrated displacement are displayed. Again, the motion is only quasi-periodic and never quite returns to the original configuration.

Remark: As we learned in Section 14.4, the natural frequencies of vibration a (homogeneous) one-dimensional medium, e.g., a violin string or a column of air in a flute, are integer multiples of each other. This has the important consequence that any resulting vibration is periodic in time. Musically, the overtones in such a one-dimensional instrument are integer multiples of each other, and so the music sounds harmonic to our ear. On the other hand, the natural frequencies of circular and rectangular drums are irrationally related, and the vibrations are only quasi-periodic. As a result, we hear a percussive sound! Thus, for some reason, our appreciation of music is psychologically attuned to the differences between rationally related/periodic and irrationally related/quasi-periodic vibrations.

Scaling and Symmetry

Symmetry methods can be effectively employed in the analysis of the wave equation. Let us consider the simultaneous rescaling

$$t \longmapsto \alpha\, t, \qquad x \longmapsto \beta\, x, \qquad y \longmapsto \beta\, y, \tag{17.105}$$

of time and space, whose effect is to change the function $u(t, x, y)$ into a rescaled version

$$U(t, x, y) = u(\alpha\, t, \beta\, x, \beta\, y). \tag{17.106}$$

The chain rule is employed to relate their derivatives:

$$\frac{\partial^2 U}{\partial t^2} = \alpha^2\, \frac{\partial^2 u}{\partial t^2}\,, \qquad \frac{\partial^2 U}{\partial x^2} = \beta^2\, \frac{\partial^2 u}{\partial x^2}\,, \qquad \frac{\partial^2 U}{\partial y^2} = \beta^2\, \frac{\partial^2 u}{\partial y^2}\,.$$

Therefore, if u satisfies the wave equation

$$u_{tt} = c^2\, \Delta u,$$

then U satisfies the rescaled wave equation

$$U_{tt} = \frac{\alpha^2\, c^2}{\beta^2}\, \Delta U = \widetilde{c}^{\,2}\, \Delta U, \quad \text{where the rescaled wave speed is} \quad \widetilde{c} = \frac{\alpha\, c}{\beta}\,. \tag{17.107}$$

In particular, rescaling only time by setting $\alpha = 1/c$, $\beta = 1$, results in a unit wave speed $\widetilde{c} = 1$. In other words, we are free to choose our unit of time measurement so as to fix the wave speed equal to 1.

If we set $\alpha = \beta$, scaling space and time in the same proportion, then the wave speed does not change, $\widetilde{c} = c$, and so

$$t \longmapsto \beta\, t, \qquad x \longmapsto \beta\, x, \qquad y \longmapsto \beta\, y, \tag{17.108}$$

defines a *symmetry transformation* for the wave equation: If $u(t, x, y)$ is any solution to the wave equation, then so is its rescaled version

$$U(t, x, y) = u(\beta\, t, \beta\, x, \beta\, y) \tag{17.109}$$

for any choice of scale parameter $\beta \neq 0$. Observe that if $u(t,x,y)$ is defined on a domain Ω, then the rescaled solution $U(t,x,y)$ will be defined on the rescaled domain

$$\widetilde{\Omega} = \frac{1}{\beta}\,\Omega = \left\{ \left(\frac{x}{\beta}\,, \frac{y}{\beta} \right) \;\middle|\; (x,y) \in \Omega \right\} = \{\, (x,y) \mid\; (\beta x,\; \beta y) \in \Omega \,\}\,. \tag{17.110}$$

For instance, the scaling parameter $\beta = 2$ has the effect of halving the size of the domain. The normal modes for the rescaled domain have the form

$$U_n(t,x,y) = u_n(\beta t, \beta x, \beta y) = \cos(\beta\,\omega_n\, t)\; v_n(\beta x, \beta y),$$
$$\widetilde{U}_n(t,x,y) = \widetilde{u}_n(\beta t, \beta x, \beta y) = \sin(\beta\,\omega_n\, t)\; v_n(\beta x, \beta y),$$

and hence the vibrational frequencies $\widetilde{\omega}_n = \beta\,\omega_n$ are scaled by the same overall factor. Thus, when $\beta < 1$, the rescaled membrane is larger by a factor $1/\beta$, and its vibrations are slowed down by the same factor β. For instance, a drum that is twice as large will vibrate twice as slowly, and hence have an octave lower overall tone. Musically, this means that all drums of a similar shape have the same pattern of overtones, differing only in their overall pitch, which is a function of their size, tautness and density.

In particular, choosing $\beta = 1/R$ will rescale the unit disk into a disk of radius R. The fundamental frequencies of the rescaled disk are

$$\widetilde{\omega}_{m,n} = \beta\,\omega_{m,n} = \frac{c}{R}\,\zeta_{m,n}, \tag{17.111}$$

where c is the wave speed and $\zeta_{m,n}$ are the Bessel roots, defined in (17.58). Observe that the ratios $\omega_{m,n}/\omega_{m',n'}$ between vibrational frequencies remain the same, independent of the size of the disk R and the wave speed c. We define the *relative vibrational frequencies*

$$\rho_{m,n} = \frac{\omega_{m,n}}{\omega_{0,1}} = \frac{\zeta_{m,n}}{\zeta_{0,1}}\,, \qquad \text{in proportion to} \qquad \omega_{0,1} = \frac{c\,\zeta_{0,1}}{R} \approx 2.4\,\frac{c}{R}\,, \tag{17.112}$$

which is the drum's dominant or lowest vibrational frequency. The relative frequencies $\rho_{m,n}$ are independent of the size, stiffness or composition of the drum membrane. In the following table, we display a list of all relative vibrational frequencies (17.112) that are < 6. Once the lowest frequency $\omega_{0,1}$ has been determined — either theoretically, numerically or experimentally — all the higher overtones $\omega_{m,n} = \rho_{m,n}\,\omega_{0,1}$ are obtained by rescaling.

Relative Vibrational Frequencies of a Circular Disk

n \ m	0	1	2	3	4	5	6	7	8	9	...
1	1.000	1.593	2.136	2.653	3.155	3.647	4.132	4.610	5.084	5.553	...
2	2.295	2.917	3.500	4.059	4.601	5.131	5.651	⋮	⋮	⋮	
3	3.598	4.230	4.832	5.412	5.977	⋮	⋮				
4	4.903	5.540	⋮	⋮	⋮						
⋮	⋮	⋮									

17.6. Nodal Curves.

When a membrane vibrates, the individual points move up and down in a quasi-periodic manner. As such, correlations between the motions of different points are not immediately evident. However, if the membrane is set to vibrate in a pure eigenmode, say

$$u_n(t, x, y) = \cos(\omega_n t)\, v_n(x, y),$$

then all points move up and down at a common frequency $\omega_n = \sqrt{\lambda_n}$, which is the square root of the eigenvalue corresponding to the eigenfunction $v_n(x, y)$. The exceptions are the points where the eigenfunction vanishes:

$$v_n(x, y) = 0. \tag{17.113}$$

Such points remain stationary. The set of all points $(x, y) \in \Omega$ that satisfy (17.113) is known as the n^{th} *nodal set* of the domain Ω. Scattering small particles (e.g., fine sand) over the membrane performing such a pure vibration, will enable us to see the nodal set, because the particles will, though random movement over the oscillating regions of the membrane, tend to accumulate along the stationary nodal curves.

It can be shown that, in general, each nodal set consists of a finite system of *nodal curves*. The nodal curves intersect at critical points of the eigenfunction, where $\nabla v_n = \mathbf{0}$, and thereby partition the membrane into *nodal regions*. Points lying in a common nodal region all vibrate in tandem, so that all points in a common nodal region are either up or down, except, momentarily, when the *entire* membrane has zero displacement. Adjacent nodal regions, lying on the opposite sides of a nodal curve, vibrate in opposing directions — when one side is up, the other is down, and then, as the membrane becomes momentarily flat, simultaneously switch directions.

Example 17.5. *Circular Drums.* Since the eigenfunctions (17.101) for a disk are products of trigonometric functions in the angular variable and Bessel functions of the radius, the nodal curves for the normal modes of vibrations of a circular membrane are rays emanating from and circles centered at the origin. Thus, the nodal regions are annular sectors. Pictures of the nodal curves for the first nine normal modes indexed by their relative frequencies are plotted in Figure 17.7. Representative displacements of the membrane in each of the first twelve modes can be found in Figure 17.3. The dominant (lowest frequency) mode is the only one that has no nodal curves; it has the form of a radially symmetric bump where the entire membrane flexes up and down. Every other mode has at least one nodal curve. For instance, the next lowest modes vibrate proportionally faster at a relative frequency $\rho_{1,1} \approx 1.593$. The most general solution with this vibrational frequency is a linear combination $\alpha\, u_{1,1} + \beta\, \widetilde{u}_{1,1}$ of the two eigensolutions. Each combination has a single diameter as a nodal curve, whose slope depends upon the coefficients α, β. Here, the two semicircular halves of the drum vibrate in opposing directions — when the top half is up, the bottom half is down and vice versa. The next set of modes have two perpendicular diameters as nodal curves; the four quadrants of the drum vibrate in tandem, with opposite quadrants having the same displacements. Next in increasing order of vibrational frequency is a single mode, with a circular nodal curve whose (relative) radius $\zeta_{0,2}/\zeta_{0,1} \approx .6276$ equals the ratio of the first two roots of the order zero Bessel function;

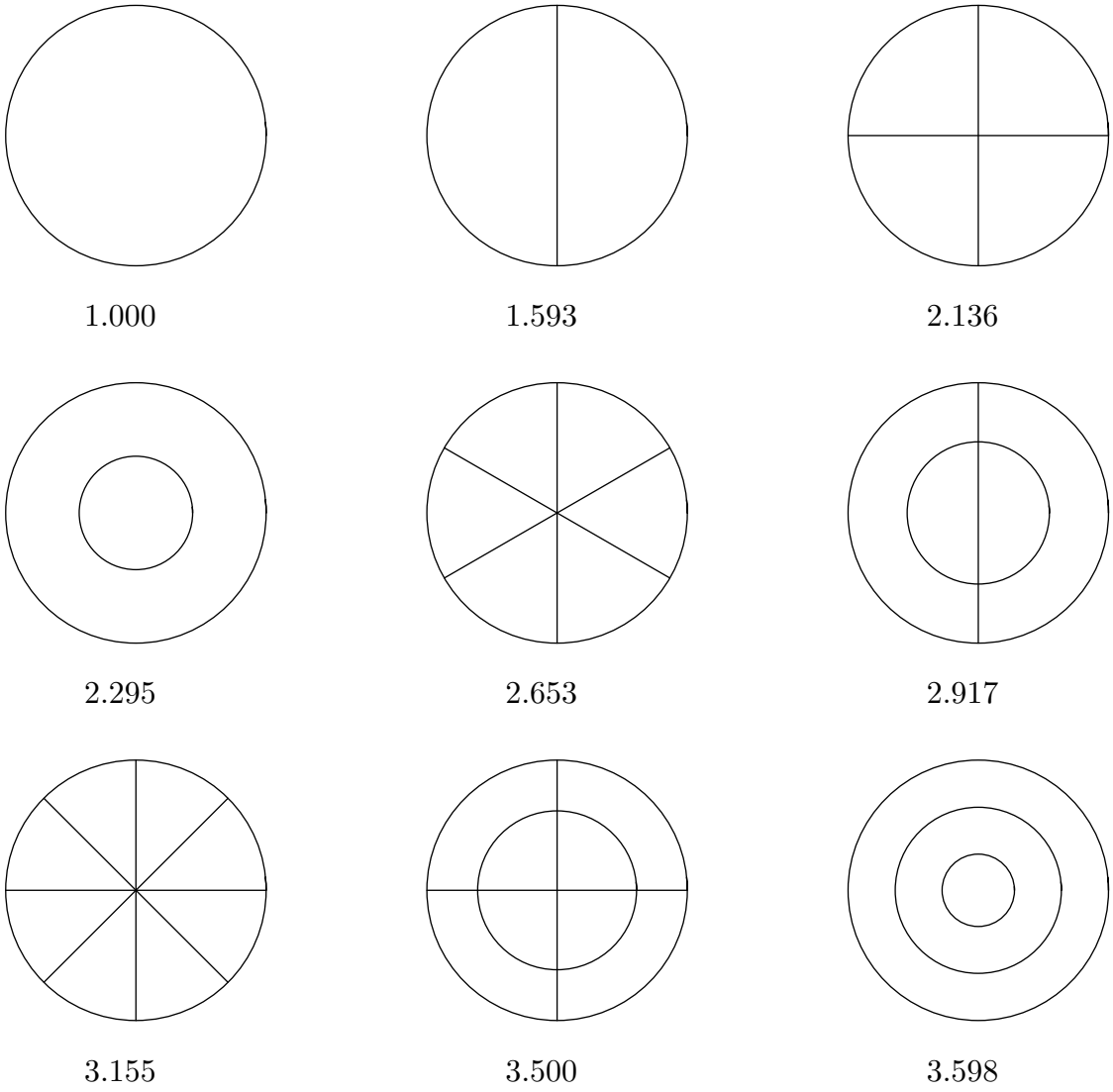


Figure 17.7. Nodal Curves and Relative Vibrational Frequencies of a Circular Membrane.

see Exercise ■ for a justification. In this case, the inner disk and the outer annulus vibrate in opposing directions.

Example 17.6. *Rectangular Drums.* For a general rectangular drum, the nodal curves are relatively uninteresting. Since the normal modes (17.99) are separable products of trigonometric functions in the coordinate variables x, y, the nodal curves are regularly equi-spaced straight lines parallel to the sides of the rectangle. The internodal regions are small rectangles, all of the same size and shape, with adjacent rectangles vibrating in opposite directions.

A more interesting collection of nodal curves occurs when the rectangle admits multiple eigenvalues — so-called *accidental degeneracies*. Two of the eigenvalues (17.97) co-

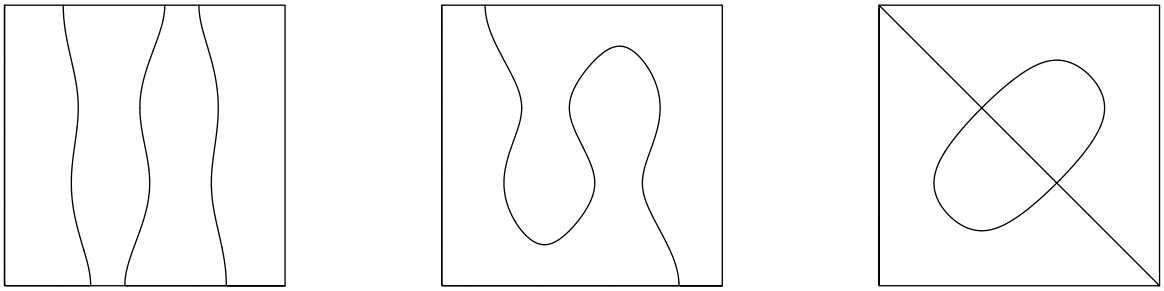


Figure 17.8. Some Nodal Curves for a Square Membrane.

incide, $\lambda_{m,n} = \lambda_{k,l}$, if and only if

$$\frac{m^2}{a^2} + \frac{n^2}{b^2} = \frac{k^2}{a^2} + \frac{l^2}{b^2} \tag{17.114}$$

where $(m,n) \neq (k,l)$ are distinct pairs of positive integers. In such situations, the distinct eigenmodes happen to vibrate with a common frequency $\omega = \omega_{m,n} = \omega_{k,l}$. Consequently, any linear combination of the eigenmodes, e.g.,

$$\cos \omega\, t \left(\alpha \sin \frac{m\,\pi\, x}{a} \, \sin \frac{n\,\pi\, y}{b} + \beta \sin \tfrac{k\,\pi\, x}{a} \, \sin \tfrac{l\,\pi\, y}{b} \right), \qquad \alpha, \beta \in \mathbb{R},$$

is a pure vibration, and hence also qualifies as a normal mode. The associated nodal curves

$$\alpha \sin \frac{m\,\pi\, x}{a} \, \sin \frac{n\,\pi\, y}{b} + \beta \sin \tfrac{k\,\pi\, x}{a} \, \sin \tfrac{l\,\pi\, y}{b} = 0 \tag{17.115}$$

have a more intriguing geometry, which can change dramatically as α, β vary.

For example, on a unit square $R = \big\{\, 0 < x, y < 1 \,\big\}$, an accidental degeneracy occurs whenever

$$m^2 + n^2 = k^2 + l^2 \tag{17.116}$$

for distinct pairs of positive integers $(m,n) \neq (k,l)$. The simplest possibility arises whenever $m \neq n$, in which case we can merely reverse the order, setting $k = n, \;\; l = m$. In Figure 17.8 we plot three sample nodal curves

$$\sin 4\,\pi\, x \, \sin \pi\, y + \beta \sin \pi\, x \, \sin 4\,\pi\, y = 0, \qquad \text{with} \qquad \beta = .2, .5, 1,$$

corresponding to the three different linear combinations of the eigenfunctions with $m = l = 4$, $n = k = 1$. The associated vibrational frequency is $\omega_{4,1} = \pi\, c\, \sqrt{17}\,$.

Remark: Classifying such accidental degeneracies takes us into the realm of number theory, [**10**, **39**]. The simplest (square) case (17.116) asks one to determine all integer points that lie on a common circle.

Remark: An interesting question is whether a circular disk has accidental degeneracies, which would occur if two different Bessel roots were to coincide. However, it is known, [**180**; p. 129], that $\zeta_{m,n} \neq \zeta_{k,l}$ whenever $(m,n) \neq (k,l)$. Thus, a disk has no such degeneracies, and all nodal curves are circles and rays around the origin.

Chapter 18

Partial Differential Equations in Space

At last we have climbed the dimensional ladder to its ultimate rung (at least for those of us living in a three-dimensional universe): partial differential equations in physical space. As in the one and two-dimensional contexts developed in the preceding chapters, the three key examples are the three-dimensional Laplace equation, modeling equilibrium configurations of solid bodies, the three-dimensional wave equation, governing vibrations of solids, liquids, gasses, and electromagnetic waves, and the three-dimensional heat equation, modeling basic spatial diffusion processes.

Fortunately, almost everything of importance has already appeared in the one- and two-dimensional situations, and appending a third dimension is, for the most part, simply a matter of appropriately adapting the constructions. We have already seen the basic underlying solution techniques: separation of variables and Green's functions or fundamental solutions. (Unfortunately, despite the best efforts of mathematicians, the most powerful of our planar tools, conformal mapping, does *not* carry over to higher dimensions.) In three-dimensional problems, separation of variables is applicable in rectangular, cylindrical and spherical coordinates. The first two do not produce anything fundamentally new, and are therefore relegated to the exercises. Separation in spherical coordinates leads to spherical harmonics and spherical Bessel functions, whose properties are investigated in some detail. These new special functions play important roles in a number of physical systems, including the quantum theory of atomic structure that underlies the spectral and chemical properties of atoms.

The Green's function for the three-dimensional Poisson equation in space can be identified as the classic Newtonian (and Coulomb) $1/r$ potential. The fundamental solution for the three-dimensional heat equation can be easily guessed from its one- and two-dimensional versions. The three-dimensional wave equation, surprisingly, has an explicit, although more intricate, solution formula of d'Alembert form, due to Poisson. Paradoxically, the best way to treat the two-dimensional version is by "descending" from the simpler three-dimensional formula. This result highlights a remarkable difference between waves in planar and spacial media. In three-dimensions, Huygens' principle states that waves emanating from a localized initial disturbance remain localized as they propagate through space. In contrast, in two dimensions, initially concentrated pulses leave a slowly decaying remnant that never entirely disappears.

18.1. The Laplace and Poisson Equations.

We begin our investigations, as usual, with systems in equilibrium, deferring dynamics

until later. The prototypical equilibrium system is the three-dimensional Laplace equation

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0, \tag{18.1}$$

in which $\mathbf{x} = (\, x, y, z \,)^T$ represents rectangular coordinates on $\mathbb{R}^3$. The solutions $u(x, y, z)$ continue to be known as *harmonic functions*. The Laplace equation models unforced equilibria; *Poisson's equation* is the inhomogeneous version

$$-\Delta u = f(x, y, z), \tag{18.2}$$

where the inhomogeneity f represents some form of external forcing.

The basic boundary value problem for the Laplace or the Poisson equation seeks a solution inside a bounded domain $\Omega \subset \mathbb{R}^3$ subject to either *Dirichlet boundary conditions*, prescribing the function values

$$u = h \qquad \text{on} \qquad \partial\Omega, \tag{18.3}$$

or *Neumann boundary conditions* prescribing its normal derivative or flux

$$\frac{\partial u}{\partial \mathbf{n}} = k \qquad \text{on} \qquad \partial\Omega, \tag{18.4}$$

or *mixed boundary conditions* in which one imposes Dirichlet conditions on part of the boundary and Neumann conditions on the remainder. Keep in mind that the boundary of the solid domain Ω consists of one or more piecewise smooth closed surfaces, which will be oriented by the outwards unit normal $\mathbf{n}$;

The boundary value problems for the three-dimensional Laplace and Poisson equations govern a wide variety of physical systems, including:

(a) *Heat conduction*: In this application, u represents the equilibrium temperature in a solid body. Dirichlet conditions correspond to fixing the temperature on the bounding surface(s), whereas homogeneous Neumann conditions correspond to an insulated boundary, i.e., one which does not allow any heat flux. The inhomogeneity f represents some form of internal heat source.

(b) *Ideal fluid flow*: Here u represents the velocity potential for an incompressible, irrotational steady state fluid flow inside a container, Ω, with velocity vector field $\mathbf{v} = \nabla u$. Homogeneous Neumann boundary conditions correspond to a solid boundary which the fluid cannot penetrate.

(c) *Elasticity*: In certain restricted situations, u represents an equilibrium deformation of a solid body, e.g., the radial deformation of a solid ball. Fully three-dimensional elasticity is governed by a more complicated system of partial differential equations, which can be found in Example 21.9.

(d) *Electrostatics*: In applications to electromagnetism, u represents the electric potential in a conducting medium; its gradient ∇u prescribes the electromotive force on a charged particle. The inhomogeneity represents an electostatic force field.

(e) *Gravitation*: The Newtonian gravitational potential in flat empty space is also prescribed by the Laplace equation. (In contrast, general relativity requires a vastly more complicated nonlinear system of partial differential equations, [**129**].)

Self–Adjoint Formulation and Minimum Principle

The Laplace and Poisson equations naturally fit into the general self-adjoint equilibrium framework summarized in Section 14.7. The construction is a straightforward adaptation of the planar version of Section 15.4. We introduce the L^2 inner products

$$\begin{aligned} \langle\, u\,, \widetilde{u}\,\rangle &= \iiint_\Omega u(x,y,z)\, \widetilde{u}(x,y,z)\, dx\, dy\, dz, \\ \langle\, \mathbf{v}\,, \widetilde{\mathbf{v}}\,\rangle &= \iiint_\Omega \mathbf{v}(x,y,z)\cdot \widetilde{\mathbf{v}}(x,y,z)\, dx\, dy\, dz, \end{aligned} \tag{18.5}$$

between scalar fields $u, \widetilde{u}$, and between vector fields $\mathbf{v}, \widetilde{\mathbf{v}}$ defined on the domain $\Omega \subset \mathbb{R}^3$. We assume that the functions in question are sufficiently nice that these inner products are well-defined; if Ω is unbounded, this requires that, at large distances, they decay reasonably rapidly to zero.

When subject to suitable homogeneous boundary conditions, the three-dimensional Laplace equation can be placed in our standard self-adjoint form

$$-\Delta u = -\nabla\cdot\nabla u = \nabla^* \circ \nabla u. \tag{18.6}$$

This relies on the fact that the adjoint of the gradient operator with respect to the L^2 inner products (18.5) is minus the divergence operator:

$$\nabla^* \mathbf{v} = -\nabla\cdot\mathbf{v}. \tag{18.7}$$

As usual, the determination of the adjoint rests on an integration by parts formula, which, in three-dimensional space, follows from the Divergence Theorem B.36. The first step is to establish the three-dimensional analog of Green's formula (15.87). We apply the divergence identity (B.82) to the product $u\,\mathbf{v}$ of a scalar field u and a vector field $\mathbf{v}$, leading to

$$\iiint_\Omega \big(\, u\,\nabla\cdot\mathbf{v} + \nabla u\cdot\mathbf{v}\,\big)\, dx\, dy\, dz = \iiint_\Omega \nabla\cdot(u\,\mathbf{v})\, dx\, dy\, dz = \iint_{\partial\Omega} u\,(\mathbf{v}\cdot\mathbf{n})\, dS. \tag{18.8}$$

Rearranging the terms produces the desired integration by parts formula for triple integrals:

$$\iiint_\Omega (\nabla u\cdot\mathbf{v})\, dx\, dy\, dz = \iint_{\partial\Omega} u\,(\mathbf{v}\cdot\mathbf{n})\, dS - \iiint_\Omega u\,(\nabla\cdot\mathbf{v})\, dx\, dy\, dz. \tag{18.9}$$

The boundary integral will vanish provided either $u = 0$ or $\mathbf{v}\cdot\mathbf{n} = \mathbf{0}$ at each point on $\partial\Omega$. When $u = 0$ on all of $\partial\Omega$, we have homogeneous Dirichlet conditions. Setting $\mathbf{v}\cdot\mathbf{n} = \mathbf{0}$ everywhere on $\partial\Omega$ results in the homogeneous Neumann boundary value problem; see Section 15.4 for a detailed explanation. Finally, when $u = 0$ on part of $\partial\Omega$ and $\mathbf{v}\cdot\mathbf{n} = \mathbf{0}$ on the rest leads to the mixed boundary value problem. Thus, subject to one of these choices, the integration by parts formula (18.9) reduces to

$$\langle\, \nabla u\,, \mathbf{v}\,\rangle = \langle\, u\,, -\nabla\cdot\mathbf{v}\,\rangle, \tag{18.10}$$

which suffices to prove the adjoint formula (18.7).

Remark: Adopting more general weighted inner products results in a more general elliptic boundary value problem. See Exercise ■ for details.

According to the abstract Theorem 7.60, the self-adjoint formulation (18.6) implies positive semi-definiteness of the boundary value problem, and positive definiteness provided $\ker \nabla = \{0\}$. Since, on a connected domain, only constant functions are annihilated by the gradient operator — see Theorem B.28 — both the Dirichlet and mixed boundary value problems are positive definite, while the Neumann boundary value problem is only semi-definite.

Finally, in the positive definite cases, the solution can be characterized by the three-dimensional version of the Dirichlet minimization principle (15.100).

Theorem 18.1. *The solution $u(x, y, z)$ to the Poisson equation* (18.2) *subject to homogeneous Dirichlet or mixed boundary conditions* (18.3) *is characterized as the unique function that minimizes the* Dirichlet integral

$$\tfrac{1}{2} \| \nabla u \|^2 - \langle u , f \rangle = \iiint_\Omega \big[\tfrac{1}{2} (u_x^2 + u_y^2 + u_z^2) - f u \big] \, dx \, dy \, dz \tag{18.11}$$

among all C^1 *functions that satisfy the prescribed boundary conditions.*

As in the two-dimensional version discussed in Section 15.4, he minimization principle continues to hold without modification in the case of the inhomogeneous Dirichlet boundary value problem. Modifications for the inhomogeneous mixed boundary value problem are discussed in Exercise ■. The three-dimensional finite element method for constructing numerical solutions to such boundary value problems rests on the associated minimization principle; see [**148**, **107**] for details.

18.2. Separation of Variables.

With conformal mapping no longer a viable option in three dimensional space, separation of variables reasserts its primacy for generating explicit solutions to the Laplace equation. As always, its applicability is unfortunately restricted to rather special, but important, geometrical configurations. In three-dimensional space, the simplest separable cases are problems formulated on rectangular, cylindrical or spherical domains. Since the first two are straightforward extensions of their two-dimensional counterparts, we will only discuss spherically separable solutions in any detail. The simplest domain to which the separation of variables method applies is a rectangular box:

$$B = \big\{ 0 < x < a, \ 0 < y < b, \ 0 < z < c \big\}.$$

For functions of three variables, one begins the separation process by splitting off one of them, by setting $u(x, y, z) = v(x) \, w(y, z)$, say. The function $v(x)$ satisfies a simple second order ordinary differential equation, while $w(y, z)$ solves the two-dimensional Helmholtz equation, which is then separated by writing $w(y, z) = p(y) \, q(z)$. The resulting fully separated solutions $u(x, y, z) = v(x) \, p(y) \, q(z)$ are (mostly) products of trigonometric and hyperbolic functions. Complete details of the technique and the resulting series solution are relegated to Exercise ■.

In the case when the domain is a cylinder, one passes to cylindrical coordinates r, θ, z to effect the separation. The resulting separable solutions $u(r, \theta, z) = v(r, \theta), w(z) = p(r)\, q(\theta), w(z)$ are products of Bessel functions of the cylindrical radius r, trigonometric functions of the polar angle θ, and hyperbolic functions of z. Details are outlined in Exercise ■.

The most interesting case is that of spherical coordinates, which we proceed to analyze in detail in the following subsection.

Remark: Beyond these three well-known cases, there are, in fact, a total of eleven different coordinate systems in which the three-dimensional Laplace equation separates. See [**128**, **131**, **133**] for details on the more exotic types of separation, including ellipsoidal, toroidal, and parabolic spheroidal coordinates. The resulting separable solutions lead to new classes of special functions.

Laplace's Equation in a Ball

Suppose a solid ball (e.g., the earth), is subject a specified steady temperature distribution on its spherical boundary. Our task is to determine the equilibrium temperature within the ball. To simplify matters, we assume that the body is composed of an isotropic, homogeneous medium, and shall choose units in which its radius equals 1. Then, to find the equilibrium temperature within the ball, we must solve the Dirichlet boundary value problem

$$\begin{aligned} &\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0, && x^2 + y^2 + z^2 < 1, \\ &u(x, y, z) = h(x, y, z), && x^2 + y^2 + z^2 = 1. \end{aligned} \tag{18.12}$$

Problems in spherical geometries tend to simplify when re-expressed in terms of spherical coordinates r, φ, θ, as defined by the usual formulae

$$x = r \sin\varphi\, \cos\theta, \qquad y = r \sin\varphi\, \sin\theta, \qquad z = r \cos\varphi. \tag{18.13}$$

Here $0 \le \theta < 2\,\pi$ measures the *azimuthal angle* or *longitude*, while $0 \le \varphi \le \pi$ measures the *zenith angle* or *latitude*.

Warning: We use the mathematician's convention for spherical coordinates. Physicists often interchange the notation for the azimuthal and zenith angles; see Example B.8 for a detailed discussion.

In spherical coordinates, the Laplace equation for[†] $u(r, \varphi, \theta)$ takes the form

$$\Delta u = \frac{\partial^2 u}{\partial r^2} + \frac{2}{r}\,\frac{\partial u}{\partial r} + \frac{1}{r^2}\,\frac{\partial^2 u}{\partial \varphi^2} + \frac{\cos\varphi}{r^2 \sin\varphi}\,\frac{\partial u}{\partial \varphi} + \frac{1}{r^2 \sin^2\varphi}\,\frac{\partial^2 u}{\partial \theta^2} \;=\; 0. \tag{18.14}$$

This important formula is the final result of a fairly nasty chain rule computation, whose messy details are left to the motivated reader. (Set aside lots of paper and keep an eraser handy!)

[†] *Warning*: See Section 15.2 for our convention on rewriting functions in new coordinates.

To construct separable solutions to the spherical coordinate form (18.14) of the Laplace equation, we begin by separating off the radial part of the solution, setting

$$u(r, \varphi, \theta) = v(r)\, w(\varphi, \theta). \tag{18.15}$$

Substituting this ansatz into (18.14), multiplying the resulting equation through by $\dfrac{r^2}{v\, w}$, and then placing all the terms involving r on one side yields

$$\frac{1}{v}\left(r^2\, \frac{d^2 v}{dr^2} + 2r\, \frac{dv}{dr} \right) = -\,\frac{1}{w}\, \Delta_S w = \mu, \tag{18.16}$$

where μ is the separation constant, and

$$\Delta_S w = \frac{\partial^2 w}{\partial \varphi^2} + \frac{\cos\varphi}{\sin\varphi}\, \frac{\partial w}{\partial \varphi} + \frac{1}{\sin^2\varphi}\, \frac{\partial^2 w}{\partial \theta^2}\ . \tag{18.17}$$

The second order differential operator Δ_S, which contains only the angular components of the full Laplacian operator Δ, is of particular significance. It is known as the *spherical Laplacian*, and governs the equilibrium and dynamics of thin spherical shells, cf. Example 18.13.

Returning to equation (18.16), our usual separation argument applies. The left hand side depends only on r, while the right hand side depends only on the angles φ, θ. This can only occur when both sides are equal to a common separation constant, denoted by μ. As a consequence, the radial component $v(r)$ satisfies the ordinary differential equation

$$r^2\, v'' + 2r\, v' - \mu\, v = 0, \tag{18.18}$$

which is of Euler type (7.51), and hence can be readily solved. We will put this equation aside for the time being, and concentrate our efforts on the more complicated part.

The angular components in (18.16) assume the form

$$\Delta_S[w] + \mu\, w = \frac{\partial^2 w}{\partial \varphi^2} + \frac{\cos\varphi}{\sin\varphi}\, \frac{\partial w}{\partial \varphi} + \frac{1}{\sin^2\varphi}\, \frac{\partial^2 w}{\partial \theta^2} + \mu\, w = 0. \tag{18.19}$$

This second order partial differential equation can be regarded as the eigenvalue equation for the spherical Laplacian operator Δ_S, and is known as the *spherical Helmholtz equation*. To solve it, we adopt a further separation of angular variables,

$$w(\varphi, \theta) = p(\varphi)\, q(\theta), \tag{18.20}$$

which we substitute into (18.19). Dividing the result by the product $w = p\, q$, multiplying by $\sin^2\varphi$, and then rearranging terms, we are led to the separated system

$$\frac{\sin^2\varphi}{p}\, \frac{d^2 p}{d\varphi^2} + \frac{\cos\varphi \sin\varphi}{p}\, \frac{dp}{d\varphi} + \mu \sin^2\varphi = -\,\frac{1}{q}\, \frac{d^2 q}{d\theta^2} = \nu.$$

The left hand side depends only on the zenith coordinate φ while the right hand side depends only on the azimuthal coordinate θ. Since these angles are independent, the only way this could hold is when the two sides equal a common separation constant, denoted

by ν. The spherical Helmholtz equation thereby splits into a pair of ordinary differential equations

$$\sin^2\varphi\,\frac{d^2p}{d\varphi^2} + \cos\varphi\,\sin\varphi\,\frac{dp}{d\varphi} + (\mu\sin^2\varphi - \nu)\,p = 0, \qquad \frac{d^2q}{d\theta^2} + \nu\,q = 0.$$

The equation for $q(\theta)$ is easy to solve. As one circumnavigates the sphere from west to east, the azimuthal angle θ increases from 0 to 2π, so $q(\theta)$ must be a 2π periodic function. Thus, $q(\theta)$ satisfies the well-studied periodic boundary value problem treated, for instance, in (15.33). Up to constant multiple, non-zero periodic solutions occur only when the separation constant assumes one of the values $\nu = m^2$, where $m = 0, 1, 2, \dots$ is an integer, with

$$q(\theta) = \cos m\,\theta \qquad \text{or} \qquad \sin m\,\theta, \qquad\qquad m = 0, 1, 2, \dots. \tag{18.21}$$

Each positive $\nu = m^2 > 0$ admits two linearly independent 2π periodic solutions, while when $\nu = 0$, only the constant solutions are periodic.

With this information, we endeavor to solve the zenith equation

$$\sin^2\varphi\,\frac{d^2p}{d\varphi^2} + \cos\varphi\,\sin\varphi\,\frac{dp}{d\varphi} + (\mu\sin^2\varphi - m^2)\,p = 0. \tag{18.22}$$

This is not easy, and constructing analytic formulas for its solutions requires some effort. The motivation behind the following steps will not be immediately apparent to the reader, since they are the result of a long, detailed study of this important differential equation by mathematicians over the last 200 years.

As an initial simplification, we will eliminate the trigonometric functions. To this end, we invoke the change of variables

$$t = \cos\varphi, \qquad \text{with} \qquad p(\varphi) = P(\cos\varphi) = P(t). \tag{18.23}$$

Since

$$0 \le \varphi \le \pi, \qquad \text{we have} \qquad 0 \le \sqrt{1-t^2}\, = \sin\varphi \le 1.$$

According to the chain rule,

$$\begin{aligned}
\frac{dp}{d\varphi} &= -\sin\varphi\,\frac{dP}{dt} = -\sqrt{1-t^2}\,\frac{dP}{dt}\,,\\
\frac{d^2p}{d\varphi^2} &= \sin^2\varphi\,\frac{d^2P}{dt^2} - \cos\varphi\,\frac{dP}{dt} = (1-t^2)\,\frac{d^2P}{dt^2} - t\,\frac{dP}{dt}\,.
\end{aligned}$$

Substituting these expressions into (18.22), we conclude that $P(t)$ must satisfy

$$(1-t^2)^2\,\frac{d^2P}{dt^2} - 2t\,(1-t^2)\,\frac{dP}{dt} + \big[\,\mu\,(1-t^2) - m^2\,\big]\,P = 0. \tag{18.24}$$

Unfortunately, the resulting differential equation is still not so easy to solve, but at least its coefficients are polynomials. Equation (18.24) is known as the *Legendre differential equation* of order m, and its solutions are known as *Legendre functions*, having first been employed by Legendre to study the gravitational attraction of ellipsoidal bodies.

Power series solutions to the Legendre equation can be constructed by the standard techniques presented in Appendix C. The most general solution is a new type of special function, known as a *Legendre function*, [**3**, **58**]. However, the solutions we are actually interested in can all be written in terms of elementary algebraic functions. First of all, since $t = \cos\varphi$, the solution only needs to be defined on the interval $-1 \le t \le 1$. The endpoints of this interval, $t = \pm 1$, correspond to the sphere's north pole, $\varphi = 0$, and south pole, $\varphi = \pi$. Both endpoints are singular points for the Legendre equation since the coefficient $(1 - t^2)^2$ of the leading order derivative vanishes when $t = \pm 1$. In fact, both are regular singular points, as shown in Exercise ∎. Since ultimately we need the separable solution (18.15) to be a well-defined function of x, y, z (even at points where the spherical coordinates degenerate, i.e., on the z axis), we need $p(\varphi)$ to be well-defined at $\varphi = 0$ and π, and this requires $P(t)$ to be bounded at the singular points:

$$| P(-1) | < \infty, \qquad | P(+1) | < \infty. \tag{18.25}$$

The combined boundary value problem (18.24–25) takes the form of an eigensystem, in which the separation constant μ is the eigenvalue and the non-zero solutions $P(t) \not\equiv 0$ are the associated eigenfunctions.

Mathematical justifications of the following statements can be found in Appendix C. Consider first the case $m = 0$, which assumes the simpler form

$$(1 - t^2)\, \frac{d^2P}{dt^2} - 2t\, \frac{dP}{dt} + \mu\, P = 0. \tag{18.26}$$

In this case, it turns out that the eigenfunctions, i.e., solutions to the Legendre boundary value problem (18.26, 25), are the *Legendre polynomials*

$$P_n(t) = \frac{1}{2^n\, n!}\, \frac{d^n}{dt^n}\, (t^2 - 1)^n \tag{18.27}$$

that first arose in Chapter 5 as our simplest example of orthogonal polynomials. Indeed, we can now finally comprehend the reason for the orthogonality of the Legendre polynomials: they are the common eigenfunctions of a self-adjoint boundary value problem! Explicit formulas for the first few Legendre polynomials appear in (5.46).

When $m > 0$, the eigenfunctions of the Legendre boundary value problem (18.24–25) are not always polynomials. They are known as the *associated Legendre functions*, and have the explicit formula

$$\begin{aligned} P_n^m(t) &= (-1)^m (1 - t^2)^{m/2}\, \frac{d^m}{dt^m}\, P_n(t) \\ &= (-1)^{m+n} \frac{(1 - t^2)^{m/2}}{2^n\, n!}\, \frac{d^{n+m}}{dt^{n+m}}\, (1 - t^2)^n, \end{aligned} \qquad n = m, m+1, \dots, \tag{18.28}$$

which generalizes the Rodrigues formula (5.48) for the classical Legendre polynomials. Its proof is similar, and done in Exercise ∎. Here is a list of the first few Legendre polynomials and associated Legendre functions:

$$P_0^0(t) = 1, \qquad P_1^0(t) = t, \qquad P_1^1(t) = -\sqrt{1 - t^2}\,,$$

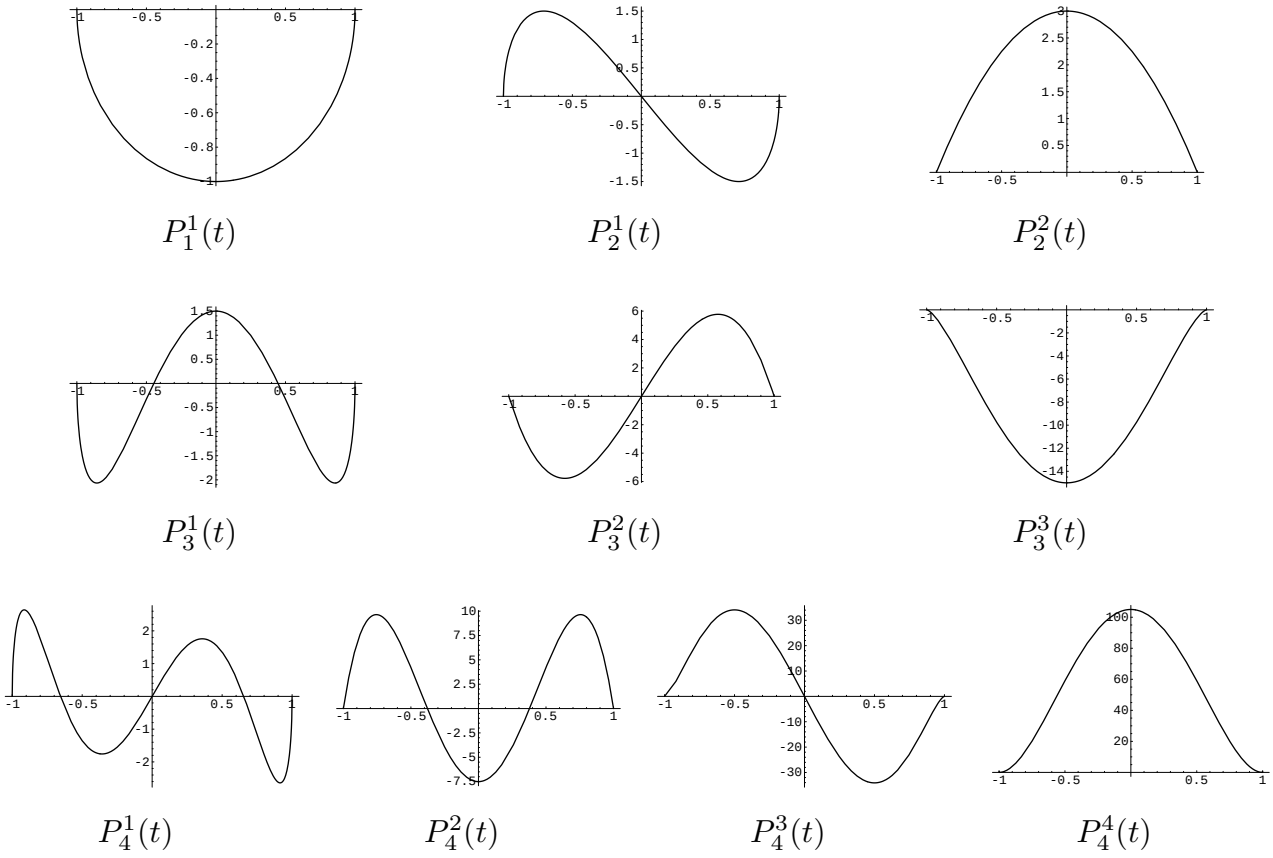


Figure 18.1. Legendre Functions.

$$
\begin{gathered}
P_2^0(t) = -\tfrac{1}{2} + \tfrac{3}{2}\,t^2, \qquad P_2^1(t) = -3t\,\sqrt{1-t^2}\,, \qquad P_2^2(t) = 3 - 3t^2, \\
P_3^0(t) = -\tfrac{3}{2}\,t + \tfrac{5}{2}\,t^3, \qquad P_3^1(t) = \left(\tfrac{3}{2} - \tfrac{15}{2}\,t^2\right)\sqrt{1-t^2}, \\
P_3^2(t) = 15t - 15t^3, \qquad P_3^3(t) = \left(-15 + 15\,t^2\right)\sqrt{1-t^2}\,, \\
P_4^0(t) = \tfrac{3}{8} - \tfrac{15}{4}\,t^2 + \tfrac{35}{8}\,t^4, \quad P_4^1(t) = \left(\tfrac{15}{2} - \tfrac{35}{2}\,t^2\right)\sqrt{1-t^2}, \quad P_4^2(t) = -\tfrac{15}{2} + 60\,t^2 - \tfrac{105}{2}\,t^4, \\
P_4^3(t) = \left(-105t + 105\,t^3\right)\sqrt{1-t^2}\,, \qquad P_4^4(t) = 105 - 210\,t^2 + 105\,t^4.
\end{gathered}
\tag{18.29}
$$

When $m = 2\,k \le n$ is an even integer, $P_n^m(t)$ is a polynomial function, while when $m = 2\,k+1 \le n$ is odd, there is a factor of $\sqrt{1-t^2}$. Keep in mind that the square root is real and positive since we are restricting our attention to the interval $-1 \le t \le 1$. If $m > n$, formula (18.28) reduces to the zero function, and is not needed in the final tally.

Warning: Even though half of the associated Legendre functions are polynomials, only those with $m = 0$, i.e., $P_n(t) = P_n^0(t)$, are called *Legendre polynomials*.

Graphs of the first few Legendre polynomials can be found in Figure 5.4. In addition, Figure 18.1 displays the graphs of the associated Legendre functions $P_n^m(t)$ for $1 \le m \le$

$n \leq 4$. Pay particular attention to the fact that, owing to the choice of normalization factor, their graphs have very different vertical scales.

The following result states that the Legendre polynomials and associated Legendre functions are a complete list of solutions to the Legendre boundary value problem (18.24–25). A proof can be found in [**26**].

Theorem 18.2. *Let $m \geq 0$ be a non-negative integer. Then the m^{th} order Legendre boundary value problem prescribed by* (18.24–25) *has eigenvalues $\mu_n = n(n+1)$ for $n = 0, 1, 2, \ldots$, and associated eigenfunctions $P_n^m(t)$ where $m = 0, \ldots, n$.*

Returning to the original variable φ via (18.23), Theorem 18.2 implies that our original boundary value problem

$$\sin^2\varphi\, \frac{d^2p}{d\varphi^2} + \cos\varphi\, \sin\varphi\, \frac{dp}{d\varphi} + (\mu\, \sin^2\varphi - m^2)\, p = 0, \qquad |\, p(0)\,|, \quad |\, p(\pi)\,| < \infty, \tag{18.30}$$

has its eigenvalues and eigenfunctions expressed in terms of the Legendre functions:

$$\mu_n = n(n+1), \qquad p_n^m(\varphi) = P_n^m(\cos\varphi), \qquad \text{for} \qquad 0 \leq m \leq n. \tag{18.31}$$

The eigenfunction $p_n^m(\varphi)$ is, in fact, a trigonometric polynomial of degree n; here are the first few, written in Fourier form:

$$\begin{aligned}
&p_0^0(\varphi) = 1, && p_1^0(\varphi) = \cos\varphi, && p_1^1(\varphi) = -\sin\varphi,\\
&p_2^0(\varphi) = \tfrac{1}{4} + \tfrac{3}{4}\cos 2\varphi, && p_2^1(\varphi) = -\tfrac{3}{2}\sin 2\varphi, && p_2^2(\varphi) = \tfrac{3}{2} - \tfrac{3}{2}\cos 2\varphi,\\
&p_3^0(\varphi) = \tfrac{3}{8}\cos\varphi + \tfrac{5}{8}\cos 3\varphi, && && p_3^1(\varphi) = -\tfrac{3}{8}\sin\varphi - \tfrac{15}{8}\sin 3\varphi,\\
&p_3^2(\varphi) = \tfrac{15}{4}\cos\varphi - \tfrac{15}{4}\cos 3\varphi, && && p_3^3(\varphi) = -\tfrac{45}{4}\sin\varphi + \tfrac{15}{4}\sin 3\varphi,\\
&p_4^0(\varphi) = \tfrac{9}{64} + \tfrac{5}{16}\cos 2\varphi + \tfrac{35}{64}\cos 4\varphi, && && p_4^1(\varphi) = -\tfrac{5}{8}\sin 2\varphi - \tfrac{35}{16}\sin 4\varphi,\\
&p_4^2(\varphi) = \tfrac{45}{16} + \tfrac{15}{4}\cos 2\varphi - \tfrac{105}{16}\cos 4\varphi, && && p_4^3(\varphi) = -\tfrac{105}{4}\sin 2\varphi + \tfrac{105}{8}\sin 4\varphi,\\
&p_4^4(\varphi) = \tfrac{315}{8} - \tfrac{105}{2}\cos 2\varphi + \tfrac{105}{8}\cos 4\varphi.
\end{aligned} \tag{18.32}$$

It is also instructive to plot the eigenfunctions in terms of the zenith angle φ; see Figure 18.2. As in Figure 18.1, the vertical scales are not the same.

At this stage, we have determined both angular components of our separable solutions (18.20). Multiplying the two parts together results in the spherical angle functions

$$\begin{aligned}
Y_n^m(\varphi, \theta) &= P_n^m(\cos\varphi)\cos m\theta, && n = 0, 1, 2, \ldots,\\
\widetilde{Y}_n^m(\varphi, \theta) &= P_n^m(\cos\varphi)\sin m\theta, && m = 0, 1, \ldots, n,
\end{aligned} \tag{18.33}$$

known as *spherical harmonics*. They satisfy the spherical Helmholtz equation

$$\Delta_S\, Y_n^m + n(n+1)\, Y_n^m \;=\; 0 \;=\; \Delta_S\, \widetilde{Y}_n^m + n(n+1)\, \widetilde{Y}_n^m, \tag{18.34}$$

and so are eigenfunctions for the spherical Laplacian operator, (18.17), with associated eigenvalues $\mu_n = n(n+1)$ for $n = 0, 1, 2, \ldots$. The n^{th} eigenvalue μ_n admits a $(2n+1)$–dimensional eigenspace, spanned by the spherical harmonics

$$Y_n^0(\varphi, \theta), \quad Y_n^1(\varphi, \theta), \quad \ldots \quad Y_n^n(\varphi, \theta), \quad \widetilde{Y}_n^1(\varphi, \theta), \quad \ldots \quad \widetilde{Y}_n^n(\varphi, \theta).$$

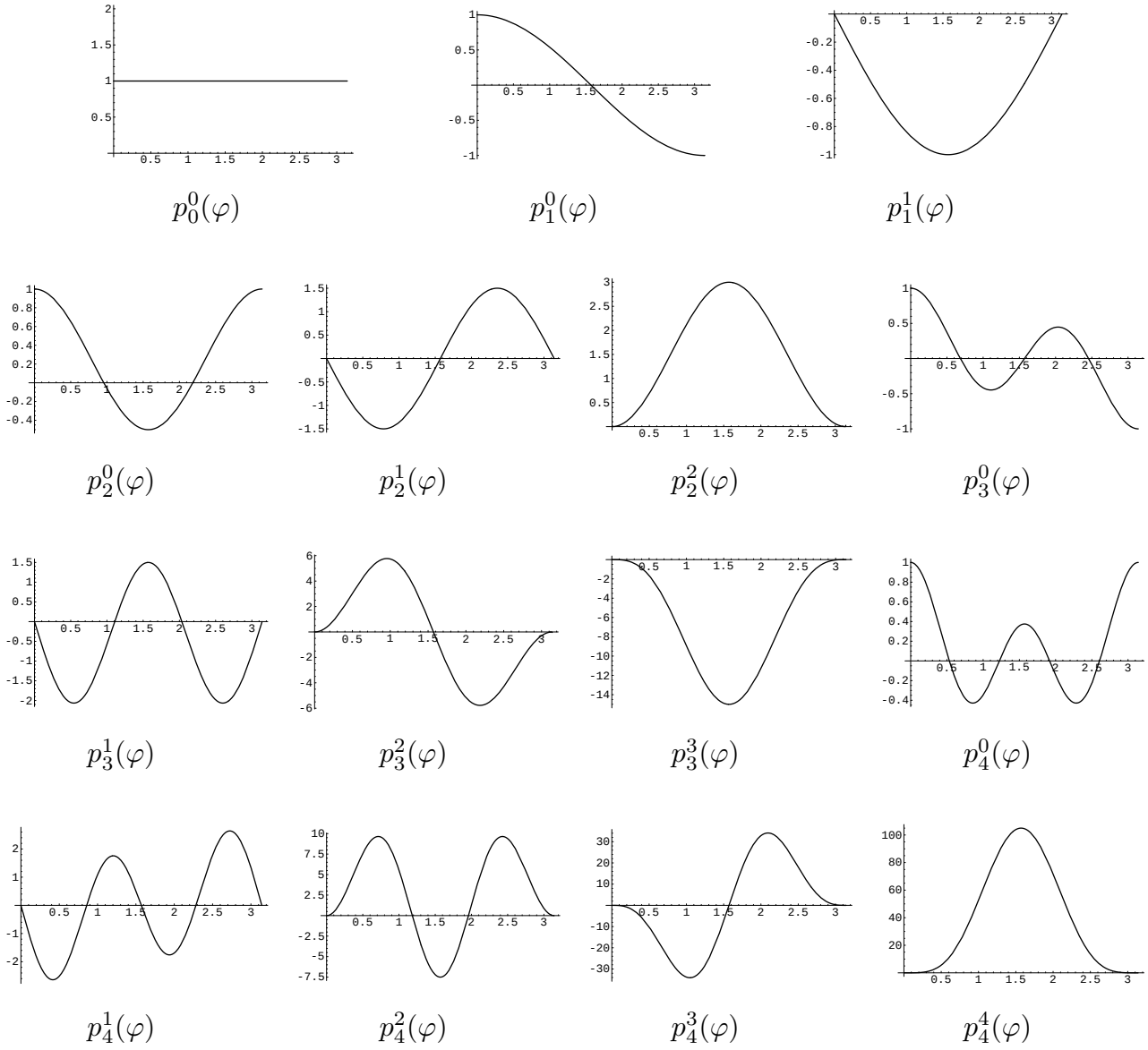


Figure 18.2. Trigonometric Legendre Functions.

(The omitted function $\widetilde{Y}_n^0(\varphi,\theta) \equiv 0$ is trivial, and so does not contribute.) In Figure 18.3 we plot the first few spherical harmonic surfaces $r = Y_n^m(\varphi,\theta)$. In these graphs, in view of the spherical coordinate formula (18.13), points with a negative value of r appear on the opposite side of the origin from the point on the unit sphere with angles φ, θ. Incidentally, the graphs of their counterparts $r = \widetilde{Y}_n^m(\varphi,\theta)$, when $m \neq 0$, are obtained by rotation around the z axis by 90°. On the other hand, the graphs of Y_n^0 are cylindrically symmetric (why?), and hence unaffected by such a rotation.

Self-adjointness of the spherical Laplacian, cf. Exercise ■, implies that the spherical

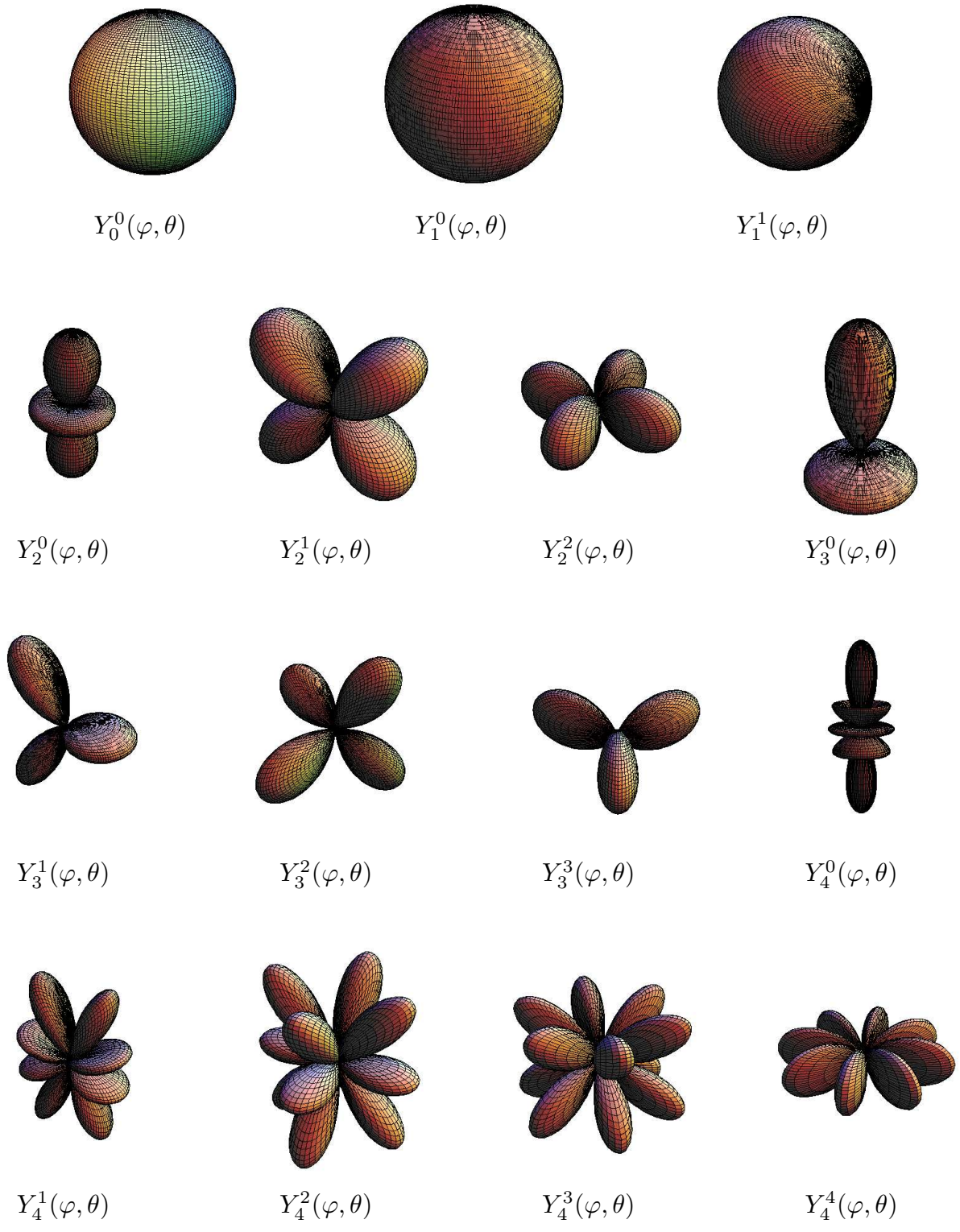


Figure 18.3. Spherical Harmonics.

harmonics are orthogonal with respect to the L^2 inner product

$$\langle f, g \rangle = \iint_{S_1} f\, g\, dS = \int_0^{\pi} \int_0^{2\pi} f(\varphi, \theta)\, g(\varphi, \theta) \sin \varphi \; d\theta \, d\varphi \tag{18.35}$$

given by integrating the product of the functions with respect to surface area over the unit sphere $S_1 = \{ \| \mathbf{x} \| = 1 \}$, cf. (B.40). More correctly, self-adjointness only guarantees orthogonality for the harmonics corresponding to distinct eigenvalues. However, the orthogonality relations

$$\begin{aligned}
\langle Y_n^m, Y_l^k \rangle &= \iint_{S_1} Y_n^m\, Y_l^k\, dS = 0, && && (m,n) \neq (k,l), \\
\langle Y_n^m, \widetilde{Y}_l^k \rangle &= \iint_{S_1} Y_n^m\, \widetilde{Y}_l^k\, dS = 0, && \text{for all} && (m,n), (k,l), \\
\langle \widetilde{Y}_n^m, \widetilde{Y}_l^k \rangle &= \iint_{S_1} \widetilde{Y}_n^m\, \widetilde{Y}_l^k\, dS = 0 && && (m,n) \neq (k,l),
\end{aligned} \tag{18.36}$$

do, in fact, hold in full generality; Exercise ■ contains the details. Their norms can be explicitly computed:

$$\| Y_n^0 \|^2 = \frac{4\pi}{2n+1}, \qquad \| Y_n^m \|^2 = \| \widetilde{Y}_n^m \|^2 = \frac{2\pi (n+m)!}{(2n+1)(n-m)!}. \tag{18.37}$$

A proof of this formula appears in Exercise ■.

With some further work, it can be shown that the spherical harmonics form a complete orthogonal system of functions on the unit sphere. This means that any reasonable function $h \colon S_1 \to \mathbb{R}$, e.g., piecewise C^1, can be expanded into a convergent *spherical Fourier series*

$$h(\varphi, \theta) = \frac{c_{0,0}}{2} + \sum_{n=1}^{\infty} \left(\frac{c_{0,n}}{2}\, Y_n^0(\varphi) + \sum_{m=1}^{n} \Big[c_{m,n} Y_n^m(\varphi, \theta) + \widetilde{c}_{m,n} \widetilde{Y}_n^m(\varphi, \theta) \Big] \right) \tag{18.38}$$

in the spherical harmonics. Applying the orthogonality relations (18.36), we find that the spherical Fourier coefficients are given by the inner products

$$c_{0,n} = \frac{2 \langle h, Y_n^0 \rangle}{\| Y_n^0 \|^2}, \qquad c_{m,n} = \frac{\langle h, Y_n^m \rangle}{\| Y_n^m \|^2}, \qquad \widetilde{c}_{m,n} = \frac{\langle h, \widetilde{Y}_n^m \rangle}{\| \widetilde{Y}_n^m \|^2}, \qquad \begin{aligned} &0 \leq n, \\ &1 \leq m \leq n, \end{aligned}$$

or, explicitly, using (18.35) and the formulae (18.37) for the norms,

$$\begin{aligned}
c_{m,n} &= \frac{(2n+1)(n-m)!}{2\pi\,(n+m)!} \int_0^{2\pi} \int_0^{\pi} h(\varphi, \theta)\, P_n^m(\cos \varphi)\, \cos m\,\theta\, \sin \varphi \; d\varphi \, d\theta, \\
\widetilde{c}_{m,n} &= \frac{(2n+1)(n-m)!}{2\pi\,(n+m)!} \int_0^{2\pi} \int_0^{\pi} h(\varphi, \theta)\, P_n^m(\cos \varphi)\, \sin m\,\theta\, \sin \varphi \; d\varphi \, d\theta.
\end{aligned} \tag{18.39}$$

As with an ordinary Fourier series, the extra $\frac{1}{2}$ was appended to the $c_{0,n}$ terms in the series (18.38) so that the formulae (18.39) are valid for all m, n. In particular, the constant term

in the spherical harmonic series is the mean of the function h over the unit sphere:

$$\frac{c_{0,0}}{2} = \frac{1}{4\,\pi} \iint_{S_1} h\, dS = \frac{1}{4\,\pi} \int_0^{2\,\pi} \int_0^{\pi} h(\varphi, \theta)\, \sin \varphi\, d\varphi\, d\theta \tag{18.40}$$

Remark: An alternative approach is to replace the real trigonometric functions by complex exponentials, and work with the *complex spherical harmonics*

$$\mathcal{Y}_n^m(\theta, \varphi) = Y_n^m(\theta, \varphi) + \mathrm{i}\, \widetilde{Y}_n^m(\theta, \varphi) = P_n^m(\cos \varphi)\, e^{\mathrm{i}\, m\, \theta}, \qquad \begin{array}{l} n = 0, 1, 2, \dots, \\ m = -n, -n+1, \dots, n. \end{array} \tag{18.41}$$

The complex orthogonality and expansion formulas are relegated to the exercises.

To complete our solution to the Laplace equation on the solid ball, we still need to analyze the ordinary differential equation (18.18) for the radial component $v(r)$. In view of our analysis of the spherical Helmholtz equation, the original separation constant is $\mu = n(n+1)$ for some non-negative integer $n \geq 0$, and so the radial equation takes the form

$$r^2\, v'' + 2\, r\, v' - n(n+1)\, v = 0. \tag{18.42}$$

As noted earlier, to solve such a second order linear equation of Euler type (3.84), we substitute the power ansatz $v(r) = r^\alpha$. The exponent α must satisfy the quadratic equation

$$\alpha^2 + \alpha - n(n+1) = 0, \qquad \text{and hence} \qquad \alpha = n \qquad \text{or} \qquad \alpha = -(n+1).$$

Therefore, the two linearly independent solutions are

$$v_1(r) = r^n \qquad \text{and} \qquad v_2(r) = r^{-n-1}. \tag{18.43}$$

Since here we are only interested in solutions that remain bounded at $r = 0$ — the center of the ball — we should just retain the first solution $v(r) = r^n$ in our subsequent analysis.

At this stage, we have solved all three ordinary differential equations for the separable solutions. We combine the results (18.21, 33, 43) together to produce the following spherically separable solutions to the Laplace equation:

$$\begin{array}{ll} H_n^m = r^n\, Y_n^m(\varphi, \theta) = r^n\, P_n^m(\cos \varphi) \cos m\, \theta & n = 0, 1, 2, \dots, \\ \widetilde{H}_n^m = r^n\, \widetilde{Y}_n^m(\varphi, \theta) = r^n\, P_n^m(\cos \varphi) \sin m\, \theta & m = 0, 1, \dots, n. \end{array} \tag{18.44}$$

Although apparently complicated, these solutions are, surprisingly, elementary polynomial functions of the rectangular coordinates x, y, z, and hence *harmonic polynomials*. The first few are

$$\begin{array}{llll} H_0^0 = 1, & H_1^0 = z, & H_2^0 = z^2 - \tfrac{1}{2}\, x^2 - \tfrac{1}{2}\, y^2 & H_3^0 = z^3 - \tfrac{3}{2}\, x^2 z - \tfrac{3}{2}\, y^2 z \\ & H_1^1 = x, & H_2^1 = 3\, x\, z, & H_3^1 = 6\, x\, z^2 - \tfrac{3}{2}\, x^3 - \tfrac{3}{2}\, x\, y^2 \\ & \widetilde{H}_1^1 = y, & \widetilde{H}_2^1 = 3\, y\, z, & \widetilde{H}_3^1 = 6\, y\, z^2 - \tfrac{3}{2}\, x^2 y - \tfrac{3}{2}\, y^3 \\ & & H_2^2 = 3\, x^2 - 3\, y^2, & H_3^2 = 15\, x^2 z - 15\, y^2 z \\ & & \widetilde{H}_2^2 = 6\, x y, & \widetilde{H}_3^2 = 30\, x\, y\, z \\ & & & H_3^3 = 15\, x^3 - 45\, x\, y^2 \\ & & & \widetilde{H}_3^3 = 45\, x^2 y - 15\, y^3. \end{array} \tag{18.45}$$

The polynomials

$$H_n^0, \;\; H_n^1, \; \dots \; , \;\; H_n^n, \;\; \widetilde{H}_n^1, \; \dots \; \widetilde{H}_n^n$$

form a basis for the vector space $\mathcal{H}^{(n)}$ of all homogeneous harmonic polynomials of degree n, which therefore has dimension $2n+1$.

The harmonic polynomials form a complete orthogonal system, and therefore the general solution to the Laplace equation inside the unit ball can be written as a harmonic polynomial series:

$$u(x,y,z) = \frac{c_{0,0}}{2} + \sum_{n=1}^{\infty} \left(\frac{c_{0,n}}{2}\, H_n^0(x,y,z) + \sum_{m=1}^{n} \Big[c_{m,n} H_n^m(x,y,z) + \widetilde{c}_{m,n} \widetilde{H}_n^m(x,y,z) \Big] \right), \tag{18.46}$$

or, equivalently in spherical coordinates,

$$u(r,\varphi,\theta) = \frac{c_{0,0}}{2} + \sum_{n=1}^{\infty} \left(\frac{c_{0,n}}{2}\, r^n\, Y_n^0(\varphi) + \sum_{m=1}^{n} \Big[c_{m,n}\, r^n\, Y_n^m(\varphi,\theta) + \widetilde{c}_{m,n}\, r^n\, \widetilde{Y}_n^m(\varphi,\theta) \Big] \right). \tag{18.47}$$

The coefficients $c_{m,n}, \widetilde{c}_{m,n}$ are uniquely prescribed by the boundary conditions. Indeed, substituting (18.47) into the Dirichlet boundary conditions on the unit sphere $r = 1$ yields

$$u(1,\varphi,\theta) = \frac{c_{0,0}}{2} + \sum_{n=1}^{\infty} \left(\frac{c_{0,n}}{2}\, Y_n^0(\varphi) + \sum_{m=1}^{n} \Big[c_{m,n} Y_n^m(\varphi,\theta) + \widetilde{c}_{m,n} \widetilde{Y}_n^m(\varphi,\theta) \Big] \right) = h(\varphi,\theta). \tag{18.48}$$

Thus, the coefficients $c_{m,n}, \widetilde{c}_{m,n}$ are given by the orthogonality formulae (18.39). If the terms in the resulting series are uniformly bounded — which occurs for all integrable functions h, as well as also certain generalized functions including the delta function — then the harmonic polynomial series (18.47) converges everywhere, and, in fact, uniformly on any smaller ball $\| \mathbf{x} \| = r \le r_0 < 1$.

In rectangular coordinates, the n^{th} summand of the series (18.46) is a homogeneous polynomial of degree n. Therefore, repeating the argument used on the two-dimensional polar coordinate solution (15.39), we conclude that the harmonic polynomial series is, in fact, a power series, and hence hence provides the *Taylor expansion for the harmonic function $u(x,y,z)$ at the origin*! In particular, this implies that the harmonic function $u(x,y,z)$ is analytic at $\mathbf{0}$.

The constant term in such a Taylor series can be identified with the value of the function at the origin: $u(0,0,0) = \frac{1}{2}\, c_{0,0}$. On the other hand, since $u = h$ on $S_1 = \partial\Omega$, the coefficient formula (18.40) tells us that

$$u(0,0,0) = \frac{c_{0,0}}{2} = \frac{1}{4\,\pi} \iint_{S_1} u \, dS. \tag{18.49}$$

Therefore, we have established the three-dimensional counterpart of Theorem 15.8: the value of the harmonic function at the center of the sphere is equal to the average of its values on the sphere's surface. In addition, the higher order coefficients $c_{m,n}, \widetilde{c}_{m,n}$ serve

to prescribe the partial derivatives $\dfrac{\partial^{i+j+k} u}{\partial x^i \partial y^j \partial z^k}(0,0,0)$. In this way, the orthogonality formulae (18.39) can be re-interpreted as three-dimensional counterparts of the Cauchy formulae (16.139) for the derivatives of the real and imaginary parts of a complex analytic function; see Exercise ■ for details.

So far, we have restricted our attention to the sphere of unit radius. A simple scaling argument serves to establish the general result.

Theorem 18.3. *If $u(\mathbf{x})$ is a harmonic function defined on a domain $\Omega \subset \mathbb{R}^3$, then u is analytic inside Ω. Moreover, its value at any point $\mathbf{x}_0 \in \Omega$ is obtained by averaging its values on any sphere centered at $\mathbf{x}_0$, so*

$$u(\mathbf{x}_0) \;=\; \frac{1}{4\,\pi\,a^2} \iint_{\| \mathbf{x}-\mathbf{x}_0 \| = a} u \; dS, \tag{18.50}$$

provided the enclosed ball $\{ \| \mathbf{x} - \mathbf{x}_0 \| \leq a \} \subset \Omega$ lies entirely within the domain of definition.

Proof: It is easily checked that, under the hypothesis of the theorem, the rescaled and translated function

$$U(\mathbf{y}) = u(a\,\mathbf{y} + \mathbf{x}_0) = u(\mathbf{x}), \qquad \text{where} \qquad \mathbf{y} = \frac{\mathbf{x} - \mathbf{x}_0}{a}, \tag{18.51}$$

is harmonic on the unit ball $\| \mathbf{y} \| \leq 1$, and hence solves the boundary value problem (18.12) with boundary values $h(\mathbf{y}) = U(\mathbf{y}) = u(a\,\mathbf{y} + \mathbf{x}_0)$ on $\| \mathbf{y} \| = 1$. By the preceding remarks, $U(\mathbf{y})$ is analytic at $\mathbf{y} = \mathbf{0}$, and so $u(\mathbf{x}) = U\left(\dfrac{\mathbf{x} - \mathbf{x}_0}{a} \right)$ is analytic at $\mathbf{x} = \mathbf{x}_0$. Since $\mathbf{x}_0$ can be any point inside Ω, this proves analyticity of u everywhere in Ω. Moreover, according to the integral formula (18.49),

$$u(\mathbf{x}_0) = U(\mathbf{0}) = \frac{1}{4\,\pi} \iint_{\| \mathbf{y} \| = 1} U \, dS = \frac{1}{4\,\pi\,a^2} \iint_{\| \mathbf{x}-\mathbf{x}_0 \| = a} u \; dS,$$

since the change of variables (18.51) has the effect of rescaling the spherical surface integral. *Q.E.D.*

Arguing as in the planar case of Theorem 15.9, we readily establish the corresponding Maximum Principle for harmonic functions of three variables.

Theorem 18.4. *A non-constant harmonic function cannot have a local maximum or minimum at any interior point of its domain of definition. Moreover, its global maximum or minimum (if any) can only occur on the boundary of the domain.*

For instance, the Maximum Principle implies that the maximum and minimum temperatures in a solid body in thermal equilibrium are to be found only on its boundary. In physical terms, since heat energy must flow away from any internal maximum and towards any internal minimum, any local temperature extremum inside the body would preclude it from being in thermal equilibrium.

Example 18.5. In this example, we shall determine the electrostatic potential inside a hollow sphere when the upper and lower hemispheres are held at different constant potentials. This device is called a *spherical capacitor* and is realized experimentally by separating the two charged conducting hemispherical shells by a thin insulating ring at the equator. A straightforward scaling argument allows us to choose our units so that the sphere has radius 1, while the potential is set equal to 1 on the upper hemisphere and equal to 0, i.e., grounded, on the lower hemisphere. The resulting electrostatic potential ssatisfies the Laplace equation

$$\Delta u = 0 \qquad \text{inside a solid ball} \qquad \| \mathbf{x} \| < 1,$$

and is subject to Dirichlet boundary conditions

$$u(x,y,z) = \begin{cases} 1, & z > 0, \\ 0, & z < 0, \end{cases} \qquad \text{on the unit sphere} \qquad \| \mathbf{x} \| = 1. \tag{18.52}$$

The solution will be prescribed by a harmonic polynomial series (18.46) whose coefficients are fixed by the boundary values (18.52). Before taking on the required computation, let us first note that since the boundary data does not depend upon the azimuthal angle θ, the solution $u = u(r,\varphi)$ will also be independent of θ. Therefore, we need only consider the θ-independent spherical harmonic polynomials (18.33), which are those with $m = 0$, and hence

$$u(r,\varphi) \;=\; \frac{1}{2} \sum_{n=0}^{\infty} c_n \, H_n^0(x,y,z) \;=\; \frac{1}{2} \sum_{n=0}^{\infty} c_n \, r^n \, P_n(\cos\varphi),$$

where we abbreviate $c_n = \frac{1}{2} c_{0,n}$. The boundary conditions (18.52) require

$$u(1,\varphi) = \frac{1}{2} \sum_{n=0}^{\infty} c_n P_n(\cos\varphi) = h(\varphi) = \begin{cases} 1, & 0 \le \varphi < \frac{1}{2}\pi, \\ 0, & \frac{1}{2}\pi < \varphi \le \pi. \end{cases}$$

The coefficients are given by (18.39), which, in the case $m = 0$, reduce to

$$c_n = \frac{2n+1}{2\pi} \iint_S f \, Y_n^0 \, dS = (2n+1) \int_0^{\pi/2} P_n(\cos\varphi) \sin\varphi \; d\varphi = (2n+1) \int_0^1 P_n(t) \, dt, \tag{18.53}$$

since $f = 0$ when $\frac{1}{2}\pi < \varphi \le \pi$. The first few are

$$c_0 = \tfrac{1}{4}, \qquad c_1 = \tfrac{3}{8}, \qquad c_2 = 0, \qquad c_3 = -\tfrac{7}{32}, \qquad c_4 = 0, \qquad \dots .$$

Therefore, the solution has the explicit Taylor expansion

$$\begin{aligned} u &= \tfrac{1}{2} + \tfrac{3}{4} \, r \cos\varphi - \tfrac{21}{128} \, r^3 \cos\varphi - \tfrac{35}{128} \, r^3 \cos 3\varphi + \; \cdots \\ &= \tfrac{1}{2} + \tfrac{3}{4} \, z - \tfrac{7}{16} \, z^3 + \tfrac{21}{32} \, (x^2 + y^2) \, z + \; \cdots \; . \end{aligned} \tag{18.54}$$

Note in particular that the value $u(0,0,0) = \frac{1}{2}$ at the center of the sphere is the average of its boundary values, in accordance with Theorem 18.3.

Remark: The same solution $u(x,y,z)$ describes the thermal equilibrium in a solid sphere whose upper hemisphere is held at temperature 1° and lower hemisphere at 0°.

Example 18.6. A closely related problem is to determine the electrostatic potential *outside* a spherical capacitor. As in the preceding example, we take our capacitor of radius 1, with electrostatic charge of 1 on the upper hemisphere and 0 on the lower hemisphere. Here, we need to solve the Laplace equation $\Delta u = 0$ in the unbounded domain $\Omega = \{ \| \mathbf{x} \| > 1 \}$ — the exterior of the unit sphere — subject to same Dirichlet boundary conditions (18.52). We anticipate that the potential will be vanishingly small at large distances away from the capacitor: $r = \| \mathbf{x} \| \gg 1$. Therefore, the harmonic polynomial solutions (18.44) will not help us solve this problem, since (except for the constant case) they become unboundedly large far away from the origin.

However, reconsideration of our original separation of variables argument will produce a different class of solutions having the desired decay properties. When we solved the radial equation (18.42), we discarded the solution $v_2(r) = r^{-n-1}$ because it had a singularity at the origin. In the present situation, the behavior of the function at $r = 0$ is irrelevant; our requirement is that the solution decays as $r \to \infty$, and $v_2(r)$ has this property. Therefore, we will utilize the complementary harmonic functions

$$\begin{aligned} K_n^m(x,y,z) &= r^{-2n-1} H_n^m(x,y,z) = r^{-n-1} Y_n^m(\varphi,\theta) = r^{-n-1} P_n^m(\cos\varphi) \cos m\theta, \\ \widetilde{K}_n^m(x,y,z) &= r^{-2n-1} \widetilde{H}_n^m(x,y,z) = r^{-n-1} \widetilde{Y}_n^m(\varphi,\theta) = r^{-n-1} P_n^m(\cos\varphi) \sin m\theta, \end{aligned} \tag{18.55}$$

for solving such exterior problems. For the capacitor problem, we need only those that are independent of θ, which have $m = 0$. We write the resulting solution as a series

$$u(r,\varphi) = \frac{1}{2} \sum_{n=0}^{\infty} c_n K_n^0(x,y,z) = \frac{1}{2} \sum_{n=0}^{\infty} c_n r^{-n-1} P_n(\cos\varphi).$$

The boundary conditions

$$u(1,\varphi) = \frac{1}{2} \sum_{n=0}^{\infty} c_n P_n(\cos\varphi) = f(\varphi) = \begin{cases} 1, & 0 \le \varphi < \frac{1}{2}\pi, \\ 0, & \frac{1}{2}\pi < \varphi \le \pi, \end{cases}$$

are identical with those in the previous example. Therefore, the coefficients are given by (18.53), leading to the series expansion

$$u = \frac{1}{2r} + \frac{3\cos\varphi}{4r^2} - \frac{21\cos\varphi + 35\cos 3\varphi}{128\, r^4} + \cdots = \frac{1}{2r} + \frac{3z}{4r^3} - \frac{14z^3 - 21(x^2+y^2)z}{32\, r^7} + \cdots, \tag{18.56}$$

where $r = \sqrt{x^2+y^2+z^2}$. Interestingly, at large distances, the higher order terms become negligible, and the potential looks like that associated with a point charge of magnitude $\frac{1}{2}$ — the average of the potential over the sphere — that is concentrated at the origin. This is indicative of a general fact: see Exercise ∎.

18.3. The Green's Function.

We now turn to the inhomogeneous form of the three-dimensional Laplace equation: the *Poisson equation*

$$-\Delta u = f \qquad \text{for all} \qquad \mathbf{x} \in \Omega \tag{18.57}$$

on a solid domain $\Omega \subset \mathbb{R}^3$. In order to uniquely specify the solution, we must impose appropriate boundary conditions: Dirichlet or mixed. We only need to discuss the case of homogeneous boundary conditions, since, by linearity, an inhomogeneous boundary value problem can be split up into a homogeneous boundary value problem for the inhomogeneous Poisson equation and an inhomogeneous boundary value problem for the homogeneous Laplace equation.

As in Chapters 11 and 15, we begin by analyzing the case of a delta function inhomogeneity that is concentrated at a single point in the domain. Thus, for each $\boldsymbol{\xi} = (\xi, \eta, \zeta) \in \Omega$, the *Green's function* $G(\mathbf{x}; \boldsymbol{\xi}) = G(x, y, z; \xi, \eta, \zeta)$ is the unique solution to the Poisson equation

$$-\Delta u = \delta(\mathbf{x} - \boldsymbol{\xi}) = \delta(x - \xi)\, \delta(y - \eta)\, \delta(z - \zeta) \qquad \text{for all} \qquad \mathbf{x} \in \Omega, \tag{18.58}$$

subject to the chosen homogeneous boundary conditions. The solution to the general Poisson equation (18.57) is then obtained by superposition: We write the forcing function

$$f(x, y, z) = \iiint_{\Omega} f(\xi, \eta, \zeta)\, \delta(x - \xi)\, \delta(y - \eta)\, \delta(z - \zeta)\, d\xi\, d\eta\, d\zeta$$

as a linear superposition of delta functions. By linearity, the solution

$$u(x, y, z) = \iiint_{\Omega} f(\xi, \eta, \zeta)\, G(x, y, z; \xi, \eta, \zeta)\, d\xi\, d\eta\, d\zeta \tag{18.59}$$

to the homogeneous boundary value problem for the Poisson equation (18.57) is then given as the corresponding superposition of the Green's function solutions.

The Green's Function in Space

Only in a few specific instances is the explicit formula for the Green's function known. Nevertheless, certain general guiding features can be readily established. The starting point is to investigate the Poisson equation (18.58) when the domain $\Omega = \mathbb{R}^3$ is all of three-dimensional space. We impose boundary constraints by seeking a solution that goes to zero, $u(\mathbf{x}) \to 0$, at large distances $\| \mathbf{x} \| \to \infty$. Since the Laplacian is invariant under translations we can, without loss of generality, place our delta impulse at the origin, and concentrate on solving the particular case

$$-\Delta u = \delta(\mathbf{x}), \qquad \mathbf{x} \in \mathbb{R}^3.$$

Since $\delta(\mathbf{x}) = 0$ for all $\mathbf{x} \neq \mathbf{0}$, the desired solution will, in fact, be a solution to the homogeneous Laplace equation

$$\Delta u = 0, \qquad \mathbf{x} \neq \mathbf{0},$$

save, possibly, for a singularity at the origin.

The Laplace equation models the equilibria of a homogeneous, isotropic medium, and so is also invariant under three-dimensional rotations; details can be found in Exercise ■. This suggests that, in any radially symmetric configuration, the solution should only depend upon the distance $r = \| \mathbf{x} \|$ from the origin. Referring to the spherical coordinate

form (18.14) of the Laplacian operator, if u only depends upon r, its derivatives with respect to the angular coordinates φ, θ are zero, and so $u(r)$ solves the ordinary differential equation

$$\frac{d^2u}{dr^2} + \frac{2}{r}\,\frac{du}{dr} = 0. \tag{18.60}$$

This equation is, in effect, a first order linear ordinary differential equation for $v = du/dr$ and hence is particularly easy to solve:

$$\frac{du}{dr} = v(r) = -\,\frac{b}{r^2}\,, \qquad \text{and hence} \qquad u(r) = a + \frac{b}{r}\,,$$

where a, b are arbitrary constants. The constant solution $u(r) = a$ does not die away at large distances, nor does it have a singularity at the origin. Therefore, if our intuition is valid, the desired solution should be of the form

$$u = \frac{b}{r} = \frac{b}{\| \mathbf{x} \|} = \frac{b}{\sqrt{x^2 + y^2 + z^2}}\,. \tag{18.61}$$

Indeed, this function is harmonic — solves Laplace's equation — everywhere away from the origin, and has a singularity at $\mathbf{x} = \mathbf{0}$.

The solution (18.61) is, up to constant multiple, the three-dimensional Newtonian gravitational potential due to a point mass at the origin. Its gradient

$$\mathbf{g}(\mathbf{x}) \;=\; \nabla \left(\frac{b}{\| \mathbf{x} \|} \right) \;=\; -\,\frac{b\,\mathbf{x}}{\| \mathbf{x} \|^3}\,. \tag{18.62}$$

defines the gravitational force vector at the point $\mathbf{x}$. When $b > 0$, the force $\mathbf{g}(\mathbf{x})$ points towards the mass at the origin. Its magnitude

$$\| \mathbf{g} \| \;=\; \frac{b}{\| \mathbf{x} \|^2} \;=\; \frac{b}{r^2}$$

is proportional to one over the squared distance, which is the well-known inverse square law of three-dimensional Newtonian gravity. Thus, (18.61) can also be interpreted as the electrostatic Coulomb potential on a charged mass at position $\mathbf{x}$ due to a concentrated electric charge at the origin, with (18.62) the corresponding electrostatic force. The constant b is positive when the charges are of opposite signs, leading to an attractive force, and negative in the repulsive case of like charges.

Returning to our problem, the remaining task is to fix the multiple b such that the Laplacian of our candidate solution (18.61) has a delta function singularity at the origin; equivalently, we must determine $c = 1/b$ such that

$$-\,\Delta\, r^{-1} = c\,\delta(\mathbf{x}). \tag{18.63}$$

This equation is certainly valid away from the origin, since $\delta(\mathbf{x}) = 0$ when $\mathbf{x} \neq \mathbf{0}$. To investigate near the singularity, we integrate both sides of (18.63) over a small solid ball $B_\varepsilon = \{ \| \mathbf{x} \| \leq \varepsilon \}$ of radius ε:

$$-\iiint_{B_\varepsilon} \Delta\, r^{-1}\; dx\, dy\, dz = \iiint_{B_\varepsilon} c\,\delta(\mathbf{x})\, dx\, dy\, dz = c, \tag{18.64}$$

where we used the definition of the delta function to evaluate the right hand side. On the other hand, since $\Delta\, r^{-1} = \nabla \cdot \nabla\, r^{-1}$, we can use the divergence theorem (B.82) to evaluate the left hand integral, whence

$$\iiint_{B_\varepsilon} \Delta\, r^{-1}\, dx\, dy\, dz = \iiint_{B_\varepsilon} \nabla \cdot \nabla\, r^{-1}\, dx\, dy\, dz = \iint_{S_\varepsilon} \frac{\partial}{\partial \mathbf{n}}\left(\frac{1}{r}\right) dS,$$

where the surface integral is over the bounding sphere $S_\varepsilon = \partial B_\varepsilon = \{\, \| \mathbf{x} \| = \varepsilon \,\}$. The sphere's unit normal $\mathbf{n}$ points in the radial direction, and hence the normal derivative coincides with differentiation with respect to r; in particular,

$$\frac{\partial}{\partial \mathbf{n}}\left(\frac{1}{r}\right) = \frac{\partial}{\partial r}\left(\frac{1}{r}\right) = -\,\frac{1}{r^2}\,.$$

The surface integral can now be explicitly evaluated:

$$\iint_{S_\varepsilon} \frac{\partial}{\partial \mathbf{n}}\left(\frac{1}{r}\right) dS = -\iint_{S_\varepsilon} \frac{1}{r^2}\, dS = -\iint_{S_\varepsilon} \frac{1}{\varepsilon^2}\, dS = -\,4\pi,$$

since S_ε has surface area $4\pi\varepsilon^2$. Substituting this result back into (18.64), we conclude that

$$c = 4\pi, \qquad \text{and hence} \qquad -\,\Delta\, r^{-1} = 4\pi\,\delta(\mathbf{x}). \tag{18.65}$$

This is our desired formula! We conclude that the solution to Poisson's equation for a delta function impulse at the origin is

$$G(x,y,z) = \frac{1}{4\pi\, r} = \frac{1}{4\pi\, \| \mathbf{x} \|} = \frac{1}{4\pi\, \sqrt{x^2+y^2+z^2}}\,, \tag{18.66}$$

which is the three-dimensional Newtonian potential due to a unit point mass situated at the origin.

If the singularity is concentrated at some other point $\boldsymbol{\xi} = (\xi, \eta, \zeta)$, then we merely translate the preceding solution. This leads immediately to the Green's function

$$G(\mathbf{x}; \boldsymbol{\xi}) = G(\mathbf{x} - \boldsymbol{\xi}) = \frac{1}{4\pi\, \| \mathbf{x} - \boldsymbol{\xi} \|} = \frac{1}{4\pi\, \sqrt{(x-\xi)^2 + (y-\eta)^2 + (z-\zeta)^2}}\,. \tag{18.67}$$

The superposition principle (18.59) implies the following integral formula for the solutions to the Poisson equation on all of three-dimensional space.

Theorem 18.7. *A particular solution to the Poisson equation*

$$-\,\Delta u = f \qquad \text{for} \qquad \mathbf{x} \in \mathbb{R}^3 \tag{18.68}$$

is given by

$$u_\star(\mathbf{x}) = \frac{1}{4\pi} \iiint_{\mathbb{R}^3} \frac{f(\boldsymbol{\xi})}{\| \mathbf{x} - \boldsymbol{\xi} \|}\, d\boldsymbol{\xi} = \frac{1}{4\pi} \iiint_{\mathbb{R}^3} \frac{f(\xi,\eta,\zeta)\; d\xi\, d\eta\, d\zeta}{\sqrt{(x-\xi)^2 + (y-\eta)^2 + (z-\zeta)^2}}\,. \tag{18.69}$$

The general solution is $u(x,y,z) = u_\star(x,y,z) + w(x,y,z)$, where $w(x,y,z)$ is an arbitrary harmonic function.

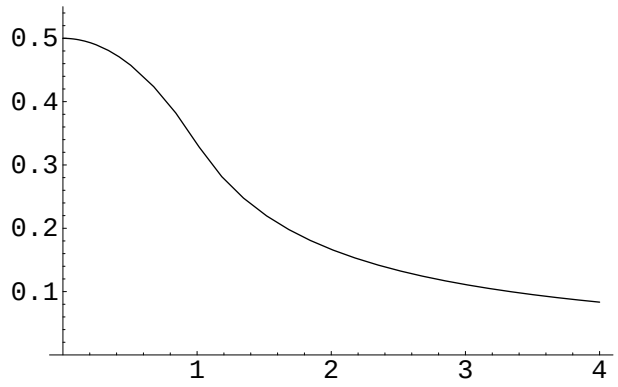


Figure 18.4. Solution to Poisson's Equation in a Solid Ball.

Example 18.8. In this example, we compute the gravitational (or electrostatic) potential in three-dimensional space due to a uniform solid ball, e.g., a spherical planet such as the earth. By rescaling, it suffices to consider the case when the forcing function

$$f(\mathbf{x}) = \begin{cases} 1, & \| \mathbf{x} \| < 1, \\ 0, & \| \mathbf{x} \| > 1, \end{cases}$$

is equal to 1 inside a solid ball of radius 1 and zero outside. The particular solution to the resulting Poisson equation (18.68) is given by the integral

$$u(\mathbf{x}) = \frac{1}{4\pi} \iiint_{\| \boldsymbol{\xi} \| < 1} \frac{1}{\| \mathbf{x} - \boldsymbol{\xi} \|} \, d\xi \, d\eta \, d\zeta. \tag{18.70}$$

Clearly, since the forcing function is radially symmetric, the solution $u = u(r)$ is also radially symmetric. To evaluate the integral, then, we can take $\mathbf{x} = (0, 0, z)$ to lie on the z axis, so that $r = \| \mathbf{x} \| = | z |$. We use cylindrical coordinates $\boldsymbol{\xi} = (\rho \cos\theta, \rho \sin\theta, \zeta)$, so that

$$\| \mathbf{x} - \boldsymbol{\xi} \| = \sqrt{\rho^2 + (z - \zeta)^2} \, .$$

The integral in (18.70) can then be explicitly computed:

$$\begin{aligned} \frac{1}{4\pi} \int_{-1}^{1} \int_{0}^{\sqrt{1-\zeta^2}} \int_{0}^{2\pi} \frac{\rho \, d\theta \, d\rho \, d\zeta}{\sqrt{\rho^2 + (z-\zeta)^2}} \; &= \\ = \frac{1}{2} \int_{-1}^{1} \Big(\sqrt{1 + z^2 - 2z\zeta} \; - \; | z - \zeta | \Big) \, d\zeta &= \begin{cases} \dfrac{1}{3 \, | z |} \, , & | z | \geq 1, \\ \dfrac{1}{2} - \dfrac{z^2}{6} \, , & | z | \leq 1. \end{cases} \end{aligned}$$

Therefore, by radial symmetry, the solution is

$$u(\mathbf{x}) = \begin{cases} \dfrac{1}{3r} \, , & r = \| \mathbf{x} \| \geq 1, \\ \dfrac{1}{2} - \dfrac{r^2}{6} \, , & r = \| \mathbf{x} \| \leq 1, \end{cases} \tag{18.71}$$

plotted, as a function of $r = \| \mathbf{x} \|$ in Figure 18.4. Note that, outside the solid ball, the solution is a Newtonian potential corresponding to a concentrated point mass of magnitude

$\frac{4}{3}\pi$ — the total mass of the planet. We have thus demonstrated a well-known result in gravitation and electrostatics: the exterior potential due to a spherically symmetric mass (or electric charge) is the same as if all the mass (charge) were concentrated at its center. In outer space if you can't see a spherical planet, you can only determine its mass, not its size, by measuring its external gravitational force.

Bounded Domains and the Method of Images

Suppose we now wish to solve the inhomogeneous Poisson equation (18.57) on a bounded domain $\Omega \subset \mathbb{R}^3$. To construct the desired Green's function, we proceed as follows. The Newtonian potential (18.67) is a particular solution to the underlying inhomogeneous equation

$$-\Delta u = \delta(\mathbf{x} - \boldsymbol{\xi}), \qquad \mathbf{x} \in \Omega, \tag{18.72}$$

but it almost surely does not have the proper boundary values on $\partial\Omega$. By linearity, the general solution to such an inhomogeneous linear equation is of the form

$$u(\mathbf{x}) = \frac{1}{4\pi \,\|\, \mathbf{x} - \boldsymbol{\xi} \,\|} \;-\; v(\mathbf{x}), \tag{18.73}$$

where the first summand is a particular solution, which we now know, while[†] $v(\mathbf{x})$ is an arbitrary solution to the homogeneous equation $\Delta v = 0$, i.e., an arbitrary harmonic function. The solution (18.73) satisfies the homogeneous boundary conditions provided the boundary values of $v(\mathbf{x})$ match those of the Green's function. Let us explicitly state the result in the Dirichlet case.

Theorem 18.9. *The Green's function for the homogeneous Dirichlet boundary value problem*

$$-\Delta u = f, \qquad \mathbf{x} \in \Omega, \qquad\qquad u = 0, \qquad \mathbf{x} \in \partial\Omega,$$

for the Poisson equation in a domain $\Omega \subset \mathbb{R}^3$ has the form

$$G(\mathbf{x}; \boldsymbol{\xi}) = \frac{1}{4\pi \,\|\, \mathbf{x} - \boldsymbol{\xi} \,\|} \;-\; v(\mathbf{x}; \boldsymbol{\xi}), \qquad \mathbf{x}, \boldsymbol{\xi} \in \Omega, \tag{18.74}$$

where $v(\mathbf{x}; \boldsymbol{\xi})$ is the harmonic function of $\mathbf{x}$ that satisfies

$$v(\mathbf{x}; \boldsymbol{\xi}) = \frac{1}{4\pi \,\|\, \mathbf{x} - \boldsymbol{\xi} \,\|} \qquad \text{for all} \qquad \mathbf{x} \in \partial\Omega.$$

In this manner, we have reduced the determination of the Green's function to the solution to a particular family of Laplace boundary value problems, which are parametrized by the point $\boldsymbol{\xi} \in \Omega$.

In certain domains with simple geometry, the Method of Images can be used to produce an explicit formula for the Green's function. As in Section 15.3, the idea is to match the boundary values of the free space Green's function due to a delta impulse at a point inside

[†] The minus sign is for later convenience.

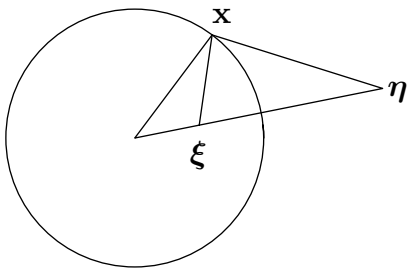


Figure 18.5. Method of Images for the Unit Ball.

the domain with one or more additional Green's functions corresponding to impulses at points outside the domain — the "image points".

The case of a solid ball of radius 1 with Dirichlet boundary conditions is the easiest to handle. Indeed, the *same* geometrical construction that we used for a planar disk, redrawn in in Figure 18.5, applies here. Although the same as Figure 15.8, we are re-interpreting it as a three-dimensional diagram, with the circle representing the unit sphere, while the lines remain lines. The required image point is given by *inversion*:

$$\boldsymbol{\eta} = \frac{\boldsymbol{\xi}}{\| \boldsymbol{\xi} \|^2}, \qquad \text{whereby} \qquad \| \boldsymbol{\xi} \| = \frac{1}{\| \boldsymbol{\eta} \|}.$$

By the similar triangles argument used before, we find

$$\frac{\| \boldsymbol{\xi} \|}{\| \mathbf{x} \|} = \frac{\| \mathbf{x} \|}{\| \boldsymbol{\eta} \|} = \frac{\| \mathbf{x} - \boldsymbol{\xi} \|}{\| \mathbf{x} - \boldsymbol{\eta} \|}, \qquad \text{and therefore} \qquad \| \mathbf{x} \| = 1.$$

As a result, the function

$$v(\mathbf{x}, \boldsymbol{\xi}) = \frac{1}{4\pi} \, \frac{\| \boldsymbol{\eta} \|}{\| \mathbf{x} - \boldsymbol{\eta} \|} = \frac{1}{4\pi} \, \frac{\| \boldsymbol{\xi} \|}{\| \boldsymbol{\xi} - \| \boldsymbol{\xi} \|^2 \, \mathbf{x} \|}$$

has the same boundary values on the unit sphere as the Newtonian potential:

$$\frac{1}{4\pi} \, \frac{\| \boldsymbol{\eta} \|}{\| \mathbf{x} - \boldsymbol{\eta} \|} = \frac{1}{4\pi \| \mathbf{x} - \boldsymbol{\xi} \|} \qquad \text{whenever} \qquad \| \mathbf{x} \| = 1.$$

We conclude that their difference

$$G(\mathbf{x}; \boldsymbol{\xi}) = \frac{1}{4\pi} \left(\frac{1}{\| \mathbf{x} - \boldsymbol{\xi} \|} - \frac{\| \boldsymbol{\xi} \|}{\| \boldsymbol{\xi} - \| \boldsymbol{\xi} \|^2 \, \mathbf{x} \|} \right) \tag{18.75}$$

has the required properties of the Green's function: it satisfies the Laplace equation inside the unit ball except at the delta function singularity $\mathbf{x} = \boldsymbol{\xi}$, and, moreover, $G(\mathbf{x}; \boldsymbol{\xi}) = 0$ has homogeneous Dirichlet conditions on the spherical boundary $\| \mathbf{x} \| = 1$.

With the Green's function in hand, we can apply the general superposition formula (18.59) to arrive at a solution to the Dirichlet boundary value problem for the Poisson equation in the unit ball.

Theorem 18.10. *The solution to the homogeneous Dirichlet boundary value problem*

$$-\Delta u = f, \qquad \text{for} \qquad \| \mathbf{x} \| < 1, \qquad\qquad u = 0, \qquad \text{for} \qquad \| \mathbf{x} \| = 1$$

is

$$u(\mathbf{x}) = \frac{1}{4\pi} \iiint_{\| \boldsymbol{\xi} \| \leq 1} \left(\frac{1}{\| \mathbf{x} - \boldsymbol{\xi} \|} - \frac{\| \boldsymbol{\xi} \|}{\| \boldsymbol{\xi} - \| \boldsymbol{\xi} \|^2 \, \mathbf{x} \|} \right) f(\boldsymbol{\xi}) \, d\xi \, d\eta \, d\zeta. \tag{18.76}$$

The Green's function can also be used to solve the inhomogeneous boundary value problem

$$-\Delta u = 0, \qquad \mathbf{x} \in \Omega, \qquad\qquad u = h, \qquad \mathbf{x} \in \partial\Omega. \tag{18.77}$$

The same argument used in the two-dimensional situation produces the solution

$$u(\mathbf{x}) = - \iint_{\partial\Omega} \frac{\partial G(\mathbf{x}; \boldsymbol{\xi})}{\partial \mathbf{n}} \, h(\boldsymbol{\xi}) \, dS. \tag{18.78}$$

In the case when Ω is a solid ball, this integral formula effectively sums the spherical harmonic series (18.46).

18.4. The Heat Equation in Three-Dimensional Media.

Thermal diffusion in a homogeneous, isotropic solid body $\Omega \subset \mathbb{R}^3$ is governed by the three-dimensional *heat equation*

$$\frac{\partial u}{\partial t} = \gamma \, \Delta u = \gamma \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right), \qquad\qquad (x, y, z) \in \Omega. \tag{18.79}$$

The positivity of the body's thermal diffusivity $\gamma > 0$ is reaquired on both physical and mathematical grounds. The physical derivation is exactly the same as the two-dimensional version (17.1), and does not need to be repeated in detail. Briefly, the heat flux vector is proportional to the temperature gradient, $\mathbf{w} = -\kappa \nabla u$, while its divergence is proportional to the rate of change of temperature: $\nabla \cdot \mathbf{w} = -\sigma u_t$. Combining these two physical laws and assuming homogeneity, whereby κ and σ are constant, produces (18.79) with $\gamma = \kappa/\sigma$.

As always, we must impose suitable boundary conditions: Dirichlet conditions $u = h$ that specify the boundary temperature; (homogeneous) Neumann conditions $\partial u/\partial \mathbf{n} = 0$ corresponding to an insulated boundary; or a mixture of the two. Given the initial temperature of the body

$$u(t_0, x, y, z) = f(x, y, z) \tag{18.80}$$

at the initial time t_0, it can be proved, [**47**], that the resulting initial-boundary value problem is well-posed, and so there is a unique solution $u(t, x, y, z)$ that is defined for all subsequent times $t \geq t_0$ and depends continuously on the initial data.

As in the one- and two-dimensional versions, we do not lose generality by restricting our attention to homogeneous boundary conditions. Separation of variables method works as usual, and we quickly review the basic ideas. One begins by imposing an exponential

ansatz $u(t, \mathbf{x}) = e^{-\lambda t}\, v(\mathbf{x})$. Substituting into the differential equation and canceling the exponentials, it follows that v satisfies the Helmholtz eigenvalue problem

$$\gamma\, \Delta v + \lambda\, v = 0,$$

subject to the relevant boundary conditions. For Dirichlet and mixed boundary conditions, the Laplacian is a positive definite operator, and hence the eigenvalues are all strictly positive,

$$0 < \lambda_1 \le \lambda_2 \le \cdots, \qquad \text{with} \qquad \lambda_n \longrightarrow \infty, \qquad \text{as} \qquad n \to \infty.$$

Moreover, on a bounded domain, the Helmholtz eigenfunctions are complete, and so linear superposition implies that the solution can be written as an eigenfunction series

$$u(t, \mathbf{x}) = \sum_{n=1}^{\infty} c_n\, e^{-\lambda_n\, t}\, v_n(\mathbf{x}). \tag{18.81}$$

The coefficients c_n are uniquely prescribed by the initial condition (18.80):

$$u(t_0, \mathbf{x}) = \sum_{n=1}^{\infty} c_n\, e^{-\lambda_n\, t_0}\, v_n(\mathbf{x}) = f(\mathbf{x}). \tag{18.82}$$

Self-adjointness of the boundary value problem implies orthogonality of the eigenfunctions, and hence the coefficients are given by the usual orthogonality formulae

$$c_n = e^{\lambda_n\, t_0}\, \frac{\langle\, f\,, v_n\,\rangle}{\|\, v_n\,\|^2} = e^{-\lambda_n\, t_0}\, \frac{\displaystyle\iiint_\Omega f(\mathbf{x})\, v_n(\mathbf{x})\, dx\, dy\, dz}{\displaystyle\iiint_\Omega v_n(\mathbf{x})^2\, dx\, dy\, dz}\,. \tag{18.83}$$

The resulting solution $u(t, \mathbf{x}) \to 0$ decays exponentially fast to thermal equilibrium, at a rate equal to the smallest positive eigenvalue $\lambda_1 > 0$. Since the higher modes — the terms with $n \gg 0$ — go to zero extremely rapidly with increasing t, the solution can be well approximated by the first few terms in its Fourier expansion. As a consequence, the heat equation rapidly smoothes out discontinuities and eliminates high frequency noise in the initial data, and so can be used to process three-dimensional images and video — although better nonlinear techniques are now available, [**156**].

Unfortunately, the explicit formulae for the eigenfunctions and eigenvalues few and far between, [**131**]. Most explicit eigensolutions of the Helmholtz boundary value problem require a further separation of variables. In a rectangular box, one separates into a product of functions depending upon the individual Cartesian coordinates, and the eigenfunctions are written as products of trigonometric and hyperbolic functions; see Exercise ■ for details. In a cylindrical domain, the separation is effected in cylindrical coordinates, and leads to eigensolutions involving trigonometric and Bessel functions, as outlined in Exercise ■. The most interesting and enlightening case is a spherical domain, and we treat this particular problem in complete detail.

Heating of a Ball

Our goal is to study heat propagation in a solid spherical body, e.g., the earth[†]. For simplicity, we take the diffusivity $\gamma = 1$, and consider the heat equation on a solid spherical ball $B_1 = \{ \| \mathbf{x} \| < 1 \}$ of radius 1, subject to homogeneous Dirichlet boundary conditions. Once we know how to solve this particular case, an easy scaling argument, as outlined in Exercise ■, will allow us to find the solution for a ball of arbitrary radius and with a general diffusion coefficient.

As usual, when dealing with a spherical geometry, we adopt spherical coordinates r, φ, θ as in (18.13), in terms of which the heat equation takes the form

$$\frac{\partial u}{\partial t} = \Delta u = \frac{\partial^2 u}{\partial r^2} + \frac{2}{r}\,\frac{\partial u}{\partial r} + \frac{1}{r^2}\,\frac{\partial^2 u}{\partial \varphi^2} + \frac{\cos\varphi}{r^2 \sin\varphi}\,\frac{\partial u}{\partial \varphi} + \frac{1}{r^2 \sin^2\varphi}\,\frac{\partial^2 u}{\partial \theta^2}\,, \tag{18.84}$$

where we have used our handy formula (18.14) for the Laplacian in spherical coordinates. The standard diffusive separation of variables ansatz

$$u(t, r, \varphi, \theta) = e^{-\lambda t}\, v(r, \varphi, \theta)$$

requires us to analyze the spherical coordinate form of the Helmholtz equation

$$\Delta v + \lambda\, v = \frac{\partial^2 v}{\partial r^2} + \frac{2}{r}\,\frac{\partial v}{\partial r} + \frac{1}{r^2}\,\frac{\partial^2 v}{\partial \varphi^2} + \frac{\cos\varphi}{r^2 \sin\varphi}\,\frac{\partial v}{\partial \varphi} + \frac{1}{r^2 \sin^2\varphi}\,\frac{\partial^2 v}{\partial \theta^2} + \lambda\, v = 0 \tag{18.85}$$

on the unit ball $\Omega = \{ r < 1 \}$ with homogeneous Dirichlet boundary conditions. To make further progress, we invoke a second variable separation, splitting off the radial coordinate by setting

$$v(r, \varphi, \theta) = p(r)\, w(\varphi, \theta).$$

The function w must be 2π periodic in θ and well-defined at the poles $\varphi = 0, \pi$. Substituting our ansatz into (18.85), and separating all the r-dependent terms from those terms depending upon the angular variables φ, θ leads to a pair of differential equations: The first is an ordinary differential equation

$$r^2\,\frac{d^2 p}{dr^2} + 2\,r\,\frac{dp}{dr} + (\lambda\, r^2 - \mu) p = 0, \tag{18.86}$$

for the radial component $p(r)$, while the second is a familiar partial differential equation

$$\Delta_S\, w + \mu\, w = \frac{\partial^2 w}{\partial \varphi^2} + \frac{\cos\varphi}{\sin\varphi}\,\frac{\partial w}{\partial \varphi} + \frac{1}{\sin^2\varphi}\,\frac{\partial^2 w}{\partial \theta^2} + \mu\, w = 0, \tag{18.87}$$

for its angular counterpart $w(\varphi, \theta)$. The operator Δ_S is the *spherical Laplacian* from (18.17). In Section 18.2, we showed that its eigenvalues are

$$\mu_m = m(m+1) \qquad \text{for} \qquad m = 0, 1, 2, 3, \ldots.$$

[†] In this simplified model, we are assuming that the earth is composed of a completely homogeneous and isotropic solid material.

The m^{th} eigenvalue admits $2\,m+1$ linearly independent eigenfunctions — the spherical harmonics $Y_m^0,\dots,Y_m^m,\widetilde{Y}_m^1,\dots,\widetilde{Y}_m^m$ defined in (18.33).

The radial ordinary differential equation (18.86) can be solved by setting

$$p(r)=\sqrt{r}\,q(r).$$

We use the product rule to relate their derivatives

$$p=\frac{1}{\sqrt{r}}\,q,\qquad \frac{dp}{dr}=\frac{1}{\sqrt{r}}\,\frac{dq}{dr}-\frac{1}{2\,r^{3/2}}\,q,\qquad \frac{d^2p}{dr^2}=\frac{1}{\sqrt{r}}\,\frac{d^2q}{dr^2}-\frac{1}{r^{3/2}}\,\frac{dq}{dr}+\frac{3}{4\,r^{5/2}}\,q.$$

Substituting these expressions back into (18.86) with $\mu=\mu_m=m(m+1)$, and multiplying the resulting equation by $\sqrt{r}$, we discover that $q(r)$ must solve the differential equation

$$r^2\,\frac{d^2q}{dr^2}+r\,\frac{dq}{dr}+\Big[\,\lambda\,r^2-\big(\,m+\tfrac{1}{2}\,\big)^2\,\Big]\,q=0, \tag{18.88}$$

which turns out to be the rescaled Bessel equation (17.52) of half integer order $m+\frac{1}{2}$. As a result, the solution to (18.88) that remains bounded at $r=0$ is (up to scalar multiple) the rescaled Bessel function

$$q(r)=J_{m+1/2}\big(\,\sqrt{\lambda}\,r.\,\big)$$

The corresponding solution

$$p(r)=r^{-1/2}\,J_{m+1/2}\big(\,\sqrt{\lambda}\,r\,\big) \tag{18.89}$$

to (18.86) is important enough to warrant a special name.

Definition 18.11. The *spherical Bessel function* of order $m\ge 0$ is defined by the formula

$$S_m(x)=\sqrt{\frac{\pi}{2\,x}}\;J_{m+1/2}(x). \tag{18.90}$$

Remark: The multiplicative factor $\sqrt{\pi/2}$ is included in the definition so as to avoid annoying factors of $\sqrt{\pi}$ and $\sqrt{2}$ in subsequent formulae.

Surprisingly, unlike the Bessel functions of integer order, the spherical Bessel functions are elementary functions! According to formula (C.64), the spherical Bessel function of order 0 is

$$S_0(x)=\frac{\sin x}{x}\,. \tag{18.91}$$

The higher order spherical Bessel functions can be obtained by use of the general recurrence relation

$$S_{m+1}(x)=-\,\frac{dS_m}{dx}+\frac{m}{x}\,S_m(x), \tag{18.92}$$

which is a consequence of Proposition C.13. The next few are, therefore,

$$\begin{aligned}
S_1(x)&=-\,\frac{dS_0}{dx} &&=-\,\frac{\cos x}{x}+\frac{\sin x}{x^2}\,,\\
S_2(x)&=-\,\frac{dS_1}{dx}+\frac{S_1}{x} &&=-\,\frac{\sin x}{x}-\frac{3\cos x}{x^2}+\frac{3\sin x}{x^3}\,,\\
S_3(x)&=-\,\frac{dS_2}{dx}+\frac{2\,S_1}{x} &&=\frac{\cos x}{x}-\frac{6\sin x}{x^2}-\frac{15\cos x}{x^3}+\frac{15\sin x}{x^4}\,,
\end{aligned} \tag{18.93}$$

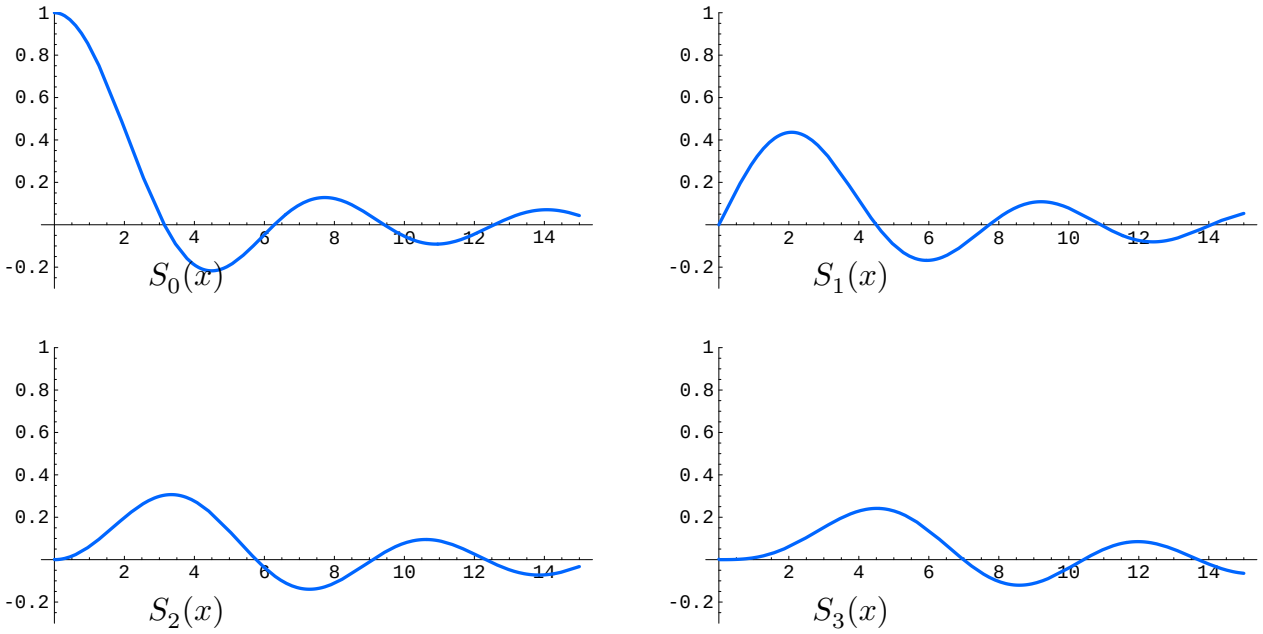


Figure 18.6. Spherical Bessel Function.

and so on. Graphs can be found in Figure 18.6. Our radial solution (18.89) is, apart from an inessential constant multiple, a rescaled spherical Bessel function of order m:

$$v_m(r) = S_m\big(\sqrt{\lambda}\, r\,\big)\,.$$

So far, we have not taken into account the (homogeneous) Dirichlet boundary condition at $r = 1$. This requires

$$p(1) = 0, \qquad \text{and hence} \qquad S_m\big(\sqrt{\lambda}\,\big) = 0.$$

Therefore, $\sqrt{\lambda}$ must be a root of the m^{th} order spherical Bessel function. We introduce the notation

$$0 < \sigma_{m,1} < \sigma_{m,2} < \sigma_{m,e} < \ \cdots$$

to denote the successive (positive) *spherical Bessel roots*, satisfying

$$S_m(\sigma_{m,n}) = 0 \qquad \text{for} \qquad n = 1, 2, \ \dots\ . \tag{18.94}$$

In particular the roots of the zero$^{\text{th}}$ order spherical Bessel function $S_0(x) = x^{-1}\sin x$ are just the integer multiples of π:

$$\sigma_{0,n} = n\,\pi \qquad \text{for} \qquad n = 1, 2, \ \dots\ .$$

A table of all spherical Bessel roots that are < 13 appears above. The columns of the table are indexed by m, the order, while the rows are indexed by n, the root number.

Re-assembling the individual constituents, we have now demonstrated that the separable eigenfunctions of the Helmholtz equation on a solid ball of radius 1, when subject

Spherical Bessel Roots $\sigma_{m,n}$

n \ m	0	1	2	3	4	5	6	7	
1	3.1416	4.4934	5.7635	8.1826	9.3558	10.5128	11.6570	12.7908	...
2	6.2832	7.7253	9.0950	11.7049	12.9665	⋮	⋮	⋮	
3	9.4248	10.9041	12.3229	⋮	⋮				
4	12.5664	⋮	⋮						
⋮	⋮								

to homogeneous Dirichlet boundary conditions, are products of spherical Bessel functions and spherical harmonics,

$$v_{k,m,n}(r,\varphi,\theta) = S_m(\sigma_{m,n}\, r)\, Y_m^k(\varphi,\theta), \qquad \widetilde{v}_{k,m,n}(r,\varphi,\theta) = S_m(\sigma_{m,n}\, r)\, \widetilde{Y}_m^k(\varphi,\theta). \quad (18.95)$$

The corresponding eigenvalues

$$\lambda_{m,n} = \sigma_{m,n}^2, \qquad m = 0, 1, 2, \dots, \qquad n = 1, 2, 3, \dots\,, \quad (18.96)$$

are given by the squared spherical Bessel roots. Since there are $2m+1$ independent spherical harmonics of order m, the eigenvalue $\lambda_{m,n}$ admits $2m+1$ linearly independent eigenfunctions, namely $v_{0,m,n}, \dots, v_{m,m,n}, \widetilde{v}_{1,m,n}, \dots, \widetilde{v}_{m,m,n}$. In particular, the radially symmetric solutions are the eigenfunctions with $k = m = 0$, namely

$$v_n(r) = v_{0,0,n}(r) = S_0(\sigma_{n,0}\, r) = \frac{\sin n\pi\, r}{n\pi\, r}, \qquad n = 1, 2, \dots\,. \quad (18.97)$$

Further analysis demonstrates that the separable solutions (18.95) form a complete system of eigenfunctions for the Helmholtz equation on the unit ball subject ot homogeneous Dirichlet boundary conditions, cf. [**46**].

We have thus completely determined the basic separable solutions to the heat equation on a solid unit ball subject to homogeneous Dirichlet boundary conditions. They are products of exponential functions of time, spherical Bessel functions of the radius and spherical harmonics:

$$\begin{aligned} u_{k,m,n}(t,r,\varphi,\theta) &= e^{-\sigma_{m,n}^2\, t}\; S_m(\sigma_{m,n}\, r)\, Y_m^k(\varphi,\theta), \\ \widetilde{u}_{k,m,n}(t,r,\varphi,\theta) &= e^{-\sigma_{m,n}^2\, t}\; S_m(\sigma_{m,n}\, r)\, \widetilde{Y}_m^k(\varphi,\theta). \end{aligned} \quad (18.98)$$

The general solution can be written as an infinite "Fourier–Bessel–spherical harmonic"

series in these fundamental modes

$$u(t,r,\theta,\varphi) = \sum_{m=0}^{\infty} \sum_{n=1}^{\infty} e^{-\sigma_{m,n}^2\, t}\, S_m(\sigma_{m,n}\, r) \left(\frac{c_{0,m,n}}{2}\, Y_m^0(\varphi,\theta) \;+ \right. \\ \left. + \sum_{k=1}^{m} \Big[c_{k,m,n} Y_m^k(\varphi,\theta) + \widetilde{c}_{k,m,n} \widetilde{Y}_m^k(\varphi,\theta) \Big] \right). \tag{18.99}$$

The series' coefficients $c_{k,m,n}, \widetilde{c}_{k,m,n}$ are uniquely prescribed by the initial data; explicit formulae follow from the usual orthogonality relations among the eigenfunctions. Detailed formulae are relegated to the exercises. In particular, the slowest decaying mode is the spherically symmetric function

$$u_{0,0,1}(t,r) = \frac{e^{-\pi^2\, t}\, \sin \pi\, r}{r} \tag{18.100}$$

corresponding to the smallest eigenvalue $\lambda_{0,1} = \sigma_{0,1}^2 = \pi^2$. Therefore, typically, the decay to thermal equilibrium of a unit sphere is at an exponential rate of $\pi^2 \approx 9.8696$, or, to a very rough approximation, 10.

The Fundamental Solution to the Heat Equation

For the heat equation (as well as more general diffusion equations), the fundamental solution measures the response of the body to a concentrated unit heat source. Thus, given a point $\boldsymbol{\xi} = (\xi, \eta, \zeta) \in \Omega$ within the body, the fundamental solution

$$u(t,\mathbf{x}) = F(t,\mathbf{x};\boldsymbol{\xi}) = F(t,x,y,z;\xi,\eta,\zeta)$$

solves the initial-boundary value problem

$$u_t = \Delta u, \qquad u(0,\mathbf{x}) = \delta(\mathbf{x} - \boldsymbol{\xi}), \qquad \text{for} \qquad \mathbf{x} \in \Omega, \quad t > 0, \tag{18.101}$$

subject to the selected homogeneous boundary conditions — Dirichlet, Neumann or mixed.

In general, the fundamental solution has no explicit formula, although in certain domains it is possible to construct it as an eigenfunction series. The one case amenable to a complete analysis is when the heat is distributed over all of three-dimensional space, so $\Omega = \mathbb{R}^3$. We recall that Lemma 17.1 showed how to construct solutions of the two-dimensional heat equation as products of one-dimensional solutions. In a similar manner, if $v(t,x)$, $w(t,x)$ and $q(t,x)$ are any three solutions to the one-dimensional heat equation $u_t = \gamma\, u_{xx}$, then their product

$$u(t,x,y,z) = p(t,x)\, q(t,y)\, r(t,z) \tag{18.102}$$

is a solution to the three-dimensional heat equation

$$u_t = \gamma\, (u_{xx} + u_{yy} + u_{zz}).$$

In particular, choosing

$$p(t,x) = \frac{e^{-(x-\xi)^2/4\gamma t}}{2\sqrt{\pi\, \gamma\, t}}, \qquad q(t,y) = \frac{e^{-(y-\eta)^2/4\gamma t}}{2\sqrt{\pi\, \gamma\, t}}, \qquad r(t,z) = \frac{e^{-(z-\zeta)^2/4\gamma t}}{2\sqrt{\pi\, \gamma\, t}},$$

to all be one-dimensional fundamental solutions, we are immediately led to the fundamental solution in the form of a three-dimensional Gaussian kernel.

Theorem 18.12. *The fundamental solution*

$$F(t, \mathbf{x}; \boldsymbol{\xi}) = F(t, \mathbf{x} - \boldsymbol{\xi}) = \frac{e^{-\| \mathbf{x} - \boldsymbol{\xi} \|^2 / 4\gamma t}}{8\,(\pi \gamma t)^{3/2}} \tag{18.103}$$

solves the three-dimensional heat equation $u_t = \gamma\, \Delta u$ *on* $\mathbb{R}^3$ *with an initial temperature equal to a delta function concentrated at the point* $\mathbf{x} = \boldsymbol{\xi}$.

Thus, the initially concentrated heat energy immediately begins to spread out in a radially symmetric manner, with a minuscule, but nonzero effect felt at arbitrarily large distances away from the initial concentration. At each individual point $\mathbf{x} \in \mathbb{R}^3$, after an initial warm-up, the temperature decays back to zero at a rate proportional to $t^{-3/2}$ — even more rapidly than in two dimensions because, intuitively, there are more directions for the heat energy to disperse.

To solve a more general initial value problem with the initial temperature $u(0, x, y, z) = f(x, y, z)$ distributed over all of space, we first write

$$f(x, y, z) = \iiint f(\boldsymbol{\xi})\, \delta(\mathbf{x} - \boldsymbol{\xi})\, d\xi\, d\eta\, d\zeta$$

as a linear superposition of delta functions. By linearity, the solution to the initial value problem is given by the corresponding superposition

$$u(t, \mathbf{x}) = \frac{1}{8\,(\pi \gamma t)^{3/2}} \iiint f(\boldsymbol{\xi})\, e^{-\| \mathbf{x} - \boldsymbol{\xi} \|^2 / 4\gamma t}\, d\xi\, d\eta\, d\zeta. \tag{18.104}$$

of the fundamental solutions. Since the fundamental solution has exponential decay as $\| \mathbf{x} \| \to \infty$, the superposition formula is valid even for initial temperature distributions which are moderately increasing at large distances. We remark that the integral (18.104) has the form of a three-dimensional convolution

$$u(t, \mathbf{x}) = F(t, \mathbf{x}) * f(\mathbf{x}) = \iiint f(\boldsymbol{\xi})\, F(t, \mathbf{x} - \boldsymbol{\xi})\, d\xi\, d\eta\, d\zeta \tag{18.105}$$

of the initial data with a one-parameter family of increasingly spread out Gaussian filters. Consequently, convolution with the Gaussian kernel has a smoothing effect on the initial temperature distribution.

18.5. The Wave Equation in Three-Dimensional Media.

The three-dimensional wave equation

$$u_{tt} = c^2 \Delta u = c^2 (u_{xx} + u_{yy} + u_{zz}), \tag{18.106}$$

in which c denotes the velocity of light, governs the propagation of electromagnetic waves (light, radio, X-rays, etc.) in a homogeneous medium, including (in the absence of gravitational effects) empty space. While the electric and magnetic vector fields $\mathbf{E}, \mathbf{B}$ are intrinsically coupled by the more complicated system of Maxwell's equations, each individual component satisfies the wave equation; see Exercise ∎ for details.

The wave equation also models certain classes[†] of vibrations of a uniform solid body. The solution $u(t, \mathbf{x}) = u(t, x, y, z)$ represents a scalar-valued displacement of the body at time t and position $\mathbf{x} = (x, y, z) \in \Omega \subset \mathbb{R}^3$. For example, $u(t, \mathbf{x})$ might represent the radial displacement of the body. One imposes suitable boundary conditions, e.g., Dirichlet, Neumann or mixed, on $\partial\Omega$, along with a pair of initial conditions

$$u(0, \mathbf{x}) = f(\mathbf{x}), \qquad \frac{\partial u}{\partial t}\,(0, \mathbf{x}) = g(\mathbf{x}), \qquad \mathbf{x} \in \Omega, \tag{18.107}$$

that specify the body's initial displacement and initial velocity. As long as the initial and boundary data are reasonably nice, there exists a unique solution to the initial-boundary value problem for all $-\infty < t < \infty$, cf. [**46**]. Thus, in contrast to the heat equation, one can follow solutions to the wave equation both forwards and backwards in time; see also Exercise ■.

Let us fix our attnetion on the homogeneous boundary value problem. The fundamental vibrational modes are found by imposing our usual trigonometric ansatz

$$u(t, x, y, z) = \cos(\omega\, t)\; v(x, y, z).$$

Substituting into the wave equation (18.106), we discover (yet again) that $v(x, y, z)$ must be an eigenfunction solving the associated Helmholtz eigenvalue problem

$$\Delta v + \lambda\, v = 0, \qquad \text{where} \qquad \lambda = \frac{\omega^2}{c^2}\,, \tag{18.108}$$

coupled to the relevant boundary conditions. In the positive definite cases, i.e., Dirichlet and mixed boundary conditions, the eigenvalues $\lambda_k = \omega_k^2/c^2 > 0$ are all positive; each eigenfunction $v_k(x, y, z)$ yields two normal vibrational modes

$$u_k(t, x, y, z) = \cos(\omega_k\, t)\; v_k(x, y, z), \qquad \widetilde{u}_k(t, x, y, z) = \sin(\omega_k\, t)\; v_k(x, y, z),$$

of frequency $\omega_k = c\,\sqrt{\lambda_k}$ equal to the square root of the corresponding eigenvalue multiplied by the wave speed. The general solution is a quasi-periodic linear combination

$$u(t, x, y, z) = \sum_{k=1}^{\infty} \big(\, a_k \cos \omega_k\, t + b_k \sin \omega_k\, t \,\big)\; v_k(x, y, z) \tag{18.109}$$

of these fundamental vibrational modes. The coefficients a_k, b_k are uniquely prescribed by the initial conditions (18.107). Thus,

$$u(0, x, y, z) = \sum_{k=1}^{\infty} a_k\, v_k(x, y, z) = f(x, y, z),$$
$$\frac{\partial u}{\partial t}\,(0, x, y, z) = \sum_{k=1}^{\infty} \omega_k\, b_k\, v_k(x, y, z) = g(x, y, z).$$

[†] Since the solution $u(t, \mathbf{x})$ to the wave equation is scalar-valued, it cannot measure the full range of possible three-dimensional motions of a solid body. The more complicated dynamical systems governing the elastic motions of solids are discussed in Exercise ■.

The explicit formulas follow immediately from the orthogonality of the eigenfunctions:

$$a_k = \frac{\langle\, f\,, v_k\,\rangle}{\|\, v_k\,\|^2} = \frac{\displaystyle\iiint_\Omega f\, v_k\; dx\, dy\, dz}{\displaystyle\iiint_\Omega v_k^2\; dx\, dy\, dz}\,, \qquad b_k = \frac{1}{\omega_k}\,\frac{\langle\, g\,, v_k\,\rangle}{\|\, v_k\,\|^2} = \frac{\displaystyle\iiint_\Omega g\, v_k\, dx\, dy\, dz}{\omega_k \displaystyle\iiint_\Omega v_k^2\; dx\, dy\, dz}\,. \tag{18.110}$$

In the positive semi-definite Neumann boundary value problem, there is an additional zero eigenvalue $\lambda_0 = 0$ corresponding to the constant null eigenfunction $v_0(x, y, z) \equiv 1$. This results in two additional terms in the eigenfunction expansion — a constant term

$$a_0 = \frac{1}{\operatorname{vol}\Omega} \iiint_\Omega f(x, y, z)\, dx\, dy\, dz$$

that equals the average initial displacement, and an unstable mode $b_0\, t$ that grows linearly in time, whose speed

$$b_0 = \frac{1}{\operatorname{vol}\Omega} \iiint_\Omega g(x, y, z)\, dx\, dy\, dz$$

is the average of the initial velocity over the entire body. The unstable mode will be excited if and only if there is a non-zero net initial velocity: $b_0 \neq 0$.

Most of the basic solution techniques we learned in the two-dimensional case apply here, and we will not dwell on the details. The case of a rectangular box is a particularly straightforward application of the method of separation of variables, and is outlined in the exercises. A similar analysis, now in cylindrical coordinates, can be applied to the case of a vibrating cylinder. The most interesting case is that of a solid spherical ball, which is the subject of the next subsection.

Vibrations of a Ball

Let us focus on the radial vibrations of a solid ball, as modeled by the three-dimensional wave equation (18.106). The solution $u(t, x, y, z)$ represents the radial displacement of the "atom" that is situated at position (x, y, z) when the ball is at rest.

For simplicity, we look at the Dirichlet boundary value problem on the unit ball $B_1 = \{\, \|\, \mathbf{x}\,\| < 1\,\}$. The normal modes of vibration are governed by the Helmholtz equation (18.108) subject to homogeneous Dirichlet boundary conditions. According to (18.95), the eigenfunctions are

$$\begin{aligned} &v_{k,m,n}(r, \varphi, \theta) = S_n(\sigma_{n,m}\, r)\, Y_m^k(\varphi, \theta), \\ &\widetilde{v}_{k,m,n}(r, \varphi, \theta) = S_m(\sigma_{m,n}\, r)\, \widetilde{Y}_m^k(\varphi, \theta), \end{aligned} \qquad \text{for} \qquad \begin{aligned} &n = 1, 2, 3, \dots\,, \\ &m = 0, 1, 2, \dots\,, \\ &k = 0, 1, \dots, m. \end{aligned} \tag{18.111}$$

Here S_m denotes the m^{th} order spherical Bessel function (18.90), $\sigma_{m,n}$ is its n^{th} root, while $Y_n^m, \widetilde{Y}_n^m$ are the spherical harmonics (18.33). Each eigenvalue

$$\lambda_{m,n} = \sigma_{m,n}^2, \qquad m = 0, 1, 2, \dots, \qquad n = 1, 2, 3, \dots,$$

corresponds to $2m+1$ independent eigenfunctions, namely

$$v_{k,m,0}(r,\varphi,\theta), \quad v_{k,m,1}(r,\varphi,\theta), \quad \dots \quad v_{k,m,m}(r,\varphi,\theta), \quad \widetilde{v}_{k,m,1}(r,\varphi,\theta), \quad \dots \quad \widetilde{v}_{k,m,m}(r,\varphi,\theta).$$

Consequently, the fundamental vibrational frequencies of a solid ball

$$\omega_{m,n} = c\sqrt{\lambda_{m,n}} = c\,\sigma_{m,n}, \qquad m = 0,1,2,\dots, \qquad n = 1,2,3,\dots, \tag{18.112}$$

are equal to the spherical Bessel roots $\sigma_{m,n}$ multiplied by the wave speed. There are a total of $2(2m+1)$ independent vibrational modes associated with each distinct frequency (18.112), namely

$$\begin{aligned}
u_{k,m,n}(t,r,\varphi,\theta) &= \cos(c\,\sigma_{m,n}\,t)\; S_m(\sigma_{m,n}\,r)\, Y_m^k(\varphi,\theta),\\
\widehat{u}_{k,m,n}(t,r,\varphi,\theta) &= \sin(c\,\sigma_{m,n}\,t)\; S_m(\sigma_{m,n}\,r)\, Y_m^k(\varphi,\theta),\\
\widetilde{u}_{k,m,n}(t,r,\varphi,\theta) &= \cos(c\,\sigma_{m,n}\,t)\; S_m(\sigma_{m,n}\,r)\, \widetilde{Y}_m^k(\varphi,\theta),\\
\widehat{\widetilde{u}}_{k,m,n}(t,r,\varphi,\theta) &= \sin(c\,\sigma_{m,n}\,t)\; S_m(\sigma_{m,n}\,r)\, \widetilde{Y}_m^k(\varphi,\theta).
\end{aligned} \qquad \begin{array}{l} n = 1,2,3,\dots,\\ m = 0,1,2,\dots,\\ k = 0,1,\dots,m. \end{array} \tag{18.113}$$

In particular, the radially symmetric modes of vibration have, according to (18.91), the elementary form

$$\begin{aligned}
u_{0,0,n}(r,\varphi,\theta) &= \cos(c\,n\,\pi\,t)\; S_0(n\,\pi\,r) = \frac{\cos c\,n\,\pi\,t\;\sin n\,\pi\,r}{r},\\
\widehat{u}_{0,0,n}(r,\varphi,\theta) &= \sin(c\,n\,\pi\,t)\; S_0(n\,\pi\,r) = \frac{\sin c\,n\,\pi\,t\;\sin n\,\pi\,r}{r},
\end{aligned} \qquad k = 1,2,3,\dots. \tag{18.114}$$

Their vibrational frequencies, $\omega_{0,n} = c\,n\,\pi$, are integral multiples of the lowest frequency $\omega_{0,1} = \pi$. Therefore, interestingly, if you only excite the radially symmetric modes, the resulting motion of the ball is periodic.

More generally, adopting the same scaling argument as in (17.111), we conclude that the fundamental frequencies for a solid ball of radius R and wave speed c are given by $\omega_{m,n} = c\,\sigma_{m,n}/R$. The relative vibrational frequencies

$$\frac{\omega_{m,n}}{\omega_{1,0}} = \frac{\sigma_{m,n}}{\sigma_{1,0}} = \frac{\sigma_{m,n}}{\pi} \tag{18.115}$$

are independent of the size of the ball R or the wave speed c. In the accompanying table, we display all relative vibrational frequencies that are less than 4 in magnitude.

The purely radial modes of vibration (18.114) have individual frequencies

$$\omega_{0,n} = \frac{n\,\pi\,c}{R}, \qquad \text{so} \qquad \frac{\omega_{0,n}}{\omega_{0,1}} = n,$$

and appear in the first column of the table. The lowest frequency is $\omega_{0,1} = \pi\,c/R$, corresponding to a vibration with period $2\pi/\omega_{0,1} = 2\,R/c$. In particular, for the earth, the radius $R \approx 6,000$ km and the wave speed in rock is, on average, $c \approx 5$ km/sec, so that the fundamental mode of vibration has period $2\,R/c \approx 2400$ seconds, or 40 minutes. Vibrations of the earth are also known as *seismic waves* and, of course, earthquakes are their

Relative Spherical Bessel Roots $\sigma_{k,m}/\sigma_{1,0}$

n \ m	0	1	2	3	4	6	7	8	...
1	1.0000	1.4303	1.8346	2.2243	2.6046	2.9780	3.3463	3.7105	...
2	2.0000	2.4590	2.8950	3.3159	3.7258	⋮	⋮	⋮	
3	3.0000	3.4709	3.9225	⋮	⋮				
4	4.0000	⋮	⋮						
⋮	⋮								

most severe manifestation. Understanding the modes of vibration is an issue of critical importance in geophysics and civil engineering, including the design of structures, buildings and bridges and the avoidance of resonant frequencies.

Of course, we have suppressed almost all interesting terrestrial geology in this very crude approximation, which has been based on the assumption that the earth is a uniform body, vibrating only in its radial direction. A more realistic modeling of the vibrations of the earth requires an understanding of the basic partial differential equations of linear and nonlinear elasticity, [**85**]. Nonuniformities in the earth lead to scattering of the vibrational waves, which are then used to locate subterranean geological structures, e.g., oil and gas deposits. We refer the interested reader to [**6**] for a comprehensive introduction to mathematical seismology.

Example 18.13. The radial vibrations of a hollow spherical shell (e.g., an elastic balloon) are governed by the differential equation

$$u_{tt} = c^2\, \Delta_S[u] = c^2 \left(\frac{\partial^2 u}{\partial \varphi^2} + \frac{\cos\varphi}{\sin\varphi}\, \frac{\partial u}{\partial \varphi} + \frac{1}{\sin^2 \varphi}\, \frac{\partial^2 u}{\partial \theta^2} \right), \tag{18.116}$$

where Δ_S denotes the spherical Laplacian (18.17). The radial displacement $u(t, \varphi, \theta)$ of a point on the sphere only depends on time t and the angular coordinates φ, θ. The solution $u(t, \varphi, \theta)$ is required to be $2\,\pi$ periodic in the azimuthal angle θ and bounded at the poles $\varphi = 0, \pi$.

According to (18.33), the n^{th} eigenvalue $\lambda_n = n(n+1)$ of the spherical Laplacian possesses $2\,n + 1$ linearly independent eigenfunctions, namely, the spherical harmonics

$$Y_n^0(\varphi, \theta), \quad Y_n^1(\varphi, \theta), \quad \dots, \quad Y_n^n(\varphi, \theta), \quad \widetilde{Y}_n^1(\varphi, \theta), \quad \dots, \quad \widetilde{Y}_n^n(\varphi, \theta).$$

As a consequence, the fundamental frequencies of vibration for a spherical shell are

$$\omega_n = c\, \sqrt{\lambda_n} \, = c\, \sqrt{n(n+1)}\,, \qquad\qquad n = 1, 2, \dots\,. \tag{18.117}$$

The vibrational solutions are quasi-periodic combinations of the fundamental spherical

harmonic modes

$$\begin{aligned} &\cos\bigl(\sqrt{n(n+1)}\,t\,\bigr)\;Y_n^m(\varphi,\theta), \qquad &&\sin\bigl(\sqrt{n(n+1)}\,t\,\bigr)\;Y_n^m(\varphi,\theta),\\ &\cos\bigl(\sqrt{n(n+1)}\,t\,\bigr)\;\widetilde{Y}_n^m(\varphi,\theta), \qquad &&\sin\bigl(\sqrt{n(n+1)}\,t\,\bigr)\;\widetilde{Y}_n^m(\varphi,\theta). \end{aligned} \tag{18.118}$$

Representative graphs can be seen in Figure 18.3. The smallest positive eigenvalue is $\lambda_1 = 2$, yielding a lowest tone of frequency $\omega_1 = c\sqrt{2}$. The higher order frequencies are irrational multiples of the fundamental frequency, implying that a vibrating spherical bell sounds percussive to our ears.

One further remark is in order. The spherical Laplacian operator is only positive semi-definite, since the lowest mode has eigenvalue $\lambda_0 = 0$, which corresponds to the constant null eigenfunction $v_0(\varphi,\theta) = Y_0^0(\varphi,\theta) \equiv 1$. Therefore, the wave equation (18.116) admits an unstable mode $b_{0,0}\,t$, corresponding to a uniform radial inflation; its coefficient

$$b_{0,0} = \frac{3}{4\pi}\iint_{S_1} \frac{\partial u}{\partial t}\,(0,\varphi,\theta)\,dS$$

represents the sphere's average initial velocity. The existence of such an unstable mode is an artifact of the simplified linear model we are using, that fails to account for nonlinearly elastic effects that serve to constrain the inflation of a spherical balloon.

18.6. Spherical Waves and Huygens' Principle.

For any dynamical (time-varying) partial differential equation, the fundamental solution measures the effect of applying an instantaneous concentrated unit impulse at a single point. Two representative physical effects to keep in mind are the light waves emanating from a sudden concentrated blast, e.g., a lightning bolt or a stellar supernova, and the sound waves due to an explosion or thunderclap, propagating in air at a much slower speed. Linear superposition utilizes the fundamental solution to build up more general solutions to initial value problems. For the wave and other second order vibrational equations, the impulse can be applied either to the initial displacement or to the initial velocity, resulting in two types of fundamental solution.

In a uniform isotropic medium, an initial concentrated blast results in a spherically expanding wave, moving away at the speed of light (or sound) in all directions. Invoking translation invariance, we will assume that the source of the disturbance is at the origin, and so the solution $u(t,\mathbf{x})$ should only depend on the distance $r = \|\,\mathbf{x}\,\|$ from the source. We adopt spherical coordinates and look for a solution $u = u(t,r)$ to the three-dimensional wave equation with no angular dependence. Substituting the formula (18.14) for the spherical Laplacian and setting both angular derivatives to 0, we are led to the following partial differential equation

$$\frac{\partial^2 u}{\partial t^2} = c^2\left(\frac{\partial^2 u}{\partial r^2} + \frac{2}{r}\,\frac{\partial u}{\partial r}\right), \tag{18.119}$$

that governs the propagation of spherically symmetric waves in three-dimensional space. Surprisingly, we can explicitly solve this partial differential equation. The secret is to

multiply both sides of the equation by r:

$$\frac{\partial^2 (r\,u)}{\partial t^2} = r\,\frac{\partial^2 u}{\partial t^2} = c^2\left(\,r\,\frac{\partial^2 u}{\partial r^2} + 2\,\frac{\partial u}{\partial r}\,\right) = c^2\,\frac{\partial^2}{\partial r^2}\,(r\,u),$$

Thus, the function

$$w(t,r) = r\,u(t,r)$$

solves the one-dimensional wave equation

$$\frac{\partial^2 w}{\partial t^2} = c^2\,\frac{\partial^2 w}{\partial r^2}\,. \tag{18.120}$$

According to Theorem 14.9, the general solution to (18.120) can be written in d'Alembert form

$$w(t,r) = p(r - c\,t) + q(r + c\,t),$$

where $p(\xi)$ and $q(\eta)$ are arbitrary functions of a single characteristic variable. Therefore, spherically symmetric solutions to the three-dimensional wave equation assume the form

$$u(t,r) \;=\; \frac{p(r - c\,t)}{r} \;+\; \frac{q(r + c\,t)}{r}\,. \tag{18.121}$$

The first term

$$u(t,r) = \frac{p(r - c\,t)}{r} \tag{18.122}$$

represents a wave moving at speed c in the direction of increasing r, and so describes the effect of a variable light source that is concentrated at the origin, e.g., a pulsating quasar in interstellar space. To highlight this interpretation, let us concentrate on the case when $p(\xi) = \delta(\xi - a)$ is a delta function, keeping in mind that more general solutions can then be assembled by linear superposition. The induced solution

$$u(t,r) = \frac{\delta(r - c\,t - a)}{r} = \frac{\delta\big(\,r - c(t - t_0)\,\big)}{r}\,, \qquad \text{where} \qquad t_0 = -\,\frac{a}{c}\,. \tag{18.123}$$

represents a spherical wave propagating through space. At the instant $t = t_0$, the light is entirely concentrated at the origin $r = 0$. The signal then moves away from the origin in all directions at speed c. At each later time $t > t_0$, the wave is concentrated on the surface of a sphere of radius $r = c\,(t - t_0)$. Its intensity at each point on the sphere, however, has decreased by a factor $1/r$, and so, the farther from the source, the dimmer the light. A stationary observer sitting at a fixed point in space will only see an instantaneous flash of light of intensity $1/r$ as the spherical wave passes by at time $t = t_0 + r/c$, where r is the observer's distance from the light source. A similar statement holds for sound waves — the sound of the explosion will only last momentarily. Thunder and lightning are the most familiar examples of this everyday phenomenon.

On the other hand, for $t < t_0$, the impulse is concentrated at a negative radius $r = c\,(t - t_0) < 0$. To interpret this, note that, for a given value of the spherical angles φ, θ, the point

$$x = r\,\sin\varphi\,\cos\theta, \qquad y = r\,\sin\varphi\,\sin\theta, \qquad z = r\,\cos\varphi,$$

for $r < 0$ lies on the antipodal point of the sphere of radius $|\, r\,|$, so that replacing r by $-r$ has the same effect as changing $\mathbf{x}$ to $-\mathbf{x}$. Thus, the solution (18.123) represents a concentrated spherically symmetric light wave arriving from the edges of the universe at speed c, that strengthens in intensity as it collapses into the origin at $t = t_0$. After collapse, it immediately reappears and expands back out into the universe.

The second solution in the d'Alembert formula (18.121) has, in fact, exactly the same physical form. Indeed, if we set

$$\widetilde{r} = -r, \qquad \widetilde{p}(\xi) = -\,q(-\xi), \qquad \text{then} \qquad \frac{q(r + c\,t)}{r} = \frac{\widetilde{p}(\widetilde{r} - c\,t)}{\widehat{r}}\,.$$

Thus, the second d'Alembert solution is redundant, and we only need to consider solutions of the form (18.122) from now on.

To effectively utilize such spherical wave solutions, we need to understand the nature of their originating singularity. For simplicity, we set $a = 0$ in (18.123) and concentrate on the particular solution

$$u(t, r) = \frac{\delta(r - c\,t)}{r}\,, \tag{18.124}$$

which has a singularity at the origin $r = 0$ when $t = 0$. We need to pin down precisely which sort of distribution this solution represents. Invoking the limiting definition is tricky, and it will be easier to work with the dual characterization of a distribution as a linear functional. Thus, at a fixed time $t \geq 0$, we must evaluate the inner product

$$\langle\, u\,,\, f\,\rangle = \iiint u(t, x, y, z)\, f(x, y, z)\, dx\, dy\, dz.$$

of the solution with a smooth test function $f(\mathbf{x}) = f(x, y, z)$. We rewrite the triple integral in spherical coordinates using the change of variables formula (B.66), whereby

$$\langle\, u\,,\, f\,\rangle = \int_0^{2\,\pi} \int_0^{\pi} \int_0^{\infty} \frac{\delta(r - c\,t)}{r}\; f(r, \varphi, \theta)\, r^2 \sin\varphi\; dr\, d\varphi\, d\theta$$

When $t \neq 0$, the r integration can be immediately computed, and so

$$\langle\, u\,,\, f\,\rangle = c\,t \int_0^{2\,\pi} \int_0^{\pi} f(c\,t, \varphi, \theta)\, \sin\varphi\; d\varphi\, d\theta = 4\,\pi\, c\,t\; \mathrm{M}^{\mathbf{0}}_{c\,t}\,[\, f\,]\,, \tag{18.125}$$

where, according to (B.41),

$$\mathrm{M}^{\mathbf{0}}_{c\,t}\,[\, f\,] = \frac{1}{4\,\pi} \int_0^{2\,\pi} \int_0^{\pi} f(c\,t, \varphi, \theta)\, \sin\varphi\; d\varphi\, d\theta = \frac{1}{4\,\pi\, c^2\, t^2} \iint_{S_{c\,t}} f\; dS \tag{18.126}$$

is the mean or average value of the function f on the sphere $S_{c\,t} = \big\{\, \|\, \mathbf{x}\,\| = c\,t\,\big\}$ of radius $r = c\,t$. In particular, the mean over the limiting sphere of radius $r = 0$ reduces to the value of the function at the origin:

$$\mathrm{M}^{\mathbf{0}}_{0}\,[\, f\,] = f(\mathbf{0}). \tag{18.127}$$

Thus, in the limit as $t \to 0$, (18.125) implies that

$$\langle\, u\,,f\,\rangle = 0 \qquad \text{for } \textit{all} \text{ functions } f,$$

and hence $u(0,r) \equiv 0$ represents a zero initial displacement.

In the absense of any intial diplacement, how, then, can the solution (18.124) be nonzero? Clearly, this must be the result of a nonzero initial velocity. To find $u_t(0,r)$, we differentiate (18.125) with respect to t, whereby

$$\begin{aligned}\left\langle \frac{\partial u}{\partial t}\,,\, f\,\right\rangle &= \frac{\partial}{\partial t}\,\langle\, u\,,f\,\rangle = \frac{\partial}{\partial t}\left(\, c\,t\,\int_0^{2\,\pi}\!\!\int_0^{\pi} f(c\,t,\varphi,\theta)\,\sin\varphi\;d\varphi\,d\theta\,\right)\\ &= c\,\int_0^{2\,\pi}\!\!\int_0^{\pi} f(c\,t,\varphi,\theta)\,\sin\varphi\;d\varphi\,d\theta + c^2\,t\,\int_0^{2\,\pi}\!\!\int_0^{\pi} \frac{\partial f}{\partial r}\,(c\,t,\varphi,\theta)\,\sin\varphi\;d\varphi\,d\theta\\ &= 4\,\pi\,c\,\mathrm{M}^{\mathbf{0}}_{c\,t}\,[\,f\,] + 4\,\pi\,c^2\,t\,\mathrm{M}^{\mathbf{0}}_{c\,t}\left[\,\frac{\partial f}{\partial r}\,\right]. \end{aligned} \tag{18.128}$$

The result is a linear combination of the mean of f and of its radial derivative f_r over the sphere of radius $c\,t$. In particular,

$$\lim_{t\,\to\,0}\,\langle\, u_t\,,f\,\rangle = 4\,\pi\,c\,\mathrm{M}^{\mathbf{0}}_{0}\,[\,f\,] = 4\,\pi\,c\,f(\mathbf{0}).$$

Since this holds for all test functions, we conclude that the initial velocity

$$u_t(0,r) = 4\,\pi\,c\,\delta(\mathbf{x})$$

is a multiple of a delta function at the origin! Dividing through by $4\,\pi\,c$, we conclude that the spherical expanding wave

$$u(t,r) = \frac{\delta(r - c\,t)}{4\,\pi\,c\,r} \tag{18.129}$$

solves the initial value problem

$$u(0,\mathbf{x}) \equiv 0, \qquad \frac{\partial u}{\partial t}\,(0,\mathbf{x}) = \delta(\mathbf{x}),$$

corresponding to an initial unit velocity impulse concentrated at the origin. This solution can be viewed as the three-dimensional version of the hammer-blow solution (14.125) to the one-dimensional wave equation. A significant difference is that, in three dimensions, there is no residual effect after the wave passes by.

More generally, if the unit impulse is concentrated at the point $\boldsymbol{\xi}$, we invoke translational symmetry to conclude that the function

$$G(t,\mathbf{x};\boldsymbol{\xi}) = \frac{\delta\big(\,\|\,\mathbf{x} - \boldsymbol{\xi}\,\| - c\,t\,\big)}{4\,\pi\,c\,\|\,\mathbf{x} - \boldsymbol{\xi}\,\|}\,, \qquad t \ge 0, \tag{18.130}$$

is the *fundamental solution* to the wave equation resulting from a concentrated unit velocity at the initial time $t = 0$:

$$G(0,\mathbf{x};\boldsymbol{\xi}) = 0, \qquad \frac{\partial G}{\partial t}\,(0,\mathbf{x};\boldsymbol{\xi}) = \delta(\mathbf{x} - \boldsymbol{\xi}). \tag{18.131}$$

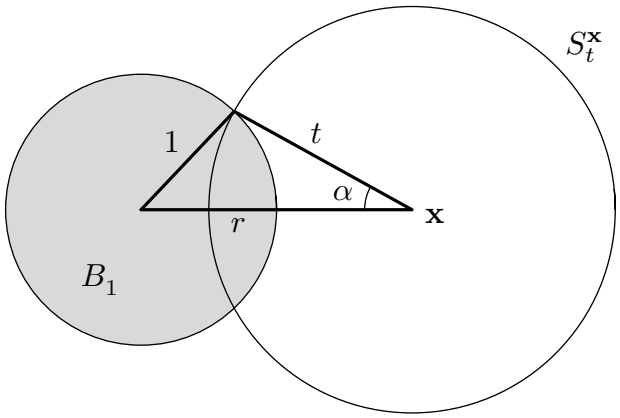


Figure 18.7. A Sphere Intersecting a Ball.

We can then apply linear superposition to solve the initial value problem

$$u(0,x,y,z) = 0, \qquad \frac{\partial u}{\partial t}\,(0,x,y,z) = g(x,y,z), \tag{18.132}$$

with zero initial displacement. Namely, we write the initial velocity

$$g(\mathbf{x}) = \iiint g(\boldsymbol{\xi})\,\delta(\mathbf{x}-\boldsymbol{\xi})\,d\xi\,d\eta\,d\zeta$$

as a superposition of impulses, and immediately conclude that the relevant solution is the self-same superposition of spherical waves:

$$u(t,\mathbf{x}) = \frac{1}{4\pi c}\iiint g(\boldsymbol{\xi})\,\frac{\delta\big(\,\|\,\mathbf{x}-\boldsymbol{\xi}\,\|-c\,t\,\big)}{\|\,\mathbf{x}-\boldsymbol{\xi}\,\|}\,d\xi\,d\eta\,d\zeta = \frac{1}{4\pi c^2\,t}\iint_{\|\,\boldsymbol{\xi}-\mathbf{x}\,\|=c\,t} g(\boldsymbol{\xi})\,dS. \tag{18.133}$$

Its value

$$u(t,\mathbf{x}) = t\,\mathrm{M}_{ct}^{\mathbf{x}}[\,g\,], \tag{18.134}$$

at a point $\mathbf{x}$ and time $t \geq 0$, is t times the average of the initial velocity function g on a sphere of radius $r = c\,t$ centered at the point $\mathbf{x}$.

Example 18.14. Let us set the wave speed $c = 1$ for simplicity. Suppose that the initial velocity

$$g(\mathbf{x}) = \begin{cases} 1, & \|\,\mathbf{x}\,\| < 1, \\ 0, & \|\,\mathbf{x}\,\| > 1 \end{cases}$$

is 1 inside the unit ball B_1 centered at the origin, and 0 outside. To solve the initial value problem, we must compute the average value of g over a sphere $S_t^{\mathbf{x}}$ of radius $t > 0$ centered at a point $\mathbf{x} \in \mathbb{R}^3$. Since $g = 0$ outside the unit ball, its average will be equal to the surface area of that part of the sphere that is contained inside the unit ball, $S_t^{\mathbf{x}} \cap B_1$, divided by the total surface area of $S_t^{\mathbf{x}}$, namely $4\pi t^2$.

To compute this quantity, let $r = \|\,\mathbf{x}\,\|$. If $t > r+1$ or $0 < t < r-1$, then the sphere of radius t lies entirely outside the unit ball, and so the average is 0; if $0 < t < 1-r$, then the sphere lies entirely within the unit ball and so the average is 1. Otherwise, referring to Figure 18.7, and referring to Exercise ∎, we see that the area of the spherical cap $S_t^{\mathbf{x}} \cap B_1$

is, by the Law of Cosines,

$$2\,\pi\,t^2(1-\cos\alpha) = 2\,\pi\,t^2\left(1-\frac{r^2+t^2-1}{2\,r\,t}\right) = \frac{\pi\,t}{r}\,[\,1-(t-r)^2\,]\,,$$

where α denotes the angle between the line joining the centers of the two spheres and the circle formed by their intersection. Assembling the different subcases, we conclude that

$$\mathrm{M}^{\mathbf{x}}_{ct}[\,g\,] = \begin{cases} 1, & 0 \le t \le 1-r, \\ \dfrac{1-(t-r)^2}{4\,r\,t}\,, & |\,r-1\,| \le t \le r+1, \\ 0, & 0 \le t \le r-1 \quad \text{or} \quad t \ge r+1. \end{cases} \tag{18.135}$$

The solution (18.134) is obtained by multiplying by t, and hence for $t \ge 0$,

$$u(t,\mathbf{x}) = \begin{cases} t, & 0 \le t \le 1-\|\,\mathbf{x}\,\|, \\ \dfrac{1-\big(t-\|\,\mathbf{x}\,\|\big)^2}{4\,\|\,\mathbf{x}\,\|}, & \big|\,\|\,\mathbf{x}\,\|-1\,\big| \le t \le \|\,\mathbf{x}\,\|+1, \\ 0, & 0 \le t \le \|\,\mathbf{x}\,\|-1 \quad \text{or} \quad t \ge \|\,\mathbf{x}\,\|+1. \end{cases} \tag{18.136}$$

Figure 18.8 plots the solution as a function of time for several fixed values of $r = \|\,\mathbf{x}\,\|$. An observer sitting at the origin will see a linearly increasing light intensity followed by a sudden decrease to 0. At other points inside the sphere, the decrease follows a parabolic arc; if the observer is closer to the edge than the center, the parabolic portion will continue to increase for a while before eventually tapering off. On the other hand, an observer sitting outside the sphere will experience, after an initially dark period, a symmetrical, parabolic increase to a maximal intensity and then decrease back to dark after a total time lapse of 2. We also show a plot of u as a function of r for various fixed times in Figure 18.9. Note that, up until time $t = 1$, the light spreads out while increasing in intensity near the origin, after which the solution is of gradually decreasing magnitude, supported within the domain lying between two concentric spheres of respective radii $t-1$ and $t+1$.

The solution described by formula (18.133) only handles initial velocities. What about solutions reuslting from a nonzero initial displacement? Surprisingly, the answer is differentiation! The key observation is that if $u(t,\mathbf{x})$ is any (sufficiently smooth) solution to the wave equation, so is its time derivative

$$v(t,\mathbf{x}) = \frac{\partial u}{\partial t}\,(t,\mathbf{x}).$$

This follows at once from differentiating both sides of the wave equation with respect to t and using the equality of mixed partial derivatives. Physically, this implies that the velocity of a wave obeys the same evolutionary principle as the wave itself, which is a manifestation of the linearity and time-independence (autonomy) of the equation. Suppose u has initial conditions

$$u(0,\mathbf{x}) = f(\mathbf{x}), \qquad\qquad u_t(0,\mathbf{x}) = g(\mathbf{x}).$$

What are the initial conditions for its derivative $v = u_t$? Clearly, its initial displacement

$$v(0,\mathbf{x}) = u_t(0,\mathbf{x}) = g(\mathbf{x})$$

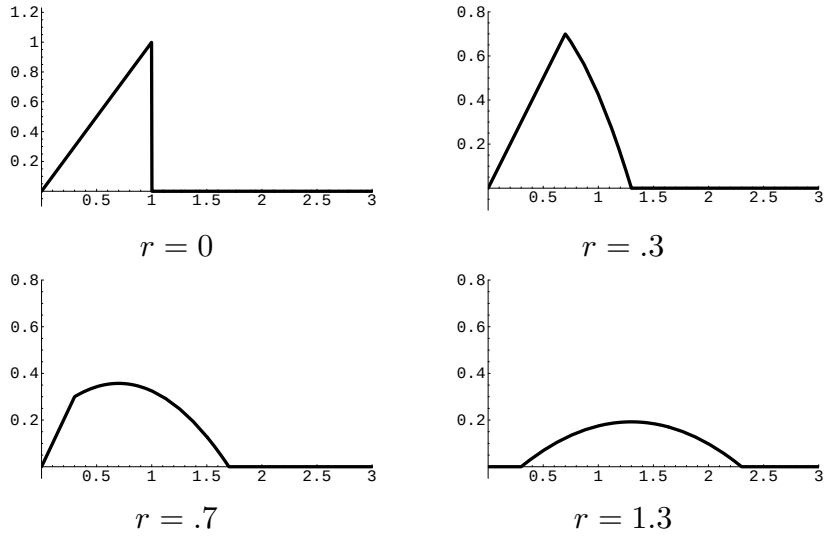


Figure 18.8. Solution to the Wave Equation Solution due to an Initial Concentrated Velocity.

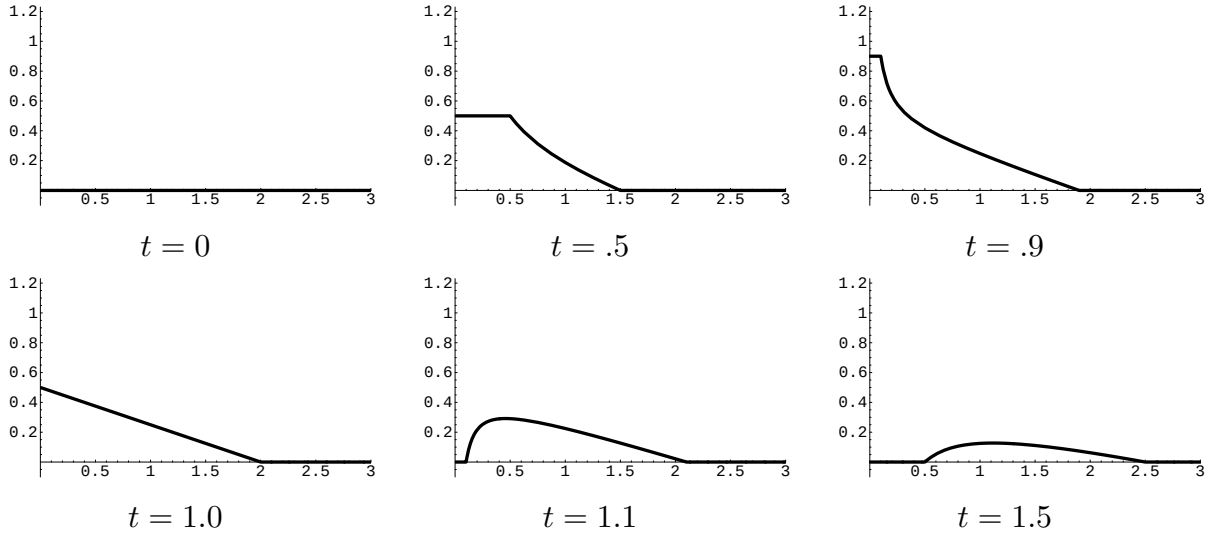


Figure 18.9. Solution to the Wave Equation Solution due to an Initial Concentrated Velocity.

equals the initial velocity of u. As for its initial velocity, we have

$$\frac{\partial v}{\partial t} = \frac{\partial^2 u}{\partial t^2} = c^2\, \Delta u$$

because we are assuming that u solves the wave equation. Thus, at the initial time

$$\frac{\partial v}{\partial t}\,(0, \mathbf{x}) = c^2\, \Delta u(0, \mathbf{x}) = c^2\, \Delta f(\mathbf{x})$$

equals c^2 times the Laplacian of the initial displacement[†]. In particular, if u satisfies the initial conditions

$$u(0, \mathbf{x}) = 0, \qquad u_t(0, \mathbf{x}) = g(\mathbf{x}), \tag{18.137}$$

then $v = u_t$ satisfies the initial conditions

$$v(0, \mathbf{x}) = g(\mathbf{x}), \qquad v_t(0, \mathbf{x}) = 0. \tag{18.138}$$

Thus, paradoxically, to solve the initial displacement problem we differentiate the initial velocity solution (18.133) with respect to t, and hence

$$v(t, \mathbf{x}) = \frac{\partial u}{\partial t}\,(t, \mathbf{x}) = \frac{\partial}{\partial t}\,\big(\, t\, \mathrm{M}^{\mathbf{x}}_{ct}\,[\, g\,]\,\big) = \mathrm{M}^{\mathbf{x}}_{ct}\,[\, g\,] + c\, t\, \mathrm{M}^{\mathbf{x}}_{ct}\left[\, \frac{\partial g}{\partial \mathbf{n}}\,\right], \tag{18.139}$$

using our computation in (18.128). Therefore, $v(t, \mathbf{x})$ is a linear combination of the mean of the function g and the mean of its normal or radial derivative $\partial g/\partial \mathbf{n} = \partial g/\partial r$, taken over a sphere of radius $c\, t$ centered at the point $\mathbf{x}$. In particular, to obtain the solution corresponding to a concentrated initial displacement,

$$F(0, \mathbf{x}; \boldsymbol{\xi}) = \delta(\mathbf{x} - \boldsymbol{\xi}), \qquad \frac{\partial F}{\partial t}\,(0, \mathbf{x}; \boldsymbol{\xi}) = 0, \tag{18.140}$$

we differentiate the solution (18.130), resulting in

$$F(t, \mathbf{x}; \boldsymbol{\xi}) = \frac{\partial G}{\partial t}\,(t, \mathbf{x}; \boldsymbol{\xi}) = -\,\frac{\delta'\big(\,\|\,\mathbf{x} - \boldsymbol{\xi}\,\| - c\, t\,\big)}{4\,\pi\,\|\,\boldsymbol{\xi} - \mathbf{x}\,\|}\,, \tag{18.141}$$

which represents a spherically expanding doublet, cf. Figure 11.10. Thus, interestingly, a concentrated initial displacement spawns an expanding spherical doublet wave, whereas a concentrated initial velocity spawns a spherical singlet or delta wave.

Example 18.15. Let $c = 1$ for simplicity. Consider the initial displacement

$$u(0, \mathbf{x}) = f(\mathbf{x}) = \begin{cases} 1, & \|\,\mathbf{x}\,\| < 1, \\ 0, & \|\,\mathbf{x}\,\| > 1 \end{cases}$$

along with zero initial velocity, modeling the effect of an instantaneously illuminated solid ball. To obtain the solution, we differentiate (18.136) with respect to t, leading to

$$u(t, \mathbf{x}) = \begin{cases} 1, & 0 \le t < 1 - \|\,\mathbf{x}\,\|, \\ \dfrac{\|\,\mathbf{x}\,\| - t}{2\,\|\,\mathbf{x}\,\|}, & \big|\,\|\,\mathbf{x}\,\| - 1\,\big| \le t \le \|\,\mathbf{x}\,\| + 1, \\ 0, & 0 \le t < \|\,\mathbf{x}\,\| - 1 \quad \text{or} \quad t > 1 + \|\,\mathbf{x}\,\|. \end{cases} \tag{18.142}$$

As illustrated in Figure 18.10, an observer sitting at the center of the ball will see a constant light intensity until $t = 1$, at which time the solution suddenly goes dark. At

[†] A similar device is used to initiate the numerical solution method for the wave equation; see Section 14.6.

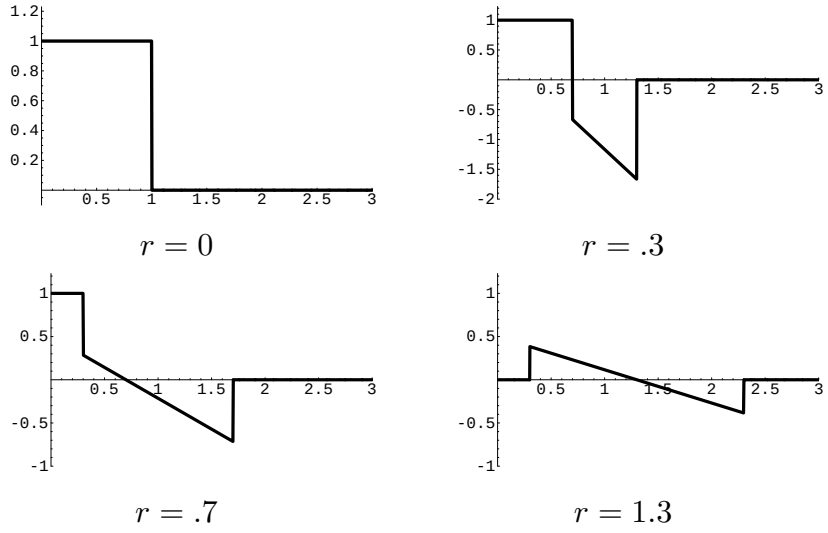


Figure 18.10. Solution to the Wave Equation due to an Initial Concentrated Displacement.

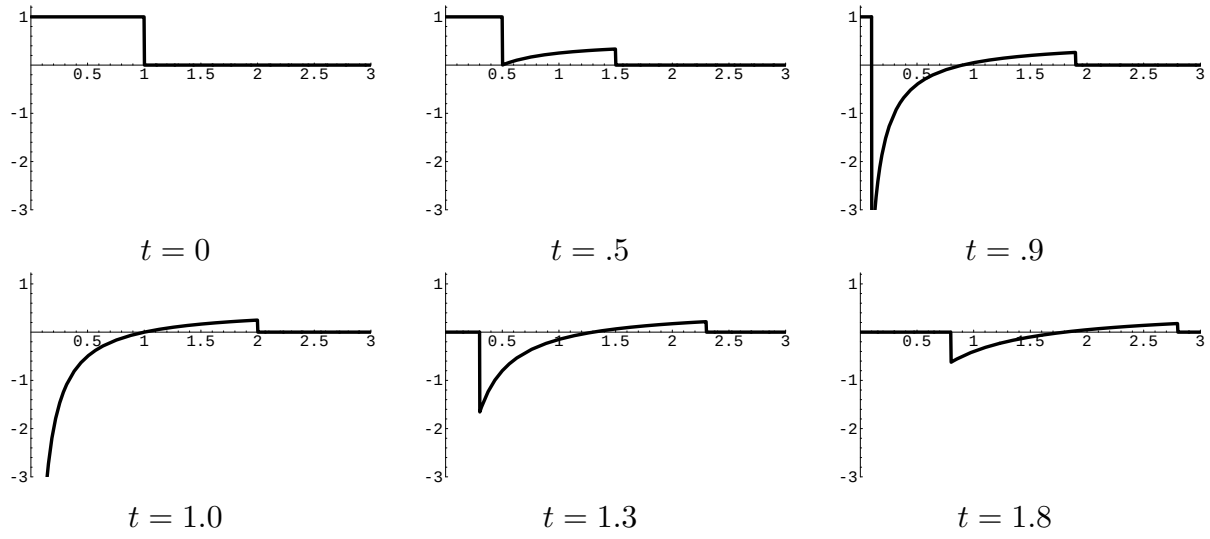


Figure 18.11. Solution to the Wave Equation due to

other points inside the ball, $0 < r < 1$, the downwards jump in intensity arrives sooner, and even goes below 0, followed by a further linear decrease, and finally a jump back to quiescent. An observer placed outside the ball will experience, after an initially dark period, a sudden increase in the light intensity, followed by a linear decrease to negative, followed by a jump back up to darkness. The farther away from the source, the fainter the light. In Figure 18.11 we plot the same solution as a function of r for different values of t. Note the sudden appearnace of a $1/r$ singularity at the origin at time $t = 1$, due to the focussing of the initial discontinuities in u over the entire unit sphere. Afterwards, the residual disturbance moves off to ∞ while gradually decreasing in intensity.

Linearly combining the two solutions (18.134, 139) establishes *Kirchhoff's formula* — although it was first discovered by Poisson — which is the three-dimensional counterpart

to the d'Alembert's solution formula for the wave equation.

Theorem 18.16. *The solution to the initial value problem*

$$u_{tt} = c^2\,\Delta u, \qquad u(0,\mathbf{x}) = f(\mathbf{x}), \qquad \frac{\partial u}{\partial t}\,(0,\mathbf{x}) = g(\mathbf{x}), \qquad \mathbf{x} \in \mathbb{R}^3, \tag{18.143}$$

for the wave equation in three-dimensional space is given by

$$u(t,\mathbf{x}) = \frac{\partial}{\partial t}\left(t\,\mathrm{M}^{\mathbf{x}}_{ct}\,[\,f\,]\right) \;+\; t\,\mathrm{M}^{\mathbf{x}}_{ct}\,[\,g\,] = \mathrm{M}^{\mathbf{x}}_{ct}\,[\,f\,] + c\,t\,\mathrm{M}^{\mathbf{x}}_{ct}\left[\,\frac{\partial f}{\partial \mathbf{n}}\,\right] \;+\; t\,\mathrm{M}^{\mathbf{x}}_{ct}\,[\,g\,]\,. \tag{18.144}$$

Here, $\mathrm{M}^{\mathbf{x}}_{ct}\,[\,f\,]$ *denotes the average of the function* f *over a sphere of radius* $c\,t$ *centered at position* $\mathbf{x}$*.*

A crucially important consequence of the Kirchhoff solution formula is the celebrated *Huygens' Principle*, which was first highlighted the pioneering seventeenth century Dutch scientist Christiaan Huygens. Roughly, Huygens' Principle states that, in three-dimensional space, localized solutions to. the wave equation remain localized. More concretely, (18.144) implis that the value of the solution at a point $\mathbf{x}$ and time t only depends upon the values of the initial displacements and velocities at a distance $c\,t$ away. Thus, all signals propagate along the light cone

$$c^2t^2 = x^2 + y^2 + z^2$$

in four-dimensional Minkowski space-time. For electromagnetic waves, this fact lies at the foundation of special relativity. Physically, Huygens' Principle means that the light that we see at a given time t arrived from points at a distance exactly $d = c\,t$ away at time $t = 0$. In particular, a sharp, localized initial signal — whether initial displacement or initial velocity — that is concentrated near a point produces a sharp, localized response that remains concentrated on an ever expanding sphere surrounding the point. In our three-dimensional universe, we only witness the light from an explosion for a brief moment, after which if there is no subsequent light source, the view returns to darkness. Similarly, a sharp sound remains sharply concentrated, with diminishing magnitude, as it propagates through space. Remarkably, as we will show next, Huygens' Principle does not hold in a two dimensional universe! In the plane, concentrated impulses will be spread out as time progresses.

Descent to Two Dimensions

So far, we have explicitly determined the solution to the wave equation in one- and three-dimensional space. The two-dimensional case

$$u_{tt} = c^2\Delta u = c^2(u_{xx} + u_{yy}). \tag{18.145}$$

is, counter-intuitively, more complicated! For instance, seeking a radially symmetric solution $u(t,r)$ requires solving the partial differential equation

$$\frac{\partial^2 u}{\partial t^2} = c^2\left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\,\frac{\partial u}{\partial r}\right) \tag{18.146}$$

which, unlike its three-dimensional cousin (18.119), is not so easily integrated.

However, our solution to the three-dimensional problem can be easily adapted to construct a solution using the so-called *Method of Descent*. Any solution $u(t,x,y)$ to the two-dimensional wave equation (18.145) can be viewed as a solution to the three-dimensional wave equation (18.106) that does not depend upon the vertical z coordinate, whence $\partial u/\partial z = 0$. Clearly, if the three-dimensional initial data does not depend on z, then the resulting solution $u(t,x,y)$ will also be independent of z.

Consider first the zero initial displacement initial conditions

$$u(0,x,y) = 0, \qquad \frac{\partial u}{\partial t}\,(0,x,y) = g(x,y). \tag{18.147}$$

We rewrite the solution formula (18.133) in the form of a surface integral over the sphere $S_{c\,t} = \big\{\, \| \boldsymbol{\xi} \| = c\,t \,\big\}$ centered at the origin:

$$u(t,\mathbf{x}) = \frac{1}{4\,\pi\,c^2\,t} \iint_{S_{c\,t}} g(\boldsymbol{\xi})\,dS = \frac{1}{4\,\pi\,c^2\,t} \iint_{\| \boldsymbol{\xi} \| = c\,t} g(\mathbf{x}+\boldsymbol{\xi})\,dS. \tag{18.148}$$

Imposing the condition that $g(x,y)$ does not depend upon the z coordinate, we see that the integrals over the upper and lower hemispheres

$$S^+_{c\,t} = \big\{\, \| \boldsymbol{\xi} \| = c\,t, \quad \zeta \geq 0 \,\big\}, \qquad\qquad S^-_{c\,t} = \big\{\, \| \boldsymbol{\xi} \| = c\,t, \quad \zeta \leq 0 \,\big\},$$

are identical. As in (B.45), to evaluate the upper hemispherical integral, we parametrize the upper hemisphere as the graph of

$$\zeta = \sqrt{c^2\,t^2 - \xi^2 - \eta^2} \qquad \text{over the disk} \qquad D_{c\,t} = \big\{\, \xi^2 + \eta^2 \leq c^2\,t^2 \,\big\},$$

We conclude that

$$u(t,x,y) = \frac{1}{2\,\pi\,c^2\,t} \iint_{S^+_{c\,t}} g(\mathbf{x}+\boldsymbol{\xi})\,dS = \frac{1}{2\,\pi\,c} \iint_{D_{c\,t}} \frac{g(x+\xi,y+\eta)}{\sqrt{c^2\,t^2 - \xi^2 - \eta^2}}\; d\xi\,d\eta \tag{18.149}$$

solves the initial value problem (18.147). In particular, if we take the initial velocity

$$g(x,y) = \delta(x-\xi)\,\delta(y-\eta)$$

to be a concentrated impulse, then the resulting solution is

$$G(t,x,y;\xi,\eta) = \begin{cases} \dfrac{1}{2\,\pi\,c\sqrt{c^2\,t^2 - (x-\xi)^2 - (y-\eta)^2}}\,, & (x-\xi)^2 + (y-\eta)^2 < c\,t, \\ 0, & (x-\xi)^2 + (y-\eta)^2 > c^2\,t^2. \end{cases} \tag{18.150}$$

An observer placed at position $\mathbf{x}$ will first experience a concentrated displacement singularity at time $t = \| \mathbf{x} - \boldsymbol{\xi} \|/c$. However, in contrast to the three-dimensional solution, even after the impulse passes by, the observer will continue to experience a decreasing, but non-zero signal of magnitude roughly proportional to $1/t$. In Figure 18.12, we plot the solution corresponding to a concentrated impulse at the origin, with unit wave speed $c = 1$. The first line shows the displacement at three different times as a function of $r = \| \mathbf{x} \|$; note the

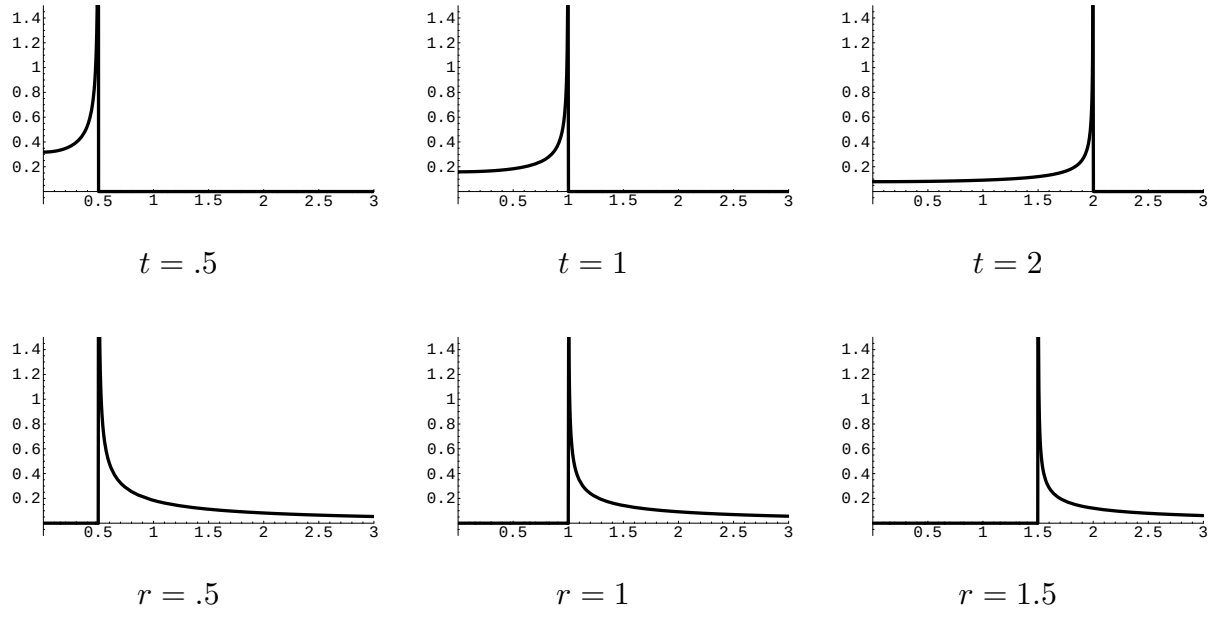


Figure 18.12. Solution to the Two-Dimensional Wave Equation for a Concentrated Impulse.

initial singularity, indicated by a spike in the graph, is followed by a progressively smaller residual displacement. The second line plots intensity as a function of t at three different radii; the further away from the initial impulse, the faster the residual displacement decays back to 0 — although it never entirely disappears.

Similarly, the solution to the initial displacement conditions

$$u(0, x, y) = f(x, y), \qquad \frac{\partial u}{\partial t}\,(0, x, y) = 0, \tag{18.151}$$

can be obtained by differentiation with respect to t, and so

$$u(t, x, y) = \frac{\partial}{\partial t}\left(\frac{1}{2\,\pi\,c} \iint_{D_{c\,t}} \frac{f(x+\xi, y+\eta)}{\sqrt{c^2\,t^2 - \xi^2 - \eta^2}}\; d\xi\, d\eta \right). \tag{18.152}$$

Again, for a concentrated impulse in the initial displacement, an observer will witness, after a certain time lapse, an abrupt impulse passing by that is followed by a progressively decaying residual effect. The general solution to the two-dimensional wave equation on all of $\mathbb{R}^2$ is a linear combination of these two types of solutions (18.149, 152).

Thus, Huygens' Principle is *not* valid in a two-dimensional universe. The solution to the two-dimensional wave equation at a point $\mathbf{x}$ at time t depends upon the initial displacement and velocity on the entire disk of radius $r\,t$ centered at the point, and not just on the points a distance $c\,t$ away. So a two-dimensional creature would experience not only a initial effect of any sound or light wave but also an "afterglow" with slowly diminishing magnitude. It would be like living in a permanent echo chamber, and so

understanding and acting upon sensory phenomena would more challenging. In general, Huygens' principle is only valid in odd-dimensional spaces; see also [**17**] for recent advances in the classification of partial differential equations that admit a Huygens' principle.

Remark: Since the solutions to the two-dimensional wave equation can be interpreted as three-dimensional solutions with no z dependence, a concentrated delta impulse in the two-dimensional wave equation would correspond to a concentrated impulse along an entire verical line in three dimensions. If light starts propagating from the line at $t = 0$, after the initial signal reaches us, we will continue to receive light from points that lie progressively farther away along the line, and this accounts for the two-dimensional afterglow.

18.7. The Schrödinger Equation and the Hydrogen Atom.

A *hydrogen atom* consists of a single electron, of mass m and charge e, circling an atomic nucleus in three-dimensional space. As a result of quantization, the Schrödinger equation governing the dynamical behavior of the electron around the nucleus takes the explicit form

$$\mathrm{i}\,\hbar\,\frac{\partial u}{\partial t} = -\,\frac{\hbar^2}{2\,m}\,\Delta u - \frac{e^2}{r}\,u = -\,\frac{\hbar^2}{2\,m}\,\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial y^2}\right) - \frac{e^2}{\sqrt{x^2+y^2+z^2}}\,u, \tag{18.153}$$

corresponding to the (attractive) potential $V(x, y, z) = e^2/r$. (This assumes the nucleus consists of a single proton; otherwise, one needs to multiply the potential $V(r)$ by Z, the atomic number of the nucleus, but with a silngloe electon. Incidentally, the Hamiltonians for multi-electron atoms are not hard to write down, but solving the associated Schrödinger equation is *much* more complicated, and still today a challenge for numerical algorithms on supercomputers.

According to the preceding analysis, the normal mode solutions have the form

$$u(t, x, y, z) = e^{\,\mathrm{i}\,\lambda\,t/\hbar} v(x),$$

where v is an eigenfunction of the Hamiltonian operator with eigenvalue λ, and hence

$$\frac{\hbar^2}{2\,m}\,\Delta u + \left(\lambda + \frac{e^2}{r}\right)\,u = 0. \tag{18.154}$$

The *bound states* of th atom correspond to the non-zero solutions to the eigenvalue problem with bounded L^2 norm:

$$\| u \|^2 =< \infty.$$

The eigenvalue λ specifies the energy of the bound states, and it can be shown that these only occur at negative energy levels: $\lambda < 0$. The bound states or eigenfunctions are *not* complete; we will leave the discussion of the continuous spectrum or scattering states to a more advanced treatment, [**127**, **149**]. This is related to the fact that the Hamiltonian is *not* positive definite!

We begin by rewriting the eigenvalue problem (18.154) in spherical coordinates:

$$\frac{\hbar^2}{2\,m}\,\left(\frac{\partial^2 u}{\partial r^2} + \frac{2}{r}\,\frac{\partial u}{\partial r} + \frac{1}{r^2}\,\frac{\partial^2 u}{\partial \varphi^2} + \frac{\cos\varphi}{r^2\sin\varphi}\,\frac{\partial u}{\partial \varphi} + \frac{1}{r^2\sin^2\varphi}\,\frac{\partial^2 u}{\partial \theta^2}\right) + \left(\lambda + \frac{e^2}{r}\right)\,u = 0. \tag{18.155}$$

We then separate off the radial coordinate, setting

$$u(r, \varphi, \theta) = p(r)\, w(\varphi, \theta).$$

The angular component satisfies the spherical Helmholtz equation

$$\Delta_S w + \mu\, w = \frac{\partial^2 w}{\partial \varphi^2} + \frac{\cos\varphi}{\sin\varphi}\, \frac{\partial w}{\partial \varphi} + \frac{1}{\sin^2 \varphi}\, \frac{\partial^2 w}{\partial \theta^2} + \mu\, w = 0,$$

that we already solved. The eigensolutions are spherical harmopnics which, because the quantum mechanical solutions are intrinsically complex-valued, we take in the complex form (18.41). Thus, the eigenvalue

$$\mu = l\,(l+1), \qquad \text{where the integer} \qquad l = 0, 1, 2, \dots \tag{18.156}$$

is known as the *angular quantum number*, has a total of $2l + 1$ linearly independent eigenfunctions

$$\mathcal{Y}_l^m(\theta, \varphi) = P_l^m(\cos\varphi)\, e^{\,\mathrm{i}\, m\, \theta}, \qquad \begin{array}{l} n = 0, 1, 2, \dots\,, \\ m = -l, -l+1, \dots, l. \end{array} \tag{18.157}$$

The radial eqaution corresponding to the separation constant (18.156) is

$$\frac{\hbar^2}{2\,m} \left(\frac{d^2 p}{dr^2} + \frac{2}{r}\, \frac{dp}{dr} \right) + \left(\lambda + \frac{e^2}{r} - \frac{l\,(l+1)}{r^2} \right) p = 0. \tag{18.158}$$

To eliminate the physical parameters, let's rescale the radial coordinate by setting

$$s = \alpha\, r, \qquad \text{where} \qquad \alpha = \sqrt{-\,\frac{8\,m\,\lambda}{\hbar^2}}\,, \tag{18.159}$$

where we use the fact that $\lambda < 0$. The resulting ordinary differential equation for

$$q(s) = p\left(\frac{s}{\alpha} \right)$$

is

$$\frac{d^2 q}{ds^2} + \frac{2}{s}\, \frac{dq}{ds} - \left(\frac{1}{4} - \frac{n}{s} + \frac{l\,(l+1)}{s^2} \right) q = 0, \tag{18.160}$$

where

$$n = \frac{2\,m\,e^2}{\alpha\,\hbar^2} = \frac{2\,m\,e^2}{\alpha\,\hbar^2} \sqrt{-\,\frac{m}{2\,\lambda}}\,. \tag{18.161}$$

Since we are searching for bound states, the solutions are define on $0 \le s < \infty$, and must be bounded at $s = 0$ and go to zero as $s \to \infty$. The proof of the following resulot is outlined in the exercises in Chapter C.

Theorem 18.17. *The bound state solutions of* (18.160) *only occur when* $n \geq l+1$ *is an ilnteger, and are given by*

$$q(s) = s^l\, e^{-\,s/2}\, L^{2\,l+1}_{n-l-1}(s), \tag{18.162}$$

where

$$L^k_j(s) = \sum_{i=0}^{k} \frac{(-1)^i}{i\,!} \binom{j+k}{j-i} x^i, \qquad j,k = 0,1,2,\dots, \tag{18.163}$$

are the associated Laguerre polynomials.

The resulting integer n, whose physical value was given in (18.161), is known as the *principle quantum number*. We further note that the scaling factor in (18.159) can be written as

$$\alpha = \frac{2\,m\,e^2}{n\,\hbar^2} = \frac{2}{n\,a} \qquad \text{where} \qquad a = \frac{\hbar^2}{m\,e^2} \approx .529 \times 10^{-\,10}\text{meter}$$

is called the *Bohr radius*, in honor of the pioneering Danish quantum physicist Niels Bohr. Reverting to physical coordinates, the bound states (18.162) become, up to an inessential constant multiple, the *radial wave functions*

$$p^n_l(r) = \left(\frac{2\,r}{n\,a}\right)^l e^{-\,r/(n\,a)}\, L^{2\,l+1}_{n-l-1}\left(\frac{2\,r}{n\,a}\right). \tag{18.164}$$

Comining them with the spherical harmonics (18.157) results in

$$U_{lmn}(r,\varphi,\theta) = p^n_l(r)\, \mathcal{Y}^m_l(\theta,\varphi) \tag{18.165}$$

that serve to span the eigenstates of the atom. These eigenstates depend upon three integers:

- $l = 0,1,2,3,\dots$: the angular quantum number;
- $n = l+1, l+2, l+3, \dots$: the principle quantum number;
- $m = -l, -l+1, \dots, l-1, l$: the magnetic quantum number.

The energy is the eigenvalue

$$\lambda_n = -\,\frac{e^4\,m}{2\,\hbar^2}\,\frac{1}{n^2}, \qquad n = 1,2,3,\dots.$$

The fact that the ratios between the higher energy levels and the lowest level are inverse squares of integers, $\lambda_n/\lambda_1 = 1/n^2$, was first established by Bohr. The n^{th} energy level has a total of

$$\sum_{l=0}^{n-1} (2l+1) = n^2$$

bound states with that energy.

The number of spherical harmonics governs the energy levels or orbital shells occupied by electrons in the hydrogen atom. In chemistry, the electron levels are indexed by the angular quantum number, i.e., the order l of the spherical harmonic, and traditionally

labeled by a letter in the sequence $p, s, d, f, \dots$. Thus, the two order $l = 0$ spherical harmonics correspond to the p shells; the six harmonics of order $l = 1$ are the s shells, and so on. Since electrons are allowed to have one of two possible spins, the Pauli exclusion principle tells us that each energy shell can be occupied by at most two electrons. Thus, the number of electrons that can reside in the n^{th} energy level of an atom is $2(2l + 1)$, the same as the number of linearly independent solutions to the wave equation associated with a given energy level. The configuration of energy shells and electrons in atoms are responsible for the periodic table. Thus, hydrogen has a single electron in the p shell. Helium has two electrons in the p shell. Lithium has 3 electrons, with two of them filling the first p shell and the third in the second p shell. Neon has 10 electrons filling the two p and first three s shells. And so on. The chemical properties of the elements are, to a very large extent, determined by the placement of the electrons within the different shells. See [**Chem**] for further details.

Continuous spectrum and scattering states.

Chapter 19

Nonlinear Systems

Nonlinearity is ubiquitous in physical phenomena. Fluid and plasma mechanics, gas dynamics, elasticity, relativity, chemical reactions, combustion, ecology, biomechanics, and many, many other phenomena are all governed by inherently nonlinear equations. (The one notable exception is quantum mechanics, which is a fundamentally linear theory. Recent attempts at grand unification of all fundamental physical theories, such as string theory and conformal field theory, [**83**], do venture into the nonlinear wilderness.) For this reason, an ever increasing proportion of modern mathematical research is devoted to the analysis of nonlinear systems.

Why, then, have we devoted the overwhelming majority of this text to linear mathematics? The facile answer is that nonlinear systems are vastly more difficult to analyze. In the nonlinear regime, many of the most basic questions remain unanswered: existence and uniqueness of solutions are not guaranteed; explicit formulae are difficult to come by; linear superposition is no longer available; numerical approximations are not always sufficiently accurate; etc., etc. A more intelligent answer is that a thorough understanding of linear phenomena and linear mathematics is an essential prerequisite for progress in the nonlinear arena. Therefore, in this introductory text on applied mathematics, we have no choice but to first develop the proper linear foundations in sufficient depth before we can realistically confront the untamed nonlinear wilderness. Moreover, many important physical systems are "weakly nonlinear", in the sense that, while nonlinear effects do play an essential role, the linear terms tend to dominate the physics, and so, to a first approximation, the system is essentially linear. As a result, such nonlinear phenomena are best understood as some form of perturbation of their linear approximations. The truly nonlinear regime is, even today, only sporadically modeled and even less well understood.

The advent of powerful computers has fomented a veritable revolution in our understanding of nonlinear mathematics. Indeed, many of the most important modern analytical techniques drew their inspiration from early computer forays into the uncharted nonlinear wilderness. However, despite dramatic advances in both hardware capabilities and sophisticated mathematical algorithms, many nonlinear systems — for instance, Einsteinian gravitation — still remain beyond the capabilities of today's computers.

Space limitations restrict us to providing just a brief overview of some of the most important ideas, mathematical techniques, and new physical phenomena that arise when venturing into the nonlinear realm. In this chapter, we start with iteration of nonlinear functions. Building on our experience with iterative linear systems, as developed in Chapter 10, we will discover that functional iteration, when it converges, provides a powerful mechanism for solving equations and for optimization. On the other hand, even very

simple nonconvergent nonlinear iterative systems may admit remarkably complex, chaotic behavior. The second section is devoted to basic solution techniques for nonlinear systems, and includes bisection, general iteration, and the very powerful Newton Method. The third section is devoted to finite-dimensional optimization principles, i.e., the minimization or maximization of nonlinear functions. Numerical optimization procedures rely on iterative procedures, and we concentrate on those associated with gradient descent.

19.1. Iteration of Functions.

As first noted in Chapter 10, iteration, meaning repeated application of a function, can be viewed as a *discrete dynamical system* in which the continuous time variable has been "quantized" to assume integer values. Even iterating a very simple quadratic scalar function can lead to an amazing variety of dynamical phenomena, including multiply-periodic solutions and genuine chaos. Nonlinear iterative systems arise not just in mathematics, but also underlie the growth and decay of biological populations, predator-prey interactions, spread of communicable diseases such as AIDS, and host of other natural phenomena. Moreover, many numerical solution methods — for systems of algebraic equations, ordinary differential equations, partial differential equations, and so on — rely on iteration, and so the theory underlies the analysis of convergence and efficiency of such numerical approximation schemes.

In general, an iterative system has the form

$$\mathbf{u}^{(k+1)} = \mathbf{g}(\mathbf{u}^{(k)}), \tag{19.1}$$

where $\mathbf{g}\colon \mathbb{R}^n \to \mathbb{R}^n$ is a real vector-valued function. (One can similarly treat iteration of complex-valued functions $\mathbf{g}\colon \mathbb{C}^n \to \mathbb{C}^n$, but, for simplicity, we only deal with real systems here.) A solution is a discrete collection of points[†] $\mathbf{u}^{(k)} \in \mathbb{R}^n$, in which the index $k = 0, 1, 2, 3, \dots$ takes on non-negative integer values. Chapter 10 dealt with the case when $\mathbf{g}(\mathbf{u}) = A\,\mathbf{u}$ is a linear function, necessarily given by multiplication by an $n \times n$ matrix A. In this chapter, we enlarge our scope to the nonlinear case.

Once we specify the initial iterate,

$$\mathbf{u}^{(0)} = \mathbf{c}, \tag{19.2}$$

then the resulting solution to the discrete dynamical system (19.1) is easily computed:

$$\mathbf{u}^{(1)} = \mathbf{g}(\mathbf{u}^{(0)}) = \mathbf{g}(\mathbf{c}), \quad \mathbf{u}^{(2)} = \mathbf{g}(\mathbf{u}^{(1)}) = \mathbf{g}(\mathbf{g}(\mathbf{c})), \quad \mathbf{u}^{(3)} = \mathbf{g}(\mathbf{u}^{(2)}) = \mathbf{g}(\mathbf{g}(\mathbf{g}(\mathbf{c}))), \quad \dots$$

and so on. Thus, unlike continuous dynamical systems, the existence and uniqueness of solutions is not an issue. As long as each successive iterate $\mathbf{u}^{(k)}$ lies in the domain of definition of $\mathbf{g}$ one merely repeats the process to produce the solution,

$$\mathbf{u}^{(k)} = \overbrace{\mathbf{g} \circ \cdots \circ \mathbf{g}}^{k \text{ times}}(\mathbf{c}), \qquad k = 0, 1, 2, \dots, \tag{19.3}$$

[†] The superscripts on $\mathbf{u}^{(k)}$ refer to the iteration number, and do *not* denote derivatives.

which is obtained by composing the function $\mathbf{g}$ with itself a total of k times. In other words, the solution to a discrete dynamical system corresponds to repeatedly pushing the $\mathbf{g}$ key on your calculator. For example, entering 0 and then repeatedly hitting the `cos` key corresponds to solving the iterative system

$$u^{(k+1)} = \cos u^{(k)}, \qquad u^{(0)} = 0. \tag{19.4}$$

The first 10 iterates are displayed in the following table:

k	0	1	2	3	4	5	6	7	8	9
$u^{(k)}$	0	1	.540302	.857553	.65429	.79348	.701369	.76396	.722102	.750418

For simplicity, we shall always assume that the vector-valued function $\mathbf{g}\colon \mathbb{R}^n \to \mathbb{R}^n$ is defined on all of $\mathbb{R}^n$; otherwise, we must always be careful that the successive iterates $\mathbf{u}^{(k)}$ never leave its domain of definition, thereby causing the iteration to break down. To avoid technical complications, we will also assume that $\mathbf{g}$ is at least continuous; later results rely on additional smoothness requirements, e.g., continuity of its first and second order partial derivatives.

While the solution to a discrete dynamical system is essentially trivial, understanding its behavior is definitely not. Sometimes the solution converges to a particular value — the key requirement for numerical solution methods. Sometimes it goes off to ∞, or, more precisely, the norms[†] of the iterates are unbounded: $\| \mathbf{u}^{(k)} \| \to \infty$ as $k \to \infty$. Sometimes the solution repeats itself after a while. And sometimes the iterates behave in a seemingly random, chaotic manner — all depending on the function $\mathbf{g}$ and, at times, the initial condition $\mathbf{c}$. Although all of these cases may arise in real-world applications, we shall mostly concentrate upon understanding convergence.

Definition 19.1. A *fixed point* or *equilibrium* of a discrete dynamical system (19.1) is a vector $\mathbf{u}^\star \in \mathbb{R}^n$ such that

$$\mathbf{g}(\mathbf{u}^\star) = \mathbf{u}^\star. \tag{19.5}$$

We easily see that every fixed point provides a constant solution to the discrete dynamical system, namely $\mathbf{u}^{(k)} = \mathbf{u}^\star$ for all k. Moreover, it is not hard to prove that any convergent solution necessarily converges to a fixed point.

Proposition 19.2. *If a solution to a discrete dynamical system converges,*

$$\lim_{k\to\infty} \mathbf{u}^{(k)} = \mathbf{u}^\star,$$

then the limit $\mathbf{u}^\star$ is a fixed point.

Proof: This is a simple consequence of the continuity of $\mathbf{g}$. We have

$$\mathbf{u}^\star = \lim_{k\to\infty} \mathbf{u}^{(k+1)} = \lim_{k\to\infty} \mathbf{g}(\mathbf{u}^{(k)}) = \mathbf{g}\left(\lim_{k\to\infty} \mathbf{u}^{(k)} \right) = \mathbf{g}(\mathbf{u}^\star),$$

the last two equalities following from the continuity of $\mathbf{g}$. *Q.E.D.*

[†] In view of the equivalence of norms on finite-dimensional vector spaces, cf. Theorem 3.17, any norm will do here.

For example, continuing the cosine iteration (19.4), we find that the iterates gradually converge to the value $u^\star \approx .739085$, which is the unique solution to the fixed point equation

$$\cos u = u.$$

Later we will see how to rigorously prove this observed behavior.

Of course, not every solution to a discrete dynamical system will necessarily converge, but Proposition 19.2 says that if it does, then it must converge to a fixed point. Thus, a key goal is to understand when a solution converges, and, if so, to which fixed point — if there is more than one. (In the linear case, only the actual convergence is a significant issues since most linear systems admit exactly one fixed point, namely $\mathbf{u}^\star = \mathbf{0}$.)

Fixed points are roughly divided into three classes:

- *asymptotically stable*, with the property that all nearby solutions converge to it,
- *stable*, with the property that all nearby solutions stay nearby, and
- *unstable*, almost all of whose nearby solutions diverge away from the fixed point.

Thus, from a practical standpoint, convergence of the iterates of a discrete dynamical system requires asymptotic stability of the fixed point. Examples will appear in abundance in the following sections.

Scalar Functions

As always, the first step is to thoroughly understand the scalar case, and so we begin with a discrete dynamical system

$$u^{(k+1)} = g(u^{(k)}), \qquad u^{(0)} = c, \tag{19.6}$$

in which $g\colon \mathbb{R} \to \mathbb{R}$ is a continuous, scalar-valued function. As noted above, we will assume, for simplicity, that g is defined everywhere, and so we do not need to worry about whether the iterates $u^{(0)}, u^{(1)}, u^{(2)}, \dots$ are all well-defined.

The linear case $g(u) = a\,u$ was treated in Section 10.1 — following equation (10.2). The simplest "nonlinear" case is that of an affine function

$$g(u) = a\,u + b, \tag{19.7}$$

leading to an *affine discrete dynamical system*

$$u^{(k+1)} = a\,u^{(k)} + b. \tag{19.8}$$

The only fixed point is the solution to

$$u^\star = g(u^\star) = a\,u^\star + b, \qquad \text{namely,} \qquad u^\star = \frac{b}{1-a}\,. \tag{19.9}$$

The formula for $u^\star$ requires that $a \neq 1$, and, indeed, the case $a = 1$ has no fixed point, as the reader can easily confirm; see Exercise ■.

Since we already know the value of $u^\star$, we can readily analyze the differences

$$e^{(k)} = u^{(k)} - u^\star, \tag{19.10}$$

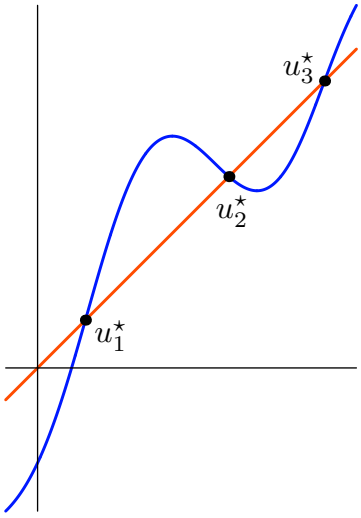


Figure 19.1. Fixed Points.

between successive iterates and the fixed point. Observe that, the smaller $e^{(k)}$ is, the closer $u^{(k)}$ is to the desired fixed point. In many applications, the iterate $u^{(k)}$ is viewed as an approximation to the fixed point $u^\star$, and so $e^{(k)}$ is interpreted as the *error* in the k^{th} iterate. Subtracting the fixed point equation (19.9) from the iteration equation (19.8), we find

$$u^{(k+1)} - u^\star = a\,(u^{(k)} - u^\star).$$

Therefore the errors $e^{(k)}$ are related by a *linear iteration*

$$e^{(k+1)} = a\,e^{(k)}, \qquad \text{and hence} \qquad e^{(k)} = a^k e^{(0)}. \tag{19.11}$$

Therefore, as we already demonstrated in Section 10.1, the solutions to this scalar linear iteration converge:

$$e^{(k)} \longrightarrow 0 \qquad \text{and hence} \qquad u^{(k)} \longrightarrow u^\star, \qquad \text{if and only if} \qquad |\,a\,| < 1.$$

This is the criterion for *asymptotic stability* of the fixed point, or, equivalently, convergence of the affine iterative system (19.8). The magnitude of a determines the rate of convergence, and the closer it is to 0, the faster the iterates approach the fixed point.

Example 19.3. The affine function

$$g(u) = \tfrac{1}{4}\,u + 2$$

leads to the iterative scheme

$$u^{(k+1)} = \tfrac{1}{4}\,u^{(k)} + 2.$$

Starting with the initial condition $u^{(0)} = 0$, the ensuing values are

k	1	2	3	4	5	6	7	8
$u^{(k)}$	2.0	2.5	2.625	2.6562	2.6641	2.6660	2.6665	2.6666

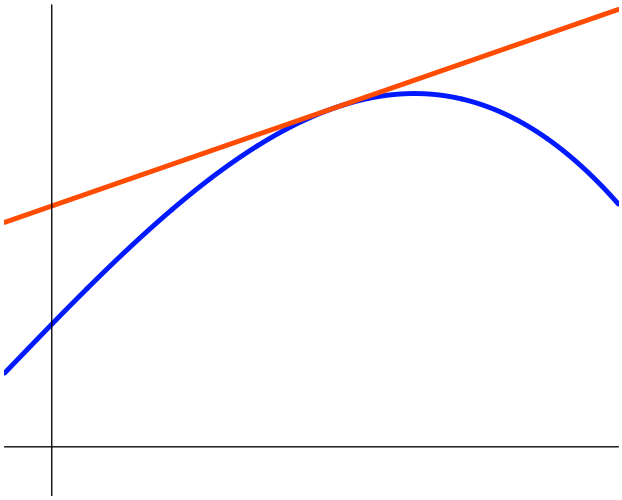


Figure 19.2. Tangent Line Approximation.

Thus, after 8 iterations, the iterates have produced the fixed point $u^\star = \frac{8}{3}$ to 4 decimal places. The rate of convergence is $\frac{1}{4}$, and indeed

$$| e^{(k)} | = | u^{(k)} - u^\star | = \left(\tfrac{1}{4}\right)^k | u^{(0)} - u^\star | = \tfrac{8}{3} \left(\tfrac{1}{4}\right)^k \longrightarrow 0 \qquad \text{as} \qquad k \longrightarrow \infty.$$

Let us now turn to the fully nonlinear case. First note that the fixed points of $g(u)$ correspond to the intersections of its graph with the graph of the function $i(u) = u$. For instance Figure 19.1 shows the graph of a function that has 3 fixed points, labeled $u_1^\star, u_2^\star, u_3^\star$.

In general, near any point in its domain, a (smooth) nonlinear function can be well approximated by its tangent line, which repre4sents the graph of an affine function; see Figure 19.2. Therefore, if we are close to a fixed point $u^\star$, then we might expect the iterative system based on the nonlinear function $g(u)$ to behave very much like that of its affine tangent line approximation. And, indeed, this intuition turns out to be essentially correct. This result forms our first concrete example of *linearization*, in which the analysis of a nonlinear system is based on its linear (or, more precisely, affine) approximation.

The explicit formula for the tangent line to $g(u)$ near the fixed point $u = u^\star = g(u^\star)$ is

$$g(u) \approx g(u^\star) + g'(u^\star)(u - u^\star) \equiv a\, u + b, \tag{19.12}$$

where

$$a = g'(u^\star), \qquad b = g(u^\star) - g'(u^\star)\, u^\star = \big(1 - g'(u^\star)\big)\, u^\star.$$

Note that $u^\star = b\,/(1 - a)$ remains a fixed point for the affine approximation: $a\, u^\star + b = u^\star$. According to the preceding discussion, the convergence of the iterates for the affine approximation is governed by the size of the coefficient $a = g'(u^\star)$. This observation inspires the basic stability criterion for fixed points of scalar iterative systems.

Theorem 19.4. *Let $g(u)$ be a continuously differentiable scalar function. Suppose $u^\star = g(u^\star)$ is a fixed point. If $| g'(u^\star) | < 1$, then $u^\star$ is an asymptotically stable fixed point, and hence any sequence of iterates $u^{(k)}$ which starts out sufficiently close to $u^\star$ will converge to $u^\star$. On the other hand, if $| g'(u^\star) | > 1$, then $u^\star$ is an unstable fixed point, and*

the only iterates which converge to it are those that land exactly on it, i.e., $u^{(k)} = u^\star$ for some $k \geq 0$.

Proof: The goal is to prove that the errors $e^{(k)} = u^{(k)} - u^\star$ between the iterates and the fixed point tend to 0 as $k \to \infty$. To this end, we try to estimate $e^{(k+1)}$ in terms of $e^{(k)}$. According to (19.6) and the Mean Value Theorem C.2 from calculus,

$$e^{(k+1)} = u^{(k+1)} - u^\star = g(u^{(k)}) - g(u^\star) = g'(v)\,(u^{(k)} - u^\star) = g'(v)\,e^{(k)}, \tag{19.13}$$

for some v lying between $u^{(k)}$ and $u^\star$. By continuity, if $|\,g'(u^\star)\,| < 1$ at the fixed point, then we can choose $\delta > 0$ and $|\,g'(u^\star)\,| < \sigma < 1$ such that the estimate

$$|\,g'(v)\,| \leq \sigma < 1 \qquad \text{whenever} \qquad |\,v - u^\star\,| < \delta \tag{19.14}$$

holds in a (perhaps small) interval surrounding the fixed point. Suppose

$$|\,e^{(k)}\,| = |\,u^{(k)} - u^\star\,| < \delta.$$

Then the point v in (19.13), which is closer to $u^\star$ than $u^{(k)}$, satisfies (19.14). Therefore,

$$|\,u^{(k+1)} - u^\star\,| \;\leq\; \sigma\,|\,u^{(k)} - u^\star\,|, \qquad \text{and hence} \qquad |\,e^{(k+1)}\,| \;\leq\; \sigma\,|\,e^{(k)}\,|. \tag{19.15}$$

In particular, since $\sigma < 1$, we have $|\,u^{(k+1)} - u^\star\,| < \delta$, and hence the subsequent iterate $u^{(k+1)}$ also lies in the interval where (19.14) holds. Repeating the argument, we conclude that, provided the initial iterate satisfies

$$|\,e^{(0)}\,| = |\,u^{(0)} - u^\star\,| < \delta,$$

the subsequent errors are bounded by

$$e^{(k)} \;\leq\; \sigma^k\,e^{(0)}, \qquad \text{and hence} \qquad e^{(k)} = |\,u^{(k)} - u^\star\,| \;\longrightarrow\; 0 \quad \text{as} \quad k \to \infty,$$

which completes the proof of the theorem in the stable case.

The proof in unstable case is left as Exercise ■ for the reader. *Q.E.D.*

Remark: The constant σ governs the rate of convergence of the iterates to the fixed point. The closer the iterates are to the fixed point, the smaller we can choose δ in (19.14), and hence the closer we can choose σ to $|\,g'(u^\star)\,|$. Thus, roughly speaking, $|\,g'(u^\star)\,|$ governs the speed of convergence, once the iterates get close to the fixed point. This observation will be developed more fully in the following subsection.

Remark: The cases when $g'(u^\star) = \pm 1$ are *not* covered by the theorem. For a linear system, such fixed points are stable, but not asymptotically stable. For nonlinear systems, more detailed knowledge of the nonlinear terms is required in order to resolve the status — stable or unstable — of the fixed point. Despite their importance in certain applications, we will not try to analyze such borderline cases any further here.

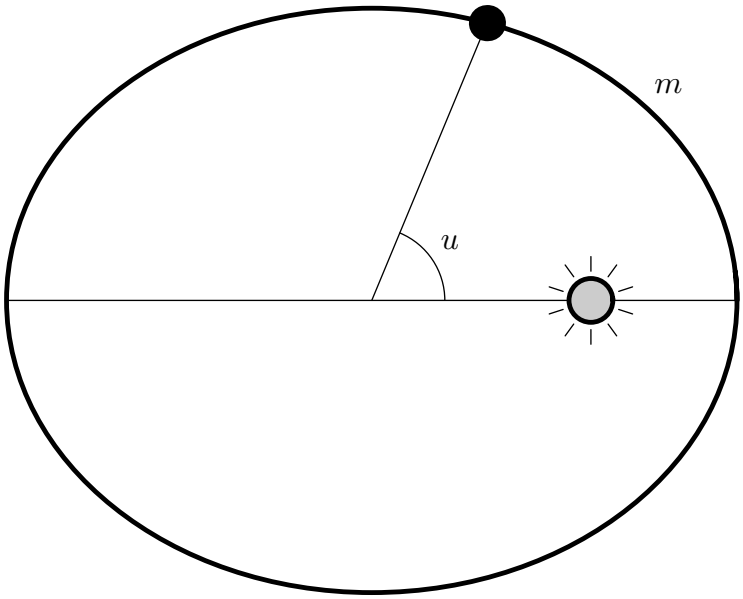


Figure 19.3. Planetary Orbit.

Example 19.5. Given constants ϵ, m, the trigonometric equation

$$u = m + \epsilon \sin u \tag{19.16}$$

is known as *Kepler's equation*. It arises in the study of planetary motion, in which $0 < \epsilon < 1$ represents the *eccentricity* of an elliptical planetary orbit, u is the *eccentric anomaly*, defined as the angle formed at the center of the ellipse by the planet and the major axis, and $m = 2\pi t\,/T$ is its *mean anomaly*, which is the time, measured in units of $T/(2\pi)$ where T is the period of the orbit, i.e., the length of the planet's year, since perihelion or point of closest approach to the sun; see Figure 19.3.

The solutions to Kepler's equation are the fixed points of the discrete dynamical system based on the function

$$g(u) = m + \epsilon \sin u.$$

Note that

$$|\, g'(u)\,| = |\, \epsilon \cos u\,| = |\,\epsilon\,| < 1, \tag{19.17}$$

which automatically implies that the as yet unknown fixed point is stable. Indeed, Exercise ■ implies that condition (19.17) is enough to prove the existence of a unique stable fixed point. In the particular case $m = \epsilon = \frac{1}{2}$, the result of iterating $u^{(k+1)} = \frac{1}{2} + \frac{1}{2}\sin u^{(k)}$ starting with $u^{(0)} = 0$ is

k	1	2	3	4	5	6	7	8	9
$u^{(k)}$	.5	.7397	.8370	.8713	.8826	.8862	.8873	.8877	.8878

After 13 iterations, we have converged sufficiently close to the solution (fixed point) $u^\star = .887862$ to have computed its value to 6 decimal places. *Remark*: In Exercise ■, we outline

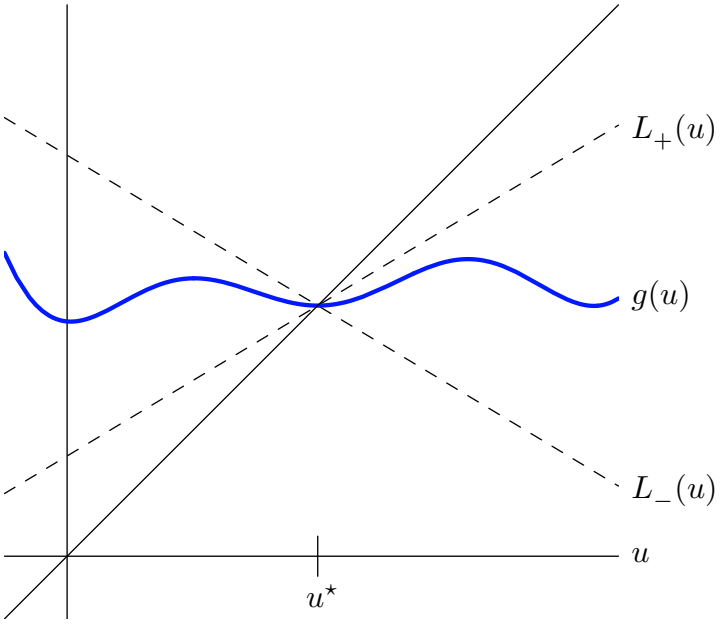


Figure 19.4. Graph of a Contraction.

Bessel's explicit Fourier series solution to Kepler's equation that led him to the original definition of Bessel functions.

Inspection of the proof of Theorem 19.4 reveals that we never really used the differentiability of g, except to verify the inequality

$$| g(u) - g(u^\star) | \leq \sigma \, | u - u^\star | \qquad \text{for some fixed } \sigma < 1. \tag{19.18}$$

A function that satisfies (19.18) for all nearby u is called a *contraction* at the point $u^\star$. *Any* function $g(u)$ whose graph lies between the two lines

$$L_\pm(u) = g(u^\star) \pm \sigma \, (u - u^\star) \qquad \text{for some} \quad \sigma < 1,$$

for all u sufficiently close to $u^\star$, i.e., such that $| u - u^\star | < \delta$ for some $\delta > 0$, defines a contraction, and hence fixed point iteration starting with $| u^{(0)} - u^\star | < \delta$ will converge to $u^\star$; see Figure 19.4. In particular, Exercise ■ asks you to prove that any function that is differentiable at $u^\star$ with $| g'(u^\star) | < 1$ defines a contraction at $u^\star$.

Example 19.6. The simplest truly nonlinear example is a quadratic polynomial. The most important case is the so-called *logistic map*

$$g(u) = \lambda \, u(1 - u), \tag{19.19}$$

where $\lambda \neq 0$ is a fixed non-zero parameter. (The case $\lambda = 0$ is completely trivial. Why?) In fact, an elementary change of variables can make any quadratic iterative system into one involving a logistic map; see Exercise ■.

The fixed points of the logistic map are the solutions to the quadratic equation

$$u = \lambda \, u(1 - u), \qquad \text{or} \qquad \lambda \, u^2 - \lambda \, u + 1 = 0.$$

Using the quadratic formula, we conclude that $g(u)$ has two fixed points:

$$u_1^\star = 0, \qquad u_2^\star = 1 - \frac{1}{\lambda}\,.$$

Let us apply Theorem 19.4 to determine their stability. The derivative is

$$g'(u) = \lambda - 2\,\lambda\,u, \qquad \text{and so} \qquad g'(u_1^\star) = \lambda, \qquad g'(u_2^\star) = 2 - \lambda.$$

Therefore, if $|\,\lambda\,| < 1$, the first fixed point is stable, while if $1 < \lambda < 3$, the second fixed point is stable. For $\lambda < -1$ or $\lambda > 3$ neither fixed point is stable, and we expect the iterates to not converge at all.

Numerical experiments with this example show that it is the source of an amazingly diverse range of behavior, depending upon the value of the parameter λ. In the accompanying Figure 19.5, we display the results of iteration starting with initial point $u^{(0)} = .5$ for several different values of λ; in each plot, the horizontal axis indicates the iterate number k and the vertical axis the iterate valoue $u^{(k)}$ for $k = 0, \dots, 100$. As expected from Theorem 19.4, the iterates converge to one of the fixed points in the range $-1 < \lambda < 3$, except when $\lambda = 1$. For λ a little bit larger than $\lambda_1 = 3$, the iterates do not converge to a fixed point. But it does not take long for them to settle down, switching back and forth between two particular values. This behavior indicates the exitence of a (stable) *period* 2 *orbit* for the discrete dynamical system, in accordance with the following definition.

Definition 19.7. A *period k orbit* of a discrete dynamical system is a solution that satisfies $u^{(n+k)} = u^{(n)}$ for all $n = 0, 1, 2, \dots$. The (*minimal*) *period* is the smallest positive value of k for which this condition holds.

Thus, a fixed point

$$u^{(0)} = u^{(1)} = u^{(2)} = \cdots$$

is a period 1 orbit. A period 2 orbit satisfies

$$u^{(0)} = u^{(2)} = u^{(4)} = \cdots \qquad \text{and} \qquad u^{(1)} = u^{(3)} = u^{(5)} = \cdots\,,$$

but $u^{(0)} \neq u^{(1)}$, as otherwise the minimal period would be 1. Similarly, a period 3 orbit has

$$u^{(0)} = u^{(3)} = u^{(6)} = \cdots\,, \qquad u^{(1)} = u^{(4)} = u^{(7)} = \cdots\,, \qquad u^{(2)} = u^{(5)} = u^{(8)} = \cdots\,,$$

with $u^{(0)}, u^{(1)}, u^{(2)}$ distinct. Stability of a period k orbit implies that nearby iterates converge to this periodic solution.

For the logistic map, the period 2 orbit persists until $\lambda = \lambda_2 \approx 3.4495$, after which the iterates alternate between four values — a period 4 orbit. This again changes at $\lambda = \lambda_3 \approx 3.5441$, after which the iterates end up alternating between eight values. In fact, there is an increasing sequence of values

$$3 = \lambda_1 < \lambda_2 < \lambda_3 < \lambda_4 < \cdots\,,$$

where, for any $\lambda_n < \lambda \leq \lambda_{n+1}$, the iterates eventually follow a period 2^n orbit. Thus, as λ passes through each value λ_n the period of the orbit goes from 2^n to $2 \cdot 2^n = 2^{n+1}$, and the

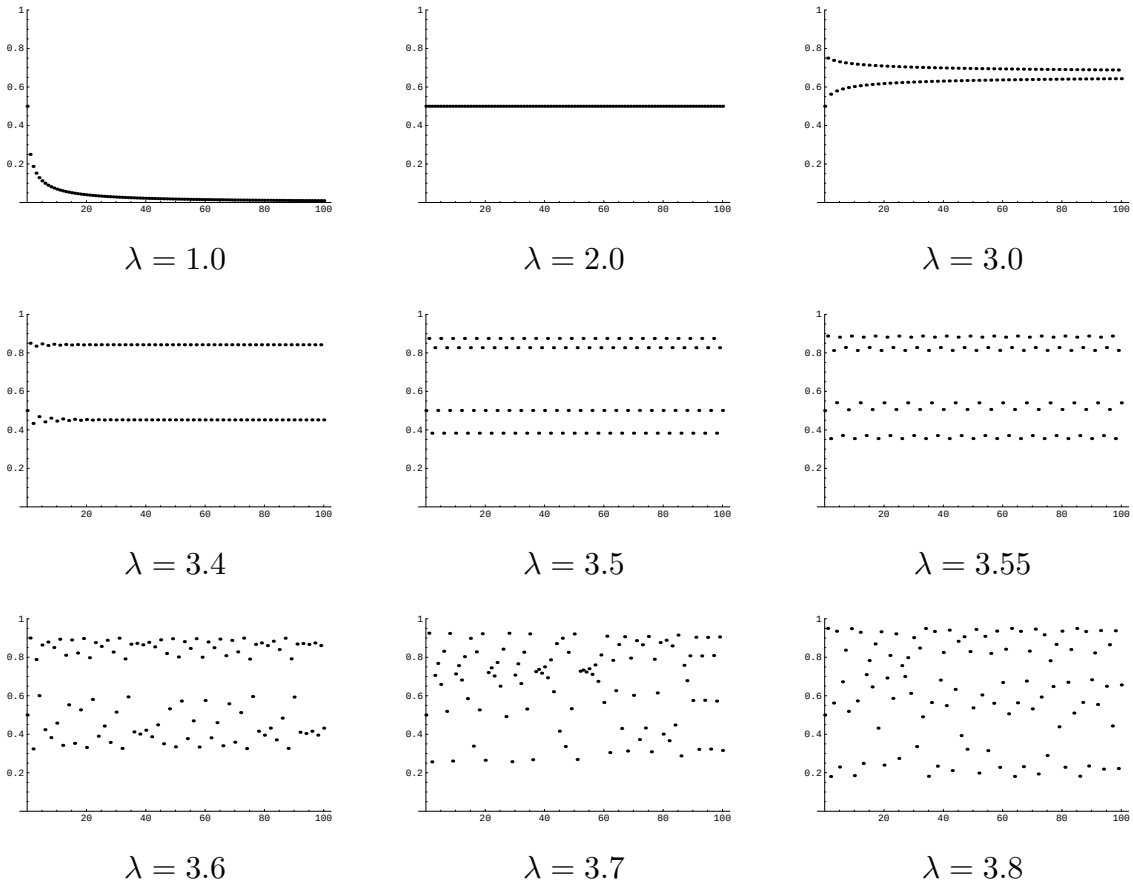


Figure 19.5. Logistic Iterates.

discrete dynamical system experiences a *bifurcation*. The bifurcation values λ_n are packed closer and closer together as n increases, piling up on an eventual limiting value

$$\lambda_\star = \lim_{n \to \infty} \lambda_n \approx 3.5699,$$

at which point the orbit's period has, so to speak, become infinitely large. The entire phenomena is known as a *period doubling cascade*.

Interestingly, the ratios of the distances between successive bifurcation points approaches a well-defined limit,

$$\frac{\lambda_{n+2} - \lambda_{n+1}}{\lambda_{n+1} - \lambda_n} \quad \longrightarrow \quad 4.6692\ldots\,, \tag{19.20}$$

known as *Feigenbaum's constant*. In the 1970's, the American physicist Mitchell Feigenbaum, [**64**], discovered that similar period doubling cascades appear in a broad range of discrete dynamical systems. Even more remarkably, in almost all cases, the corresponding ratios of distances between bifurcation points has the *same* limiting value. Feigenbaum's experimental observations were rigorously proved by Oscar Lanford in 1982, [**123**].

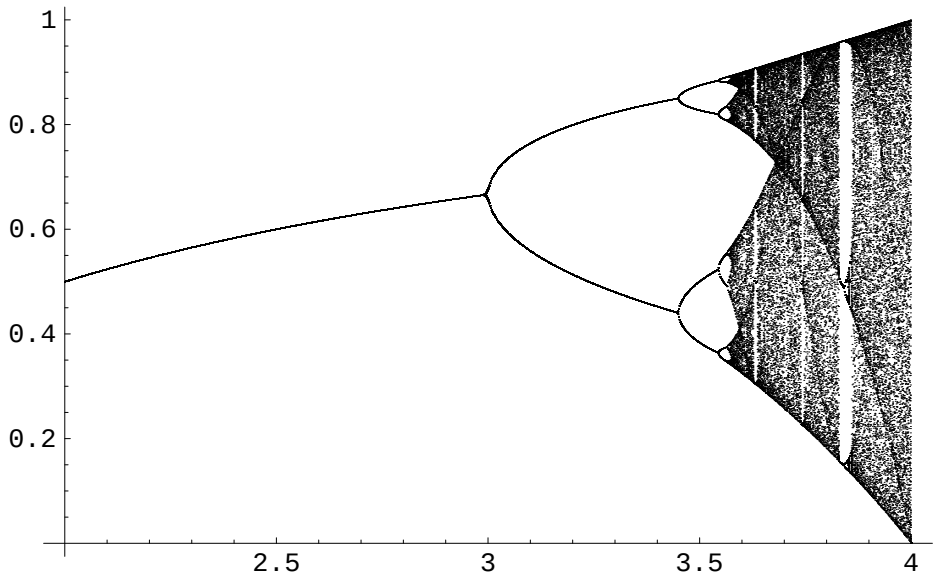


Figure 19.6. The Logistic Map.

After λ passes the limiting value $\lambda_\star$, all hell breaks loose. The iterates become completely chaotic†, moving at random over the interval $[0, 1]$. But this is not the end of the story. Embedded within this chaotic regime are certain small ranges of λ where the system settles down to a stable orbit, whose period is no longer necessarily a power of 2. In fact, there exist values of λ for which the iterates settle down to a stable orbit of period k for *any* positive integer k. For instance, as λ increases past $\lambda_{3,\star} \approx 3.83$, a period 3 orbit appears over a small range of values, after which, as λ increses slightly further, there is a period doubling cascade where period 6, 12, 24, ... orbits successively appear, each persisting on a shorter and shorter range of parameter values, until λ passes yet another critical value where chaos breaks out yet again. There is a well-prescribed order in which the periodic orbits make their successive appearance, and each odd period k orbit is followed by a very closely spaced sequence of period doubling bifurcations, of periods $2^n k$ for $n = 1, 2, 3, \ldots$, after which the iterates revert to completely chaotic behavior until the next periodic case emerges. The ratios of distances between bifurcation points always have the same Feigenbaum limit (19.20). Finally, these periodic and chaotic windows all pile up on the ultimate parameter value $\lambda^\star_\star = 4$. And then, when $\lambda > 4$, all the iterates go off to ∞, and the system ceases to be interesting.

The reader is encouraged to write a simple computer program and perform some numerical experiments. In particular, Figure 19.6 shows the asymptotic behavior of the iterates for values of the parameter in the interesting range $2 < \lambda < 4$. The horizontal axis is λ, and the marked points show the ultimate fate of the iteration for the given value of λ. For instance, each point the single curve lying above the smaller values of λ represents a stable fixed point; this bifurcates into a pair of curves representing stable

† The term "chaotic" does have a precise mathematical definition, but the reader can take it more figuratively for the purposes of this elementary exposition.

period 2 orbits, which then bifurcates into 4 curves representing period 4 orbits, and so on. Chaotic behavior is indicated by a somewhat random pattern of points lying above the value of λ. To plot this figure, we ran the logistic iteration $u^{(n)}$ for $0 \le n \le 100$, discarded the first 50 points, and then plotted the next 50 iterates $u^{(51)}, \dots, u^{(100)}$. Investigation of the fine detailed structure of the logistic map requires yet more iterations with increased numerical accuracy. In addition one should discard more of the initial iterates so as to give the system enough time to settle down to a stable periodic orbit or, alternatively, continue in a chaotic manner.

Remark: So far, we have only looked at real scalar iterative systems. Complex discrete dynamical systems display yet more remarkable and fascinating behavior. The complex version of the logistic iteration equation leads to the justly famous Julia and Mandelbrot sets, [**125**], with their stunning, psychedelic fractal structure, [**147**].

The rich range of phenomena in evidence, even in such extremely simple nonlinear iterative systems, is astounding. While intimations first appeared in the late nineteenth century research of the influential French mathematician Henri Poincaré, serious investigations were delayed until the advent of the computer era, which precipitated an explosion of research activity in the area of dynamical systems. Similar period doubling cascades and chaos are found in a broad range of nonlinear systems, [**7**], and are often encountered in physical applications, [**130**]. A modern explanation of fluid turbulence is that it is a (very complicated) form of chaos, [**7**].

Quadratic Convergence

Let us now return to the more mundane case when the iterates converge to a stable fixed point of the discrete dynamical system. In applications, we use the iterates to compute a precise[†] numerical value for the fixed point, and hence the efficiency of the algorithm depends on the speed of convergence of the iterates.

According to the remark following the proof Theorem 19.4, the convergence rate of an iterative system is essentially governed by the magnitude of the derivative $|\, g'(u^\star)\,|$ at the fixed point. The basic inequality (19.15) for the errors $e^{(k)} = u^{(k)} - u^\star$, namely

$$|\, e^{(k+1)}\,| \le \sigma\, |\, e^{(k)}\,|,$$

is known as a *linear convergence estimate*. It means that, once the iterates are close to the fixed point, the error decreases by a factor of (at least) $\sigma \approx |\, g'(u^\star)\,|$ at each step. If the k^{th} iterate $u^{(k)}$ approximates the fixed point $u^\star$ correctly to m decimal places, so its error is bounded by

$$|\, e^{(k)}\,| < .5 \times 10^{-m},$$

then the $(k+1)^{\text{st}}$ iterate satisfies the error bound

$$|\, e^{(k+1)}\,| \le \sigma\, |\, e^{(k)}\,| < .5 \times 10^{-m}\, \sigma = .5 \times 10^{-m + \log_{10} \sigma}.$$

[†] The degree of precision is to be specified by the user and the application.

More generally, for any $j > 0$,

$$| e^{(k+j)} | \leq \sigma^j \, | e^{(k)} | < .5 \times 10^{-m} \sigma^j = .5 \times 10^{-m + j \, \log_{10} \sigma},$$

which means that the $(k+j)^{\text{th}}$ iterate $u^{(k+j)}$ has at least[‡]

$$m - j \, \log_{10} \sigma = m + j \, \log_{10} \sigma^{-1}$$

correct decimal places. For instance, if $\sigma = .1$ then each new iterate produces one new decimal place of accuracy (at least), while if $\sigma = .9$ then it typically takes $22 \approx -1/\log_{10} .9$ iterates to produce just one additional accurate digit!

This means that there is a huge advantage — particularly in the application of iterative methods to the numerical solution of equations — to arrange that $| g'(u^\star) |$ be as small as possible. The fastest convergence rate of all will occur when $g'(u^\star) = 0$. In fact, in such a happy situation, the rate of convergence is not just slightly, but dramatically faster than linear.

Theorem 19.8. *Suppose that $g \in \mathrm{C}^2$, and $u^\star = g(u^\star)$ is a fixed point such that $g'(u^\star) = 0$. Then, for all iterates $u^{(k)}$ sufficiently close to $u^\star$, the errors $e^{(k)} = u^{(k)} - u^\star$ satisfy the* quadratic convergence estimate

$$| e^{(k+1)} | \;\leq\; \tau \, | e^{(k)} |^2 \tag{19.21}$$

for some constant $\tau > 0$.

Proof: Just as that of the linear convergence estimate (19.15), the proof relies on approximating $g(u)$ by a simpler function near the fixed point. For linear convergence, an affine approximation sufficed, but here we require a higher order approximation. Thus, we replace the mean value formula (19.13) by the first order Taylor expansion

$$g(u) = g(u^\star) + g'(u^\star) \, (u - u^\star) + \tfrac{1}{2} \, g''(w) \, (u - u^\star)^2, \tag{19.22}$$

where the final error term depends on an (unknown) point w that lies between u and $u^\star$; see (C.10) for details. At a fixed point, the constant term is $g(u^\star) = u^\star$. Furthermore, under our hypothesis $g'(u^\star) = 0$, and so (19.22) reduces to

$$g(u) - u^\star = \tfrac{1}{2} \, g''(w) \, (u - u^\star)^2.$$

Therefore,

$$| g(u) - u^\star | \leq \tau \, | u - u^\star |^2, \tag{19.23}$$

where τ is chosen so that

$$\tfrac{1}{2} \, | g''(w) | \leq \tau \tag{19.24}$$

for all w sufficiently close to $u^\star$. Therefore, the magnitude of τ is governed by the size of the *second derivative* of the iterative function $g(u)$ near the fixed point. We use the inequality (19.23) to estimate the error

$$| e^{(k+1)} | = | u^{(k+1)} - u^\star | = | g(u^{(k)}) - g(u^\star) | \;\leq\; \tau \, | u^{(k)} - u^\star |^2 = \tau \, | e^{(k)} |^2,$$

which establishes the quadratic convergence estimate (19.21). *Q.E.D.*

[‡] Note that since $\sigma < 1$, the logarithm $\log_{10} \sigma^{-1} = -\log_{10} \sigma > 0$ is positive.

Let us see how the quadratic estimate (19.21) speeds up the convergence rate. Following our earlier argument, suppose $u^{(k)}$ is correct to m decimal places, so

$$| e^{(k)} | < .5 \times 10^{-m}.$$

Then (19.21) implies that

$$| e^{(k+1)} | < .5 \times (10^{-m})^2 \, \tau = .5 \times 10^{-2m + \log_{10} \tau},$$

and so $u^{(k+1)}$ has $2m - \log_{10} \tau$ accurate decimal places. If $\tau \approx | g''(u^\star) |$ is of moderate size, we have essentially *doubled* the number of accurate decimal places in just a single iterate! A second iteration will double the number of accurate digits yet again. Thus, the convergence of a quadratic iteration scheme is *extremely* rapid, and, barring round-off errors, one can produce any desired number of digits of accuracy in a very short time. For example, if we start with an initial guess that is accurate in the first decimal digit, then a linear iteration with $\sigma = .1$ will require 49 iterations to obtain 50 decimal place accuracy, whereas a quadratic iteration (with $\tau = 1$) will only require 6 iterations to obtain $2^6 = 64$ decimal places of accuracy!

Example 19.9. Consider the function

$$g(u) = \frac{2\,u^3 + 3}{3\,u^2 + 3}\,.$$

There is a unique (real) fixed point $u^\star = g(u^\star)$, which is the real solution to the cubic equation

$$\tfrac{1}{3}\,u^3 + u - 1 = 0.$$

Note that

$$g'(u) = \frac{2\,u^4 + 6\,u^2 - 6\,u}{3\,(u^2+1)^2} = \frac{6\,u\left(\tfrac{1}{3}\,u^3 + u - 1\right)}{3\,(u^2+1)^2}\,,$$

and hence $g'(u^\star) = 0$ vanishes at the fixed point. Theorem 19.8 implies that the iterations will exhibit quadratic convergence to the root. Indeed, we find, starting with $u^{(0)} = 0$, the following values:

k	1	2	3
$u^{(k)}$	1.00000000000000	.833333333333333	.817850637522769

4	5	6
.817731680821982	.817731673886824	.817731673886824

The convergence rate is dramatic: after only 5 iterations, we have produced the first 15 decimal places of the fixed point. In contrast, the linearly convergent scheme based on $\widetilde{g}(u) = 1 - \frac{1}{3}\,u^3$ takes 29 iterations just to produce the first 5 decimal places of the same solution.

In practice, the appearance of a quadratically convergent fixed point is a matter of luck. The construction of quadratically convergent iterative methods for solving equations will be the focus of the following Section 19.2.

Vector–Valued Iteration

Extending the preceding analysis to vector-valued iterative systems is not especially difficult. We will build on our experience with linear iterative systems, and so the reader is advised to review the basic concepts and results from Chapter 10 before proceeding to the nonlinear systems presented here.

We begin by fixing a norm $\| \cdot \|$ on $\mathbb{R}^n$. Since we will also be computing the associated matrix norm $\| A \|$, as defined in Theorem 10.20, it may be more convenient for computations to adopt either the 1 or the ∞ norms rather than the standard Euclidean norm.

We begin by defining the vector-valued counterpart of the basic linear convergence condition (19.18).

Definition 19.10. A function $\mathbf{g}\colon \mathbb{R}^n \to \mathbb{R}^n$ is a *contraction* at a point $\mathbf{u}^\star \in \mathbb{R}^n$ if there exists a constant $0 \leq \sigma < 1$ such that

$$\| \mathbf{g}(\mathbf{u}) - \mathbf{g}(\mathbf{u}^\star) \| \;\leq\; \sigma \, \| \mathbf{u} - \mathbf{u}^\star \| \tag{19.25}$$

for all $\mathbf{u}$ sufficiently close to $\mathbf{u}^\star$, i.e., $\| \mathbf{u} - \mathbf{u}^\star \| < \delta$ for some fixed $\delta > 0$.

Remark: The notion of a contraction depends on the underlying choice of matrix norm. Indeed, the linear function $\mathbf{g}(\mathbf{u}) = A\mathbf{u}$ if and only if $\| A \| < 1$, which implies that A is a convergent matrix. While every convergent matrix satisfies $\| A \| < 1$ in *some* matrix norm, and hence defines a contraction relative to that norm, it may very well have $\| A \| > 1$ in a particular norm, violating the contaction condition; see (rowsumexA■) for an explicit example.

Theorem 19.11. *If $\mathbf{u}^\star = \mathbf{g}(\mathbf{u}^\star)$ is a fixed point for the discrete dynamical system* (19.1) *and $\mathbf{g}$ is a contraction at $\mathbf{u}^\star$, then $\mathbf{u}^\star$ is an asymptotically stable fixed point.*

Proof: The proof is a copy of the last part of the proof of Theorem 19.4. We write

$$\| \mathbf{u}^{(k+1)} - \mathbf{u}^\star \| = \| \mathbf{g}(\mathbf{u}^{(k)}) - \mathbf{g}(\mathbf{u}^\star) \| \leq \sigma \, \| \mathbf{u}^{(k)} - \mathbf{u}^\star \|,$$

using the assumed estimate (19.25). Iterating this basic inequality immediately demonstrates that

$$\| \mathbf{u}^{(k)} - \mathbf{u}^\star \| \;\leq\; \sigma^k \, \| \mathbf{u}^{(0)} - \mathbf{u}^\star \| \qquad \text{for} \qquad k = 0, 1, 2, 3, \dots \, . \tag{19.26}$$

Since $\sigma < 1$, the right hand side tends to 0 as $k \to \infty$, and hence $\mathbf{u}^{(k)} \to \mathbf{u}^\star$. *Q.E.D.*

In most interesting situations, the function $\mathbf{g}$ is differentiable, and so can be approximated by its first order Taylor polynomial (C.14):

$$\mathbf{g}(\mathbf{u}) \approx \mathbf{g}(\mathbf{u}^\star) + \mathbf{g}'(\mathbf{u}^\star)\,(\mathbf{u} - \mathbf{u}^\star) = \mathbf{u}^\star + \mathbf{g}'(\mathbf{u}^\star)\,(\mathbf{u} - \mathbf{u}^\star). \tag{19.27}$$

Here

$$\mathbf{g}'(\mathbf{u}) = \begin{pmatrix} \dfrac{\partial g_1}{\partial u_1} & \dfrac{\partial g_1}{\partial u_2} & \cdots & \dfrac{\partial g_1}{\partial u_n} \\ \dfrac{\partial g_2}{\partial u_1} & \dfrac{\partial g_2}{\partial u_2} & \cdots & \dfrac{\partial g_2}{\partial u_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial g_n}{\partial u_1} & \dfrac{\partial g_n}{\partial u_2} & \cdots & \dfrac{\partial g_n}{\partial u_n} \end{pmatrix}, \tag{19.28}$$

denotes the $n \times n$ *Jacobian matrix* of the vector-valued function $\mathbf{g}$, whose entries are the partial derivatives of its individual components. Since $\mathbf{u}^\star$ is fixed, the the right hand side of (19.27) is an affine function of $\mathbf{u}$. Moreover, $\mathbf{u}^\star$ remains a fixed point of the affine approximation. Proposition 10.44 tells us that iteration of the affine function will converge to the fixed point if and only if its coefficient matrix, namely $\mathbf{g}'(\mathbf{u}^\star)$, is a convergent matrix, meaning that its spectral radius $\rho(\mathbf{g}'(\mathbf{u}^\star)) < 1$. This observation motivates the following theorem and corollary.

Theorem 19.12. *Let $\mathbf{u}^\star$ be a fixed point for the discrete dynamical system $\mathbf{u}^{(k+1)} = \mathbf{g}(\mathbf{u}^{(k)})$. If the Jacobian matrix norm $\| \mathbf{g}'(\mathbf{u}^\star) \| < 1$, then $\mathbf{g}$ is a contraction at $\mathbf{u}^\star$, and hence the fixed point $\mathbf{u}^\star$ is asymptotically stable.*

Proof: The first order Taylor expansion (C.14) of $\mathbf{g}(\mathbf{u})$ at the fixed point $\mathbf{u}^\star$ takes the form

$$\mathbf{g}(\mathbf{u}) = \mathbf{g}(\mathbf{u}^\star) + \mathbf{g}'(\mathbf{u}^\star)\,(\mathbf{u} - \mathbf{u}^\star) + R(\mathbf{u} - \mathbf{u}^\star), \tag{19.29}$$

where the remainder term satisfies

$$\lim_{\mathbf{u} \to \mathbf{u}^\star} \frac{R(\mathbf{u} - \mathbf{u}^\star)}{\| \mathbf{u} - \mathbf{u}^\star \|} = 0.$$

Let $\varepsilon > 0$ be such that

$$\sigma = \| \mathbf{g}'(\mathbf{u}^\star) \| + \varepsilon < 1.$$

Choose $0 < \delta < 1$ such that $\| R(\mathbf{u} - \mathbf{u}^\star) \| \le \varepsilon \| \mathbf{u} - \mathbf{u}^\star \|$ whenever $\| \mathbf{u} - \mathbf{u}^\star \| \le \delta$. For such $\mathbf{u}$, we have, by the Triangle Inequality,

$$\begin{aligned} \| \mathbf{g}(\mathbf{u}) - \mathbf{g}(\mathbf{u}^\star) \| &\le \| \mathbf{g}'(\mathbf{u}^\star)\,(\mathbf{u} - \mathbf{u}^\star) \| + \| R(\mathbf{u} - \mathbf{u}^\star) \| \\ &\le \big(\| \mathbf{g}'(\mathbf{u}^\star) \| + \varepsilon \big) \| \mathbf{u} - \mathbf{u}^\star \| = \sigma \| \mathbf{u} - \mathbf{u}^\star \|, \end{aligned}$$

which establishes the contraction inequality (19.25). *Q.E.D.*

Corollary 19.13. *If the Jacobian matrix $\mathbf{g}'(\mathbf{u}^\star)$ is a convergent matrix, meaning that its spectral radius satisfies $\rho\big(\mathbf{g}'(\mathbf{u}^\star)\big) < 1$, then $\mathbf{u}^\star$ is an asymptotically stable fixed point.*

Proof: Corollary 10.32 assures us that $\| \mathbf{g}'(\mathbf{u}^\star) \| < 1$ in some matrix norm. Using this norm, the result immediately follows from the theorem. *Q.E.D.*

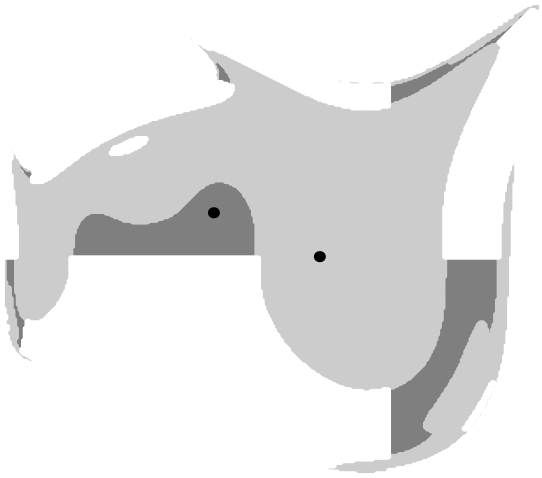


Figure 19.7. Basins of Attraction.

Example 19.14. Consider the function

$$\mathbf{g}(u,v) = \begin{pmatrix} -\frac{1}{4}\,u^3 + \frac{9}{8}\,u + \frac{1}{4}\,v^3 \\ \frac{3}{4}\,v - \frac{1}{2}\,uv \end{pmatrix}.$$

There are four (real) fixed points; stability is determined by the size of the eigenvalues of the Jacobian matrix

$$\mathbf{g}'(u,v) = \begin{pmatrix} \frac{9}{8} - \frac{3}{4}\,u^2 & -\frac{1}{2}\,v \\ \frac{3}{4}\,v^2 & \frac{3}{4} - \frac{1}{2}\,u \end{pmatrix}$$

at each of the fixed points. The results are summarized in the following table:

fixed point	$\mathbf{u}_1^\star = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$	$\mathbf{u}_2^\star = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ 0 \end{pmatrix}$	$\mathbf{u}_3^\star = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ 0 \end{pmatrix}$	$\mathbf{u}_4^\star = \begin{pmatrix} -\frac{1}{2} \\ \frac{1}{2} \end{pmatrix}$
Jacobian matrix	$\begin{pmatrix} \frac{9}{8} & 0 \\ 0 & \frac{3}{4} \end{pmatrix}$	$\begin{pmatrix} \frac{3}{4} & 0 \\ 0 & \frac{3}{4} - \frac{1}{2\sqrt{2}} \end{pmatrix}$	$\begin{pmatrix} \frac{3}{4} & 0 \\ 0 & \frac{3}{4} + \frac{1}{2\sqrt{2}} \end{pmatrix}$	$\begin{pmatrix} \frac{15}{16} & -\frac{1}{4} \\ \frac{3}{16} & 1 \end{pmatrix}$
eigenvalues	1.125, .75	.75, .396447	1.10355, .75	$.96875 \pm .214239\,\mathrm{i}$
spectral radius	1.125	.75	1.10355	.992157

Thus, $\mathbf{u}_2^\star$ and $\mathbf{u}_4^\star$ are stable fixed points, whereas $\mathbf{u}_1^\star$ and $\mathbf{u}_3^\star$ are both unstable. Indeed, starting with $\mathbf{u}^{(0)} = (\,.5, .5\,)^T$, it takes 24 iterates to converge to $\mathbf{u}_2^\star$ with 4 significant decimal digits, whereas starting with $\mathbf{u}^{(0)} = (\,-.7, .7\,)^T$, it takes 1049 iterates to converge to within 4 digits of $\mathbf{u}_4^\star$; the slower convergence rate is predicted by the larger Jacobian spectral radius. The two basins of attraction are plotted in Figure 19.7. The stable fixed points are indicated by black dots. The light gray region contains $\mathbf{u}_2^\star$ and indicates all the

points that converge to it; the darker gray indicates points converging, more slowly, to $\mathbf{u}_4^\star$. All other initial points, except $\mathbf{u}_1^\star$ and $\mathbf{u}_3^\star$, have rapidly unbounded iterates: $\| \mathbf{u}^{(k)} \| \to \infty$.

Theorem 19.12 tells us that initial values $\mathbf{u}^{(0)}$ that are sufficiently near a stable fixed point $\mathbf{u}^\star$ are guaranteed to converge to it. In the linear case, closeness of the initial data to the fixed point was not, in fact, an issue; all stable fixed points are, in fact, globally stable. For nonlinear iteration, it is of critical importance, and one does not typically expect iteration starting with far away initial data to converge to the desired fixed point. An interesting (and difficult) problem is to determine the so-called *basin of attraction* of a stable fixed point, defined as the set of all initial data that ends up converging to it. As in the elementary logistic map (19.19), initial values that lie outside a basin of attraction can lead to divergent iterates, periodic orbits, or even exhibit chaotic behavior. The full range of possible phenomena is a topic of contemporary research in dynamical systems theory, [**DDSY**], and in numerical analysis, [**7**].

The smaller the spectral radius or matrix norm of the Jacobian matrix at the fixed point, the faster the nearby iterates will converge to it. As in the scalar case, quadratic convergence will occur when the Jacobian matrix $\mathbf{g}'(\mathbf{u}^\star) = \mathrm{O}$ is the zero matrix[†], i.e., *all* first order partial derivatives of the components of $\mathbf{g}$ vanish at the fixed point. The quadratic convergence estimate

$$\| \mathbf{u}^{(k+1)} - \mathbf{u}^\star \| \;\le\; \tau \, \| \mathbf{u}^{(k)} - \mathbf{u}^\star \|^2 \tag{19.30}$$

is a consequence of the second order Taylor expansion at the fixed point. Details of the proof are left as an exercise.

Of course, in practice we don't know the norm or spectral radius of the Jacobian matrix $\mathbf{g}'(\mathbf{u}^\star)$ because we don't know where the fixed point is. This apparent difficulty can be easily circumvented by requiring that $\| \mathbf{g}'(\mathbf{u}) \| < 1$ for all $\mathbf{u}$ — or, at least, for all $\mathbf{u}$ in a domain Ω containing the fixed point. In fact, this hypothesis can be used to prove the exitence and uniqueness of asymptotically stable fixed points. Rather than work with the Jacobian matrix, let us return to the contraction condition (19.25), but now imposed uniformly on an entire domain.

Definition 19.15. A function $\mathbf{g}\colon \mathbb{R}^n \to \mathbb{R}^n$ is called a *contraction mapping* on a domain $\Omega \subset \mathbb{R}^n$ if

(*a*) it maps Ω to itself, so $\mathbf{g}(\mathbf{u}) \in \Omega$ whenever $\mathbf{u} \in \Omega$, and

(*b*) there exists a *constant* $0 \le \sigma < 1$ such that

$$\| \mathbf{g}(\mathbf{u}) - \mathbf{g}(\mathbf{v}) \| \;\le\; \sigma \, \| \mathbf{u} - \mathbf{v} \| \qquad \text{for all} \qquad \mathbf{u}, \mathbf{v} \in \Omega. \tag{19.31}$$

In other words, applying a contraction mapping reduces the mutual distance between points. So, as its name indicates, a contraction mapping effectively shrinks the size of its domain; see Figure 19.8. As the iterations proceed, the successive image domains become smaller and smaller. If the original domain is closed and bounded, then it is forced to shrink down to a single point, which is the unique fixed point of the iterative system. See [**X**] for the full proof of the following *Contraction Mapping Theorem*

[†] Having zero spectral radius is not sufficient for quadratic convergence; see Exercise ■.

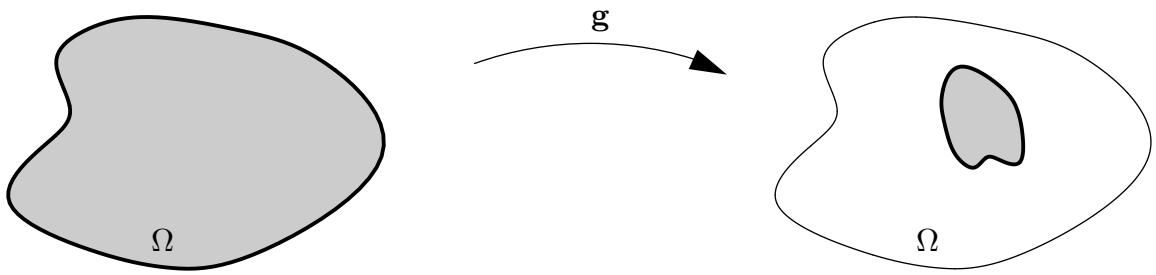


Figure 19.8. A Contraction Mapping.

Theorem 19.16. *If* $\mathbf{g}$ *is a contraction mapping on a closed bounded domain* $\Omega \subset \mathbb{R}^n$, *then* $\mathbf{g}$ *admits a unique fixed point* $\mathbf{u}^\star \in \Omega$. *Moreover, starting with any initial point* $\mathbf{u}^{(0)} \in \Omega$, *the iterates* $\mathbf{u}^{(k+1)} = \mathbf{g}(\mathbf{u}^{(k)})$ *necessarily converge to the fixed point:* $\mathbf{u}^{(k)} \to \mathbf{u}^\star$.

In particular, if $\| \mathbf{g}'(\mathbf{u}) \| < 1$ for all $\mathbf{u} \in \Omega$, then the conclusions of the Contraction Mapping Theorem 19.16 hold.

19.2. Solution of Equations and Systems.

Solving nonlinear equations and systems of equations is, of course, a problem of utmost importance in mathematics and its manifold applications. It can also be extremely difficult. Indeed, finding a complete set of (numerical) solutions to a complicated nonlinear system can be an almost insurmountable challlenge. In its most general version, we are given a collection of m functions $f_1, \dots, f_m$ depending upon n variables $u_1, \dots, u_n$, and are asked to determine all possible solutions $\mathbf{u} = (\, u_1, u_2, \dots, u_n \,)^T$ to the system

$$f_1(u_1, \dots, u_n) = 0, \qquad \dots \qquad f_m(u_1, \dots, u_n) = 0. \tag{19.32}$$

In many applications, the number of equations equals the number of unknowns, $m = n$, in which case one expects both existence and uniqueness of solutions. This point will be discussed in further detail below.

Here are some prototypical examples:

(*a*) Find the roots of the quintic polynomial equation

$$u^5 + u + 1 = 0. \tag{19.33}$$

Graphing the left hand side of the equation, as in Figure 19.9, convinces us that there is just one real root, lying somewhere between -1 and $-.5$. While there are explicit algebraic formulas for the roots of quadratic, cubic, and quartic polynomials, a famous theorem[†] due to the Norwegian mathematician Nils Henrik Abel in the early 1800's states that there is *no* such formula for generic fifth order polynomial equations.

(*b*) Any fixed point equation $u = g(u)$ has the form (19.35) where $f(u) = u - g(u)$. For example, the trigonometric Kepler equation

$$u - \epsilon \sin u = m$$

[†] A modern proof of this fact relies on Galois theory, [**74**].

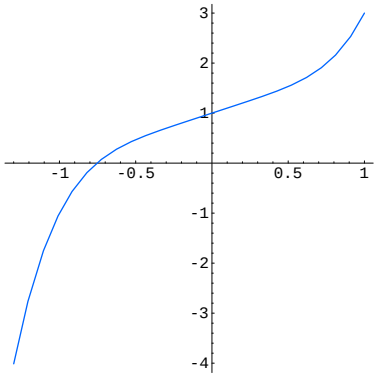


Figure 19.9. Graph of $u^5 + u + 1$.

arises in the study of planetary motion, cf. Example 19.5. Here ϵ, m are fixed constants, and we seek a corresponding solution u.

(*c*) Suppose we are given chemical compounds A, B, C that react to produce a fourth compound D according to

$$2\,A + B \;\longleftrightarrow\; D, \qquad\qquad A + 3\,C \;\longleftrightarrow\; D.$$

Let a, b, c be the initial concentrations of the reagents A, B, C injected into the reaction chamber. If u denotes the concentration of D produced by the first reaction, and v that by the second reaction, then the final equilibrium concentrations

$$a_\star = a - 2\,u - v, \qquad b_\star = b - u, \qquad c_\star = c - 3\,v, \qquad d_\star = u + v,$$

of the reagents will be determined by solving the nonlinear system

$$(a - 2\,u - v)^2(b - u) = \alpha\,(u + v), \qquad (a - 2\,u - v)(c - 3\,v)^3 = \beta\,(u + v), \tag{19.34}$$

where α, β are the known equilibrium constants of the two reactions.

Our immediate goal is to develop numerical algorithms for solving such nonlinear equations. Unfortunately, there is no direct universal solution method for nonlinear systems comparable to Gaussian elimination. As a result, numerical solution techniques rely almost exclusively on iterative algorithms. This section presents the principal methods for numerically approximating the solution(s) to a nonlinear system. We shall only discuss general purpose algorithms; specialized methods for solving particular classes of equations, e.g., polynomial equations, can be found in numerical analysis texts, e.g., [**27**, **34**, **148**]. Of course, the most important specialized methods — those designed for solving linear systems — will continue to play a critical role, even in the nonlinear regime.

The Bisection Method

We begin, as always, with the scalar case. Thus, we are given a real-valued function $f\colon \mathbb{R} \to \mathbb{R}$, and seek its *roots*, i.e., the real[†] solution(s) to the scalar equation

$$f(u) = 0. \tag{19.35}$$

[†] Complex roots to complex equations will be discussed later.

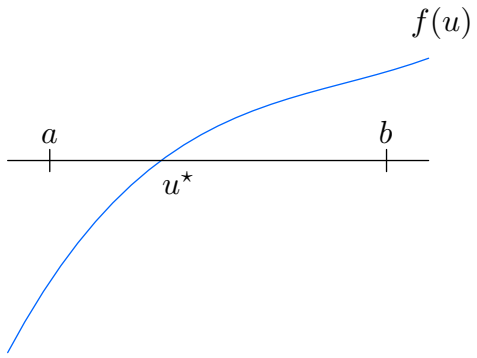


Figure 19.10. Intermediate Value Theorem.

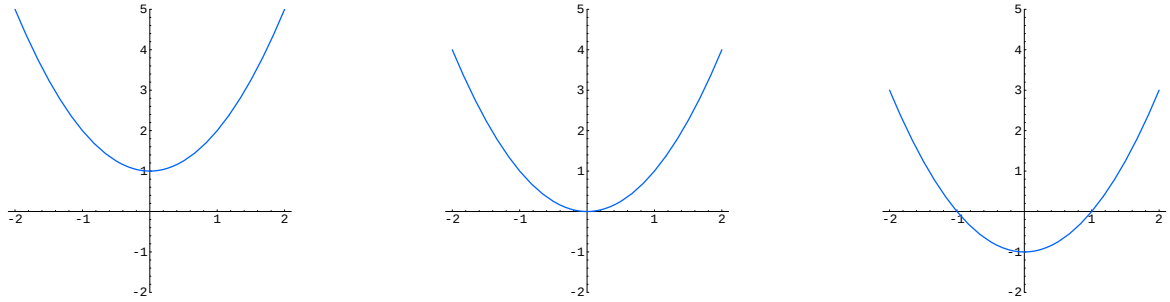


Figure 19.11. Roots of Quadratic Functions.

Our immediate goal is to develop numerical algorithms for solving such nonlinear scalar equations. The most primitive algorithm, and *the only one that is guaranteed to work in all cases*, is the Bisection Method. While it has an iterative flavor, it cannot be properly classed as a method governed by functional iteration as defined in the preceding section, and so must be studied directly in its own right.

The starting point is the Intermediate Value Theorem, which we state in simplified form. See Figure 19.10 for an illustration, and [**9**] for a proof.

Lemma 19.17. *Let $f(u)$ be a continuous scalar function. Suppose we can find two points $a < b$ where the values of $f(a)$ and $f(b)$ take* opposite *signs, so either $f(a) < 0$ and $f(b) > 0$, or $f(a) > 0$ and $f(b) < 0$. Then there exists at least one point $a < u^\star < b$ where $f(u^\star) = 0$.*

Note that if $f(a) = 0$ or $f(b) = 0$, then finding a root is trivial. If $f(a)$ and $f(b)$ have the same sign, then there may or may not be a root in between. Figure 19.11 plots the functions $u^2 + 1$, u^2 and $u^2 - 1$, on the interval $-2 \le u \le 2$. The first has two simple roots; the second has a single double root, while the third has no root. We also note that continuity of the function on the entire interval $[a, b]$ is an essential hypothesis. For example, the function $f(u) = 1/u$ satisfies $f(-1) = -1$ and $f(1) = 1$, but there is no root to the equation $1/u = 0$.

Note carefully that the Lemma 19.17 does *not* say there is a unique root between a

k	$u^{(k)}$	$v^{(k)}$	$w^{(k)} = \frac{1}{2}(u^{(k)} + v^{(k)})$	$f(w^{(k)})$
0	1	2	1.5	.75
1	1	1.5	1.25	−.1875
2	1.25	1.5	1.375	.2656
3	1.25	1.375	1.3125	.0352
4	1.25	1.3125	1.2813	−.0771
5	1.2813	1.3125	1.2969	−.0212
6	1.2969	1.3125	1.3047	.0069
7	1.2969	1.3047	1.3008	−.0072
8	1.3008	1.3047	1.3027	−.0002
9	1.3027	1.3047	1.3037	.0034
10	1.3027	1.3037	1.3032	.0016
11	1.3027	1.3032	1.3030	.0007
12	1.3027	1.3030	1.3029	.0003
13	1.3027	1.3029	1.3028	.0001
14	1.3027	1.3028	1.3028	−.0000

and b. There may be many roots, or even, in pathological examples, infinitely many. All the theorem guarantees is that, under the stated hypotheses, there is at least one root.

Once we are assured that a root exists, bisection relies on a "divide and conquer" strategy. The goal is to locate a root $a < u^\star < b$ between the endpoints. Lacking any additional evidence, one tactic would be to try the midpoint $c = \frac{1}{2}(a + b)$ as a first guess for the root. If, by some miracle, $f(c) = 0$, then we are done, since we have found a solution! Otherwise (and typically) we look at the sign of $f(c)$. There are two possibilities. If $f(a)$ and $f(c)$ are of opposite signs, then the Intermediate Value Theorem tells us that there is a root $u^\star$ lying between $a < u^\star < c$. Otherwise, $f(c)$ and $f(b)$ must have opposite signs, and so there is a root $c < u^\star < b$. In either event, we apply the same method to the interval in which we are assured a root lies, and repeat the procedure. Each iteration halves the length of the interval, and chooses the half in which a root is sure to lie. (There may, of course, be a root in the other half interval, but as we cannot be sure, we discard it from further consideration.) The root we home in on lies trapped in intervals of smaller and smaller width, and so convergence of the method is guaranteed. Figure bisect■ illustrates the steps in a particular example.

Example 19.18. The roots of the quadratic equation

$$f(u) = u^2 + u - 3 = 0$$

can be computed exactly by the quadratic formula:

$$u_1^\star = \frac{-1+\sqrt{13}}{2} \approx 1.302775\ldots, \qquad u_2^\star = \frac{-1-\sqrt{13}}{2} \approx -2.302775\ldots.$$

Let us see how one might approximate them by applying the Bisection Algorithm. We start the procedure by choosing the points $a = u^{(0)} = 1$, $b = v^{(0)} = 2$, noting that $f(1) = -1$ and $f(2) = 3$ have opposite signs and hence we are guaranteed that there is at least one root between 1 and 2. In the first step we look at the midpoint of the interval $[1, 2]$, which is 1.5, and evaluate $f(1.5) = .75$. Since $f(1) = -1$ and $f(1.5) = .75$ have opposite signs, we know that there is a root lying between 1 and 1.5. Thus, we take $u^{(1)} = 1$ and $v^{(1)} = 1.5$ as the endpoints of the next interval, and continue. The next midpoint is at 1.25, where $f(1.25) = -.1875$ has the opposite sign to $f(1.5) = .75$, and so a root lies between $u^{(2)} = 1.25$ and $v^{(2)} = 1.5$. The process is then iterated as long as desired — or, more practically, as long as your computer's precision does not become an issue.

The table displays the result of the algorithm, rounded off to four decimal places. After 14 iterations, the Bisection Method has correctly computed the first four decimal digits of the positive root $u_1^\star$. A similar bisection starting with the interval from $u^{(1)} = -3$ to $v^{(1)} = -2$ will produce the negative root.

A formal implementation of the Bisection Algorithm appears in the accompanying pseudocode program. The endpoints of the k^{th} interval are denoted by $u^{(k)}$ and $v^{(k)}$. The midpoint is $w^{(k)} = \frac{1}{2}\big(u^{(k)} + v^{(k)}\big)$, and the key decision is whether $w^{(k)}$ should be the right or left hand endpoint of the next interval. The integer n, governing the number of iterations, is to be prescribed in accordance with how accurately we wish to approximate the root $u^\star$.

The algorithm produces two sequences of approximations $u^{(k)}$ and $v^{(k)}$ that both converge monotonically to $u^\star$, one from below and the other from above:

$$a = u^{(0)} \le u^{(1)} \le u^{(2)} \le \cdots \le u^{(k)} \longrightarrow u^\star \longleftarrow v^{(k)} \le \cdots \le v^{(2)} \le v^{(1)} \le v^{(0)} = b.$$

and $u^\star$ is trapped between the two. Thus, the root is trapped inside a sequence of intervals $[\,u^{(k)}, v^{(k)}\,]$ of progressively shorter and shorter length. Indeed, the length of each interval is exactly half that of its predecessor:

$$v^{(k)} - u^{(k)} = \tfrac{1}{2}\,(v^{(k-1)} - u^{(k-1)}).$$

Iterating this formula, we conclude that

$$v^{(n)} - u^{(n)} = \big(\tfrac{1}{2}\big)^n\,(v^{(0)} - u^{(0)}) = \big(\tfrac{1}{2}\big)^n\,(b-a) \longrightarrow 0 \qquad \text{as} \qquad n \longrightarrow \infty.$$

The midpoint

$$w^{(n)} = \tfrac{1}{2}\,(u^{(n)} + v^{(n)})$$

lies within a distance

$$|\,w^{(n)} - u^\star\,| \le \tfrac{1}{2}\,(v^{(n)} - u^{(n)}) = \big(\tfrac{1}{2}\big)^{n+1}\,(b-a)$$

The Bisection Method

start
  if $f(a)\,f(b) < 0$ set $u^{(0)} = a$, $v^{(0)} = b$
    else print "Bisection Method not applicable"
  for $k = 0$ to $n - 1$
    set $w^{(k)} = \frac{1}{2}(u^{(k)} + v^{(k)})$
    if $f(w^{(k)}) = 0$, stop; print $u^\star = w^{(k)}$
    if $f(u^{(k)})\,f(w^{(k)}) < 0$, set $u^{(k+1)} = u^{(k)}$, $v^{(k+1)} = w^{(k)}$
      else set $u^{(k+1)} = w^{(k)}$, $v^{(k+1)} = v^{(k)}$
  next k
  print $u^\star = w^{(n)} = \frac{1}{2}(u^{(n)} + v^{(n)})$
end

of the root. Consequently, if we desire to approximate the root within a prescribed tolerance ε, we should choose the number of iterations n so that

$$\left(\tfrac{1}{2}\right)^{n+1}(b-a) < \varepsilon, \qquad \text{or} \qquad n > \log_2 \frac{b-a}{\varepsilon} - 1\,. \tag{19.36}$$

Summarizing:

Theorem 19.19. *If $f(u)$ is a continuous function, with $f(a)\,f(b) < 0$, then the Bisection Method starting with $u^{(0)} = a$, $v^{(0)} = b$, will converge to a solution $u^\star$ to the equation $f(u) = 0$ lying between a and b. After n steps, the midpoint $w^{(n)} = \frac{1}{2}\,(u^{(n)} + v^{(n)})$ will be within a distance of $\varepsilon = 2^{-n-1}(b-a)$ from the solution.*

For example, in the case of the quadratic equation in Example 19.18, after 14 iterations, we have approximated the positive root to within

$$\varepsilon = \left(\tfrac{1}{2}\right)^{15}(2-1) \approx 3.052 \times 10^{-5},$$

reconfirming our observation that we have accurately computed its first four decimal places. If we are in need of 10 decimal places, we set our tolerance to $\varepsilon = .5 \times 10^{-10}$, and so, according to (19.36), must perform $n = 34 > 33.22 \approx \log_2 2 \times 10^{10} - 1$ successive bisections†.

Example 19.20. As noted at the beginning of this section, the quintic equation

$$f(u) = u^5 + u + 1 = 0$$

† This assumes we have sufficient precision on the computer to avoid round-off errors.

has one real root, whose value can be readily computed by bisection. We start the algorithm with the initial points $u^{(0)} = -1,\ v^{(0)} = 0$, noting that $f(-1) = -1 < 0$ while $f(0) = 1 > 0$ are of opposite signs. In order to compute the root to 6 decimal places, we set $\varepsilon = .5 \times 10^{-6}$ in (19.36), and so need to perform $n = 20 > 19.93 \approx \log_2 2 \times 10^6 - 1$ bisections. Indeed, the algorithm produces the approximation $u^\star \approx -.754878$ to the root, and the displayed digits are guaranteed to be accurate.

Fixed Point Methods

The Bisection Method has an ironclad guarantee to converge to a root of the function — provided it can be properly started by locating two points where the function takes opposite signs. This may be tricky if the function has two very closely spaced roots and is, say, negative only for a very small interval between them, and may be impossible for multiple roots, e.g., the root $u^\star = 0$ of the quadratic function $f(u) = u^2$. When applicable, its convergence rate is completely predictable, but not especially fast. Worse, it has no immediately apparent extension to systems of equations, since there is *no* obvious counterpart to the Intermediate Value Theorem for vector-valued functions.

Most other numerical schemes for solving equations rely on some form of fixed point iteration. Thus, we seek to replace the system of equations $\mathbf{f}(\mathbf{u}) = \mathbf{0}$ with a fixed point system $\mathbf{u} = \mathbf{g}(\mathbf{u})$, that leads to the iterative solution scheme $\mathbf{u}^{(k+1)} = \mathbf{g}(\mathbf{u}^{(k)})$. For this to work, there are two key requirements:

(*a*) The solution $\mathbf{u}^\star$ to the equation $\mathbf{f}(\mathbf{u}) = \mathbf{0}$ is also a fixed point for $\mathbf{g}(\mathbf{u})$, and

(*b*) $\mathbf{u}^\star$ is, in fact a stable fixed point, meaning that the Jacobian $\mathbf{g}'(\mathbf{u}^\star)$ is a convergent matrix, or, slightly more restrictively, $\| \mathbf{g}'(\mathbf{u}^\star) \| < 1$ for a prescribed matrix norm.

If both conditions hold, then, *provided we choose the initial iterate* $\mathbf{u}^{(0)} = \mathbf{c}$ *sufficiently close to* $\mathbf{u}^\star$, the iterates $\mathbf{u}^{(k)} \to \mathbf{u}^\star$ will converge to the desired solution. Thus, the key to the practical use of functional iteration for solving equations is the proper design of an iterative system — coupled with a reasonably good initial guess for the solution. Before implementing general procedures, let us discuss a naïve example.

Example 19.21. To solve the cubic equation

$$f(u) = u^3 - u - 1 = 0 \tag{19.37}$$

we note that $f(1) = -1$ while $f(2) = 5$, and so there is a root between 1 and 2. Indeed, the Bisection Method leads to the approximate value $u^\star \approx 1.3247$ after 17 iterations.

Let us try to find the same root by fixed point iteration. As a first, naïve, guess, we rewrite the cubic equation in fixed point form

$$u = u^3 - 1 = \widetilde{g}(u).$$

Starting with the initial guess $u^{(0)} = 1.5$, successive approximations to the solution are found by iterating

$$u^{(k+1)} = \widetilde{g}(u^{(k)}) = (u^{(k)})^3 - 1, \qquad k = 0, 1, 2, \dots .$$

However, their values

$$u^{(0)} = 1.5, \qquad u^{(1)} = 2.375, \qquad u^{(2)} = 12.396,$$
$$u^{(3)} = 1904, \qquad u^{(4)} = 6.9024 \times 10^{9}, \qquad u^{(5)} = 3.2886 \times 10^{29}, \qquad \dots$$

rapidly become unbounded, and so fail to converge. This could, in fact, have been predicted by the convergence criterion in Theorem 19.4. Indeed, $\widetilde{g}\,'(u) = -3u^2$ and so $|\,\widetilde{g}\,'(u)\,| > 3$ for all $u \geq 1$, including the root $u^\star$. This means that $u^\star$ is an unstable fixed point, and the iterates cannot converge to it.

On the other hand, we can rewrite the equation (19.37) in the alternative iterative form

$$u = \sqrt[3]{1+u} \; = g(u).$$

In this case

$$0 \;\leq\; g'(u) \;=\; \frac{1}{3(1+u)^{2/3}} \;\leq\; \frac{1}{3} \qquad \text{for} \qquad u > 0.$$

Thus, the stability condition (19.14) is satisfied, and we anticipate convergence at a rate of at least $\frac{1}{3}$. (The Bisection Method converges more slowly, at rate $\frac{1}{2}$.) Indeed, the first few iterates $u^{(k+1)} = \sqrt[3]{1+u^{(k)}}$ are

$$1.5, \qquad 1.35721, \qquad 1.33086, \qquad 1.32588, \qquad 1.32494, \qquad 1.32476, \qquad 1.32473,$$

and we have converged to the root, correct to four decimal places, in only 6 iterations.

Newton's Method

Our immediate goal is to design an efficient iterative scheme $u^{(k+1)} = g(u^{(k)})$ whose iterates converge rapidly to the solution of the given scalar equation $f(u) = 0$. As we learned in Section 19.1, the convergence of the iteration is governed by the magnitude of its derivative at the fixed point. At the very least, we should impose the stability criterion $|\,g'(u^\star)\,| < 1$, and the smaller this quantity can be made, the faster the iterative scheme converges. if we are able to arrange that $g'(u^\star) = 0$, then the iterates will converge quadratically fast, leading, as noted in the discussion following Theorem 19.8, to a dramatic improvement in speed and efficiency.

Now, the first condition requires that $g(u) = u$ whenever $f(u) = 0$. A little thought will convince you that the iterative function should take the form

$$g(u) = u - h(u)\, f(u), \tag{19.38}$$

where $h(u)$ is a reasonably nice function. If $f(u^\star) = 0$, then clearly $u^\star = g(u^\star)$, and so $u^\star$ is a fixed point. The converse holds provided $h(u) \neq 0$ is never zero.

For quadratic convergence, the key requirement is that the derivative of $g(u)$ be zero at the fixed point solutions. We compute

$$g'(u) = 1 - h'(u)\, f(u) - h(u)\, f'(u).$$

Thus, $g'(u^\star) = 0$ at a solution to $f(u^\star) = 0$ if and only if

$$0 = 1 - h'(u^\star)\, f(u^\star) - h(u^\star)\, f'(u^\star) = 1 - h(u^\star)\, f'(u^\star).$$

Consequently, we should require that

$$h(u^\star) = \frac{1}{f'(u^\star)} \tag{19.39}$$

to ensure a quadratically convergent iterative scheme. This assumes that $f'(u^\star) \neq 0$, which means that $u^\star$ is a *simple root* of f. For here on, we leave aside multiple roots, which require a different approach, to be outlined in Exercise ■.

Of course, there are many functions $h(u)$ that satisfy (19.39), since we only need to specify its value at a single point. The problem is that we do not know $u^\star$ — after all this is what we are trying to compute — and so cannot compute the value of the derivative of f there. However, we can circumvent this apparent difficulty by a simple device: we impose equation (19.39) at all points, setting

$$h(u) = \frac{1}{f'(u)}\,, \tag{19.40}$$

which certainly guarantees that it holds at the solution $u^\star$. The result is the function

$$g(u) = u \,-\, \frac{f(u)}{f'(u)}\,, \tag{19.41}$$

and the resulting iteration scheme is known as *Newton's Method*, which, as the name suggests, dates back to the founder of the calculus. To this day, Newton's Method remains *the* most important general purpose algorithm for solving equations. It starts with an initial guess $u^{(0)}$ to be supplied by the user, and then successively computes

$$u^{(k+1)} = u^{(k)} - \frac{f(u^{(k)})}{f'(u^{(k)})}\,. \tag{19.42}$$

As long as the initial guess is sufficiently close, the iterates $u^{(k)}$ are guaranteed to converge, quadratically fast, to the (simple) root $u^\star$ of the equation $f(u) = 0$.

Theorem 19.22. *Suppose $f(u) \in \mathrm{C}^2$ is twice continuously differentiable. Let $u^\star$ be a solution to the equation $f(u^\star) = 0$ such that $f'(u^\star) \neq 0$. Given an initial guess $u^{(0)}$ sufficiently close to $u^\star$, the Newton iteration scheme* (19.42) *converges at a quadratic rate to the solution $u^\star$.*

Proof: By continuity, if $f'(u^\star) \neq 0$, then $f'(u) \neq 0$ for all u sufficiently close to $u^\star$, and hence the Newton iterative function (19.41) is well defined and continuously differentiable near $u^\star$. Since $g'(u) = f(u)\, f''(u)/f'(u)^2$, we have $g'(u^\star) = 0$ when $f(u^\star) = 0$, as promised by our construction. The quadratic convergence of the resulting iterative scheme is an immediate consequence of Theorem 19.8. *Q.E.D.*

Example 19.23. Consider the cubic equation

$$f(u) = u^3 - u - 1 = 0,$$

that we already solved in Example 19.21. The function used in the Newton iteration is

$$g(u) = u - \frac{f(u)}{f'(u)} = u - \frac{u^3 - u - 1}{3\,u^2 - 1}\,,$$

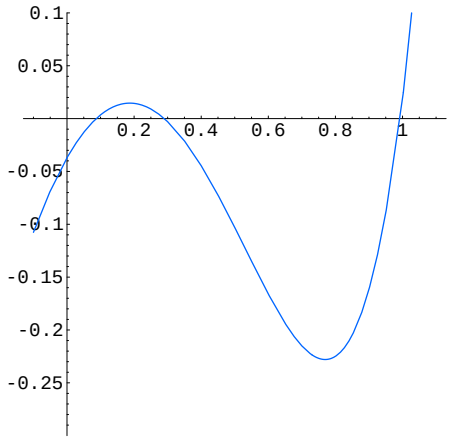


Figure 19.12. The function $f(u) = u^3 - \frac{3}{2}\,u^2 + \frac{5}{9}\,u - \frac{1}{27}$.

which is well-defined as long as $u \neq \pm \frac{1}{\sqrt{3}}$. We will try to avoid these singular points. The iterative procedure

$$u^{(k+1)} = g(u^{(k)}) = u^{(k)} - \frac{(u^{(k)})^3 - u^{(k)} - 1}{3\,(u^{(k)})^2 - 1}$$

with initial guess $u^{(0)} = 1.5$ produces the following values:

$$1.5, \qquad 1.34783, \qquad 1.32520, \qquad 1.32472,$$

and we have computed the root to 5 decimal places after only three iterations. The quadratic convergence of Newton's Method implies that, roughly, each new iterate doubles the number of correct decimal places. Thus, to compute the root accurately to 40 decimal places would only require 3 further iterations†. This underscores the tremendous advantage that the Newton algorithm offers over competing methods.

Example 19.24. Consider the cubic polynomial equation

$$f(u) = u^3 - \tfrac{3}{2}\,u^2 + \tfrac{5}{9}\,u - \tfrac{1}{27} = 0.$$

Since

$$f(0) = -\tfrac{1}{27}\,, \qquad f\big(\tfrac{1}{3}\big) = \tfrac{1}{54}\,, \qquad f\big(\tfrac{2}{3}\big) = -\tfrac{1}{27}\,, \qquad f(1) = \tfrac{1}{54}\,,$$

the Intermediate Value Lemma 19.17 guarantees that there are three roots on the interval $[0, 1]$: one between 0 and $\frac{1}{3}$, the second between $\frac{1}{3}$ and $\frac{2}{3}$, and the third between $\frac{2}{3}$ and 1. The graph in Figure 19.12 reconfirms this observation. Since we are dealing with a cubic polynomial, there are no other roots. (Why?)

† This assumes we are working in a sufficiently high precision arithmetic so as to avoid round-off errors.

It takes sixteen iterations of the Bisection Method starting with the three subintervals $\left[0\,,\tfrac{1}{3}\right]$, $\left[\tfrac{1}{3}\,,\tfrac{2}{3}\right]$ and $\left[\tfrac{2}{3}\,,1\right]$, to produce the roots to six decimal places:

$$u_1^\star \approx .085119, \qquad u_2^\star \approx .451805, \qquad u_3^\star \approx .963076.$$

Incidentally, if we start with the interval $[0,1]$ and apply bisection, we converge (perhaps surprisingly) to the largest root $u_3^\star$ in 17 iterations.

Fixed point iteration based on the formulation

$$u = g(u) = -u^3 + \tfrac{3}{2}\,u^2 + \tfrac{4}{9}\,u + \tfrac{1}{27}$$

can be used to find the first and third roots, but not the second root. For instance, starting with $u^{(0)} = 0$ produces $u_1^\star$ to 5 decimal places after 23 iterations, whereas starting with $u^{(0)} = 1$ produces $u_3^\star$ to 5 decimal places after 14 iterations. The reason we cannot produce $u_2^\star$ is due to the magnitude of the derivative

$$g'(u) = -3\,u^2 + 3\,u + \tfrac{4}{9}$$

at the roots, which is

$$g'(u_1^\star) \approx 0.678065, \qquad g'(u_2^\star) \approx 1.18748, \qquad g'(u_3^\star) \approx 0.551126.$$

Thus, $u_1^\star$ and $u_3^\star$ are stable fixed points, but $u_2^\star$ is unstable. However, because $g'(u_1^\star)$ and $g'(u_3^\star)$ are both bigger than .5, this iterative algorithm actually converges *slower* than ordinary bisection!

Finally, Newton's Method is based upon iteration of the rational function

$$g(u) = u - \frac{f(u)}{f'(u)} = u - \frac{u^3 - \tfrac{3}{2}\,u^2 + \tfrac{5}{9}\,u - \tfrac{1}{27}}{3\,u^2 - 3\,u + \tfrac{5}{9}}\,.$$

Starting with an initial guess of $u^{(0)} = 0$, the method computes $u_1^\star$ to 6 decimal places after only 4 iterations; starting with $u^{(0)} = .5$, it produces $u_2^\star$ to similar accuracy after 2 iterations; while starting with $u^{(0)} = 1$ produces $u_3^\star$ after 3 iterations — a dramatic speed up over the other two methods.

Newton's Method has a very pretty graphical interpretation, that helps us understand what is going on and why it converges so fast. Given the equation $f(u) = 0$, suppose we know an approximate value $u = u^{(k)}$ for a solution. Nearby $u^{(k)}$, we can approximate the nonlinear function $f(u)$ by its tangent line

$$y = f(u^{(k)}) + f'(u^{(k)})(u - u^{(k)}). \tag{19.43}$$

As long as the tangent line is not horizontal — which requires $f'(u^{(k)}) \neq 0$ — it crosses the axis at

$$u^{(k+1)} = u^{(k)} - \frac{f(u^{(k)})}{f'(u^{(k)})}\,,$$

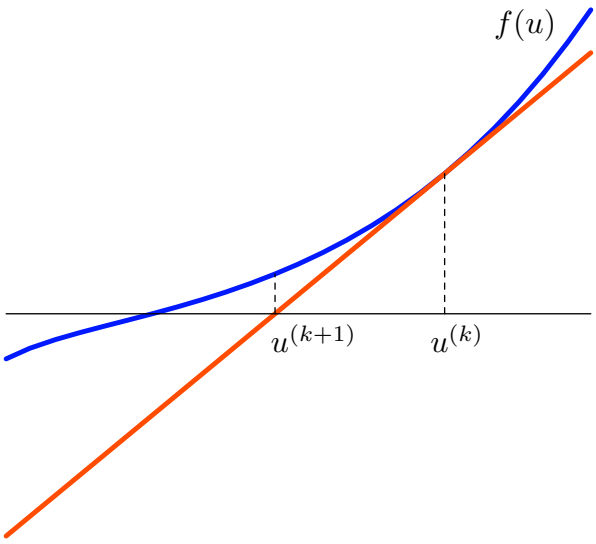


Figure 19.13. Newton's Method.

which represents a new, and, presumably more accurate, approximation to the desired root. The procedure is illustrated pictorially in Figure 19.13. Note that the passage from $u^{(k)}$ to $u^{(k+1)}$ is exactly the Newton iteration step (19.42). Thus, Newtonian iteration is the same as the approximation of function's root by those of its successive tangent lines.

Given a sufficiently accurate initial guess, Newton's Method will rapidly produce highly accurate values for the simple roots to the equation in question. In practice, barring some kind of special exploitable structure, Newton's Method is the root-finding algorithm of choice. The one caveat is that we need to start the process reasonably close to the root we are seeking. Otherwise, there is no guarantee that a particular set of iterates will converge, although if they do, the limiting value is necessarily a root of our equation. The behavior of Newton's Method as we change parameters and vary the initial guess is very similar to the simpler logistic map that we studied in Section 19.1, including period doubling bifurcations and chaotic behavior. The reader is invited to experiment with simple examples, some of which are provided in Exercise ■; further details can be found in [**147**].

Example 19.25. For fixed values of the eccentricity ϵ, Kepler's equation

$$u - \epsilon \sin u = m \tag{19.44}$$

can be viewed as a implicit equation defining the eccentric anomaly u as a function of the mean anomaly m. To solve Kepler's equation by Newton's Method, we introduce the iterative function

$$g(u) \;=\; u \;-\; \frac{u - \epsilon \sin u - m}{1 - \epsilon \cos u}\,.$$

Notice that when $|\,\epsilon\,| < 1$, the denominator never vanishes and so the iteration remains well-defined everywhere. Starting with a sufficiently close initial guess $u^{(0)}$, we are assured that the method will quickly converge to the solution.

Fixing the eccentricity ϵ, we can employ the method of *continuation* to determine how the solution $u^\star = h(m)$ depends upon the mean anomaly m. Namely, we start at

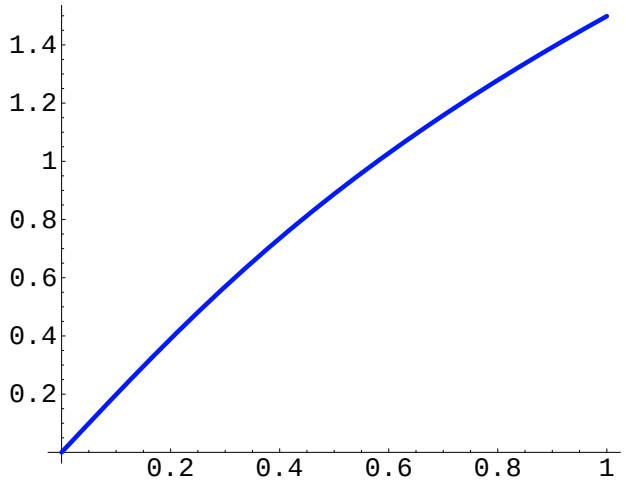


Figure 19.14. The Solution to the Kepler Equation for Eccentricity $\epsilon = .5$.

$m = m_0 = 0$ with the obvious solution $u^\star = h(0) = 0$. Then, to compute the solution at successive closely spaced values $0 < m_1 < m_2 < m_3 < \cdots$, we use the previously computed value as an initial guess $u^{(0)} = h(m_k)$ for the value of the solution at the next mesh point m_{k+1}, and run the Newton scheme until it converges to a sufficiently accurate approximation to the value $u^\star = h(m_{k+1})$. As long as m_{k+1} is reasonably close to m_k, Newton's Method will converge to the solution quite quickly.

The continuation method will quickly produce the values of u at the sample points. Intermediate values can either be determined by an interpolation scheme, e.g., a cubic spline fit of the data, or by running the Newton scheme using the closest known value as an initial condition. A plot for $0 \le m \le 1$ using the value $\epsilon = .5$ appears in Figure 19.14.

Systems of Equations

Let us now turn our attention to nonlinear systems of equations. We shall only consider the case when there are the same number of equations as unknowns:

$$f_1(u_1, \dots, u_n) = 0, \qquad \dots \qquad f_n(u_1, \dots, u_n) = 0. \tag{19.45}$$

We shall rewrite the system in vector form

$$\mathbf{f}(\mathbf{u}) = \mathbf{0}, \tag{19.46}$$

where $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ is a vector-valued function of n variables. In practice, we do not necessarily require that $\mathbf{f}$ be defined on all of $\mathbb{R}^n$, although this does simplify the exposition.

We shall only consider solutions that are separated from any others. More formally:

Definition 19.26. A solution $\mathbf{u}^\star$ to a system $\mathbf{f}(\mathbf{u}) = \mathbf{0}$ is called *isolated* if there exists $\delta > 0$ such that $\mathbf{f}(\mathbf{u}) \neq \mathbf{0}$ for all $\mathbf{u}$ satisfying $0 < \| \mathbf{u} - \mathbf{u}^\star \| < \delta$.

Example 19.27. Consider the planar equation

$$x^2 + y^2 = (x^2 + y^2)^2.$$

Rewriting the equation in polar coordinates as

$$r = r^2 \qquad \text{or} \qquad r(r-1) = 0,$$

we immediately see that the solutions consist of the origin $x = y = 0$ and all points on the unit circle $r^2 = x^2 + y^2 = 1$. Only the origin is an isolated solution, since every solution lying on the circle has plenty of other points on the circle that lie arbitrarily close to it.

Typically, solutions to a system of n equations in n unknowns are isolated, although this is not always true. For example, if A is a singular $n \times n$ matrix, then the solutions to the homogeneous linear system $A\,\mathbf{u} = \mathbf{0}$ form a nontrivial subspace, and so are not isolated. Nonlinear systems with non-isolated solutions can similarly be viewed as exhibiting some form of degeneracy. In general, the numerical computation of non-isolated solutions, e.g., solving the implicit equations for a curve or surface, is a much more difficult problem, and we will not attempt to discuss these issues in this introductory presentation. (However, our continuation approach to the Kepler equation in Example 19.25 indicates how one might proceed in such situations.)

In the case of a single scalar equation, the simple roots, meaning those for which $f'(u^\star) \neq 0$, are the easiest to compute. In higher dimensions, the role of the derivative of the function is played by the Jacobian matrix (19.28), and this motivates the following definition.

Definition 19.28. A solution $\mathbf{u}^\star$ to a system $\mathbf{f}(\mathbf{u}) = \mathbf{0}$ is called *nonsingular* if the associated Jacobian matrix is nonsingular there: $\det \mathbf{f}'(\mathbf{u}^\star) \neq 0$.

Note that the Jacobian matrix is square if and only if the system has the same number of equations as unknowns, which is thus one of the requirements for a solution to be nonsingular in our sense. Moreover, the Inverse Function Theorem from multivariable calculus, [**9**, **126**], implies that a nonsingular solution is necessarily isolated.

Theorem 19.29. *Every nonsingular solution* $\mathbf{u}^\star$ *to a system* $\mathbf{f}(\mathbf{u}) = \mathbf{0}$ *is isolated.*

Being the multivariate counterparts of simple roots also means that nonsingular solutions of systems are the most amenable to practical computation. Computing non-isolated solutions, as well as isolated solutions with a singular Jacobian matrix, is a considerable challenge, and practical algorithms remain much less well developed. For this reason, we focus exclusively on numerical solution techniques for nonsingular solutions.

Now, let us turn to numerical solution techniques. The first remark is that, unlike the scalar case, proving existence of a solution to a system of equations is often a challenging issue. There is no counterpart to the Intermediate Value Lemma 19.17 for vector-valued functions. It is not hard to find vector-valued functions whose entries take on both positive and negative values, but admit no solutions; a simple example can be found in Exercise ∎. This precludes any simple analog of the Bisection Method for nonlinear systems in more than one unknown.

On the other hand, Newton's Method can be straightforwardly adapted to compute nonsingular solutions to systems of equations, and is *the* most widely used method for this

purpose. The derivation proceeds in very similar manner to the scalar case. First, we replace the system (19.46) by a fixed point system

$$\mathbf{u} = \mathbf{g}(\mathbf{u}) \tag{19.47}$$

having the same solutions. By direct analogy with (19.38), any (reasonable) fixed point method will take the form

$$\mathbf{g}(\mathbf{u}) = \mathbf{u} - L(\mathbf{u})\,\mathbf{f}(\mathbf{u}), \tag{19.48}$$

where $L(\mathbf{u})$ is an $n \times n$ matrix-valued function. Clearly, if $\mathbf{f}(\mathbf{u}) = \mathbf{0}$ then $\mathbf{g}(\mathbf{u}) = \mathbf{u}$; conversely, if $\mathbf{g}(\mathbf{u}) = \mathbf{u}$, then $L(\mathbf{u})\,\mathbf{f}(\mathbf{u}) = \mathbf{0}$. If we further require that the matrix $L(\mathbf{u})$ be nonsingular, i.e., $\det L(\mathbf{u}) \neq 0$, then every fixed point of the iterator (19.48) will be a solution to the system (19.46) and vice versa.

According to Theorem 19.12, the speed of convergence (if any) of the iterative method

$$\mathbf{u}^{(k+1)} = \mathbf{g}(\mathbf{u}^{(k)}) \tag{19.49}$$

is governed by the matrix norm (or, more precisely, the spectral radius) of the Jacobian matrix $\mathbf{g}'(\mathbf{u}^\star)$ at the fixed point. In particular, if

$$\mathbf{g}'(\mathbf{u}^\star) = \mathrm{O} \tag{19.50}$$

is the zero matrix, then the method converges quadratically fast. Let's figure out how this can be arranged. Computing the derivative using the matrix version of the Leibniz rule for the derivative of a matrix product, (9.40), we find

$$\mathbf{g}'(\mathbf{u}^\star) = \mathrm{I} - L(\mathbf{u}^\star)\,\mathbf{f}'(\mathbf{u}^\star), \tag{19.51}$$

where I is the $n \times n$ identity matrix. (Fortunately, all the terms that involve derivatives of the entries of $L(\mathbf{u})$ go away since $\mathbf{f}(\mathbf{u}^\star) = \mathbf{0}$ by assumption; details are relegated to Exercise ■.) Therefore, the quadratic convergence criterion (19.50) holds if and only if

$$L(\mathbf{u}^\star)\,\mathbf{f}'(\mathbf{u}^\star) = \mathrm{I}\,, \qquad \text{and hence} \qquad L(\mathbf{u}^\star) = \mathbf{f}'(\mathbf{u}^\star)^{-1} \tag{19.52}$$

should be the inverse of the Jacobian matrix of $\mathbf{f}$ at the solution, which, fortuitously, was already assumed to be nonsingular.

As in the scalar case, we don't know the solution $\mathbf{u}^\star$, but we can arrange that condition (19.52) holds by setting

$$L(\mathbf{u}) = \mathbf{f}'(\mathbf{u})^{-1}$$

everywhere — or at least everywhere that $\mathbf{f}$ has a nonsingular Jacobian matrix. The resulting fixed point system

$$\mathbf{u} = \mathbf{g}(\mathbf{u}) = \mathbf{u} - \mathbf{f}'(\mathbf{u})^{-1}\,\mathbf{f}(\mathbf{u}), \tag{19.53}$$

leads to the quadratically convergent *Newton iteration scheme*

$$\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} - \mathbf{f}'(\mathbf{u}^{(k)})^{-1}\,\mathbf{f}(\mathbf{u}^{(k)}). \tag{19.54}$$

All it requires is that we guess an initial value $\mathbf{u}^{(0)}$ that is sufficiently close to the desired solution $\mathbf{u}^\star$. We are then guaranteed, by Exercise ■, that the iterates $\mathbf{u}^{(k)}$ converge quadratically fast to $\mathbf{u}^\star$.

Theorem 19.30. *Let $\mathbf{u}^\star$ be a nonsingular solution to the system $\mathbf{f}(\mathbf{u}) = \mathbf{0}$. Then, provided $\mathbf{u}^{(0)}$ is sufficiently close to $\mathbf{u}^\star$, the Newton iteration scheme* (19.54) *converges at a quadratic rate to the solution: $\mathbf{u}^{(k)} \to \mathbf{u}^\star$.*

Example 19.31. Consider the pair of simultaneous cubic equations

$$f_1(u,v) = u^3 - 3\,u\,v^2 - 1 = 0, \qquad f_2(u,v) = 3\,u^2\,v - v^3 = 0. \tag{19.55}$$

It is not difficult to prove that there are precisely three solutions:

$$\mathbf{u}_1^\star = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \mathbf{u}_2^\star = \begin{pmatrix} -.5 \\ .866025\ldots \end{pmatrix}, \qquad \mathbf{u}_3^\star = \begin{pmatrix} -.5 \\ -.866025\ldots \end{pmatrix}. \tag{19.56}$$

The Newton scheme relies on the Jacobian matrix

$$\mathbf{f}'(\mathbf{u}) = \begin{pmatrix} 3\,u^2 - 3\,v^2 & -6\,u\,v \\ 6\,u\,v & 3\,u^2 - 3\,v^2 \end{pmatrix}.$$

Since $\det \mathbf{f}'(\mathbf{u}) = 9(u^2 + v^2)$ is non-zero except at the origin, all three solutions are non-singular, and hence, for a sufficiently close initial value, Newton's Method will converge to the nearby solution. We explicitly compute the inverse Jacobian matrix:

$$\mathbf{f}'(\mathbf{u})^{-1} = \frac{1}{9(u^2+v^2)} \begin{pmatrix} 3\,u^2 - 3\,v^2 & 6\,u\,v \\ -6\,u\,v & 3\,u^2 - 3\,v^2 \end{pmatrix}.$$

Hence, in this particular example, the Newton iterator (19.53) is

$$\mathbf{g}(\mathbf{u}) = \begin{pmatrix} u \\ v \end{pmatrix} - \frac{1}{9(u^2+v^2)} \begin{pmatrix} 3\,u^2 - 3\,v^2 & 6\,u\,v \\ -6\,u\,v & 3\,u^2 - 3\,v^2 \end{pmatrix} \begin{pmatrix} u^3 - 3\,u\,v^2 - 1 \\ 3\,u^2\,v - v^3 \end{pmatrix}.$$

A complete diagram of the three basins of attraction, consisting of points whose Newton iterates converge to each of the three roots, has a remarkably complicated, fractal-like structure, as illustrated in Figure 19.15. In this plot, the x and y coordinates run from -1.5 to 1.5. The points in the black region all converge to $\mathbf{u}_1^\star$; those in the light gray region all converge to $\mathbf{u}_2^\star$; while those in the dark gray region all converge to $\mathbf{u}_3^\star$. The closer one is to the root, the sooner the iterates converge. On the interfaces between the basins of attraction are points for which the Newton iterates fail to converge, but exhibit a random, chaotic behavior. However, round-off errors will cause such iterates to fall into one of the basins, making it extremely difficult to observe such behavio over the long run.

Remark: The alert reader may notice that in this example, we are in fact merely computing the cube roots of unity, i.e., equations (19.55) are the real and imaginary parts of the complex equation $z^3 = 1$ when $z = u + \mathrm{i}\,v$.

Example 19.32. A robot arm consists of two rigid rods that are joined end-to-end to a fixed point in the plane, which we take as the origin $\mathbf{0}$. The arms are free to rotate, and the problem is to configure them so that the robot's hand ends up at the prescribed position $\mathbf{a} = (\,a, b\,)^T$. The first rod has length ℓ and makes an angle α with the horizontal, so its end is at position $\mathbf{v}_1 = (\,\ell\cos\alpha, \ell\sin\alpha\,)^T$. The second rod has length m and makes an

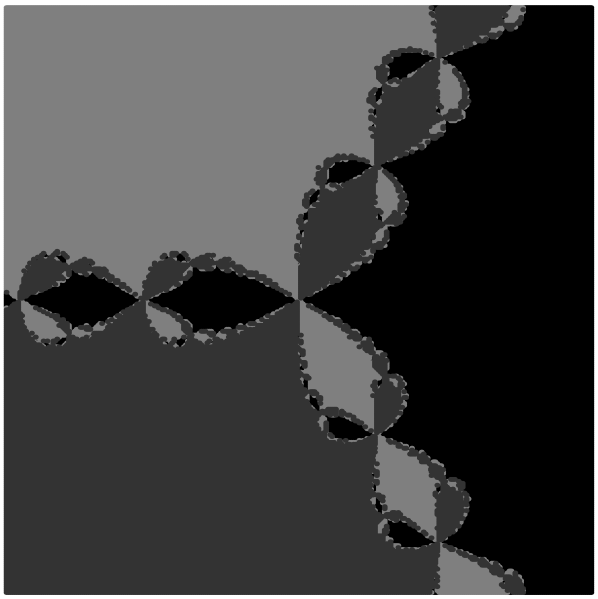


Figure 19.15. Computing the Cube Roots of Unity by Newton's Method.

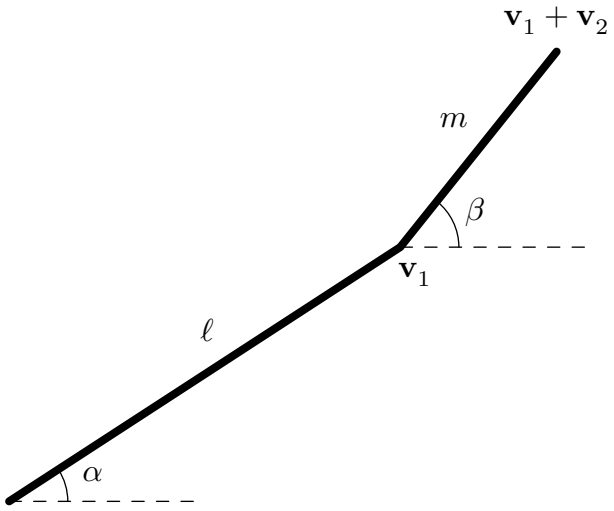


Figure 19.16. Robot Arm.

angle β with the horizontal, and so is represented by the vector $\mathbf{v}_2 = (\, m\cos\beta, m\sin\beta\,)^T$. The hand at the end of the second arm is at position $\mathbf{v}_1 + \mathbf{v}_2$, and the problem is to find values for the angles α, β so that $\mathbf{v}_1 + \mathbf{v}_2 = \mathbf{a}$; see Figure 19.16. To this end, we need to solve the system of equations

$$\ell\cos\alpha + m\cos\beta = a, \qquad \ell\sin\alpha + m\sin\beta = b, \tag{19.57}$$

for the angles α, β.

To find the solution, we shall apply Newton's Method. First, we compute the Jacobian

matrix of the system with respect to α, β, which is

$$\mathbf{f}'(\alpha, \beta) = \begin{pmatrix} -\ell \sin \alpha & -m \sin \beta \\ \ell \cos \alpha & m \cos \beta \end{pmatrix},$$

with inverse

$$\mathbf{f}'(\alpha, \beta)^{-1} = \frac{1}{\ell m \sin(\beta - \alpha)} \begin{pmatrix} -\ell \sin \alpha & m \sin \beta \\ -\ell \cos \alpha & m \cos \beta \end{pmatrix}.$$

As a result, the Newton iteration equation (19.54) has the explicit form

$$\begin{pmatrix} \alpha^{(k+1)} \\ \beta^{(k+1)} \end{pmatrix} = \begin{pmatrix} \alpha^{(k)} \\ \beta^{(k)} \end{pmatrix} - \\ - \frac{1}{\ell m \sin(\beta^{(k)} - \alpha^{(k)})} \begin{pmatrix} -\ell \cos \alpha^{(k)} & m \sin \beta^{(k)} \\ -\ell \cos \alpha^{(k)} & m \sin \beta^{(k)} \end{pmatrix} \begin{pmatrix} \ell \cos \alpha^{(k)} + m \cos \beta^{(k)} - a \\ \ell \sin \alpha^{(k)} + m \sin \beta^{(k)} - b \end{pmatrix}.$$

when running the iteration, one must be careful to avoid points at which $\alpha^{(k)} - \beta^{(k)} = 0$ or π, i.e., where the robot arm has straightened out.

As an example, let us assume that the rods have lengths $\ell = 2$, $m = 1$, and the desired location of the hand is at $\mathbf{a} = (\,1, 1\,)^T$. We start with an initial guess of $\alpha^{(0)} = 0$, $\beta^{(0)} = \frac{1}{2}\pi$, so the first rod lies along the x–axis and the second is perpendicular. The first few Newton iterates are given in the accompanying table. The first column is the iterate number k; the second and third columns indicate the angles $\alpha^{(k)}$, $\beta^{(k)}$ of the rods. The fourth and fifth give the position $(x^{(k)}, y^{(k)})^T$ of the joint or elbow, while the final two indicate the position $(z^{(k)}, w^{(k)})^T$ of the robot's hand.

k	$\alpha^{(k)}$	$\beta^{(k)}$	$x^{(k)}$	$y^{(k)}$	$z^{(k)}$	$w^{(k)}$
0	.0000	1.5708	2.0000	.0000	2.0000	1.0000
1	.0000	2.5708	2.0000	.0000	1.1585	.5403
2	.3533	2.8642	1.8765	.6920	.9147	.9658
3	.2917	2.7084	1.9155	.5751	1.0079	.9948
4	.2987	2.7176	1.9114	.5886	1.0000	1.0000
5	.2987	2.7176	1.9114	.5886	1.0000	1.0000

Observe that the robot has rapidly converged to one of the two possible configurations. (Can you figure out what the second equilibrium is?) In general, convergence depends on the choice of initial configuration, and the Newton iterates do not always settle down to a fixed point. For instance, if $\| \mathbf{a} \| > \ell + m$, there is no possible solution, since the arms are too short for the hand to reach to desired location; thus, no choice of initial conditions will lead to a convergent scheme and the robot arm flaps around in a chaotic manner.

Now that we have gained a little experience with Newton's Method for systems of equations, some supplementary remarks are in order. As we learned back in Chapter 1, except perhaps in very low-dimensional situations, one should not directly invert a matrix,

but rather use Gaussian elimination, or, in favorable situations, a linear iterative scheme, e.g., Jacobi, Gauss–Seidel or even SOR. So a better strategy is to leave the Newton system (19.54) in unsolved, implicit form

$$\mathbf{f}'(\mathbf{u}^{(k)})\,\mathbf{v}^{(k)} = -\,\mathbf{f}(\mathbf{u}^{(k)}), \qquad \mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + \mathbf{v}^{(k)}. \tag{19.58}$$

Given the iterate $\mathbf{u}^{(k)}$, we compute the Jacobian matrix $\mathbf{f}'(\mathbf{u}^{(k)})$ and the right hand side $-\,\mathbf{f}(\mathbf{u}^{(k)})$, and then use our preferred linear systems solver to find $\mathbf{v}^{(k)}$. Adding $\mathbf{u}^{(k)}$ to the result immediately yields the updated approximation $\mathbf{u}^{(k+1)}$ to the solution.

The main bottleneck in the implementation of the Newton scheme, particularly for large systems, is solving the linear system in (19.58). The coefficient matrix $\mathbf{f}'(\mathbf{u}^{(k)})$ must be recomputed at each step of the iteration, and hence knowing the solution to the k^{th} linear system does not appear to help us solve the subsequent system. Pereforming a complete Gaussian elimination at every step will tend to slow down the algorithm, particularly in high dimensional situations involving many equations in many unknowns.

One simple dodge for speeding up the computation is to note that, once we start converging, $\mathbf{u}^{(k)}$ will be very close to $\mathbf{u}^{(k-1)}$ and so we will probably not go far wrong by using $\mathbf{f}'(\mathbf{u}^{(k-1)})$ in place of the updated Jacobian matrix $\mathbf{f}'(\mathbf{u}^{(k)})$. Since we have already solved the linear system with coefficient matrix $\mathbf{f}'(\mathbf{u}^{(k-1)})$, we know its LU factorization, and hence can use Forward and Back Substitution to quickly solve the modified system

$$\mathbf{f}'(\mathbf{u}^{(k-1)})\,\mathbf{v}^{(k+1)} = -\,\mathbf{f}(\mathbf{u}^{(k)}), \qquad \mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + \mathbf{v}^{(k)}. \tag{19.59}$$

If $\mathbf{u}^{(k+1)}$ is still close to $\mathbf{u}^{(k-1)}$, we can continue to use $\mathbf{f}'(\mathbf{u}^{(k-1)})$ as the coefficient matrix when proceeding on to the next iterate $\mathbf{u}^{(k+2)}$. We proceed in this manner until there has been a notable change in the iterates, at which stage we can revert to solving the correct, unmodified linear system (19.58) by Gaussian Elimination. This strategy may dramatically reduce the total amount of computation required to approximate the solution to a prescribed accuracy. The down side is that this *quasi-Newton scheme* is only linearly convergent, and so does not home in on the root as fast as the unmodified implementation. The user needs to balance the trade-off between speed of convergence versus amount of time needed to solve the linear system at each step in the process. See [**148**] for further discussion.

19.3. Optimization.

We have already noted the importance of quadratic minimization principles for characterizing the equilibrium solutions of linear systems of physical significance. In nonlinear systems, optimization — either maximization or minimization — retains its centrality, and the wealth of practical applications has spawned an entire subdiscipline of applied mathematics. Physical systems naturally seek to minimize the potential energy function, and so determination of the possible equilibrium configurations requires solving a nonlinear minimization principle. Engineering design is guided by a variety of optimization constraints, such as performance, longevity, safety, and cost. Non-quadratic minimization principles also arise in the fitting of data by schemes that go beyond the simple linear least squares approximation method discussed in Section 4.3. Additional applications naturally

appear in economics and financial mathematics — one often wishes to minimize expenses or maximize profits, in biological and ecological systems, in pattern recognition and signal processing, in statistics, and so on. In this section, we will describe the basic mathematics underlying simple nonlinear optimization problems along with basic numerical techniques.

The Objective Function

Throughout this section, the real-valued function $F(\mathbf{u}) = F(u_1, \dots, u_n)$ to be optimized — the energy, cost, entropy, performance, etc. — will be called the *objective function*. As such, it depends upon one or more variables $\mathbf{u} = (u_1, u_2, \dots, u_n)^T$ that belong to a prescribed subset $\Omega \subset \mathbb{R}^n$.

Definition 19.33. A point $\mathbf{u}^\star \in \Omega$ is a *global minimum* of the objective function

$$F(\mathbf{u}^\star) \le F(\mathbf{u}) \qquad \text{for all} \qquad \mathbf{u} \in \Omega. \tag{19.60}$$

The minimum is called *strict* if

$$F(\mathbf{u}^\star) < F(\mathbf{u}) \qquad \text{for} \qquad \mathbf{u}^\star \neq \mathbf{u} \in \Omega. \tag{19.61}$$

The point $\mathbf{u}^\star$ is called a (*strict*) *local minimum* if the relevant inequality holds just for points $\mathbf{u} \in \Omega$ nearby $\mathbf{u}^\star$, i.e., satisfying $\| \mathbf{u} - \mathbf{u}^\star \| < \delta$ for some $\delta > 0$. In particular, strict local minima are *isolated*.

The definition of a *maximum* — local or global — is the same, but with the reversed inequality: $F(\mathbf{u}^\star) \ge F(\mathbf{u})$ or, in the strict case, $F(\mathbf{u}^\star) > F(\mathbf{u})$. Alternatively, a maximum of $F(\mathbf{u})$ is the same as a minimum of the negative $-F(\mathbf{u})$. Therefore, every result that applies to minimization of a function can easily be translated into a result on maximization, which allows us to concentrate exclusively on the minimization problem without any loss of generality. We will use *extremum* as a shorthand term for either a minimum or a maximum.

Remark: As we already noted in Section 4.1, *any* system of equations can be readily converted into a minimization principle. Given a system $\mathbf{f}(\mathbf{u}) = \mathbf{0}$, we introduce the objective function

$$F(\mathbf{u}) = \| \mathbf{f}(\mathbf{u}) \|^2, \tag{19.62}$$

where $\| \cdot \|$ is any convenient norm on $\mathbb{R}^n$. By the basic properties of the norm, the minimum value is $F(\mathbf{u}) = 0$, and this is achieved if and only if $\mathbf{f}(\mathbf{u}) = \mathbf{0}$, i.e., at a solution to the system. More generally, if there is no solution to the system, the minimizer(s) of $F(\mathbf{u})$ play the role of a least squares solution, at least for an inner product-based norm, along with the extensions to more general norms.

In contrast to the rather difficult question of existence of solutions to systems of equations, there is a general theorem that guarantees the existence of minima (and, hence, maxima) for a broad class of optimization problems.

Theorem 19.34. *If $F\colon \Omega \to \mathbb{R}$ is continuous, and $\Omega \subset \mathbb{R}^n$ is* compact, *maening closed and bounded, then F has at least one global minimum $\mathbf{u}^\star \in \Omega$.*

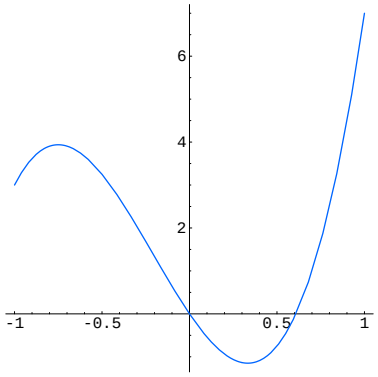


Figure 19.17. The function $8\,u^3 + 5\,u^2 - 6\,u$.

Proof: Let $m^\star = \min\{\, F(\mathbf{u}) \,|\, \mathbf{u} \in \Omega \,\}$, which may, *a priori*, be $-\infty$. Choose points $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots \in \Omega$, such that $F(\mathbf{u}^{(k)}) \to m$ as $k \to \infty$. By the basic properties of compact sets, [**152**], there is a convergent subsequence $\mathbf{u}^{(k_i)} \to \mathbf{u}^\star \in \Omega$. By continuity,

$$F(\mathbf{u}^\star) = F\left(\lim_{k_i \to \infty} \mathbf{u}^{(k_i)} \right) = \lim_{k_i \to \infty} F\left(\mathbf{u}^{(k_i)} \right) = m^\star,$$

and hence $\mathbf{u}^\star$ is a minimizer. *Q.E.D.*

Although Theorem 19.34 assures us of the existence of a global minimum of any continuous function on a bounded domain, it does not guarantee uniqueness, nor does it indicate how to go about finding it. Just as with the solution of nonlinear systems of equations, it is quite rare that one can extract explicit formulae for the minima of non-quadratic functions. Our goal, then, is to formulate practical algorithms that can accurately compute the minima of general nonlinear functions.

The most naïve algorithm, but one that is often successful in small scale problems, [**148, opt**], is to select a reasonably dense set of sample points $\mathbf{u}^{(k)}$ in the domain and choose the one that provides the smallest value for $F(\mathbf{u}^{(k)})$. If the points are sufficiently densely distributed and the function is not too wild, this will give a reasonable approximation to the minimum. The algorithm can be speeded up by appealing to more sophisticated means of selecting the sample points.

In the rest of this section, we will discuss optimization strategies that exploit the differential calculus. Let us first review the basic procedure for optimizing functions that you learned in first and second year calculus. As you no doubt remember, there are two different possible types of minima. An *interior minimum* occurs at an interior point of the domain of definition of the function, whereas a *boundary minimum* occurs on its boundary $\partial\Omega$. Interior local minima are easier to find, and, to keep the presentation simple, we shall focus our efforts on them. Let us begin with a simple scalar example.

Example 19.35. Let us optimize the scalar function

$$f(u) = 8\,u^3 + 5\,u^2 - 6\,u$$

on the domain $-1 \le u \le 1$. To locate the minimum, the first step is to look at the *critical*

points where the derivative vanishes:

$$f'(u) = 24\,u^2 + 10\,u - 6 = 0, \qquad \text{and hence} \qquad u = \tfrac{1}{3},\ -\tfrac{3}{4}.$$

To ascertain the local nature of the two critical points, we apply the second derivative test. Since $f''(u) = 48\,u + 10$, we have

$$f''\left(\tfrac{1}{3}\right) = 26 > 0, \qquad \text{whereas} \qquad f''\left(-\tfrac{3}{4}\right) = -26 < 0.$$

We conclude that $\frac{1}{3}$ is a local minimum, while $\frac{3}{4}$ is a local maximum.

To find the global minimum and maximum on the interval $[-1, 1]$, we must also take into account the boundary points ± 1. Comparing the function values at the four points,

$$f(-1) = 3, \qquad f\left(\tfrac{1}{3}\right) = -\tfrac{31}{27} \approx -1.148, \qquad f\left(-\tfrac{3}{4}\right) = \tfrac{63}{16} = 3.9375, \qquad f(1) = 7,$$

we see that $\frac{1}{3}$ is the global minimum, whereas 1 is the global maximum — which occurs on the boundary of the interval. This is borne out by the graph of the function, as displayed in Figure 19.17.

The Gradient

As you first learn in multi-variable calculus, [**9**, **126**], the interior extrema — minima and maxima — of a smooth function $F(\mathbf{u}) = F(u_1, \dots, u_n)$ are necessarily *critical points*, meaning places where the gradient of F vanishes. The standard gradient is the vector field whose entries are its first order partial derivatives:

$$\nabla F(\mathbf{u}) = \left(\frac{\partial F}{\partial u_1}\,,\ \dots\ , \frac{\partial F}{\partial u_n} \right)^T. \tag{19.63}$$

Let us, in preparation for the more general minimization problems over infinite-dimensional function spaces to be treated in Chapter 21, reformulate the definition of the gradient in a more intrinsic manner. An important but subtle point is that the gradient operator, in fact, relies upon the introduction of an inner product on the underlying vector space. The version (19.63) is, in fact, based upon on the Euclidean dot product on $\mathbb{R}^n$. Altering the inner product will change the formula for the gradient!

Definition 19.36. Let V be an inner product space. The *gradient* of a function $F\colon V \to \mathbb{R}$ at a point $\mathbf{u} \in V$ is the vector $\nabla F(\mathbf{u}) \in V$ that satisfies

$$\langle\, \nabla F(\mathbf{u})\,, \mathbf{v}\,\rangle = \frac{d}{dt}\, F(\mathbf{u} + t\,\mathbf{v}) \,\Big|_{t=0} \qquad \text{for all} \qquad \mathbf{v} \in V. \tag{19.64}$$

Remark: The function F does not have to be defined on all of the space V in order for this definition to make sense.

The quantity displayed in the preceding formula is known as the *directional derivative* of F with respect to $\mathbf{v} \in V$, and typically denoted by $\partial F/\partial \mathbf{v}$. Thus, by definition, the directional derivative equals the inner product with the gradient vector. The directional

derivative measures the rate of change of F in the direction of the vector $\mathbf{v}$, scaled in proportion to its length.

In the Euclidean case, when $F(\mathbf{u}) = F(u_1, \dots, u_n)$ is a function of n variables, defined for $\mathbf{u} = (\, u_1, u_2, \dots, u_n \,)^T \in \mathbb{R}^n$, we can use the chain rule to compute

$$\begin{aligned} \frac{d}{dt}\, F(\mathbf{u} + t\,\mathbf{v}) &= \frac{d}{dt}\, F(u_1 + t\,v_1, \dots, u_n + t\,v_n) \\ &= \frac{\partial F}{\partial u_1}(\mathbf{u} + t\,\mathbf{v})\, v_1 \;+\; \cdots \;+\; \frac{\partial F}{\partial u_n}(\mathbf{u} + t\,\mathbf{v})\, v_n. \end{aligned} \tag{19.65}$$

Setting $t = 0$, the right hand side of (19.64) reduces to

$$\frac{d}{dt}\, F(\mathbf{u} + t\,\mathbf{v}) \,\Big|_{t=0} = \frac{\partial F}{\partial u_1}\,(\mathbf{u})\, v_1 \;+\; \cdots \;+\; \frac{\partial F}{\partial u_n}\,(\mathbf{u})\, v_n = \nabla F(\mathbf{u}) \cdot \mathbf{v}.$$

Therefore, the directional derivative equals the Euclidean dot product between the usual gradient of the function (19.63) and the direction vector $\mathbf{v}$, justifying (19.64) in the Euclidean case.

Remark: In this chapter, we will only deal with the standard Euclidean dot product, which results in the usual gradient formula (19.63). If we introduce an alternative inner product on $\mathbb{R}^n$, then the notion of gradient, as defined in (19.64) will change. Details are outlined in Exercise ■.

A function $F(\mathbf{u})$ is *continuously differentiable* if and only if its gradient $\nabla F(\mathbf{u})$ is a continuously varying vector-valued function of $\mathbf{u}$. This is equivalent to the requirement that its first order partial derivatives $\partial F/\partial u_i$ are all continuous. As usual, we use $\mathrm{C}^1(\Omega)$ to denote the vector space of all continuously differentiable scalar-valued functions defined on a domain $\Omega \subset \mathbb{R}^n$. From now on, all objective functions are assumed to be continuously differentiable on their domain of definition.

If $\mathbf{u}(t)$ represents a parametrized curve contained within the domain of definition of $F(\mathbf{u})$, then the instantaneous rate of change in the scalar quantity F as we move along the curve is given by

$$\frac{d}{dt}\, F(\mathbf{u}(t)) = \left\langle\, \nabla F(\mathbf{u})\,,\, \frac{d\mathbf{u}}{dt} \,\right\rangle, \tag{19.66}$$

which is the directional derivative of F with respect to the velocity or tangent vector $\mathbf{v} = \dot{\mathbf{u}}$ to the curve. For instance, suppose $F(u_1, u_2)$ represents the elevation of a mountain range at position $\mathbf{u} = (\, u_1, u_2 \,)^T$. If we travel through the mountains along the path $\mathbf{u}(t) = (\, u_1(t), u_2(t) \,)^T$, then our instantaneous rate of ascent or descent (19.66) is equal to the dot product of our velocity vector $\dot{\mathbf{u}}(t)$ with the gradient of the elevation function. This observation leads to an important interpretation of the gradient vector.

Theorem 19.37. *The gradient $\nabla F(\mathbf{u})$ of a scalar function $F(\mathbf{u})$ points in the direction of its steepest increase at the point $\mathbf{u}$. The negative gradient, $-\nabla F(\mathbf{u})$, which points in the opposite direction, indicates the direction of steepest decrease.*

Thus, when F represents elevation, ∇F tells us the direction that is steepest uphill, while $-\nabla F$ points directly downhill — the direction water will flow. Similarly, if F represents the temperature of a solid body, then ∇F tells us the direction in which it is heating up the quickest. Heat energy (like water) will flow in the opposite, coldest direction, namely that of the negative gradient vector $-\nabla F$.

But you need to be careful in how you interpret Theorem 19.37. Clearly, the faster you move along a curve, the faster the function $F(\mathbf{u})$ will vary, and one needs to take this into account when comparing the rates of change along different curves. The easiest way to effect the comparison is to assume that the tangent vector $\mathbf{a} = \dot{\mathbf{u}}$ has unit norm, so $\| \mathbf{a} \| = 1$, which means that we are passing through the point $\mathbf{u}(t)$ with unit speed. Once this is done, Theorem 19.37 is an immediate consequence of the Cauchy–Schwarz inequality (3.18). Indeed,

$$\left| \frac{\partial F}{\partial \mathbf{a}} \right| = | \mathbf{a} \cdot \nabla F | \leq \| \mathbf{a} \| \, \| \nabla F \| = \| \nabla F \|, \qquad \text{when} \qquad \| \mathbf{a} \| = 1,$$

with equality if and only if $\mathbf{a}$ points in the same direction as the gradient. Therefore, the maximum rate of change is when $\mathbf{a} = \nabla F / \| \nabla F \|$ is the unit vector in the gradient direction, while the minimum is achieved when $\mathbf{a} = -\nabla F / \| \nabla F \|$ points in the opposite direction.

Critical Points

Thus, the only points at which the gradient fails to indicate directions of increase/decrease of the objective fulnction are where it vanishes. Such points play a critical role in the analysis, whence the following definition.

Definition 19.38. A point $\mathbf{u}^\star$ is called a *critical point* of the objective function $F(\mathbf{u})$ if

$$\nabla F(\mathbf{u}^\star) = \mathbf{0}. \tag{19.67}$$

Let us prove that all local minima are indeed critical points. The most important thing about this proof is that it only relies on the intrinsic definition of gradient, and therefore applies to any function on any inner product space. Moreover, even though the gradient will change if we alter the underlying inner product, the requirement that it vanish at a local minimum does not.

Theorem 19.39. *Every local (interior) minimum $\mathbf{u}^\star$ of a continuously differentiable function $F(\mathbf{u})$ is a critical point: $\nabla F(\mathbf{u}^\star) = \mathbf{0}$.*

Proof: Let $\mathbf{0} \neq \mathbf{v} \in \mathbb{R}^n$ be any vector. Consider the scalar function

$$g(t) = F(\mathbf{u}^\star + t\,\mathbf{v}) = F(u_1^\star + t\,v_1, \dots, u_n^\star + t\,v_n), \tag{19.68}$$

where $t \in \mathbb{R}$ is sufficiently small to ensure that $\mathbf{u}^\star + t\,\mathbf{v}$ remains strictly inside the domain of F. Note that g measures the values of F along a straight line passing through $\mathbf{u}^\star$ in the direction prescribed by $\mathbf{v}$. Since $\mathbf{u}^\star$ is a local minimum,

$$F(\mathbf{u}^\star) \leq F(\mathbf{u}^\star + t\,\mathbf{v}), \qquad \text{and hence} \qquad g(0) \leq g(t)$$

for all t sufficiently close to zero. In other words, $g(t)$, as a function of the single variable t, has a local minimum at $t = 0$. By the basic calculus result on minima of functions of one variable, the derivative of $g(t)$ must vanish at $t = 0$. Therefore, by the definition (19.64) of gradient,

$$0 = g'(0) = \frac{d}{dt} F(\mathbf{u}^\star + t\,\mathbf{v}) \Big|_{t=0} = \langle\, \nabla F(\mathbf{u}^\star)\,, \mathbf{v}\,\rangle.$$

We conclude that the gradient vector $\nabla F(\mathbf{u}^\star)$ at the critical point must be orthogonal to *every* vector $\mathbf{v} \in \mathbb{R}^n$, which is only possible if $\nabla F(\mathbf{u}^\star) = \mathbf{0}$. *Q.E.D.*

Thus, provided the objective function is continuously differentiable, every interior minimum, both local and global, is necessarily a critical point. The converse is not true; critical points can be maxima; they can also be saddle points or of some degenerate form. The basic analytical method[†] for determining the (interior) minima of a given function is to first find all its critical points by solving the system of equations (19.67). Each critical point then needs to be examined more closely — as it could be either a minimum, or a maximum, or neither.

Example 19.40. Consider the function

$$F(u, v) = u^4 - 2\,u^2 + v^2,$$

which is defined and continuously differentiable on all of $\mathbb{R}^2$. Since $\nabla F = \big(\,4\,u^3 - 4\,u, 2\,v\,\big)^T$, its critical points are obtained by solving the pair of equations

$$4\,u^3 - 4\,u = 0, \qquad 2\,v = 0.$$

The solutions to the first equation are $u = 0, \pm 1$, while the second equation requires $v = 0$. Therefore, F has three critical points:

$$\mathbf{u}_1^\star = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \mathbf{u}_2^\star = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \mathbf{u}_3^\star = \begin{pmatrix} -1 \\ 0 \end{pmatrix}. \tag{19.69}$$

Inspecting its graph in Figure 19.18, we suspect that the first critical point $\mathbf{u}_1^\star$ is a saddle point, whereas the other two appear to be local minima, having the same value $F(\mathbf{u}_2^\star) = F(\mathbf{u}_3^\star) = -1$. This will be confirmed once we learn how to rigorously distinguish critical points.

The student should also pay attention to the distinction between local minima and global minima. In the absence of theoretical justification, the only practical way to determine whether or not a minimum is global is to find all the different local minima, including those on the boundary, and see which one gives the smallest value. If the domain is unbounded, one must also worry about the asymptotic behavior of the objective function for large $\mathbf{u}$.

[†] Numerical methods are discussed below.

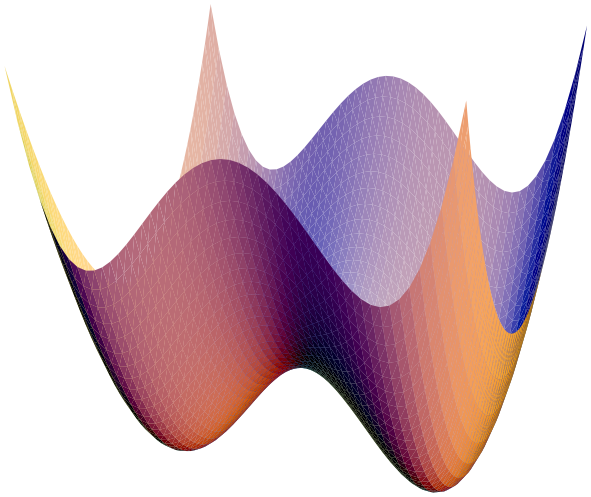


Figure 19.18. The Function $u^4 - 2u^2 + v^2$.

The Second Derivative Test

The status of critical point — minimum, maximum, or neither — can often be resolved by analyzing the second derivative of the objective function at the critical point. Let us first review the one variable second derivative test you learned in first year calculus.

Proposition 19.41. *Let $g(t) \in \mathrm{C}^2$ be a scalar function, and suppose $t^\star$ a critical point: $g'(t^\star) = 0$. If $t^\star$ is a local minimum, then $g''(t^\star) \geq 0$. Conversely, if $g''(t^\star) > 0$, then $t^\star$ is a strict local minimum. Similarly, $g''(t^\star) \leq 0$ is required at a local maximum, while $g''(t^\star) < 0$ implies that $t^\star$ is a strict local maximum.*

Remark: If $g''(t^\star) \neq 0$, then the quadratic Taylor polynomial has a minimum or maximum at $t^\star$ according to the sign of the second derivative. In the borderline case, when $g''(t^\star) = 0$, the second derivative test is inconclusive, and the point could be either maximum or minimum or neither. One must look at the higher order terms in the Taylor expansion to resolve the issue; see Exercise ■.

In multi-variate calculus, the "second derivative" of a function $F(\mathbf{u}) = F(u_1, \ldots, u_n)$ is represented by the $n \times n$ *Hessian*[†] *matrix*, whose entries are its second order partial

[†] Named after the early eighteenth century German mathematician Ludwig Otto Hesse.

derivatives:

$$\nabla^2 F(\mathbf{u}) = \begin{pmatrix} \dfrac{\partial^2 F}{\partial u_1^2} & \dfrac{\partial^2 F}{\partial u_1\, \partial u_2} & \cdots & \dfrac{\partial^2 F}{\partial u_1\, \partial u_n} \\ \dfrac{\partial^2 F}{\partial u_2\, \partial u_1} & \dfrac{\partial^2 F}{\partial u_2^2} & \cdots & \dfrac{\partial^2 F}{\partial u_2\, \partial u_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 F}{\partial u_n\, \partial u_1} & \dfrac{\partial^2 F}{\partial u_n\, \partial u_2} & \cdots & \dfrac{\partial^2 F}{\partial u_n^2} \end{pmatrix}, \tag{19.70}$$

We will always assume that $F(\mathbf{u}) \in \mathrm{C}^2$ has continuous second order partial derivatives. In this case, its mixed partial derivatives are equal: $\partial^2 F/\partial u_i\, \partial u_j = \partial^2 F/\partial u_j\, \partial u_i$, cf. [**9**, **126**]. As a result, the Hessian is a symmetric matrix: $\nabla^2 F(\mathbf{u}) = \nabla^2 F(\mathbf{u})^T$.

The second derivative test for a local minimum of scalar function relies on the positivity of its second derivative. For a function of several variables, the corresponding condition is that the Hessian matrix be positive definite, as in Definition 3.20. More specifically:

Theorem 19.42. *Let $F(\mathbf{u}) = F(u_1, \dots, u_n) \in \mathrm{C}^2(\Omega)$ be a real-valued, twice continuously differentiable function defined on an open domain $\Omega \subset \mathbb{R}^n$. If $\mathbf{u}^\star \in \Omega$ is a (local, interior) minimum for F, then it is necessarily a critical point, so $\nabla F(\mathbf{u}^\star) = \mathbf{0}$. Moreover, the Hessian matrix* (19.70) *must be positive semi-definite at the minimum, so $\nabla^2 F(\mathbf{u}^\star) \geq 0$. Conversely, if $\mathbf{u}^\star$ is a critical point with positive definite Hessian matrix $\nabla^2 F(\mathbf{u}^\star) > 0$, then $\mathbf{u}^\star$ is a strict local minimum of F.*

A maximum requires a negative semi-definite Hessian matrix. If, moreover, the Hessian at the critical point is negative definite, then the critical point is a strict local maximum. If the Hessian matrix is indefinite, then the critical point is a saddle point — neither minimum nor maximum. In the borderline case, when the Hessian is only positive or negative semi-definite at the critical point, the second derivative test is inconclusive. Resolving the nature of the critical point requires more detailed knowledge of the objective function, e.g., its higher order derivatives.

We defer the proof of Theorem 19.42 until the end of this section.

Example 19.43. As a first, elementary example, consider the quadratic function

$$F(u, v) = u^2 - 2\,u\,v + 3\,v^2.$$

To minimize F, we begin by computing its gradient

$$\nabla F(u, v) = \begin{pmatrix} 2\,u - 2\,v \\ -2\,u + 6\,v \end{pmatrix}.$$

Solving the pair of equations $\nabla F = \mathbf{0}$, namely

$$2\,u - 2\,v = 0, \qquad -2\,u + 6\,v = 0,$$

we see that the only critical point is the origin $u = v = 0$. To test whether the origin is a maximum or minimum, we further compute the Hessian matrix

$$H = \nabla^2 F(u,v) = \begin{pmatrix} F_{uu} & F_{uv} \\ F_{uv} & F_{vv} \end{pmatrix} = \begin{pmatrix} 2 & -2 \\ -2 & 6 \end{pmatrix}.$$

Using the methods of Section 3.5, we easily prove that the Hessian matrix is positive definite. Therefore, by Theorem 19.42, $\mathbf{u}^\star = \mathbf{0}$ is a strict local minimum of F.

Indeed, we recognize $F(u,v)$ to be, in fact, a homogeneous positive definite quadratic form, which can be written in the form

$$F(u,v) = \mathbf{u}^T K\,\mathbf{u}, \qquad \text{where} \qquad K = \begin{pmatrix} 1 & -1 \\ -1 & 3 \end{pmatrix} = \tfrac{1}{2}\,H, \qquad \mathbf{u} = \begin{pmatrix} u \\ v \end{pmatrix}.$$

Positive definiteness of the coefficient matrix K implies that $F(u,v) > 0$ for all $\mathbf{u} = (\,u, v\,)^T \neq \mathbf{0}$, and hence $\mathbf{0}$ is, in fact, a global minimum.

In general, any quadratic function $Q(\mathbf{u}) = Q(u_1, \dots, u_n)$ can be written in the form

$$Q(\mathbf{u}) = \mathbf{u}^T K\,\mathbf{u} - 2\,\mathbf{b}^T\mathbf{u} + c = \sum_{i,j=1}^{m} k_{ij}\,u_i\,u_j - 2\sum_{i=1}^{n} b_i\,u_i + c, \tag{19.71}$$

where $K = K^T$ is a symmetric $n \times n$ matrix, $\mathbf{b} \in \mathbb{R}^n$ is a fixed vector, and $c \in \mathbb{R}$ is a scalar. A straightforward computation produces the formula for its gradient and Hessian matrix:

$$\nabla Q(\mathbf{u}) = 2\,K\,\mathbf{u} - 2\,\mathbf{b}, \qquad \nabla^2 Q(\mathbf{u}) = 2\,K. \tag{19.72}$$

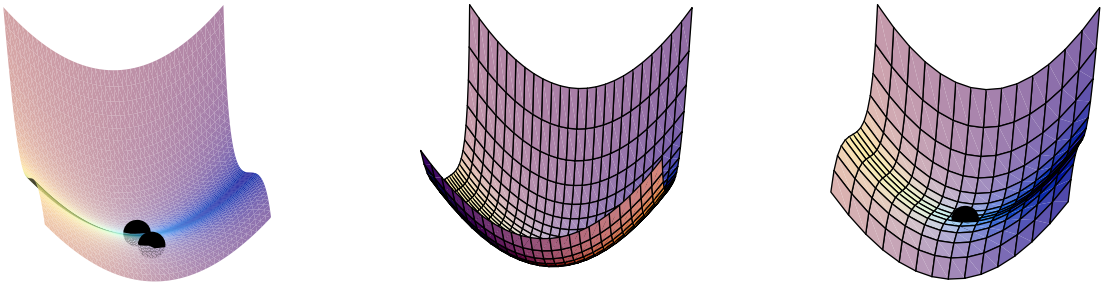
As a result, the critical points of the quadratic function are the solutions to the linear system $K\,\mathbf{u} = \mathbf{b}$. If K is nonsingular, there is a unique critical point $\mathbf{u}^\star$, which is a strict local minimum if and only if $K > 0$ is positive definite. In fact, Theorem 4.1 tells us that, in the positive definite case, $\mathbf{u}^\star$ is a strict *global* minimum for $Q(\mathbf{u})$. Thus, the algebraic approach of Chapter 4 provides additional, global information that cannot be gleaned directly from the local, multivariable calculus Theorem 19.42. But algebra is only able to handle quadratic minimization problems with ease. The analytical classification of minima and maxima of more complicated objective functions necessarily relies the gradient and Hessian criteria of Theorem 19.42.

Example 19.44. The function

$$F(u,v) = u^2 + v^2 - v^3 \qquad \text{has gradient} \qquad \nabla F(u,v) = \begin{pmatrix} 2\,u \\ 2\,v - 3\,v^2 \end{pmatrix}.$$

The critical point equation $\nabla F = \mathbf{0}$ has two solutions: $\mathbf{u}_1^\star = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $\mathbf{u}_2^\star = \begin{pmatrix} 0 \\ \frac{2}{3} \end{pmatrix}$. The Hessian matrix of the objective function is

$$\nabla^2 F(u,v) = \begin{pmatrix} 2 & 0 \\ 0 & 2 - 6\,v \end{pmatrix}.$$



$u^2 + v^2 - v^3$ $\qquad$ $u^2 + v^4$ $\qquad$ $u^2 + v^3$

Figure 19.19. Critical Points.

At the first critical point, the Hessian $\nabla^2 F(0,0) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$ is positive definite. Therefore, the origin is a strict local minimum. On the other hand, $\nabla^2 F\left(0, \frac{2}{3}\right) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}$ is indefinite, and hence $\mathbf{u}_2^\star = \begin{pmatrix} 0 \\ \frac{2}{3} \end{pmatrix}$ a saddle point. The function is graphed in Figure 19.19, with the critical points indicated by the small solid balls. The origin is, in fact, only a local minimum, since $F(0,0) = 0$, whereas $F(0,v) < 0$ for all $v > 1$. Thus, this particular function has no global minimum or maximum on $\mathbb{R}^2$.

Next, consider the function

$$F(u,v) = u^2 + v^4, \qquad \text{with gradient} \qquad \nabla F(u,v) = \begin{pmatrix} 2\,u \\ 4\,v^3 \end{pmatrix}.$$

The only critical point is the origin $u = v = 0$. The origin is a strict global minimum because $F(u,v) > 0 = F(0,0)$ for all $(u,v) \neq (\,0,0\,)^T$. However, its Hessian matrix

$$\nabla^2 F(u,v) = \begin{pmatrix} 2 & 0 \\ 0 & 12\,v^2 \end{pmatrix}$$

is only positive semi-definite at the origin, $\nabla^2 F(0,0) = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$.

On the other hand, the origin $u = v = 0$ is also the only critical point for the function

$$F(u,v) = u^2 + v^3 \qquad \text{with} \qquad \nabla F(u,v) = \begin{pmatrix} 2\,u \\ 3\,v^2 \end{pmatrix}.$$

The Hessian matrix is

$$\nabla^2 F(u,v) = \begin{pmatrix} 2 & 0 \\ 0 & 6\,v \end{pmatrix}, \qquad \text{and so} \qquad \nabla^2 F(0,0) = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}$$

is the same positive semi-definite matrix at the critical point. However, in this case $(0,0)$ is not a local minimum; indeed

$$F(0,v) < 0 = F(0,0) \qquad \text{whenever} \qquad v < 0,$$

and so there exist points arbitrarily close to the origin where F takes on smaller values. As illustrated in Figure 19.19, the origin is, in fact, a degenerate saddle point.

Finally, the function

$$F(u,v) = u^2 - 2\,u\,v + v^2 \qquad \text{has gradient} \qquad \nabla F(u,v) = \begin{pmatrix} 2\,u - 2\,v \\ -2\,u + 2\,v \end{pmatrix},$$

and so every point $u = v$ is a critical point. The Hessian matrix

$$\nabla^2 F(u,v) = \begin{pmatrix} F_{uu} & F_{uv} \\ F_{uv} & F_{vv} \end{pmatrix} = \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix}$$

is positive semi-definite everywhere. Since $F(u,u) = 0$, while $F(u,v) = (u-v)^2 > 0$ when $u \neq v$, each of these critical points is a non-isolated (and hence non-strict) local minimum. Thus, comparing the three preceding examples, we see that a semi-definite Hessian is unable to distinguish between different types of degenerate critical points.

Finally, the reader should always keep in mind that first and second derivative tests only determine the local behavior of the function near the critical point. They cannot be used to determine whether or not we are at a global minimum. This requires some additional analysis, and, often, a fair amount of ingenuity.

Proof of Theorem 19.42: We return to the proof of Theorem 19.39. Given a local minimum $\mathbf{u}^\star$, the scalar function $g(t) = F(\mathbf{u}^\star + t\,\mathbf{v})$ in (19.68) has a local minimum at $t = 0$. As noted above, basic calculus tells us that its derivatives at $t = 0$ must satisfy

$$g'(0) = 0, \qquad\qquad g''(0) \geq 0. \tag{19.73}$$

The first condition leads to the critical point equation $\nabla F(\mathbf{u}^\star) = \mathbf{0}$. A straightforward chain rule calculation produces the formula

$$g''(0) = \sum_{i,j=1}^{n} \frac{\partial^2 F}{\partial u_i\,\partial u_j}(\mathbf{u}^\star)\,v_i\,v_j = \mathbf{v}^T\,\nabla^2 F(\mathbf{u}^\star)\,\mathbf{v}.$$

As a result, the second condition in (19.73) requires that

$$\mathbf{v}^T\,\nabla^2 F(\mathbf{u}^\star)\,\mathbf{v} \geq 0.$$

Since this condition is required for every direction $\mathbf{v} \in \mathbb{R}^n$, the Hessian matrix $\nabla^2 F(\mathbf{u}^\star) \geq 0$ satisfies the criterion for positive semi-definiteness, proving the first part of the theorem.

The proof of the converse relies[†] on the second order Taylor expansion (C.16) of the function:

$$\begin{aligned} F(\mathbf{u}) &= F(\mathbf{u}^\star) + \nabla F(\mathbf{u}^\star)\cdot\mathbf{v} + \tfrac{1}{2}\,\mathbf{v}^T\nabla^2 F(\mathbf{u}^\star)\,\mathbf{v} + S(\mathbf{v},\mathbf{u}^\star) \\ &= F(\mathbf{u}^\star) + \tfrac{1}{2}\,\mathbf{v}^T\nabla^2 F(\mathbf{u}^\star)\,\mathbf{v} + S(\mathbf{v},\mathbf{u}^\star), \end{aligned} \qquad \text{where} \qquad \mathbf{v} = \mathbf{u} - \mathbf{u}^\star, \tag{19.74}$$

[†] Actually, it is not hard to prove the first part using the first order Taylor expansion without resorting to the scalar function g; see Exercise ∎. On the other hand, when we look at infinite-dimensional minimization problems arising in the calculus of variations, we will no longer have the luxury of appealilng to the finite-dimensional Taylor expansion, whereas the previous argument continues to apply in general contextx.

at the critical point, whence $\nabla F(\mathbf{u}^\star) = \mathbf{0}$. The remainder term in the Taylor formaul goes to 0 as $\mathbf{u} \to \mathbf{u}^\star$ at a rate faster than quadratic:

$$\frac{S(\mathbf{v}, \mathbf{u}^\star)}{\| \mathbf{v} \|^2} \longrightarrow 0 \qquad \text{as} \qquad \mathbf{v} \longrightarrow \mathbf{0}. \tag{19.75}$$

Assuming $\nabla^2 F(\mathbf{u}^\star)$, there is a constant $C > 0$ such that

$$\mathbf{v}^T \nabla^2 F(\mathbf{u}^\star)\, \mathbf{v} \geq C \| \mathbf{v} \|^2 \qquad \text{for all} \qquad \mathbf{v} \in \mathbb{R}^n.$$

This is a consequence of the Theorem 3.17 on the equivalence of norms, coupled with the fact that every positive definite matrix defines a norm. By (19.75), we can find $\delta > 0$ such that

$$| S(\mathbf{v}, \mathbf{u}^\star) | < \tfrac{1}{2} C \| \mathbf{v} \|^2 \qquad \text{whenever} \qquad 0 < \| \mathbf{v} \| = \| \mathbf{u} - \mathbf{u}^\star \| < \delta.$$

But then the Taylor formula (19.74) implies that, for all $\mathbf{u}$ satisfying the preceding inequality,

$$0 < \tfrac{1}{2} \mathbf{v}^T \nabla^2 F(\mathbf{u}^\star)\, \mathbf{v} + S(\mathbf{v}, \mathbf{u}^\star) = F(\mathbf{u}) - F(\mathbf{u}^\star),$$

which implies $\mathbf{u}^\star$ is a strict local minimum of $F(\mathbf{u})$. *Q.E.D.*

Constrained Optimization and Lagrange Multipliers

In many applications, the function to be minimized is subject to constraints. For instance, finding boundary minima requires constraining the minima to the boundary of the domain. Another example would be to find the minimum temperature on the surface of the earth. Assuming the earth is a perfect sphere of radius R, the temperature function $T(u, v, w)$ is then to be minimized subject to the constraint $u^2 + v^2 + w^2 = R^2$.

Let us focus on finding the minimum value of an objective function $F(\mathbf{u}) = F(u, v, w)$ when its arguments (u, v, w) are constrained to lie on a regular surface $S \subset \mathbb{R}^3$. Suppose $\mathbf{u}^\star = (\, u^\star, v^\star, w^\star \,)^T \in S$ is a (local) minimum for the constrained objective function. Let $\mathbf{u}(t) = (\, u(t), v(t), w(t) \,)^T \subset S$ be any curve contained within the surface that passes through the minimum, with $\mathbf{u}(0) = \mathbf{u}^\star$. Then the scalar function $g(t) = F(\mathbf{u}(t))$ must have a local minimum at $t = 0$, and hence, in view of (19.66),

$$0 = g'(0) = \frac{d}{dt} F(\mathbf{u}(t)) \,\Big|_{t=0} = \nabla F(\mathbf{u}(0)) \cdot \dot{\mathbf{u}}(0) = \nabla F(\mathbf{u}^\star) \cdot \dot{\mathbf{u}}(0)\,. \tag{19.76}$$

Thus, the gradient of the objective function at the surface minimum must be orthogonal to the tangent vector to the curve. Since the curve was constrained to lies entirely in S, its tangent vector $\dot{\mathbf{u}}(0)$ is tangent to the surface at the point $\mathbf{u}^\star$. Since every tangent vector to the surface is tangent to some curve contained in the surface, $\nabla F(\mathbf{u}^\star)$ must be orthogonal to every tangent vector, and hence point in the normal direction to the surface. Thus, a *contrained critical point* $\mathbf{u}^\star \in S$ of a function on a surface is defined so that

$$\nabla F(\mathbf{u}^\star) = \lambda\, \mathbf{n}, \tag{19.77}$$

where $\mathbf{n}$ denotes the normal to the surface at the point $\mathbf{u}^\star$. The scalar factor λ is known as the *Lagrange multiplier* in honor or Lagrange, one of the pioneers of constrained optimization. The value of the Lagrange multiplier is not fixed a priori., but must be determined by solving the critical point system (19.77). The same reasoning applies to local maxima, which are also constrained critical points. The nature of a constrained critical point — local minimum, local maximum, local saddle point, etc. — is fixed by a constrained second derivative test; details are discussed in Exercise ∎.

Example 19.45. Our problem is to find the minimum value of the objective function $F(u, v, w) = u^2 - 2w^3$ when u, v, w are restricted to the unit sphere $S = (u^2 + v^2 + w^2 = 1)$. The radial vector $\mathbf{n} = (\, u, v, w \,)^T$ is normal to the sphere, and so the critical point condition (19.77) is

$$\nabla F = \begin{pmatrix} 2u \\ 0 \\ -6w^2 \end{pmatrix} = \lambda \begin{pmatrix} u \\ v \\ w \end{pmatrix}.$$

Thus, we must solve the system of equations

$$2u = \lambda u, \qquad 0 = \lambda v, \qquad -6w^2 = \lambda w, \qquad \text{subject to} \qquad u^2 + v^2 + w^2 = 1,$$

for the unknowns u, v, w and λ. This needs to be done carefully to avoid missing any cases. First, if $u \neq 0$, then $\lambda = 2$, $v = 0$, and either $w = 0$ whence $u = \pm 1$, or $w = -\frac{1}{3}$ and so $u = \pm\sqrt{1 - w^2} = \pm\frac{2\sqrt{2}}{3}$. On the other hand, if $u = 0$, then either $\lambda = 0$, $w = 0$ and so $v = \pm 1$, or $v = 0$, $w = \pm 1$, and $\lambda = \mp 6$. Collecting these together, we discover that there are a total of 8 critical points on the unit sphere:

$$\mathbf{u}_1^\star = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \qquad \mathbf{u}_2^\star = \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix}, \qquad \mathbf{u}_3^\star = \begin{pmatrix} \frac{2\sqrt{2}}{3} \\ 0 \\ -\frac{1}{3} \end{pmatrix}, \qquad \mathbf{u}_4^\star = \begin{pmatrix} -\frac{2\sqrt{2}}{3} \\ 0 \\ -\frac{1}{3} \end{pmatrix},$$

$$\mathbf{u}_5^\star = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \qquad \mathbf{u}_6^\star = \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix}, \qquad \mathbf{u}_7^\star = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \qquad \mathbf{u}_8^\star = \begin{pmatrix} 0 \\ 0 \\ -1 \end{pmatrix}.$$

Since the unit sphere is closed and bounded, we are assured that F has a global maximum and a global minimum when restricted to S, which are both to be found among our candidate critical points. Thus, we merely compute the value of the objective function at each critical point,

$$\begin{aligned} &F(\mathbf{u}_1^\star) = 1, \qquad F(\mathbf{u}_2^\star) = 1, \qquad F(\mathbf{u}_3^\star) = \tfrac{22}{27}, \qquad F(\mathbf{u}_4^\star) = \tfrac{26}{27}, \\ &F(\mathbf{u}_5^\star) = 0, \qquad F(\mathbf{u}_6^\star) = 0, \qquad F(\mathbf{u}_7^\star) = -2, \qquad F(\mathbf{u}_8^\star) = 2. \end{aligned}$$

Therefore, $\mathbf{u}_7^\star$ must be the global minimum and $\mathbf{u}_8^\star$ the global maximum of the objective function restricted to the unit sphere. The status of the other six critical points — constrainned local maximum, minimum, or neither — is less evident, and a full classification requires the second derivative test outlined in Exercise ∎.

If the surface is given as the level set of a function

$$G(u, v, w) = c, \tag{19.78}$$

then at any point $\mathbf{u}^\star \in S$, the gradient vector $\nabla G(\mathbf{u}^\star)$ points in the normal direction to the surface, and hence, *provided* $\mathbf{n} = \nabla G(\mathbf{u}^\star) \neq \mathbf{0}$, the surface critical point condition can be rewritten as

$$\nabla F(\mathbf{u}^\star) = \lambda \nabla G(\mathbf{u}^\star), \tag{19.79}$$

or, in full detail, the critical point $(u^\star, v^\star, w^\star)^T$ must satisfy

$$\begin{aligned} \frac{\partial F}{\partial u}\,(u, v, w) &= \lambda\, \frac{\partial G}{\partial u}\,(u, v, w), \\ \frac{\partial F}{\partial v}\,(u, v, w) &= \lambda\, \frac{\partial G}{\partial v}\,(u, v, w), \\ \frac{\partial F}{\partial w}\,(u, v, w) &= \lambda\, \frac{\partial G}{\partial w}\,(u, v, w). \end{aligned} \tag{19.80}$$

Thus, to find the contrained critical points, one needs to solve the combined system (19.78, 80) of 4 equations for the four unknowns u, v, w and the Lagrange multiplier λ.

Formally, one can reformulate the problem as an unconstrained optimization problem by introducing the *augmented objective function*

$$H(u, v, w, \lambda) = F(u, v, w) - \lambda\big(G(u, v, w) - c \big). \tag{19.81}$$

The critical points of the augmented function are where its gradient, with respect to all four arguments, vanishes. Setting the partial derivatives with respect to u, v, w to 0 reproduces the system (19.80), while it partial derivative with respect to λ reporduces the constraint (19.78).

Example 19.46. Consider the problem of minimizing $F(u, v, w) =$ ■ when constrained to the surface ■ *torus*?

If $F(\mathbf{u})$ is defined on a closed subdomain $\Omega \subset \mathbb{R}^n$, then its minima may also occur at boundary points $\mathbf{u} \in \partial\Omega$. When the boundary is smooth, there is an analogous critical point condition for local boundary extrema.

Theorem 19.47. *Let $\Omega \subset \mathbb{R}^n$ be a domain with smooth boundary $\partial\Omega$. Suppose $F(\mathbf{u})$ is co;ntinuously differentiable at all points in $\overline{\Omega} = \Omega \cup \partial\Omega$. If the boundary point $\mathbf{u}_0 \in \partial\Omega$ is a (local) minimum for F when restricted to the closed domain $\overline{\Omega}$, then the gradient vector $\nabla F(\mathbf{u}_0)$ is either $\mathbf{0}$ or points inside the domain in the normal direction to $\partial\Omega$. See Figure bmin■ for a sketch of the geometrical configuration.*

Proof: Let $\mathbf{u}(t) \subset \partial\Omega$ be any curve that is entirely contained in the boundary, with $\mathbf{u}(0) = \mathbf{u}_0$. Then the scalar function $g(t) = F(\mathbf{u}(t))$ must have a local minimum at $t = 0$, and hence, in view of (19.66),

$$0 = g'(0) = \frac{d}{dt}\, F(\mathbf{u}(t)) \,\Big|_{t=0} = \big\langle\, \nabla F(\mathbf{u}(0))\,,\, \dot{\mathbf{u}}(0) \,\big\rangle\ .$$

Since the curve lies entirely in $\partial\Omega$, its tangent vector $\dot{\mathbf{u}}(0)$ is tangent to the boundary at the point $\mathbf{u}_0$; moreover, we can realize any such tangent vector by a suitable choice of curve. We conclude that $\nabla F(\mathbf{u}_0)$ is orthogonal to every tangent vector to $\partial\Omega$ at the point $\mathbf{u}_0$, and hence must point in the normal direction to the boundary. Moreover, if non-zero, it cannot point outside Ω since then $-\nabla F(\mathbf{u}_0)$, which is the direction of decrease in the objective function, would point inside the domain, which would preclude $\mathbf{u}_0$ from being a local minimum. *Q.E.D.*

The same ideas can be applied to optimization problems involving functions of several varaibles subject to one or more constraints. Suppose the objective function $F(\mathbf{u}) = F(u_1, \dots, u_n)$ is subject to the constraints

$$G_1(\mathbf{u}) = c_1, \qquad \cdots \qquad G_k(\mathbf{u}) = c_k. \tag{19.82}$$

A point $\mathbf{u}$ satisfying the constraints is termed *regular* if the corresponding gradient vectors $\nabla G_1(\mathbf{u}), \dots, \nabla G_k(\mathbf{u})$ are linearly independent. (Irregular points are more tricky, and must be handled separately.) A regular constrained critical point necessarily satisfies the vector equation

$$\nabla F(\mathbf{u}) = \lambda_1 \nabla G_1(\mathbf{u}) + \ \cdots \ + \lambda_1 \nabla G_1(\mathbf{u}), \tag{19.83}$$

where the unspecified scalars $\lambda_1, \dots, \lambda_k$ are called the Lagrange multipliers for the constrained optimization problem. The critical points are thus found by solving the combined system (19.82–83) for the $n + k$ varaibles $u_1, \dots, u_n$ and $\lambda_1, \dots, \lambda_k$. As in (19.81) we can reformulate this as an uncontrained optimization problem for the *augmented objective function*

$$H(\mathbf{u}, \boldsymbol{\lambda}) = F(\mathbf{u}) - \sum_{i=1}^{k} \lambda_i \big(G_i(\mathbf{u}) - c_i \big). \tag{19.84}$$

The gradient with respect to $\mathbf{u}$ reproduces the critical point system (19.83), while its gradient with respect to to $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_k)$ recovers the constraints (19.82).

Theorem 19.48. *Every regular constrained local minimum and maximum is a constrained critical point.*

Numerical Minimization of Scalar Functions

In practical optimization, one typically bypasses the preliminary characterization of minima as critical points, and instead implements a direct iterative procedure that constructs a sequence of successively better approximations to the desired minimum. As the computation progresses, the approximations are adjusted so that the objective function is made smaller and smaller, which, we hope, will ensure that we are converging to some form of minimum.

As always, to understand the issues involved, it is essential to consider the simplest scalar situation. Thus, we are given the problem of minimizing a scalar function $F(u)$ on a bounded interval $a \le u \le b$. The minimum value can either be at an endpoint or an interior minimum. Let us first state a result that plays a similar role to the Intermediate Value Lemma 19.17 that formed the basis of the Bisection Method for locating roots.

Lemma 19.49. *Suppose that $F(u)$ is defined and continuous for all $a \leq u \leq b$. Suppose that we can find a point $a < c < b$ such that $F(c) < F(a)$ and $F(c) < F(b)$. Then $F(u)$ has a minimum at some point $a < u^\star < b$.*

The proof is an easy consequence of Theorem 19.34. Therefore, if we find three points $a < c < b$ satisfying the conditions of the lemma, we are assured of the existence of a local minimum for the function between the two endpoints. Once this is done, we can design an algorithm to home in on the minimum $u^\star$. We choose another point, say d between a and c and evaluate $F(d)$. If $F(d) < F(c)$, then $F(d) < F(a)$ also, and so the points $a < d < c$ satisfy the hypotheses of Lemma 19.49. Otherwise, if $F(d) > F(c)$ then the points $d < c < b$ satisfy the hypotheses of the lemma. In either case, a local minimum has been narrowed down to a smaller interval, either $[a, c]$ or $[d, b]$. In the unlikely even that $F(d) = F(c)$, one can try another point instead — unless the objective function is constant, one will eventually find a suitable value of d. Iterating the method will produce a sequence of progressively smaller and smaller intervals in which the minimum is trapped, and, just like the Bisection Method, the endpoints of the intervals get closer and closer to $u^\star$.

The one question is how to choose the point d. We described the algorithm when it was selected to lie between a and c, but one could equally well try a point between c and b. To speed up the algorithm, it makes sense to place d in the larger of the two subintervals $[a, c]$ and $[c, b]$. One could try placing d in the midpoint of the interval, but a more inspired choice is to place it at a fraction $\theta = \frac{5}{\sqrt{2}} - \frac{1}{2} \approx .61803$ of the way along the interval, i.e., at $\theta\, a + (1 - \theta)\, c$ if$[a, c]$ is the longer interval. (One could equally well take the point $(1-\theta)\, a + \theta\, c$.) The result is the *Golden Section Method*. At each stage, the length of the interval has been reduced by a factor of θ, so the convergence rate is linear, although a bit slower than bisection.

Another strategy is to use an interpolating polynomial passing through the three points on the graph of $F(u)$ and use its minimum value as the next approximation to the minimum. According to Exercise 4.4.23, the minimizing value occurs at

$$d = \frac{m\, s - n\, t}{s - t}\,,$$

where

$$s = \frac{F(c) - F(a)}{c - a}\,, \qquad t = \frac{F(b) - F(c)}{b - c}\,, \qquad m = \frac{a + c}{2}\,, \qquad n = \frac{c + b}{2}\,.$$

As long as $a < c < b$ satisfy the hypothesis of Lemma 19.49, we are assured that the quadratic interpolant has a minimum (and not a maximum!), and that the minimum remains between the endpoints of the interval: $a < d < b$. If the length of the interval is small, the minimum value should be a good approximation to the minimizer $u^\star$ of $F(u)$ itself. Once d is determined, the algorithm proceeds as before. In this case, convergence is not quite guaranteed, or, in unfavorable situations, could be much slower than the Golden Section Method. One can even try using the method when the function values do not satisfy the hypothesis of Lemma 19.49, although now the new point d will not necessarily lie between a and b. Worse, the quadratic interpolant may have a maximum at d, and one

ends up going in the wrong direction, which can even happen in the minimizing case due to the discrepancy between it and the objective function $F(u)$. Thus, this method must be handled with care.

A final idea is to focus not on the objective function $F(u)$ but rather its derivative $f(u) = F'(u)$. The critical points of F are the roots of $f(u) = 0$, and so one can use one of the solution methods, e.g., Bisection or Newton's Method, to find the critical points. Of course, one must then take care that the critical point $u^\star$ is indeed a minimum, as it could equally well be a maximum of the original objective function. (It will probably not be an inflection point, as these do not correspond to simple roots of $f(u)$.) The status of the critical point can be checked by looking at the sign of $F''(u^\star) = f'(u^\star)$; indeed, if we use Newton's Method we will be computing the derivative at each stage of the algorithm, and can stop looking if the derivative turns out to be of the wrong sign.

Gradient Descent

Now, let us turn our attention to multi-dimensional optimization problems. We are seeking to minimize a (smooth) scalar objective function $F(\mathbf{u}) = F(u_1, \dots, u_n)$. According to Theorem 19.37, at any given point $\mathbf{u}$ in the domain of definition of F, the negative gradient vector $-\nabla F(\mathbf{u})$, if nonzero, points in the direction of the steepest decrease in F. Thus, to minimize F, an evident strategy is to "walk downhill", and, to be efficient, walk downhill as fast as possible, namely in the direction $-\nabla F(\mathbf{u})$. After walking in this direction for a little while, we recompute the gradient, and this tells us the new direction to head downhill. With luck, we will eventually end up at the bottom of the valley, i.e., at a (local) minimum value of the objective function.

This simple idea forms the basis of the *Gradient Descent Method* for minimizing the objective function $F(\mathbf{u})$. In a numerical implementation, we start the iterative procedure with an initial guess $\mathbf{u}^{(0)}$, and let $\mathbf{u}^{(k)}$ denote the k^{th} approximation to the minimum $\mathbf{u}^\star$. To compute the next approximation, we set out from $\mathbf{u}^{(k)}$ in the direction of the negative gradient, and set

$$\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} - t_k \, \nabla F(\mathbf{u}^{(k)}). \tag{19.85}$$

for some positive scalar $t_k > 0$. We are free to adjust t_k so as to optimize our descent path, and this is the key to the success of the method.

If $\nabla F(\mathbf{u}^{(k)}) \neq \mathbf{0}$, then, at least when $t_k > 0$ is sufficiently small,

$$F(\mathbf{u}^{(k+1)}) < F(\mathbf{u}^{(k)}), \tag{19.86}$$

and so $\mathbf{u}^{(k+1)}$ is, presumably, a better approximation to the desired minimum. Clearly, we cannot choose t_k too large or we run the risk of overshooting the minimum and reversing the inequality (19.86). Think of walking downhill in the Swiss Alps. If you walk too far in a straight line, which is what happens as t_k increases, then you might very well miss the valley and end up higher than you began — not a good strategy for descending to the bottom! On the other hand, if we choose t_k too small, taking very tiny steps, then the method may end up converging to the minimum much too slowly to be of practical use.

How should we choose an optimal value for the factor t_k? Keep in mind that the goal is to minimize $F(\mathbf{u})$. Thus, a good strategy would be to set t_k equal to the value of $t > 0$

that minimizes the scalar objective function

$$g(t) = F\big(\, \mathbf{u}^{(k)} - t\, \nabla F(\mathbf{u}^{(k)}) \,\big) \tag{19.87}$$

obtained by restricting $F(\mathbf{u})$ to the ray emanating from $\mathbf{u}^{(k)}$ that lies in the negative gradient direction. Physically, this corresponds to setting off in a straight line in the direction of steepest decrease, and continuing on until we cannot go down any further. Barring luck, we will not have reached the actual bottom of the valley, but must then readjust our direction and continue on down the hill in a series of straight line paths.

In practice, one can rarely compute the minimizing value $t^\star$ of (19.87) exactly. Instead, we employ one of the scalar minimization algorithms presented in the previous subsection. Note that we only need to look for a minimum among positive values of $t > 0$, since our choice of the negative gradient direction assures us that, at least for t sufficiently small and positive, $g(t) < g(0)$.

A more sophisticated approach is to employ the second order Taylor polynomial to approximate the function and then use its minimum (assuming such exists) as the next approximation to the desired minimizer. Specifically, if $\mathbf{u}^{(k)}$ is the current approximation to the minimum, then we approximate

$$F(\mathbf{u}) \approx c^{(k)} + (\mathbf{u} - \mathbf{u}^{(k)})^T \mathbf{g}^{(k)} + \tfrac{1}{2}\, (\mathbf{u} - \mathbf{u}^{(k)})^T H^{(k)} (\mathbf{u} - \mathbf{u}^{(k)}) \tag{19.88}$$

near $\mathbf{u}^{(k)}$, where

$$c^{(k)} = F(\mathbf{u}^{(k)}), \qquad \mathbf{g}^{(k)} = \nabla F(\mathbf{u}^{(k)}), \qquad H^{(k)} = \nabla^2 F(\mathbf{u}^{(k)}), \tag{19.89}$$

are, respectively, the function value, the gradient and the Hessian at the current iterate. If $\mathbf{u}\star$ is a stric local minimum, then $\nabla^2 F(\mathbf{u}^\star)$ is positive definite, and hence, assuming $\mathbf{u}^{(k)}$ is close, so is $H^{(k)} = \nabla^2 F(\mathbf{u}^{(k)})$. Thus, by Theorem 4.1, the quadratic Taylor approximation has a unique minimum value $\mathbf{u}^{(k+1)}$ which satisfies the linear system

$$H^{(k)}(\mathbf{u}^{(k+1)} - \mathbf{u}^{(k)}) = -\,\mathbf{g}^{(k)}. \tag{19.90}$$

The solution serves to define the next approximation $\mathbf{u}^{(k+1)}$. While not guaranteed to converge, the method does perform well in all reasonable situations.

Chapter 20

Nonlinear Ordinary Differential Equations

This chapter is concerned with initial value problems for systems of ordinary differential equations. We have already dealt with the linear case in Chapter 9, and so here our emphasis will be on nonlinear phenomena and properties, particularly those with physical relevance. Finding a solution to a differential equation may not be so important if that solution never appears in the physical model represented by the system, or is only realized in exceptional circumstances. Thus, equilibrium solutions, which correspond to configurations in which the physical system does not move, only occur in everyday situations if they are stable. An unstable equilibrium will not appear in practice, since slight perturbations in the system or its physical surroundings will immediately dislodge the system far away from equilibrium.

Of course, very few nonlinear systems can be solved explicitly, and so one must typically rely on a numerical scheme to accurately approximate the solution. Basic methods for initial value problems, beginning with the simple Euler scheme, and working up to the extremely popular Runge–Kutta fourth order method, will be the subject of the final section of the chapter. However, numerical schemes do not always give accurate results, and we breifly discuss the class of stiff differential equations, which present a more serious challenge to numerical analysts.

Without some basic theoretical understanding of the nature of solutions, equilibrium points, and stability properties, one would not be able to understand when numerical solutions (even those provided by standard well-used packages) are to be trusted. Moreover, when testing a numerical scheme, it helps to have already assembled a repertoire of nonlinear problems in which one already knows one or more explicit analytic solutions. Further tests and theoretical results can be based on first integrals (also known as conservation laws) or, more generally, Lyapunov functions. Although we have only space to touch on these topics briefly, but, we hope, this will whet the reader's appetite for delving into this subject in more depth. The references [**19**, **56**, **89**, **99**, **105**] can be profitably consulted.

20.1. First Order Systems of Ordinary Differential Equations.

Let us begin by introducing the basic object of study in discrete dynamics: the initial value problem for a first order system of ordinary differential equations. Many physical applications lead to higher order systems of ordinary differential equations, but there is a simple reformulation that will convert them into equivalent first order systems. Thus, we do not lose any generality by restricting our attention to the first order case throughout. Moreover, numerical solution schemes for higher order initial value problems are entirely based on their reformulation as first order systems.

As always, when confronted with a new problem, it is essential to fully understand the simplest case first. Thus, we begin with a single scalar, first order ordinary differential equation

$$\frac{du}{dt} = F(t,u). \tag{20.1}$$

In many applications, the independent variable t represents time, and the unknown function $u(t)$ is some dynamical physical quantity. Throughout this chapter, all quantities are assumed to be real. (Results on complex ordinary differential equations can be found in [**98**].) Under appropriate conditions on the right hand side (to be formalized in the following section), the solution $u(t)$ is uniquely specified by its value at a single time,

$$u(t_0) = u_0. \tag{20.2}$$

The combination (20.1–2) is referred to as an *initial value problem*, and our goal is to devise both analytical and numerical solution strategies.

A differential equation is called *autonomous* if the right hand side does not explicitly depend upon the time variable:

$$\frac{du}{dt} = F(u). \tag{20.3}$$

All autonomous scalar equations can be solved by direct integration. We divide both sides by $F(u)$, whereby

$$\frac{1}{F(u)}\,\frac{du}{dt} = 1,$$

and then integrate with respect to t; the result is

$$\int \frac{1}{F(u)}\,\frac{du}{dt}\,dt = \int dt = t + k,$$

where k is the constant of integration. The left hand integral can be evaluated by the change of variables that replaces t by u, whereby $du = (du/dt)\,dt$, and so

$$\int \frac{1}{F(u)}\,\frac{du}{dt}\,dt = \int \frac{du}{F(u)} = G(u),$$

where $G(u)$ indicates a convenient anti-derivative[†] of the function $1/F(u)$. Thus, the solution can be written in implicit form

$$G(u) = t + k. \tag{20.4}$$

If we are able to solve the implicit equation (20.4), we may thereby obtain the explicit solution

$$u(t) = H(t + k) \tag{20.5}$$

[†] Technically, a second constant of integration should appear here, but this can be absorbed into the previous constant k, and so proves to be unnecessary.

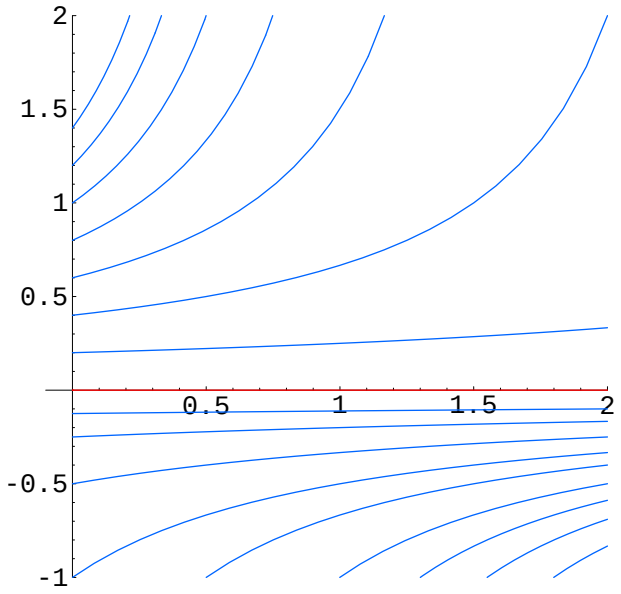


Figure 20.1. Solutions to $\dot{u} = u^2$.

in terms of the inverse function $H = G^{-1}$. Finally, to satisfy the initial condition (20.2), we set $t = t_0$ in the implicit solution formula (20.4), whereby $G(u_0) = t_0 + k$. Therefore, the solution to our initial value problem is

$$G(u) - G(u_0) = t - t_0, \qquad \text{or, explicitly,} \qquad u(t) = H\big(t - t_0 + G(u_0)\big). \tag{20.6}$$

Remark: A more direct version of this solution technique is to rewrite the differential equation (20.3) in the "separated form"

$$\frac{du}{F(u)} = dt,$$

in which all terms involving u, including its differential du, are collected on the left hand side of the equation, while all terms involving t and its differential are placed on the right, and then formally integrate both sides, leading to the same implicit solution formula:

$$G(u) = \int \frac{du}{F(u)} = \int dt = t + k. \tag{20.7}$$

Before completing our analysis of this solution method, let us run through a couple of elementary examples.

Example 20.1. Consider the autonomous initial value problem

$$\frac{du}{dt} = u^2, \qquad u(t_0) = u_0. \tag{20.8}$$

To solve the differential equation, we rewrite it in the separated form

$$\frac{du}{u^2} = dt, \qquad \text{and then integrate both sides:} \qquad -\frac{1}{u} = \int \frac{du}{u^2} = t + k.$$

Solving the resulting algebraic equation for u, we deduce the solution formula

$$u = -\frac{1}{t+k}\,. \tag{20.9}$$

To specify the integration constant k, we evaluate u at the initial time t_0; this implies

$$u_0 = -\frac{1}{t_0+k}\,, \qquad \text{so that} \qquad k = -\frac{1}{u_0} - t_0.$$

Therefore, the solution to the initial value problem is

$$u = \frac{u_0}{1 - u_0(t-t_0)}\,. \tag{20.10}$$

Figure 20.1 shows the graphs of some typical solutions.

As t approaches the critical value $t^\star = t_0 + 1/u_0$ from below, the solution "blows up", meaning $u(t) \to \infty$ as $t \to t^\star$. The blow-up time $t^\star$ depends upon the initial data — the larger $u_0 > 0$ is, the sooner the solution goes off to infinity. If the initial data is negative, $u_0 < 0$, the solution is well-defined for all $t > t_0$, but has a singularity in the past, at $t^\star = t_0 + 1/u_0 < t_0$. The only solution that exists for all positive and negative time is the constant solution $u(t) \equiv 0$, corresponding to the initial condition $u_0 = 0$.

In general, the constant *equilibrium solutions* to an autonomous ordinary differential equation, also known as its *fixed points*, play a distinguished role. If $u(t) \equiv u^\star$ is a constant solution, then $du/dt \equiv 0$, and hence the differential equation (20.3) implies that $F(u^\star) = 0$. Therefore, the equilibrium solutions coincide with the *roots* of the function $F(u)$. In point of fact, since we divided by $F(u)$, the derivation of our formula for the solution (20.7) assumed that we were *not* at an equilibrium point. In the preceding example, our final solution formula (20.10) happens to include the equilibrium solution $u(t) \equiv 0$, corresponding to $u_0 = 0$, but this is a lucky accident. Indeed, the equilibrium solution does *not* appear in the "general" solution formula (20.9). One must typically take extra care that equilibrium solutions do not elude us when utilizing this basic integration method.

Example 20.2. Although a population of people, animals, or bacteria consists of individuals, the aggregate behavior can often be effectively modeled by a dynamical system that involves continuously varying variables. As first proposed by the English economist Thomas Malthus in 1798, the population of a species grows, roughly, in proportion to its size. Thus, the number of individuals $N(t)$ at time t satisfies a first order differential equation of the form

$$\frac{dN}{dt} = \rho\, N, \tag{20.11}$$

where the proportionality factor $\rho = \beta - \delta$ measures the rate of growth, namely the difference between the birth rate $\beta \geq 0$ and the death rate $\delta \geq 0$. Thus, if births exceed deaths, $\rho > 0$, and the population increases, whereas if $\rho < 0$, more individuals are dying and the population shrinks.

In the very simplest model, the growth rate ρ is assumed to be independent of the population size, and (20.11) reduces to the simple linear ordinary differential equation

(8.1) that we solved at the beginning of Chapter 8. The solutions satisfy the Malthusian exponential growth law $N(t) = N_0\, e^{\rho\, t}$, where $N_0 = N(0)$ is the initial population size. Thus, if $\rho > 0$, the population grows without limit, while if $\rho < 0$, the population dies out, so $N(t) \to 0$ as $t \to \infty$, at an exponentially fast rate. The Malthusian population model provides a reasonably accurate description of the behavior of an isolated population in an environment with unlimited resources.

In a more realistic scenario, the growth rate will depend upon the size of the population as well as external environmental factors. For example, in the presence of limited resources, relatively small populations will increase, whereas an excessively large population will have insufficient resources to survive, and so its growth rate will be negative. In other words, the growth rate $\rho(N) > 0$ when $N < N^\star$, while $\rho(N) < 0$ when $N > N^\star$, where the *carrying capacity* $N^\star > 0$ depends upon the resource availability. The simplest class of functions that satifies these two inequalities are of the form $\rho(N) = \lambda(N^\star - N)$, where $\lambda > 0$ is a positive constant. This leads us to the nonlinear population model

$$\frac{dN}{dt} = \lambda\, N\,(N^\star - N). \tag{20.12}$$

In deriving this model, we assumed that the environment is not changing over time; a dynamical environment would require a more complicated non-autonomous differential equation.

Before analyzing the solutions to the nonlinear population model, let us make a preliminary change of variables, and set $u(t) = N(t)/N^\star$, so that u represents the size of the population in proportion to the *carrying capacity* $N^\star$. A straightforward computation shows that $u(t)$ satisfies the so-called *logistic differential equation*

$$\frac{du}{dt} = \lambda\, u\,(1 - u), \qquad\qquad u(0) = u_0, \tag{20.13}$$

where we assign the initial time to be $t_0 = 0$ for simplicity. The logistic differential equation can be viewed as the continuous counterpart of the logistic map (19.19). However, unlike its discrete namesake, the logistic differential equation is quite sedate, and its solutions easily understood.

First, there are two equilibrium solutions: $u(t) \equiv 0$ and $u(t) \equiv 1$, obtained by setting the right hand side of the equation equal to zero. The first represents a nonexistent population with no individuals and hence no reproduction. The second equilibrium solution corresponds to a static population $N(t) \equiv N^\star$ that is at the ideal size for the environment, so deaths exactly balance births. In all other situations, the population size will vary over time.

To integrate the logistic differential equation, we proceed as above, first writing it in the separated form

$$\frac{du}{u(1-u)} = \lambda\, dt.$$

Integrating both sides, and using partial fractions,

$$\lambda t + k = \int \frac{du}{u(1-u)} = \int \left[\, \frac{1}{u} \;+\; \frac{1}{1-u} \,\right] du = \log \left| \frac{u}{1-u} \right| ,$$

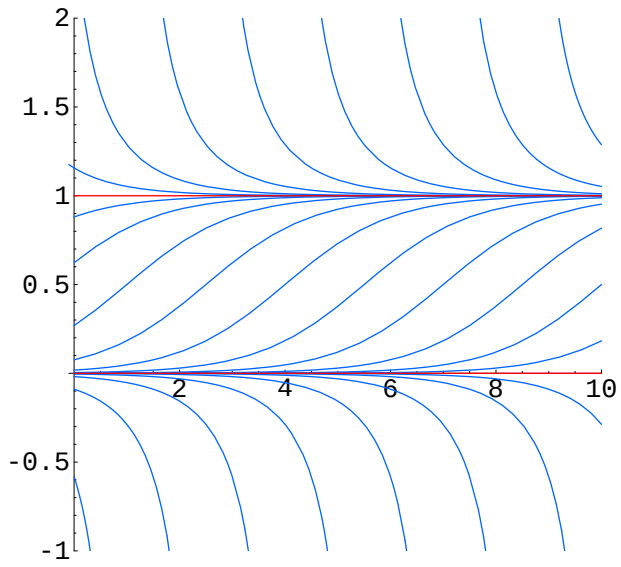


Figure 20.2. Solutions to $u' = u(1-u)$.

where k is a constant of integration. Therefore

$$\frac{u}{1-u} = c e^{\lambda t}, \qquad \text{where} \qquad c = \pm e^{k}.$$

Solving for u, we deduce the solution

$$u(t) = \frac{c e^{\lambda t}}{1 + c e^{\lambda t}}\,. \tag{20.14}$$

The constant of integration is fixed by the initial condition. Solving the algebraic equation

$$u_0 = u(0) = \frac{c}{1+c} \qquad \text{yields} \qquad c = \frac{u_0}{1-u_0}\,.$$

Substituting the result back into the solution formula (20.14) and simplifying, we find

$$u(t) = \frac{u_0\, e^{\lambda t}}{1 - u_0 + u_0\, e^{\lambda t}}\,. \tag{20.15}$$

The resulting solutions are illustrated in Figure 20.2. Interestingly, while the equilibrium solutions are not covered by the integration method, they reappear in the final solution formula, corresponding to initial data $u_0 = 0$ and $u_0 = 1$ respectively. However, this is a lucky accident, and cannot be anticipated in more complicated situations.

When using the logistic equation to model population dynamics, the initial data is assumed to be positive, $u_0 > 0$. As time $t \to \infty$, the solution (20.15) tends to the equilibrium value $u(t) \to 1$ — which corresponds to $N(t) \to N^\star$ approaching the carrying capacity in the original population model. For small initial values $u_0 \ll 1$ the solution initially grows at an exponential rate λ, corresponding to a population with unlimited resources. However, as the population increases, the gradual lack of resources tends to

slow down the growth rate, and eventually the population saturates at the equilibrium value. On the other hand, if $u_0 > 1$, the population is too large to be sustained by the available resources, and so dies off until it reaches the same saturation value. If $u_0 = 0$, then the solution remains at equilibrium $u(t) \equiv 0$. Finally, when $u_0 < 0$, the solution only exists for a finite amount of time, with

$$u(t) \longrightarrow -\infty \qquad \text{as} \qquad t \longrightarrow t^\star = \frac{1}{\lambda} \log\left(1 - \frac{1}{u_0}\right).$$

Of course, this final case does appear in the physical world, since we cannot have a negative population!

The separation of variables method used to solve autonomous equations can be straightforwardly extended to a special class of non-autonomous equations. A *separable* ordinary differential equation has the form

$$\frac{du}{dt} = a(t)\, F(u), \tag{20.16}$$

in which the right hand side is the product of a function of t and a function of u. To solve the equation, we rewrite it in the separated form

$$\frac{du}{F(u)} = a(t)\, dt.$$

Integrating both sides leads to the solution in implicit form

$$G(u) = \int \frac{du}{F(u)} = \int a(t)\, dt = A(t) + k. \tag{20.17}$$

The integration constant k is then fixed by the initial condition. And, as before, one must properly account for any equilibrium solutions, when $F(u) = 0$.

Example 20.3. Let us solve the particular initial value problem

$$\frac{du}{dt} = (1 - 2t)\, u, \qquad u(0) = 1. \tag{20.18}$$

We begin by writing the differential equation in separated form

$$\frac{du}{u} = (1 - 2t)\, dt.$$

Integrating both sides leads to

$$\log u = \int \frac{du}{u} = \int (1 - 2t)\, dt = t - t^2 + k,$$

where k is the constant of integration. We can readily solve for

$$u(t) = c e^{t - t^2},$$

where $c = \pm e^k$. The latter formula constitutes the general solution to the differential equation, and happens to include the equilibrium solution $u(t) \equiv 0$ when $c = 0$. The given initial condition requires that $c = 1$, and hence $u(t) = e^{t - t^2}$ is the unique solution to the initial value problem. The solution is graphed in Figure 20.3.

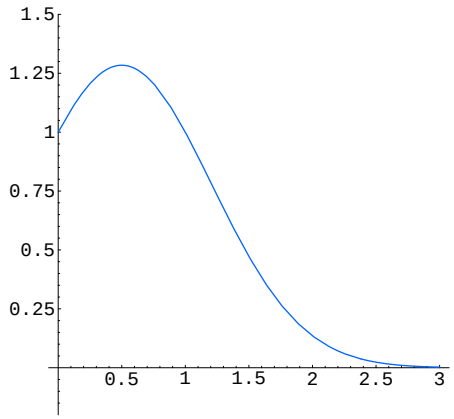


Figure 20.3. Solution to the Initial Value Problem $\dot{u} = (1 - 2t)\, u$, $u(0) = 1$.

First Order Systems

A *first order system of ordinary differential equations* has the general form

$$\frac{du_1}{dt} = F_1(t, u_1, \dots, u_n), \qquad \cdots \qquad \frac{du_n}{dt} = F_n(t, u_1, \dots, u_n). \tag{20.19}$$

The unknowns $u_1(t), \dots, u_n(t)$ are scalar functions of the real variable t, which usually represents time. We shall write the system more compactly in vector form

$$\frac{d\mathbf{u}}{dt} = \mathbf{F}(t, \mathbf{u}), \tag{20.20}$$

where $\mathbf{u}(t) = (\, u_1(t), \dots, u_n(t)\,)^T$, and $\mathbf{F}(t, \mathbf{u}) = (\, F_1(t, u_1, \dots, u_n), \dots, F_n(t, u_1, \dots, u_n)\,)^T$ is a vector-valued function of $n + 1$ variables. By a *solution* to the differential equation, we mean a vector-valued function $\mathbf{u}(t)$ that is defined and continuously differentiable on an interval $a < t < b$, and, moreover, satisfies the differential equation on its interval of definition. Each solution $\mathbf{u}(t)$ serves to parametrize a curve $C \subset \mathbb{R}^n$, also known as a *trajectory* or *orbit* of the system.

In this chapter, we shall concentrate on initial value problems for such first order systems. The general initial conditions are

$$u_1(t_0) = a_1, \qquad u_2(t_0) = a_2, \qquad \cdots \qquad u_n(t_0) = a_n, \tag{20.21}$$

or, in vectorial form,

$$\mathbf{u}(t_0) = \mathbf{a} \tag{20.22}$$

Here t_0 is a prescribed initial time, while the vector $\mathbf{a} = (\, a_1, a_2, \dots, a_n\,)^T$ fixes the initial position of the desired solution. In favorable situations, as described below, the initial conditions serve to uniquely specify a solution to the differential equations — at least for nearby times. The general issues of existence and uniquenss of solutions will be addressed in the following section.

A system of differential equations is called *autonomous* if the right hand side does not explicitly depend upon the time t, and so takes the form

$$\frac{d\mathbf{u}}{dt} = \mathbf{F}(\mathbf{u}). \tag{20.23}$$

One important class of autonomous first order systems are the steady state fluid flows. Here $\mathbf{F}(\mathbf{u}) = \mathbf{v}$ represents the fluid velocity vector field at the position $\mathbf{u}$. The solution $\mathbf{u}(t)$ to the initial value problem (20.23, 22) describes the motion of a fluid particle that starts at position $\mathbf{a}$ at time t_0. The differential equation tells us that the fluid velocity at each point on the particle's trajectory matches the prescribed vector field. Additional details can be found in Chapter 16 and Appendices A and B.

An *equilibrium solution* is constant: $\mathbf{u}(t) \equiv \mathbf{u}^\star$ for all t. Thus, its derivative must vanish, $d\mathbf{u}/dt \equiv \mathbf{0}$, and hence, every equilibrium solution arises as a solution to the system of algebraic equations

$$\mathbf{F}(\mathbf{u}^\star) = \mathbf{0} \tag{20.24}$$

prescribed by the vanishing of the right hand side of the system (20.23).

Example 20.4. A *predator-prey system* is a simplified ecological model of two species: the predators which feed on the prey. For example, the predators might be lions roaming the Serengeti and the prey zebra. We let $u(t)$ represent the number of prey, and $v(t)$ the number of predators at time t. Both species obey a population growth model of the form (20.11), and so the dynamical equations can be written as

$$\frac{du}{dt} = \rho\, u, \qquad\qquad \frac{dv}{dt} = \sigma\, v,$$

where the growth rates ρ, σ may depend upon the other species. The more prey, i.e., the larger u is, the faster the predators reproduce, while a lack of prey will cause them to die off. On the other hand, the more predators, the faster the prey are consumed and the slower their net rate of growth.

If we assume that the environment has unlimited resources for the prey, which, barring drought, is probably valid in the case of the zebras, then the simplest model that incorporates these assumptions is the *Lotka–Volterra system*

$$\frac{du}{dt} = \alpha\, u - \delta\, u\, v, \qquad\qquad \frac{dv}{dt} = -\beta\, v + \gamma\, u\, v, \tag{20.25}$$

corresponding to growth rates $\rho = \alpha - \delta\, v$, $\sigma = -\beta + \gamma\, u$. The parameters $\alpha, \beta, \gamma, \delta > 0$ are all positive, and their precise values will depend upon the species involved and how they interact, as indicated by field data, combined with, perhaps, educated guesses. In particular, α represents the unrestrained growth rate of the prey in the absence of predators, while $-\beta$ represents the rate that the predators die off in the absence of their prey. The nonlinear terms model the interaction of the two species: the rate of increase in the predators is proportional to the number of available prey, while the rate of decrese in the prey is proportional to the number of predators. The initial conditions $u(t_0) = u_0$, $v(t_0) = v_0$ represent the initial populations of the two species.

We will discuss the integration of the Lotka–Volterra system (20.25) in Section 20.3. Here, let us content ourselves with determining the possible equilibria. Setting the right hand sides of the system to zero leads to the nonlinear algebraic system

$$0 = \alpha\, u - \delta\, u\, v = u(\alpha - \delta\, v), \qquad 0 = -\beta\, v + \gamma\, u\, v = v(-\beta + \gamma\, u).$$

Thus, there are two distinct equilibria, namely

$$u_1^\star = v_1^\star = 0, \qquad u_2^\star = \beta/\gamma, \quad v_2^\star = \alpha/\delta.$$

The first is the uninteresting (or, rather catastropic) situation where there are no animals — no predators and no prey. The second is a nontrivial solution in which both populations maintain a steady value, for which the birth rate of the prey is precisely sufficient to continuously feed the predators. Is this a feasible solution? Or, to state the question more mathematically, is this a stable equilibrium? We shall develop the tools to answer this question below.

Higher Order Systems

A wide variety of physical systems are modeled by nonlinear systems of differential equations depending upon second and, occasionally, even higher order derivatives of the unknowns. But there is an easy device that will reduce any higher order ordinary differential equation or system to an equivalent first order system. "Equivalent" means that each solution to the first order system uniquely corresponds to a solution to the higher order equation and vice versa. The upshot is that, for all practical purposes, one only needs to analyze first order systems. Moreover, the vast majority of numerical solution algorithms are designed for first order systems, and so to numerically integrate a higher order equation, one must place it into an equivalent first order form.

We have already encountered the main idea in our discussion of the phase plane approach to second order scalar equations

$$\frac{d^2u}{dt^2} = F\left(t, u, \frac{du}{dt}\right). \tag{20.26}$$

We introduce a new dependent variable $v = \dfrac{du}{dt}$. Since $\dfrac{dv}{dt} = \dfrac{d^2u}{dt^2}$, the functions u, v satisfy the equivalent first order system

$$\frac{du}{dt} = v, \qquad \frac{dv}{dt} = F(t, u, v). \tag{20.27}$$

Conversely, it is easy to check that if $\mathbf{u}(t) = \big(\, u(t), v(t)\,\big)^T$ is any solution to the first order system, then its first component $u(t)$ defines a solution to the scalar equation, which establishes their equivalence. The basic initial conditions $u(t_0) = u_0$, $v(t_0) = v_0$, for the first order system translate into a pair of initial conditions $u(t_0) = u_0$, $\dot u(t_0) = v_0$, specifying the value of the solution and its first order derivative for the second order equation.

Similarly, given a third order equation

$$\frac{d^3u}{dt^3} = F\left(t, u, \frac{du}{dt}, \frac{d^2u}{dt^2}\right),$$

we set

$$v = \frac{du}{dt}\,, \qquad\qquad w = \frac{dv}{dt} = \frac{d^2u}{dt^2}\,.$$

The variables u, v, w satisfy the equivalent first order system

$$\frac{du}{dt} = v, \qquad\qquad \frac{dv}{dt} = w, \qquad\qquad \frac{dw}{dt} = F(t, u, v, w).$$

The general technique should now be clear.

Example 20.5. The forced *van der Pol equation*

$$\frac{d^2u}{dt^2} + (u^2 - 1)\,\frac{du}{dt} + u = f(t) \tag{20.28}$$

arises in the modeling of an electrical circuit with a triode whose resistance changes with the current, [**EE**]. It also arises in certain chemical reactions, [**odz**], and wind-induced motions of structures, [**odz**]. To convert the van der Pol equation into an equivalent first order system, we set $v = du/dt$, whence

$$\frac{du}{dt} = v, \qquad\qquad \frac{dv}{dt} = f(t) - (u^2 - 1)\,v - u, \tag{20.29}$$

is the equivalent phase plane system.

Example 20.6. The Newtonian equations for a mass m moving in a potential force field are a second order system of the form

$$m\,\frac{d^2\mathbf{u}}{dt^2} = -\,\nabla F(\mathbf{u})$$

in which $\mathbf{u}(t) = (\,u(t), v(t), w(t)\,)^T$ represents the position of the mass and $F(\mathbf{u}) = F(u, v, w)$ the potential function. In components,

$$m\,\frac{d^2u}{dt^2} = -\,\frac{\partial F}{\partial u}\,, \qquad m\,\frac{d^2v}{dt^2} = -\,\frac{\partial F}{\partial v}\,, \qquad m\,\frac{d^2w}{dt^2} = -\,\frac{\partial F}{\partial w}\,. \tag{20.30}$$

For example, a planet moving in the sun's gravitational field satisfies the Newtonian system for the gravitational potential

$$F(\mathbf{u}) = -\,\frac{\alpha}{\|\,\mathbf{u}\,\|} = -\,\frac{\alpha}{\sqrt{u^2 + v^2 + w^2}}\,, \tag{20.31}$$

where α depends on the masses and the universal gravitational constant. (This simplified model ignores all interplanetary forces.) Thus, the mass' motion in such a gravitational force field follows the solution to the second order Newtonian system

$$m\,\frac{d^2\mathbf{u}}{dt^2} = -\,\nabla F(\mathbf{u}) = -\,\frac{\alpha\,\mathbf{u}}{\|\,\mathbf{u}\,\|^3} = \frac{\alpha}{(u^2 + v^2 + w^2)^{3/2}}\begin{pmatrix} u \\ v \\ w \end{pmatrix}.$$

The same system of ordinary differential equations describes the motion of a charged particle in a Coulomb electric force field, where the sign of α is positive for attracting opposite charges, and negative for repelling like charges.

To convert the second order Newton equations into a first order system, we set $\mathbf{v} = \dot{\mathbf{u}}$ to be the mass' velocity vector, with components

$$p = \frac{du}{dt}\,, \qquad q = \frac{dv}{dt}\,, \qquad r = \frac{dw}{dt}\,,$$

and so

$$\begin{aligned} &\frac{du}{dt} = p, && \frac{dv}{dt} = q, && \frac{dw}{dt} = r, \\ &\frac{dp}{dt} = -\frac{1}{m}\,\frac{\partial F}{\partial u}\,(u,v,w), && \frac{dq}{dt} = -\frac{1}{m}\,\frac{\partial F}{\partial v}\,(u,v,w), && \frac{dr}{dt} = -\frac{1}{m}\,\frac{\partial F}{\partial w}\,(u,v,w). \end{aligned} \tag{20.32}$$

One of Newton's greatest acheivements was to solve this system in the case of the central gravitational potential (20.31), and thereby confirm the validity of Kepler's laws of planetary motion.

Finally, we note that there is a simple device that will convert any non-autonomous system into an equivalent autonomous system involving one additional variable. Namely, one introduces an extra coordinate $u_0 = t$ to represent the time, which satisfies the elementary differential equation $du_0/dt = 1$ with initial condition $u_0(t_0) = t_0$. Thus, the original system (20.19) can be written in the autonomous form

$$\frac{du_0}{dt} = 1, \qquad \frac{du_1}{dt} = F_1(u_0, u_1, \dots, u_n), \qquad \cdots \qquad \frac{du_n}{dt} = F_n(u_0, u_1, \dots, u_n). \tag{20.33}$$

For example, the autonomous form of the forced van der Pol system (20.29) is

$$\frac{du_0}{dt} = 1, \qquad \frac{du_1}{dt} = u_2, \qquad \frac{du_2}{dt} = f(u_0) - (u_1^2 - 1)u_2 - u_1, \tag{20.34}$$

in which u_0 represents the time variable.

20.2. Existence, Uniqueness, and Continuous Dependence.

It goes without saying that there is no general analytical method that will solve all differential equations. Indeed, even relatively simple first order, scalar, non-autonomous ordinary differential equations cannot be solved in closed form. For example, the solution to the particular *Riccati equation*

$$\frac{du}{dt} = u^2 + t \tag{20.35}$$

cannot be written in terms of elementary functions, although the method of Exercise ■ can be used to obtain a formula that relies on Airy functions, cf. (C.42, 44). The *Abel equation*

$$\frac{du}{dt} = u^3 + t \tag{20.36}$$

fares even worse, since its general solution cannot be written in terms of even standard special functions — although power series solutions can be tediously ground out term by term. Understanding when a given differential equation can be solved in terms of elementary functions or known special functions is an active area of contemporary research, [**31**]. In this vein, we cannot resist mentioning that the most important class of exact solution techniques for differential equations are those based on symmetry. An introduction can be found in the author's graduate level monograph [**141**]; see also [**38**, **104**].

Existence

Before worrying about how to solve a differential equation, either analytically, qualitatively, or numerically, it behooves us to try to resolve the core mathematical issues of existence and uniqueness. First, does a solution exist? If, not, it makes no sense trying to find one. Second, is the solution uniquely determined? Otherwise, the differential equation probably has scant relevance for physical applications since we cannot use it as a predictive tool. Since differential equations inevitably have lots of solutions, the only way in which we can deduce uniqueness is by imposing suitable initial (or boundary) conditions.

Unlike partial differential equations, which must be treated on a case-by-case basis, there are complete general answers to both the existence and uniqueness questions for initial value problems for systems of ordinary differential equations. (Boundary value problems are more subtle.) While obviously important, we will not take the time to present the proofs of these fundamental results, which can be found in most advanced textbooks on the subject, including [**19**, **89**, **99**, **105**].

Let us begin by stating the Fundamental Existence Theorem for initial value problems associated with first order systems of ordinary differential equations.

Theorem 20.7. *Let $\mathbf{F}(t, \mathbf{u})$ be a continuous function. Then the initial value problem*[†]

$$\frac{d\mathbf{u}}{dt} = \mathbf{F}(t, \mathbf{u}), \qquad \mathbf{u}(t_0) = \mathbf{a}, \tag{20.37}$$

admits a solution $\mathbf{u} = \mathbf{f}(t)$ that is, at least, defined for nearby times, i.e., when $| t - t_0 | < \delta$ for some $\delta > 0$.

Theorem 20.7 guarantees that the solution to the initial value problem exists — at least for times sufficiently close to the initial instant t_0. This may be the most that can be said, although in many cases the maximal interval $\alpha < t < \beta$ of existence of the solution might be much larger — possibly infinite, $-\infty < t < \infty$, resulting in a *global solution*. The interval of existence of a solution typically depends upon both the equation and the particular initial data. For instance, even though its right hand side is defined everywhere, the solutions to the scalar initial value problem (20.8) only exist up until time $1/u_0$, and so, the larger the initial data, the shorter the time of existence. In this example, the only global solution is the equilibrium solution $u(t) \equiv 0$. It is worth noting that this short-term

[†] If $\mathbf{F}(t, \mathbf{u})$ is only defined on a subdomain $\Omega \subset \mathbb{R}^{n+1}$, then we must assume that the point $(t_0, \mathbf{a}) \in \Omega$ specifying the initial conditions belongs to its domain of definition.

existence phenomenon does not appear in the linear regime, where, barring singularities in the equation itself, solutions to a linear ordinary differential equation are guaranteed to exist for all time.

In practice, one always extends a solutions to its maximal interval of existence. The Existence Theorem 20.7 implies that there are only two possible ways in whcih a solution cannot be extended beyond a time $t^\star$: Either

(i) the solution becomes unbounded: $\| \mathbf{u}(t) \| \to \infty$ as $t \to t^\star$, or

(ii) if the right hand side $F(t, \mathbf{u})$ is only defined on a subset $\Omega \subset \mathbb{R}^{n+1}$, then the solution $\mathbf{u}(t)$ reaches the boundary $\partial\Omega$ as $t \to t^\star$.

If neither occurs in finite time, then the solution is necessarily global. In other words, a solution to an ordinary differential equation cannot suddenly vanish into thin air.

Remark: The existence theorem can be readily adapted to any higher order system of ordinary differential equations through the method of converting it into an equivalent first order system by introducing additional variables. The appropriate initial conditions guaranteeing existence are induced from those of the corresponding first order system, as in the second order example (20.26) discussed above.

Uniqueness and Smoothness

As important as existence is the question of uniqueness. Does the initial value problem have more than one solution? If so, then we cannot use the differential equation to predict the future behavior of the system from its current state. While continuity of the right hand side of the differential equation will guarantee that a solution exists, it is not quite sufficient to ensure uniqueness of the solution to the initial value problem. The difficulty can be appreciated by looking at an elementary example.

Example 20.8. Consider the nonlinear initial value problem

$$\frac{du}{dt} = \frac{5}{3}\, u^{2/5}, \qquad u(0) = 0. \tag{20.38}$$

Since the right hand side is a continuous function, Theorem 20.7 assures us of the existence of a solution — at least for t close to 0. This autonomous scalar equation can be easily solved by the usual method:

$$\int \frac{3}{5}\, \frac{du}{u^{2/5}} = u^{3/5} = t + c, \qquad \text{and so} \qquad u = (t + c)^{5/3}.$$

Substituting into the initial condition implies that $c = 0$, and hence $u(t) = t^{5/3}$ is a solution to the initial value problem.

On the other hand, since the right hand side of the differential equation vanishes at $u = 0$, the constant function $u(t) \equiv 0$ is an equilibrium solution to the differential equation. (Here is an example where the integration method fails to recover the equilibrium solution.) Moreover, the equilibrium solution has the same initial value $u(0) = 0$. Therefore, we have constructed two different solutions to the initial value problem (20.38). Uniqueness is *not*

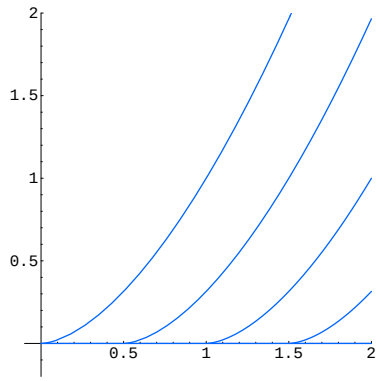


Figure 20.4. Solutions to the Differential Equation $\dot{u} = \frac{5}{3}\, u^{2/5}$.

valid! Worse yet, there are, in fact, an *infinite* number of solutions to the initial value problem. For *any* $a > 0$, the function

$$u(t) = \begin{cases} 0, & 0 \le t \le a, \\ (t-a)^{5/3}, & t \ge a, \end{cases} \tag{20.39}$$

is differentiable everywhere, even at $t = a$. (Why?) Moreover, it satisfies both the differential equation and the initial condition, and hence defines a solution to the initial value problem. Several of these solutions are plotted in Figure 20.4.

Thus, to ensure uniqueness of solutions, we need to impose a more stringent condition, beyond mere continuity. The proof of the following basic uniqueness theorem can be found in the above references.

Theorem 20.9. *If* $\mathbf{F}(t, \mathbf{u}) \in \mathrm{C}^1$ *is continuously differentiable, then there exists one and only one solution*† *to the initial value problem* (20.37).

Thus, the difficulty with the differential equation (20.38) is that the function $F(u) = \frac{5}{3}\, u^{2/5}$, although continuous everywhere, is not differentiable at $u = 0$, and hence the Uniqueness Theorem 20.9 does not apply. On the other hand, $F(u)$ is continuously differentiable away from $u = 0$, and so any nonzero initial condition $u(t_0) = u_0 \neq 0$ will produce a unique solution — for as long as it remains away from the problematic value $u = 0$.

Blanket Hypothesis: From now on, all differential equations must satisfy the uniqueness criterion that their right hand side is continuously differentiable.

While continuous differentiability is sufficient to guarantee uniqueness of solutions, the smoother the right hand side of the system, the smoother the solutions. Specifically:

Theorem 20.10. *If* $\mathbf{F} \in \mathrm{C}^n$ *for* $n \ge 1$, *then any solution to the system* $\dot{\mathbf{u}} = \mathbf{F}(t, \mathbf{u})$ *is of class* $\mathbf{u} \in \mathrm{C}^{n+1}$. *If* $\mathbf{F}(t, \mathbf{u})$ *is an analytic function, then all solutions* $\mathbf{u}(t)$ *are analytic.*

† As noted earlier, we extend all solutions to their maximal interval of existence.

The basic outline of the proof of the first result is clear: Continuity of $\mathbf{u}(t)$ (which is a basic prerequiste of any solution) implies continuity of $\mathbf{F}(t, \mathbf{u}(t))$, which means $\dot{\mathbf{u}}$ is continuous and hence $\mathbf{u} \in \mathrm{C}^1$. This in turn implies $\mathbf{F}(t, \mathbf{u}(t)) = \dot{\mathbf{u}}$ is a continuously differentiable of t, and so $\mathbf{u} \in \mathrm{C}^2$. And so on, up to order n. The proof of analyticity follows from a detailed analysis of the power series solutions, [**98**]. Indeed, the analytic result underlies the method of power series solutions of ordinary differential equations, developed in detail in Appendix C.

Uniqueness has a number of particularly important consequences for the solutions to autonomous systems, i.e., those whoe right hand side does not explicitly depend upon t. Throughout the remainder of this section, we will deal with an autonomous system of ordinary differential equations

$$\frac{d\mathbf{u}}{dt} = \mathbf{F}(\mathbf{u}), \qquad \text{where} \qquad \mathbf{F} \in \mathrm{C}^1, \tag{20.40}$$

whose right hand side is defined and continuously differentiable for all $\mathbf{u}$ in a domain $\Omega \subset \mathbb{R}^n$. As a consequence, each solution $\mathbf{u}(t)$ is, on its interval of existence, uniquely determined by its initial data. Autonomy of the differential equation is an essential hypothesis for the validity of the following properties.

The first result tells us that the solution trajectories of an autonomous system do not vary over time.

Proposition 20.11. *If $\mathbf{u}(t)$ is the solution to the autonomous system* (20.40) *with initial condition $\mathbf{u}(t_0) = \mathbf{u}_0$, then the solution to the initial value problem $\widetilde{\mathbf{u}}(t_1) = \mathbf{u}_0$ is $\widetilde{\mathbf{u}}(t) = \mathbf{u}(t - t_1 + t_0)$.*

Proof: Let $\widetilde{\mathbf{u}}(t) = \mathbf{u}(t - t_1 + t_0)$, where $\mathbf{u}(t)$ is the original solution. In view of the chain rule and the fact that t_1 and t_0 are fixed,

$$\frac{d}{dt}\,\widetilde{\mathbf{u}}(t) = \frac{d\mathbf{u}}{dt}\,(t - t_1 + t_0) = \mathbf{F}\big(\,\mathbf{u}(t - t_1 + t_0)\,\big) = \mathbf{F}\big(\,\widetilde{\mathbf{u}}(t)\,\big),$$

and hence $\widetilde{\mathbf{u}}(t)$ is also a solution to the system (20.40). Moreover,

$$\widetilde{\mathbf{u}}(t_1) = \mathbf{u}(t_0) = \mathbf{u}_0$$

has the indicated initial conditions, and hence, by uniqueness, must be the one and only solution to the latter initial value problem. *Q.E.D.*

Note that the two solutions $\mathbf{u}(t)$ and $\widetilde{\mathbf{u}}(t)$ parametrize the *same* curve in $\mathbb{R}^n$, differing only by an overall "phase shift", $t_1 - t_0$, in their parametrizations. Thus, all solutions passing through the point $\mathbf{u}_0$ follow the same trajectory, irrespective of the time they arrive there. Indeed, not only is the trajectory the same, but the solutions have identical speeds at each point along the trajectory curve. For instance, if the right hand side of (20.40) represents the velocity vector field of steady state fluid flow, Proposition 20.11 implies that the stream lines — the paths followed by the individual fluid particles — do not change in time, even though the fluid itself is in motion. This, indeed, is the meaning of the term "steady state" in fluid mechanics.

One particularly important consequence of uniqueness is that a solution $\mathbf{u}(t)$ to an autonomous system is either stuck at an equilibrium for all time, or is always in motion. In other words, either $\dot{\mathbf{u}} \equiv \mathbf{0}$, in the case of equilibrium, or, otherwise, $\dot{\mathbf{u}} \neq \mathbf{0}$ wherever defined.

Proposition 20.12. *Let $\mathbf{u}^\star$ be an equilibrium for the autonomous system* (20.40)*, so $\mathbf{F}(\mathbf{u}^\star) = \mathbf{0}$. If $\mathbf{u}(t)$ is any solution such that $\mathbf{u}(t^\star) = \mathbf{u}^\star$ at some time $t^\star$, then $\mathbf{u}(t) \equiv \mathbf{u}^\star$ is the equilibrium solution.*

Proof: We regard $\mathbf{u}(t^\star) = \mathbf{u}^\star$ as initial data for the given solution $\mathbf{u}(t)$ at the initial time $t^\star$. Since $\mathbf{F}(\mathbf{u}^\star) = \mathbf{0}$, the constant function $\mathbf{u}^\star(t) \equiv \mathbf{u}^\star$ is a solution of the differential equation that satisfies the same initial conditions. Therefore, by uniqueness, it coincides with the solution in question. *Q.E.D.*

In other words, it is mathematically impossible for a solution to reach an equilibrium position in a finite amount of time — although it may well approach equilibrium in an asymptotic fashion as $t \to \infty$. Physically, this result has the interesting and counterintuitive consequence that a system never actually attains an equilibrium position! Even at very large times, there is always some very slight residual motion. In practice, though, once the solution gets sufficiently close to equilibrium, we are unable to detect the motion, and the physical system has, in all but name, reached its stationary equilibrium configuration. And, of course, the inherent motion of the atoms and molecules not included in such a simplified model would hide any infinitesimal residual effects of the mathematical solution. Without uniqueness, the result is false. For example, the function $u(t) = (t - t^\star)^{5/3}$ is a solution to the scalar ordinary differential equation (20.38) that reaches the equilibrium point $u^\star = 0$ in a finite time $t = t^\star$.

Although a solution cannot reach equilibrium in a finite time, it may certainly have a well-defined limiting value. Using the existence and uniqueness theorems, it can be proved, [**89**], that such a limit point is necessarily an equilibrium solution.

Theorem 20.13. *If $\mathbf{u}(t)$ is any solution to an autonomous system $\dot{\mathbf{u}} = \mathbf{F}(\mathbf{u})$ such that $\lim_{t \to \infty} \mathbf{u}(t) = \mathbf{u}^\star$, then $\mathbf{u}^\star$ is an equilibrium solution, and so $\mathbf{F}(\mathbf{u}^\star) = \mathbf{0}$.*

The same conclusion holds if we run time backwards: if $\lim_{t \to -\infty} \mathbf{u}(t) = \mathbf{u}_\star$, then $\mathbf{u}_\star$ is also an equilibrium point. solutions that start and end at equilibrium points are of particular interest for understanding the dynamics, and known as *heteroclinic*, or, if the start and end equilibria are the same, *homoclinic orbits*. Of course, limiting equilibrium points are but one of the possible long term behaviors of solutions to nonlinear ordinary differential equations, which can also become unbounded in finite or infinite time, or approach periodic orbits, known as *limit cycles*, or become completely chaotic, depending upon the nature of the system and the initial conditions. Resolving the long term behavior os solutions is one of the many challenges awaiting the detailed analysis of any nonlinear ordinary differential equation.

Continuous Dependence

In a real-world applications, initial conditions are almost never known exactly. Rather,

experimental and physical errors will only allow us to say that their values are approximately equal to those in our mathematical model. Thus, to retain physical relevance, we need to be sure that small errors in our initial measurements do not induce a large change in the solution. A similar argument can be made for any physical parameters, e.g., masses, charges, spring stiffnesses, frictional coefficients, etc., that appear in the differential equation itself. A slight change in the parameters should not have a dramatic effect on the solution.

Mathematically, what we are after is a criterion of *continuous dependence* of solutions upon both initial data and parameters. Fortunately, the desired result holds without any additional assumptions, beyond requiring that the parameters appear continuously in the differential equation. We state both results in a single theorem.

Theorem 20.14. *Consider an initial value problem problem*

$$\frac{d\mathbf{u}}{dt} = \mathbf{F}(t, \mathbf{u}, \boldsymbol{\mu}), \qquad \mathbf{u}(t_0) = \mathbf{a}(\boldsymbol{\mu}), \tag{20.41}$$

in which the differential equation and/or the initial conditions depend continuously on one or more parameters $\boldsymbol{\mu} = (\mu_1, \dots, \mu_k)$. Then the unique[†] solution $\mathbf{u}(t, \boldsymbol{\mu})$ depends continuously upon the parameters.

Example 20.15. Let us look at a perturbed version

$$\frac{du}{dt} = \alpha\, u^2, \qquad u(0) = u_0 + \varepsilon,$$

of the initial value problem that we considered in Example 20.1. We regard ε as a small perturbation of our original initial data u_0, and α as a variable parameter in the equation. The solution is

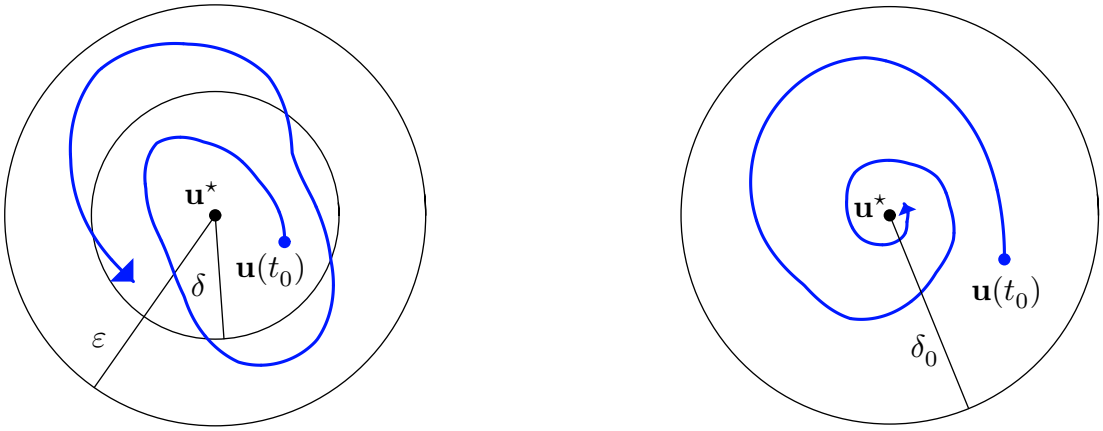
$$u(t, \varepsilon) = \frac{u_0 + \varepsilon}{1 - \alpha\,(u_0 + \varepsilon)\, t}\,. \tag{20.42}$$

Note that, where defined, this is a continuous function of both parameters α, ε. Thus, a small change in the initial data, or in the equation, produces a small change in the solution — at least for times near the initial time.

Continuous dependence *does not* preclude nearby solutions from eventually becoming far apart. Indeed, the blow-up time $t^\star = 1/\big[\,\alpha\,(u_0 + \varepsilon)\,\big]$ for the solution (20.42) depends upon both the initial data and the parameter in the equation. Thus, as we approach the singularity, solutions that started out very close to each other will get arbitrarily far apart; see Figure 20.1 for an illustration.

An even simpler example is the linear model of exponential growth $\dot u = \alpha\, u$ when $\alpha > 0$. A very tiny change in the initial conditions has a negligible short term effect upon the solution, but over longer time intervals, the differences between the two solutions will be dramatic. Thus, the "sensitive dependence" of solutions on initial conditions already appears in very simple linear equations. For similar reasons, sontinuous dependence does *not* prevent solutions from exhibiting chaotic behavior. Further development of these ideas can be found in [**7**, **54**] and elsewhere.

[†] We continue to impose our blanket uniqueness hypothesis.



Stability Asymptotic Stability

Figure 20.5. Stability of Equilibria.

20.3. Stability.

Once a solution to a system of ordinary differential equations has settled down, its limiting value is an equilibrium solution; this is the content of Theorem 20.13. However, not all equilibria appear in this fashion. The only steady state solutions that one directly observes in a physical system are the stable equilibria. Unstable equilibria are hard to sustain, and will disappear when subjected to even the tiniest perturbation, e.g., a breath of air, or outside traffic jarring the experimental apparatus. Thus, finding the equilibrium solutions to a system of ordinary differential equations is only half the battle; one must then understand their stability properties in order to characterize those that can be realized in normal physical circumstances.

We will focus our attention on autonomous systems

$$\dot{\mathbf{u}} = \mathbf{F}(\mathbf{u})$$

whose right hand sides are at least continuously differentiable, so as to ensure the uniqueness of solutions to the initial value problem. If *every* solution that starts out near a given equilibrium solution tends to it, the equilibrium is called *asymptotically stable*. If the solutions that start out nearby stay nearby, then the equilibrium is *stable*. More formally:

Definition 20.16. An equilibrium solution $\mathbf{u}^\star$ to an autonomous system of first order ordinary differential equations is called

- *stable* if for every (small) $\varepsilon > 0$, there exists a $\delta > 0$ such that every solution $\mathbf{u}(t)$ having initial conditions within distance $\delta > \| \mathbf{u}(t_0) - \mathbf{u}^\star \|$ of the equilibrium rmains within distance $\varepsilon > \| \mathbf{u}(t) - \mathbf{u}^\star \|$ for all $t \geq t_0$.
- *asymptotically stable* if it is stable and, in addition, there exists $\delta_0 > 0$ such that whenever $\delta_0 > \| \mathbf{u}(t_0) - \mathbf{u}^\star \|$, then $\mathbf{u}(t) \to \mathbf{u}^\star$ as $t \to \infty$.

Thus, although solutions nearby a stable equilibrium may drift slightly farther away, they must remain relatively close. In the case of asymptotic stability, they will eventually return to equilibrium. This is illustrated in Figure 20.5

Example 20.17. As we saw, the logistic differential equation

$$\frac{du}{dt} = \lambda\, u(1-u)$$

has two equilibrium solutions, corresponding to the two roots of the quadratic equation $\lambda\, u(1-u) = 0$. The solution graphs in Figure 20.1 illustrate the behavior of the solutions. Observe that the first equilibrium solution $u_1^\star = 0$ is unstable, since all nearby solutions go away from it at an exponentially fast rate. On the other hand, the other equilibrium solution $u_2^\star = 1$ is asymptotically stable, since any solution with initial condition $0 < u_0$ tends to it, again at an exponentially fast rate.

Example 20.18. Consider an autonomous (meaning constant coefficient) homogeneous linear planar system

$$\frac{du}{dt} = a\, u + b\, v, \qquad\qquad \frac{dv}{dt} = c\, u + d\, v,$$

with coefficient matrix $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. The origin $u^\star = v^\star = 0$ is an evident equilibrium, solution, and, moreover, is the only equilibrium provided A is nonsingular. According to the results in Section 9.3, the stability of the origin depends upon the eigenvalues of A: It is (globally) asymptotically stable if and only if both eigenvalues are real and negative, and is stable, but not asymptotically stable if and only if both eigenvalues are purely imaginary, or if 0 is a double eigenvalue and so $A = \mathrm{O}$. In all other cases, the origin is an unstable equilibrium. Later, we will see how this simple linear analysis has a direct bearing on the stability question for nonlinear planar systems.

Stability of Scalar Differential Equations

Before looking at any further examples, we need to develop some basic mathematical tools for investigating the stability of equilibria. We begin at the beginning. The stability analysis for first order scalar ordinary differential equations

$$\frac{du}{dt} = F(u) \tag{20.43}$$

is particularly easy. The first observation is that all non-equilibrium solutions $u(t)$ are *strictly monotone* functions, meaning they are either always increasing or always decreasing. Indeed, when $F(u) > 0$, then (20.43) implies that the derivative $\dot u > 0$, and hence $u(t)$ is increasing at such a point. Vice versa, solutions are decreasing at any point where $F(u) < 0$. Since $F(u(t))$ depends continuously on t, any non-monotone solution would have to pass through an equilibrium value where $F(u^\star) = 0$, in violation of Proposition 20.12. This proves the claim.

As a consequence of monotonicity, there are only three possible behaviors for a non-equilibrium solution:

(*a*) it becomes unbounded at some finite time: $|\, u(t)\,| \to \infty$ as $t \to t^\star$; or

(*b*) it exists for all $t \geq t_0$, but becomes unbounded as $t \to \infty$; or

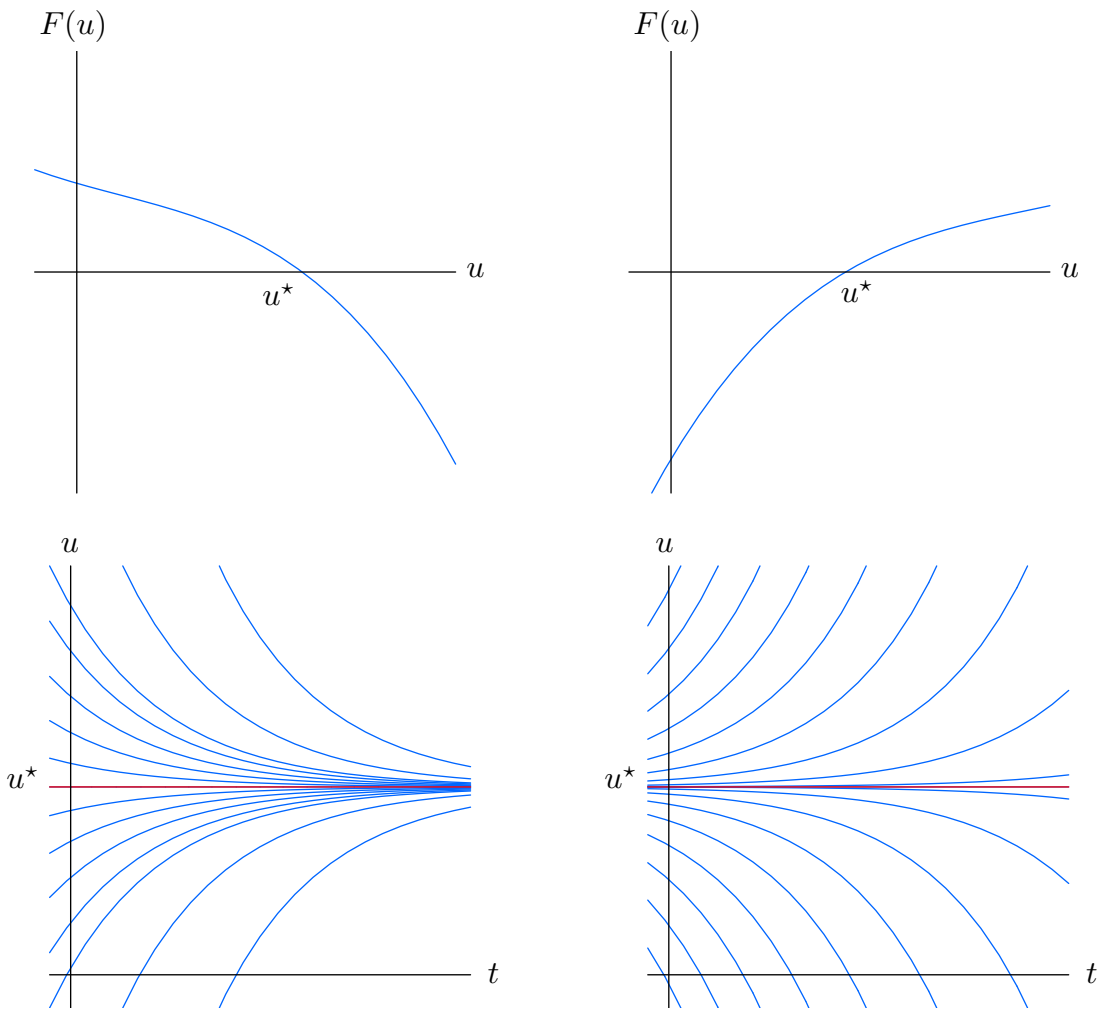


Figure 20.6. Equilibria of Scalar Ordinary Differential Equations.

(c) it exists for all $t \geq t_0$ and has a limiting value, $u(t) \to u^\star$ as $t \to \infty$, which, by Theorem 20.13 must be an equilibrium point.

Let us look more carefully at the last eventuality. Suppose $u^\star$ is an equilibrium point, so $F(u^\star) = 0$. Suppose that $F(u) > 0$ for all u lying slightly below $u^\star$, i.e., on an interval of the form $u^\star - \delta < u < u^\star$. Any solution $u(t)$ that starts out on this interval, $u^\star - \delta < u(t_0) < u^\star$ must be increasing. Moreover, $u(t) < u^\star$ for all t since, according to Proposition 20.12, the solution cannot pass through the equilibrium point. Therefore, $u(t)$ is a solution of type (c). It must have limiting value $u^\star$, since by assumption, this is the only equilibrium solution it can increase to. Therefore, in this situation, the equilibrium point $u^\star$ is *asymptotically stable from below*: solutions that start out slightly below return to it in the limit. On the other hand, if $F(u) < 0$ for all u slightly below $u^\star$, then any solution that starts out in this regime will be monotonically decreasing, and so will move downwards, away from the equilibrium point, which is thus *unstable from below*.

By the same reasoning, if $F(u) < 0$ for u slightly above $u^\star$, then solutions starting

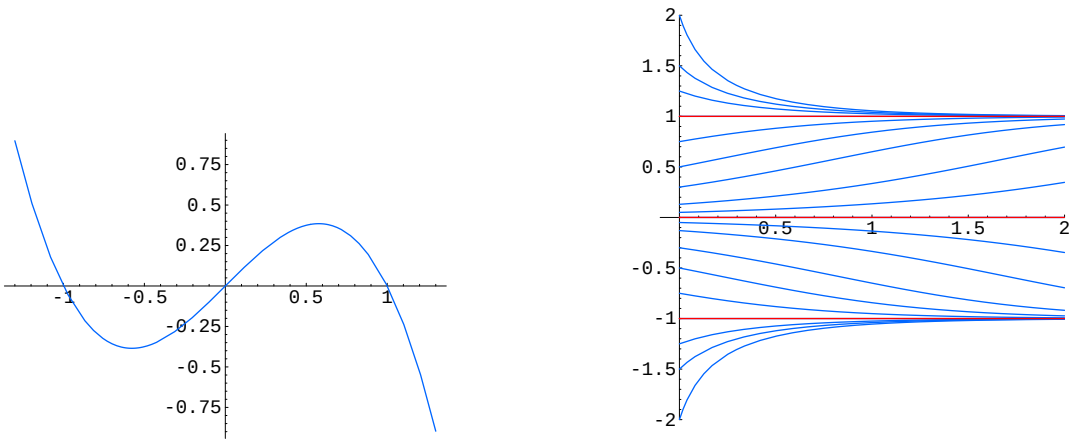


Figure 20.7. Stability of $\dot{u} = u - u^3$.

out there will be monotonically decreasing, bounded from below by $u^\star$, and hence have no choice but to tend to $u^\star$ in the limit. Under this condition, the equilibrium point is *asymptotically stable from above*. The reverse inequality, $F(u) > 0$, corresponds to solutions that increase away from $u^\star$, which is hence *unstable from above*. Combining the two stable cases produces the basic asymptotic stability criterion for scalar ordinary differential equations.

Theorem 20.19. *A equilibrium point $u^\star$ of an autonomous scalar differential equation is asymptotically stable if and only if $F(u) > 0$ for $u^\star - \delta < u < u^\star$ and $F(u) < 0$ for $u^\star < u < u^\star + \delta$, for some $\delta > 0$.*

In other words, if $F(u)$ switches sign from positive to negative as u increases through the equilibrium point, then the equilibrium is asymptotically stable. If the inequalities are reversed, and $F(u)$ goes from negative to positive, then the equilibrium point is unstable. The two cases are illustrated in Figure 20.6. An equilibrium point where $F(u)$ is of one sign on both sides, e.g., the point $u^\star = 0$ for $F(u) = u^2$, is stable from one side, and unstable from the other; in Exercise ■ you are asked to analyze such cases in detail.

Example 20.20. Consider the differential equation

$$\frac{du}{dt} = u - u^3. \tag{20.44}$$

Solving the algebraic equation $F(u) = u - u^3 = 0$, we find that the equation has three equilibria: $u_1^\star = -1$, $u_2^\star = 0$, $u_3^\star = +1$, As u increases, the graph of the function $F(u) = u - u^3$ switches from positive to negative at the first equilibrium point $u_1^\star = -1$, which proves its stability. Similarly, the graph goes back from positive to negative at $u_2^\star = 0$, establishing the instability of the second equilibrium. The final equilibrium $u_3^\star = +1$ is stable because $F(u)$ again changes from negative to positive there.

With this information in hand, we are able to completely characterize the behavior of all solutions to the system. Any solution with negative initial condition, $u_0 < 0$, will end up, asymptotically, at the first equilibrium, $u(t) \to -1$ as $t \to \infty$. Indeed, if $u_0 < -1$, then $u(t)$ is monotonically increasing to -1, while if $-1 < u_0 < 0$, the solution is decreasing

towards -1. On the other hand, if $u_0 > 0$, the corresponding solution ends up at the other stable equilibrium, $u(t) \to +1$; those with $0 < u_0 < 1$ are monotonically increasing, while those with $u_0 > 1$ are decreasing. The only solution that does not end up at either -1 or $+1$ as $t \to \infty$ is the unstable equilibrium solution $u(t) \equiv 0$. Any perturbation of it, no matter how tiny, will force the solutions to choose one of the stable equilibria. Representative solutions are plotted in Figure 20.7. Note that all the curves, with the sole exception of the horizontal axis, converge to one of the stable solutions ± 1, and diverge from the unstable solution 0 as $t \to \infty$.

Thus, the sign of the function $F(u)$ nearby an equilibrium determines its stability. In most instances, this can be checked by looking at the derivative of the function at the equilibrium. If $F'(u^\star) < 0$, then we are in the stable situation, where $F(u)$ goes from positive to negative with increasing u, whereas if $F'(u^\star) > 0$, then the equilibrium $u^\star$ unstable on both sides.

Theorem 20.21. *Let $u^\star$ be a equilibrium point for a scalar ordinary differential equation $\dot{u} = F(u)$. If $F'(u^\star) < 0$, then $u^\star$ is asymptotically stable. If $F'(u^\star) > 0$, then $u^\star$ is unstable.*

For instance, in the preceding example,

$$F'(u) = 1 - 3\,u^2,$$

and its value at the equilibria are

$$F'(-1) = -2 < 0, \qquad F'(0) = 1 > 0, \qquad F'(1) = -2 < 0.$$

The signs reconfirm our conclusion that ± 1 are stable equilibria, while 0 is unstable.

In the borderline case when $F'(u^\star) = 0$, the derivative test is inconclusive, and further analysis is needed to resolve the status of the equilibrium point. For example, the equations $\dot{u} = u^3$ and $\dot{u} = -u^3$ both satisfy $F'(0) = 0$ at the equilibrium point $u^\star = 0$. But, according to the criterion of Theorem 20.19, the former has an unstable equilibrium, while the latter's is stable. Thus, Theorem 20.21 is not as powerful as the direct algebraic test in Theorem 20.19. But it does have the advantage of being a bit easier to use. More significantly, unlike the algebraic test, it can be directly generalized to systems of ordinary differential equations.

Linearization and Stability

In higher dimensional situations, we can no longer rely on simple monotonicity properties, and a more sophisticated approach to stability issues is required. The key idea is already contained in the second characterization of stable equilibria in Theorem 20.21. The derivative $F'(u^\star)$ determines the slope of the tangent line, which is a linear approximation to the function $F(u)$ near the equilibrium point. In a similar fashion, a vector-vallued function $\mathbf{F}(\mathbf{u})$ is replaced by its linear approximation near an equilibrium point. The basic stability criteria for the resulting linearized differential equation were established in Section 9.2, and, in most situations, the linearized stability or instability carries over to the nonlinear regime.

Let us first revisit the scalar case

$$\frac{du}{dt} = F(u) \tag{20.45}$$

from this point of view. *Linearization* of a scalar function at a point means to replace it by its tangent line approximation

$$F(u) \approx F(u^\star) + F'(u^\star)(u - u^\star) \tag{20.46}$$

If $u^\star$ is an equilibrium point, then $F(u^\star) = 0$, and so the first term disappears. Therefore, we anticipate that, near the equilibrium point, the solutions to the nonlinear ordinary differential equation (20.45) will be well approximated by its linearization

$$\frac{du}{dt} = F'(u^\star)(u - u^\star).$$

Let us rewrite the linearized equation in terms of the deviation $v(t) = u(t) - u^\star$ of the solution from equilibrium. Since $u^\star$ is fixed, $dv/dt = du/dt$, and so the linearized equation takes the elementary form

$$\frac{dv}{dt} = a\, v, \qquad \text{where} \qquad a = F'(u^\star) \tag{20.47}$$

is the value of the derivative at the equilibrium point. Note that the original equilibrium point $u^\star$ corresponds to the zero equilibrium point $v^\star = 0$ of the linearized equation (20.47). We already know that the linear differential equation (20.47) has an asymptotically stable equilibrium at $v^\star = 0$ if and only if $a = F'(u^\star) < 0$, while for $a = F'(u^\star) > 0$ the origin is unstable. In this manner, the linearized stability criterion reproduces that established in Theorem 20.21.

The same linearization technique can be applied to analyze the stability of an equilibrium solution $\mathbf{u}^\star$ to a first order autonomous system

$$\dot{\mathbf{u}} = \mathbf{F}(\mathbf{u}). \tag{20.48}$$

We approximate the function $\mathbf{F}(\mathbf{u})$ near an equilibrium point, where $\mathbf{F}(\mathbf{u}^\star) = \mathbf{0}$, by its first order Taylor polynomial:

$$\mathbf{F}(\mathbf{u}) \approx \mathbf{F}(\mathbf{u}^\star) + \mathbf{F}'(\mathbf{u}^\star)(\mathbf{u} - \mathbf{u}^\star) = \mathbf{F}'(\mathbf{u}^\star)(\mathbf{u} - \mathbf{u}^\star). \tag{20.49}$$

Here, $\mathbf{F}'(\mathbf{u}^\star)$ denotes its $n \times n$ Jacobian matrix (19.28) at the equilibrium point. Thus, for nearby solutions, we expect that the deviation from equilibrium, $\mathbf{v}(t) = \mathbf{u}(t) - \mathbf{u}^\star$, will be governed by the linearized system

$$\frac{d\mathbf{v}}{dt} = A\,\mathbf{v}, \qquad \text{where} \qquad A = \mathbf{F}'(\mathbf{u}^\star). \tag{20.50}$$

Now, we already know the complete stability criteria for linear systems. According to Theorem 9.15, the zero equilibrium solution to (20.50) is asymptotically stable if and only if all the eigenvalues of the coefficient matrix $A = \mathbf{F}'(\mathbf{u}^\star)$ have negative real part. In contrast, if one or more of the eigenvalues has positive real part, then the zero solution is unstable. Indeed, it can be proved, [**89**, **99**], that these linearized stability criteria are also valid in the nonlinear case.

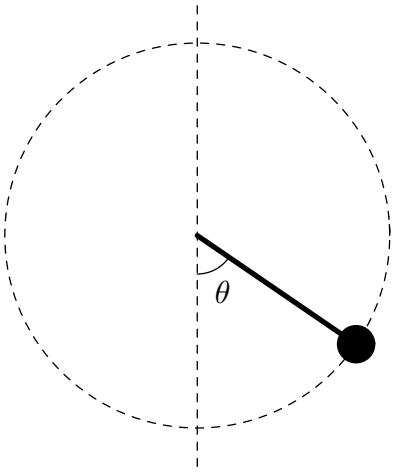


Figure 20.8. The Pendulum.

Theorem 20.22. *Let $\mathbf{u}^\star$ be an equilibrium point for the first order ordinary differential equation $\dot{\mathbf{u}} = \mathbf{F}(\mathbf{u})$. If all of the eigenvalues of the Jacobian matrix $\mathbf{F}'(\mathbf{u}^\star)$ have negative real part,* $\operatorname{Re} \lambda < 0$, *then $\mathbf{u}^\star$ is asymptotically stable. If, on the other hand, $\mathbf{F}'(\mathbf{u}^\star)$ has one or more eigenvalues with positive real part,* $\operatorname{Re} \lambda > 0$, *then $\mathbf{u}^\star$ is an unstable equilibrium.*

Intuitively, the additional nonlinear terms in the full system should only slightly perturb the eigenvalues, and hence, at least for those with nonzero real part, not alter their effect on the stability of solutions. The borderline case occurs when one or more of the eigenvalues of $\mathbf{F}'(\mathbf{u}^\star)$ is either 0 or purely imaginary, i.e., $\operatorname{Re} \lambda = 0$, while all other eigenvalues have negative real part. In such situations, the linearized stability test is inconclusive, and we need more detailed information (which may not be easy to come by) to resolve the status of the equilibrium.

Example 20.23. The second order ordinary differential equation

$$m \, \frac{d^2\theta}{dt^2} + \mu \, \frac{d\theta}{dt} + \kappa \sin \theta = 0 \tag{20.51}$$

describes the damped oscillations of a rigid pendulum that rotates on a pivot subject to a uniform gravitational force in the vertical direction. The unknown function $\theta(t)$ measures the angle of the pendulum from the vertical, as illustrated in Figure 20.8. The constant $m > 0$ is the mass of the pendulum bob, $\mu > 0$ is the coefficient of friction, assumed here to be strictly positive, and $\kappa > 0$ represents the gravitational force.

In order to study the equilibrium solutions and their stability, we must first convert the equation into a first order system. Setting $u(t) = \theta(t)$, $v(t) = \dfrac{d\theta}{dt}$, we find

$$\frac{du}{dt} = v, \qquad \frac{dv}{dt} = -\alpha \sin u - \beta \, v, \qquad \text{where} \qquad \alpha = \frac{\kappa}{m}, \qquad \beta = \frac{\mu}{m}, \tag{20.52}$$

are both positive constants. The equilibria occur where the right hand sides of the first

order system (20.52) simultaneously vanish, that is,

$$v = 0, \qquad -\alpha \sin u - \beta v = 0, \qquad \text{and hence} \qquad u = 0,\ \pm\pi,\ \pm 2\pi,\ \dots\ .$$

Thus, the system has infinitely many equilibrium points:

$$\mathbf{u}_k^\star = (k\pi, 0) \qquad \text{where} \quad k = 0, \pm 1, \pm 2, \dots \quad \text{is any integer.} \tag{20.53}$$

The equilibrium point $\mathbf{u}_0^\star = (0,0)$ corresponds to $u = \theta = 0$, $v = \dot\theta = 0$, which means that the pendulum is at rest at the bottom of its arc. Our physical intuition leads us to expect this to describe a stable configuration, as the frictional effects will eventually damp out small nearby motions. The next equilibrium $\mathbf{u}_1^\star = (\pi, 0)$ corresponds to $u = \theta = \pi$, $v = \dot\theta = 0$, which means that the pendulum is sitting motionless at the top of its arc. This is a theoretically possible equilibrium configuration, but highly unlikely to be observed in practice, and is thus expected to be unstable. Now, since $u = \theta$ is an angular variable, equilibria whose u values differ by an integer multiple of 2π define the same physical configuration, and hence should have identical stability properties. Therefore, all the remaining equilibria $\mathbf{u}_k^\star$ physically correspond to one or the other of these two possibilities: when $k = 2j$ is even, the pendulum is at the bottom, while when $k = 2j+1$ is odd, the pendulum is at the top.

Let us now confirm our intuition by applying the linearization stability criterion of Theorem 20.22. The right hand side of the system (20.52), namely

$$\mathbf{F}(u,v) = \begin{pmatrix} v \\ -\alpha \sin u - \beta v \end{pmatrix}, \qquad \text{has Jacobian matrix} \qquad \mathbf{F}'(u,v) = \begin{pmatrix} 0 & 1 \\ -\alpha \cos u & -\beta \end{pmatrix}.$$

At the bottom equilibrium $\mathbf{u}_0^\star = (0,0)$, the Jacobian matrix

$$\mathbf{F}'(0,0) = \begin{pmatrix} 0 & 1 \\ -\alpha & -\beta \end{pmatrix} \qquad \text{has eigenvalues} \qquad \lambda = \frac{-\beta \pm \sqrt{\beta^2 - 4\alpha}}{2}\,.$$

Under our assumption that $\alpha, \beta > 0$, both eigenvalues have negative real part, and hence the origin is a stable equilibrium. If $\beta^2 < 4\alpha$ — the *underdamped* case — the eigenvalues are complex, and hence, in the terminology of Section 9.3, the origin is a *stable focus*. In the phase plane, the solutions spiral in to the focus, which corresponds to a pendulum with damped oscillations of decreasing magnitude. On the other hand, if $\beta^2 > 4\alpha$, then the system is *overdamped*. Both eigenvalues are negative, and the origin is a *stable node*. In this case, the solutions decay exponentially fast to $\mathbf{0}$. Physically, this would be like a pendulum moving in a vat of molasses. The exact same analysis applies at all even equilibria $\mathbf{u}_{2j}^\star = (2j\pi, 0)$ — which really represent the same bottom equilibrium point.

On the other hand, at the top equilibrium $\mathbf{u}_1^\star = (\pi, 0)$, the Jacobian matrix

$$\mathbf{F}'(0,0) = \begin{pmatrix} 0 & 1 \\ \alpha & -\beta \end{pmatrix} \qquad \text{has eigenvalues} \qquad \lambda = \frac{-\beta \pm \sqrt{\beta^2 + 4\alpha}}{2}\,.$$

In this case, one of the eigenvalues is real and positive while the other is negative. The linearized system has an unstable saddle point, and hence the nonlinear system is also

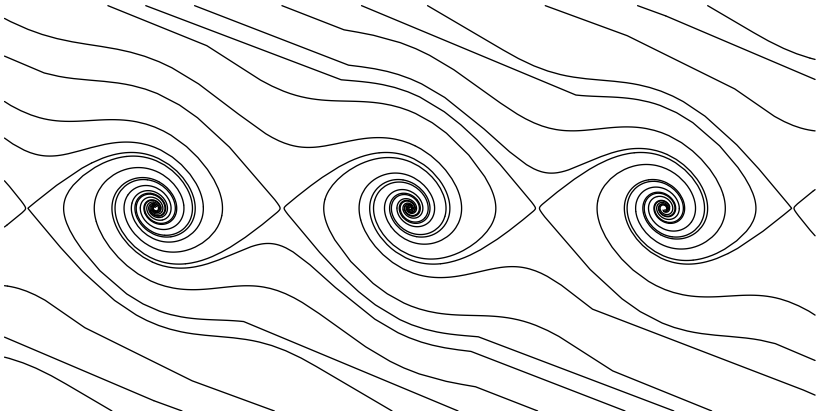


Figure 20.9. The Underdamped Pendulum.

unstable at this equilibrium point. Any tiny perturbation of an upright pendulum will dislodge it, causing it to swing down, and eventually settle into a damped oscillatory motion converging on one of the stable bottom equilibria.

The complete phase portrait of an underdamped pendulum appears in Figure 20.9. Note that, as advertised, almost all solutions end up spiraling into the stable equilibria. Solutions with a large initial velocity will spin several times around the center, but eventually the cumulative effect of frictional forces wins out and the pendulum ends up in a damped oscillatory mode. Each of the the unstable equilibria has the same saddle form as its linearizations, with two very special solutions, corresponding to the stable eigenline of the linearization, in which the pendulum spins around a few times, and, in the $t \to \infty$ limit, ends up standing upright at the unstable equilibrium position. However, like unstable equilibria, such solutions are practically impossible to achieve in a physical environment as any tiny perturbation will cause the pendulum to sightly deviate and then end up eventually decaying into the usual damped oscillatory motion at the bottom.

A deeper analysis demonstrates the local *structural stability* of any nonlinear equilibrium whose linearization is structurally stable, and hence has no eigenvalues on the imaginary axis: $\operatorname{Re} \lambda \neq 0$. Structural stability means that, not only are the stability properties of the equilibrium dictated by the linearized approximation, but, nearby the equilibrium point, all solutions to the nonlinear system are slight perturbations of solutions to the corresponding linearized system, and so, close to the equilibrium point, the two phase portraits have the same qualitative features. Thus, stable foci of the linearization remain stable foci of the nonlinear system; unstable saddle points remain saddle points, although the eigenlines become slightly curved as they depart from the equilibrium. Thus, the structural stability of linear systems, as discussed at the end of Section 9.3 also carries over to the nonlinear regime near an equilibrium. A more in depth discussion of these issues can be found, for instance, in [**89**, **99**].

,

Example 20.24. Consider the unforced *van der Pol system*

$$\frac{du}{dt} = v, \qquad \frac{dv}{dt} = -(u^2 - 1)\, v - u. \tag{20.54}$$

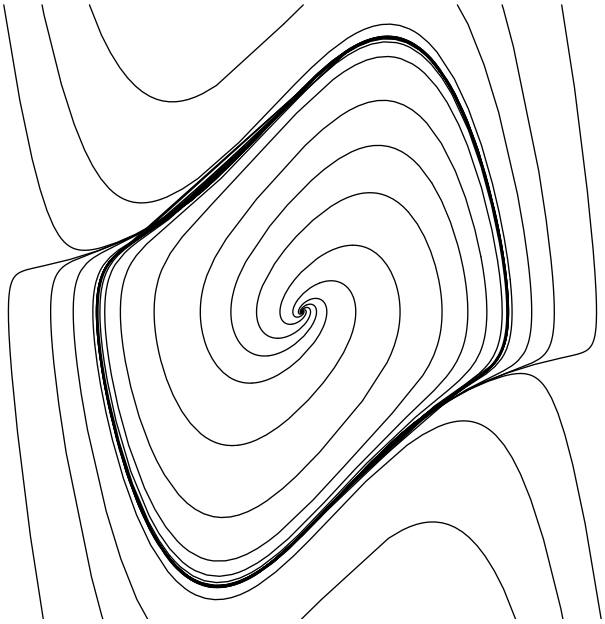


Figure 20.10. Phase Portrait of the van der Pol System.

that we derived in Example 20.5. The only equilibrium point is at the origin $u = v = 0$. Computing the Jacobian matrix of the right hand side,

$$\mathbf{F}'(u,v) = \begin{pmatrix} 0 & 1 \\ 2\,u\,v - 1 & 1 \end{pmatrix}, \qquad \text{and hence} \qquad \mathbf{F}'(0,0) = \begin{pmatrix} 0 & 1 \\ -1 & 1 \end{pmatrix}.$$

The eigenvalues of $\mathbf{F}'(0,0)$ are $\frac{1}{2} \pm \mathrm{i}\,\frac{\sqrt{3}}{2}$, and correspond to an unstable focus of the linearized system near the equilibrium point. Therefore, the origin is an unstable equilibrium for nonlinear van der Pol system, and all non-equilibrium solutions starting out near $\mathbf{0}$ eventually spiral away.

On the other hand, it can be shown that solutions that are sufficiently far away from the origin spiral in towards the center, cf. Exercise ■. So what happens to the solutions? As illustrated in the phase plane portrait sketched in Figure 20.10, all non-equilibrium solutions spiral towards a stable periodic orbit, known as a *limit cycle* for the system. Any non-zero initial data will eventually end up closely following the limit cycle orbit as it periodically circles around the origin. A rigorous proof of the existence of a limit cycle relies on the more sophisticated *Poincaré–Bendixson Theory* for planar autonomous systems, discussed in detail in [**89**].

Example 20.25. The nonlinear system

$$\frac{du}{dt} = u\,(v-1), \qquad \frac{dv}{dt} = 4 - u^2 - v^2,$$

has four equilibria: $(0, \pm 2)$ and $(\pm\sqrt{3}\,, 1)$. Its Jacobian matrix is

$$\mathbf{F}'(u,v) = \begin{pmatrix} v-1 & u \\ -2\,u & -2\,v \end{pmatrix}.$$

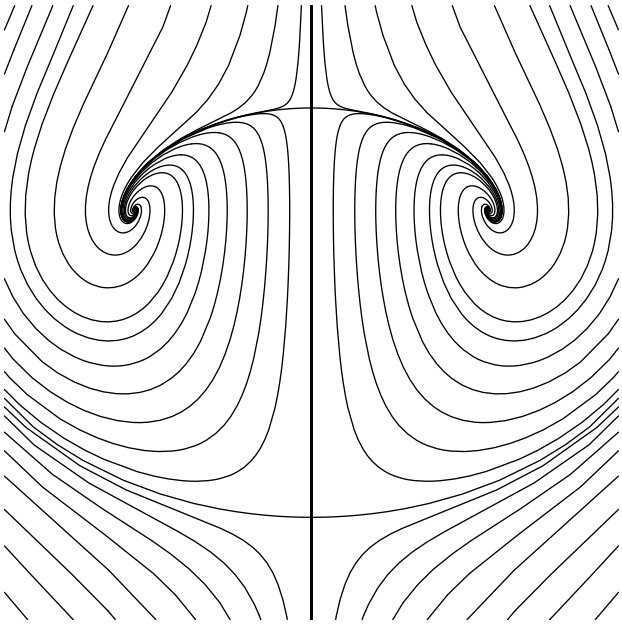


Figure 20.11. Phase Portrait for $\dot{u} = u\,(v-1),\ \dot{v} = 4 - u^2 - v^2$..

A table of the eigenvalues at the equilibrium points and their stability follows: These results are reconfirmed by the phase portrait drawn in Figure 20.11

Equilibrium Point	Jacobian matrix	Eigenvalues	Stability
$(0, 2)$	$\begin{pmatrix} 1 & 0 \\ 0 & -4 \end{pmatrix}$	$1,\ -4$	unstable saddle
$(0, -2)$	$\begin{pmatrix} -3 & 0 \\ 0 & 6 \end{pmatrix}$	$-3, 6$	unstable saddle
$(\sqrt{3}\,, 1)$	$\begin{pmatrix} 0 & -\sqrt{3} \\ 2\sqrt{3} & -2 \end{pmatrix}$	$-1 \pm \mathrm{i}\sqrt{5}$	stable focus
$(-\sqrt{3}\,, 1)$	$\begin{pmatrix} 0 & -\sqrt{3} \\ 2\sqrt{3} & -2 \end{pmatrix}$	$-1 \pm \mathrm{i}\sqrt{5}$	stable focus

Conservative Systems

When modeling a physical system that includes some form of damping — due to friction, viscosity, or dissipation — linearization will usually suffice to resolve the stability or instability of equilibria. However, when dealing with conservative systems, when damping is absent and energy is preserved, the linearization test is often inconclusive, and one must rely on more sophisticated stability criteria. In such situations, one can often exploit

conservation of energy, appealing to our general philosophy that minimizers of an energy function should be stable (but not necessarily asymptotically stable) equilibria.

By saying that energy is *conserved*, we mean that it remains constant as the solution evolves. Conserved quantities are also known as *first integrals* for the system of ordinary differential equations. Additional well-known examples include the laws of conservation of mass, and conservation of linear and angular momentum. Let us mathematically formulate the general definition.

Definition 20.26. A *first integral* of an autonomous system $\dot{\mathbf{u}} = \mathbf{F}(\mathbf{u})$ is a real-valued function $I(\mathbf{u})$ which is constant on solutions.

In other words, for each solution $\mathbf{u}(t)$ to the differential equation,

$$I(\mathbf{u}(t)) = c \qquad \text{for all} \qquad t, \tag{20.55}$$

where c is a fixed constant, which will depend upon which solution is being monitored. The value of c is fixed by the initial data since, in particular, $c = I(\mathbf{u}(t_0)) = I(\mathbf{u}_0)$. Or, to rephrase this condition in another, equivalent manner, every solution to the dynamical system is constrained to move along a single *level set* $\{\, I(\mathbf{u}) = c \,\}$ of the first integral, namely the level set that contains the initial data $\mathbf{u}_0$.

Note first that any constant function, $I(\mathbf{u}) \equiv c_0$, is trivially a first integral, but this provides no useful information whatsoever about the solutions, and so is uninteresting. We will call any autonomous system that possesses a nontrivial first integral $I(\mathbf{u})$ a *conservative system*.

How do we find first integrals? In applications, one often appeals to the underlying physical principles such as conservation of energy, momentum, or mass. Mathematically, the most convenient way to check whether a function is constant is to verify that its derivative is identically zero. Thus, differentiating (20.55) with respect to t and invoking the chain rule leads to the basic condition

$$0 = \frac{d}{dt}\, I(\mathbf{u}(t)) = \nabla I(\mathbf{u}(t)) \cdot \frac{d\mathbf{u}}{dt} = \nabla I(\mathbf{u}(t)) \cdot \mathbf{F}(\mathbf{u}(t)). \tag{20.56}$$

The final expression can be identified as the directional derivative of $I(\mathbf{u})$ with respect to the vector field $\mathbf{v} = \mathbf{F}(\mathbf{u})$ that specifies the differential equation, cf. (19.64). Writing out (20.56) in detail, we find that a first integral $I(u_1, \dots, u_n)$ must satisfy a first order linear partial differential equation:

$$F_1(u_1, \dots, u_n)\, \frac{\partial I}{\partial u_1} + \, \cdots \, + F_n(u_1, \dots, u_n)\, \frac{\partial I}{\partial u_n} = 0. \tag{20.57}$$

As such, it looks harder to solve than the original ordinary differential equation! Often, one falls back on either physical intuition, intelligent guesswork, or, as a last resort, a lucky guess. A deeper fact, due to the pioneering twentieth century mathematician Emmy Noether, cf. [**137**, **141**], is that first integrals and conservation laws are the result of underlying symmetry properties of the differential equation. Like many nonlinear methods, it remains the subject of contemporary research.

Let us specialize to planar autonomous systems

$$\frac{du}{dt} = F(u,v), \qquad \frac{dv}{dt} = G(u,v). \tag{20.58}$$

According to (20.57), any first integral $I(u,v)$ must satisfy the linear partial differential equation

$$F(u,v)\,\frac{\partial I}{\partial u} \;+\; G(u,v)\,\frac{\partial I}{\partial v} = 0. \tag{20.59}$$

This nonlinear first order partial differential equation can be solved by the method of characteristics[†]. Consider the auxiliary first order scalar ordinary differential equation[‡]

$$\frac{dv}{du} = \frac{G(u,v)}{F(u,v)} \tag{20.60}$$

for $v = h(u)$. Note that (20.60) can be formally obtained by dividing the second equation in the original system (20.58) by the first, and then canceling the time differentials dt. Suppose we can write the general solution to the scalar equation (20.60) in the implicit form

$$I(u,v) = c, \tag{20.61}$$

where c is a constant of integration. We claim that the function $I(u,v)$ is a first integral of the original system (20.58). Indeed, differentiating (20.61) with respect to u, and using the chain rule, we find

$$0 = \frac{d}{du}\,I(u,v) = \frac{\partial I}{\partial u} + \frac{dv}{du}\,\frac{\partial I}{\partial v} = \frac{\partial I}{\partial u} + \frac{G(u,v)}{F(u,v)}\,\frac{\partial I}{\partial v}\,.$$

Clearing the denominator, we conclude that $I(u,v)$ solves the partial differential equation (20.59), which justifies our claim.

Example 20.27. As an elementary example, consider the linear system

$$\frac{du}{dt} = -\,v, \qquad \frac{dv}{dt} = u. \tag{20.62}$$

To construct a first integral, we form the auxiliary equation (20.60), which is

$$\frac{dv}{du} = -\,\frac{u}{v}\,.$$

This first order ordinary differential equation can be solved by separating variables:

$$v\,dv = -\,u\,du, \qquad \text{and hence} \qquad \tfrac{1}{2}\,u^2 + \tfrac{1}{2}\,v^2 = c,$$

[†] See Section 22.1 for an alternative perspective on solving such partial differential equations.

[‡] We assume that $F(u,v) \not\equiv 0$. Otherwise, $I(u) = u$ is itself a first integral, and the system reduces to a scalar equation for v; see Exercise ■.

where c is the constant of integration. Therefore, by the preceding result,

$$I(u,v) = \tfrac{1}{2}\,u^2 + \tfrac{1}{2}\,v^2$$

is a first integral. The level sets of $I(u,v)$ are the concentric circles centered at the origin, and we recover the fact that the solutions of (20.62) go around the circles. The origin is a stable equilibrium — a center.

This simple example hints at the importance of first integrals in stability theory. The following key result confirms our general philosophy that energy minimizers, or, more generally, minimizers of first integrals, are stable equilibria. See Exercise ■ for an outline of the proof of the key result.

Theorem 20.28. *Let $I(\mathbf{u})$ be a first integral for the autonomous system $\dot{\mathbf{u}} = \mathbf{F}(\mathbf{u})$. If $\mathbf{u}^\star$ is a strict local extremum — minimum or mximum — of I, then $\mathbf{u}^\star$ is a stable equilibrium point for the system.*

Remark: At first sight, the fact that strict maxima are also stable equilibria appears to contradict our intuition. However, energy functions typically do not have local maxima. Indeed, physical energy is the sum of kinetic and potential contributions. While potential energy can admit maxima, e.g., the pendulum at the top of its arc, these are only unstable saddle points for the full energy function, since the kinetic energy can always be increased by moving a bit faster.

Example 20.29. Consider the specific predator-prey system

$$\frac{du}{dt} = 2\,u - u\,v, \qquad \frac{dv}{dt} = -9\,v + 3\,u\,v, \tag{20.63}$$

modeling populations of, say, lions and zebra, and a special case of (20.25). According to Example 20.4, there are two possible equilibria:

$$u_1^\star = v_1^\star = 0, \qquad u_2^\star = 3, \quad v_2^\star = 2.$$

Let us first try to determine their stability by linearization. The Jacobian matrix for the system is

$$\mathbf{F}'(u,v) = \begin{pmatrix} 2 - v & -u \\ 3\,v & 3\,u - 9 \end{pmatrix}.$$

At the first, trivial equilibrium,

$$\mathbf{F}'(0,0) = \begin{pmatrix} 2 & 0 \\ 0 & -9 \end{pmatrix}, \qquad \text{with eigenvalues } \ 2 \ \text{ and } \ -9.$$

Since there is one positive and one negative eigenvalue, the origin is an unstable saddle point. On the other hand, at the nonzero equilibrium, the Jacobian matrix

$$\mathbf{F}'(3,2) = \begin{pmatrix} 0 & -3 \\ 6 & 0 \end{pmatrix}, \qquad \text{has purely imaginary eigenvalues} \qquad \pm\, 3\,\sqrt{2}\ \mathrm{i}\,.$$

So the linearized system has a stable center. However, as purely imaginary eigenvalues is a borderline situation, Theorem 20.22 cannot be applied. Thus, the linearization stability test is *inconclusive*.

It turns out that the predator-prey model is a conservative system. To find a first integral, we need to solve the auxiliary equation (20.60), which is

$$\frac{dv}{du} = \frac{-9\,v + 3\,u\,v}{2\,u - u\,v} = \frac{-9/u + 3}{2/v - 1}\,.$$

Fortunately, this is a separable first order ordinary differential equation. Integrating,

$$2\,\log v - v = \int \left(\frac{2}{v} - 1\right) dv = \int \left(-\frac{9}{u} + 3\right) du = -9\log u + 3\,u + c,$$

where c is the constant of integration. Writing the solution in the form (20.61), we conclude that

$$I(u,v) = 9\,\log u - 3\,u + 2\,\log v - v = c,$$

is a first integral of the system. The solutions to (20.63) must stay on the level sets of $I(u,v)$. Note that

$$\nabla I(u,v) = \begin{pmatrix} 9/u - 3 \\ 2/v - 1 \end{pmatrix}, \qquad \text{and hence} \qquad \nabla I(3,2) = \mathbf{0},$$

which shows that the second equilibrium is a critical point. (On the other hand, $I(u,v)$ is not defined at the unstable zero equilibrium.) Moreover, the Hessian matrix at the critical point,

$$\nabla^2 I(3,2) = \begin{pmatrix} -3 & 0 \\ 0 & -1 \end{pmatrix},$$

is negative definite, and hence $\mathbf{u}_2^\star = (\,3,2\,)^T$ is a strict local maximum of the first integral $I(u,v)$. Thus, Theorem 20.28 proves that the equilibrium point is a stable center.

The first integral serves to completely characterize the qualitative behavior of the system. In the physically relevant region, i.e., the upper right quadrant $Q = \{\,u > 0,\ v > 0\,\}$ where both populations are positive, all of the level sets of the first integral are closed curves encircling the equilibrium point $\mathbf{u}_2^\star = (\,3,2\,)^T$. The solutions move in a counterclockwise direction around the closed level curves, and hence all non-equilibrium solutions in the positive quadrant are periodic. The phase portrait is illustrated in Figure 20.12, along with a typicl periodic solution. Thus, in such an idealized ecological model, for any initial conditions starting with some zebra and lions, i.e., where $u(t_0), v(t_0) > 0$, the populations will maintain a balance over the long term, each varying periodically between maximum and minimum values. Observe also that the maximum and minimum values of the two populations are not achieved simultaneously. Starting with a small number of predators, the number of prey will initially increase. The predators then have more food availbable, and so also start to increase in numbers. At a certain critical point, the predators are sufficiently numerous as to kill prey faster than they can reproduce. At this point, the prey population has reached its maximum, and begins to decline. But it takes a while

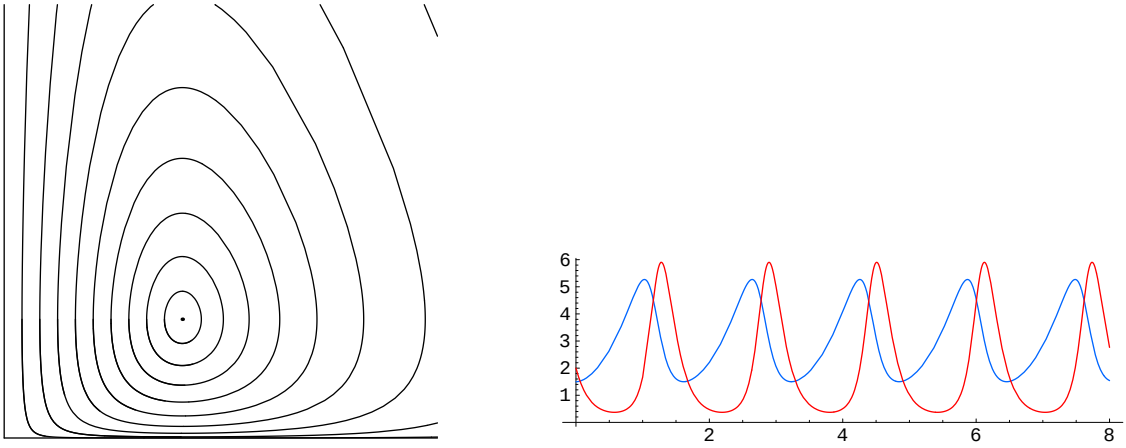


Figure 20.12. Phase Portrait and Solution of the Predator-Prey System.

for the predator population to feel the effect, and so it continues to increase. Eventually the increasingly rapid decline in the number of prey begins to affect the predators. After the predators reach their maximum number, both populations are in decline. Eventually, enough predators have died off so as to relieve the pressure on the prey, whose population bottoms out, and then slowly begins to rebound. Later, the number of predators also reaches a minimum, at which point the entire growth and decay cycle starts over again.

In contrast to a linear system, the period of the population cycle is not fixed, but depends upon how far away from the stable equilibrium the solution orbit lies. Near equilibrium, the solutions are close to those of the linearized system which, in view of its eigenvalues $\pm 3\,\mathrm{i}\,\sqrt{2}$, are periodic of frequency $3\,\sqrt{2}$ and period $\sqrt{2}\,\pi/3$. However, solutions that are far away from equilibrium have much longer periods, and so greater imbalances between predator and prey populations leads to longer periods, and more radically varying numbers. Understanding the mechanisms behind these population cycles is of increasing important in the ecological management of natural resources, [**biol**].

Example 20.30. In our next example, we look at the undamped oscillations of a pendulum. When we set the friction coefficient $\mu = 0$, the nonlinear second order ordinary differential equation (20.51) reduces to

$$m\,\frac{d^2\theta}{dt^2} + \kappa \sin\theta = 0. \tag{20.64}$$

As before, we convert this into a first order system

$$\frac{du}{dt} = v, \qquad \frac{dv}{dt} = -\,\alpha \sin u, \tag{20.65}$$

where

$$u(t) = \theta(t), \qquad v(t) = \frac{d\theta}{dt}\,, \qquad \alpha = \frac{\kappa}{m}\,.$$

The equilibria,

$$\mathbf{u}^\star_k = (k\,\pi, 0) \qquad \text{for} \qquad k = 0, \pm 1, \pm 2, \dots\,,$$

are the same as in the damped case, i.e., the pendulum is either at the top (k even) or the bottom (k odd) of the circle.

Let us see what the linearization stability test tells us. In this case, the Jacobian matrix of (20.65) is

$$\mathbf{F}'(u,v) = \begin{pmatrix} 0 & 1 \\ -\alpha \cos u & 0 \end{pmatrix}.$$

At the top equilibria

$$\mathbf{F}'(\mathbf{u}^\star_{2\,j+1}) = \mathbf{F}'\big(\,(2j+1)\,\pi, 0\,\big) = \begin{pmatrix} 0 & 1 \\ \alpha & 0 \end{pmatrix} \qquad \text{has real eigenvalues} \qquad \pm\sqrt{\alpha}\,,$$

and hence these equilibria are unstable saddle points, just as in the damped version. On the other hand, at the bottom equilibria

$$\mathbf{F}'(\mathbf{u}^\star_{2\,j}) = \mathbf{F}'(2\,j\,\pi, 0) = \begin{pmatrix} 0 & 1 \\ -\alpha & 0 \end{pmatrix}, \quad \text{has purely imaginary eigenvalues} \quad \pm\,\mathrm{i}\,\sqrt{\alpha}\,.$$

Without the benefit of damping, the linearization test is inconclusive, and the stability of the bottom equilibria remains in doubt.

Since we are dealing with a conservative system, the total energy of the pendulum, namely

$$E(u,v) = \tfrac{1}{2}\,m\,v^2 + \kappa\,(1-\cos u) = \frac{m}{2}\left(\frac{d\theta}{dt}\right)^2 + \kappa\,(1-\cos\theta) \tag{20.66}$$

should provide us with a first integral. Note that E is a sum of two terms, which represent, respectively, the kinetic energy due to the pendulum's motion, and the potential energy[†] due to the height of the pendulum bob. To verify that $E(u,v)$ is indeed a first integral, we compute

$$\frac{dE}{dt} = \frac{\partial E}{\partial u}\,\frac{du}{dt} + \frac{\partial E}{\partial v}\,\frac{dv}{dt} = (\kappa \sin u)\,v + (m\,v)(-\alpha \sin u) = 0, \qquad \text{since} \qquad \alpha = \frac{\kappa}{m}\,.$$

Therefore, E is indeed constant on solutions, reconfirming the physical basis of the model.

The phase plane solutions to the pendulum equation will move along the level sets of the energy function $E(u,v)$, which are plotted in Figure 20.13. Its critical points are the equilibria, where

$$\nabla E(\mathbf{u}) = \begin{pmatrix} \kappa \sin u \\ m\,v \end{pmatrix} = \mathbf{0}, \qquad \text{and hence} \qquad \mathbf{u} = \mathbf{u}^\star_k = (\,k\,\pi, 0\,), \quad k = 0, \pm 1, \pm 2, \dots\,.$$

To characterize the critical points, we appeal to the second derivative test, and so evaluate the Hessian

$$\nabla^2 E(u,v) = \begin{pmatrix} \kappa \cos u & 0 \\ 0 & m \end{pmatrix}.$$

[†] In a physical system, the potential energy is only defined up to an additive constant. Here we have fixed the zero energy level to be at the bottom of the pendulum's arc.

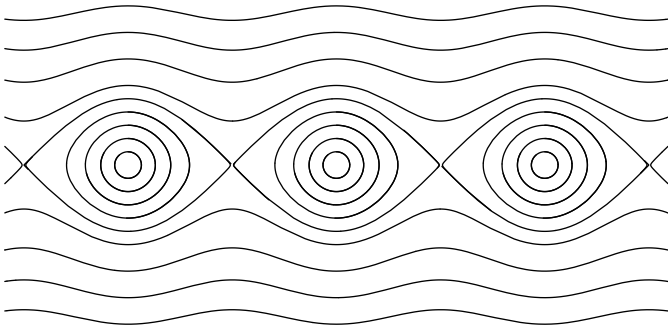


Figure 20.13. The Undamped Pendulum.

At the bottom equilibria,

$$\nabla^2 E(\mathbf{u}^\star_{2\,j}) = \nabla^2 E(2\,j\,\pi, 0) = \begin{pmatrix} \kappa & 0 \\ 0 & m \end{pmatrix}$$

is positive definite, since κ and m are positive constants. Therefore, the bottom equilibria are strict local minima of the energy, and so Theorem 20.28 guarantees their stability.

Each stable equilibrium is surrounded by a family of closed oval-shaped level curves, and hence forms a center. Each oval corresponds to a periodic solution[†] of the system, in which the pendulum oscillates back and forth symmetrically around the bottom of its arc. Near the equilibrium, the period is close to that of the linearized system, namely $2\,\pi/\sqrt{\alpha}$ as prescribed by the eigenvalues of the Jacobian matrix. This fact underlies the use of pendulum-based clocks for keeping time, first recognized by Galileo. A grandfather clock is accurate because the amplitude of its pendulum's oscillations is kept relatively small. However, moving further away from the equilibrium point in the phase plane, we find that the periodic solutions with very large amplitude oscillations, in which the pendulum becomes nearly vertical, have much longer periods, and so would lead to inaccurate time-keeping.

The large amplitude limit of the periodic solutions is of particular interest. The pair of trajectories connecting two distinct unstable equilibria are known as the *homoclinic orbits*, and play an essential role in the more advanced analysis of the pendulum under perturbations, [**130**, **X**]. Physically, a homoclinic orbit corresponds to a pendulum that starts out just shy of vertical, goes through exactly one full rotation, and eventually (as $t \to \infty$) ends up vertical again.

Finally, the level sets lying above and below the "cat's-eyes" formed by the homoclinic and periodic orbits are known as the *running orbits*. Since $u = \theta$ is a $2\,\pi$ periodic angular variable, a running orbit solution $(\,u(t), v(t)\,)^T = (\theta(t), \dot\theta(t))^T$, in fact, also corresponds to a periodic physical motion, in which the pendulum spins around and around its pivot. The larger the total energy $E(u, v)$, the farther away from the u–axis the running orbit lies, and the faster the pendulum spins.

[†] More precisely, a family of periodic solutions indexed by their initial condition on the oval, and differing only by a phase shift: $\mathbf{u}(t - \delta)$.

In summary, the qualitative behavior of a solution to the pendulum equation is almost entirely characterized by its energy:

- $E = 0$, stable equilibria,
- $0 < E < 2\kappa$, periodic oscillating orbits,
- $E = 2\kappa$, unstable equilibria and homoclinic orbits,
- $E > 2\kappa$, running orbits.

Example 20.31. The system governing the dynamical rotations of a rigid solid body around its center of mass are known as the *Euler equations* of rigid body mechanics, in honor of the prolific eighteenth century Swiss mathematician Leonhard Euler, cf. [**78**]. According to Exercise 8.4.23, the eigenvectors of the positive definite inertia tensor of the body prescribe its three mutually orthogonal *principal axes*. The corresponding eigenvalues $I_1, I_2, I_3 > 0$ are called the *principal moments of inertia*. Let $u_1(t), u_2(t), u_3(t)$ denote the angular momenta of the body around its three principal axes. In the absence of external forces, the dynamical system governing the body's rotations around its center of mass takes the symmetric form

$$\frac{du_1}{dt} = \frac{I_2 - I_3}{I_2\, I_3}\, u_2 u_3, \qquad \frac{du_2}{dt} = \frac{I_3 - I_1}{I_1\, I_3}\, u_1 u_3, \qquad \frac{du_3}{dt} = \frac{I_1 - I_2}{I_1\, I_2}\, u_1 u_2. \tag{20.67}$$

Th Euler equations model, for example, the dynamics of a satellite spinning in its orbit around the earth. The solution will prescribe the angular motions of the satellite around its center of mass, but not the overall motion of the center of mass as the satellite orbits the earth.

Let us assume that the moments of inertia are all different, which we place in increasing order $0 < I_1 < I_2 < I_3$. The equilibria of the Euler system (20.67) are where the right hand sides simultaneously vanish, which requires that either $u_2 = u_3 = 0$, or $u_1 = u_3 = 0$, or $u_1 = u_2 = 0$. In other words, every point on the three coordinate axes is an equilibrium configuration. Since the variables represent angular momenta, these equilibria correspond to the body spinning around one of its principal axes at a fixed angular velocity.

Let us analyze the stability of these equilibrium configurations. The linearization test fails completely — as it must whenever dealing with a non-isolated equilibrium. But the Euler equations turn out to admit two independent first integrals:

$$E(\mathbf{u}) = \frac{1}{2}\left(\frac{u_1^2}{I_1} + \frac{u_2^2}{I_2} + \frac{u_3^2}{I_3}\right), \qquad A(\mathbf{u}) = \frac{1}{2}\left(u_1^2 + u_2^2 + u_3^2\right). \tag{20.68}$$

The first is the total kinetic energy, while the second is the total angular momentum. The proof that $dE/dt = 0 = dA/dt$ for any solution $\mathbf{u}(t)$ to (20.67) is left as Exercise ■ for the reader.

Since both E and A are constant, the solutions to the system are constrained to move along a common level set $C = \{\, E = e,\ A = a \,\}$. Thus, the solution trajectories are the curves obtained by intersecting the sphere $S_a = \{\, A(\mathbf{u}) = a \,\}$ of radius $\sqrt{2\,a}$ with the ellipsoid $L_e = \{\, E(\mathbf{u}) = e \,\}$. In Figure rigid■, we have graphed the intersection curves

$C_{a,e} = S_a \cap L_e$ of a fixed sphere with a family of ellipsoids corresponding to different values of the kinetic energy. The six equilibria on the sphere are at its intersections with the coordinate axes. Those on the x and z axes are surrounded by closed periodic orbits, and hence are stable equilibria; indeed, they are, respectively, local minima and maxima of the energy when restricted to the sphere. On the other hand, the two equilibria on the y axis have the form of unstable saddle points. We conclude that a body that spins around either of its principal axes with the smallest or the largest moment of inertia is stable, whereas one that spins around the axis corresponding to the intermediate moment of inertia is unstable. This mathematical deduction can be demonstrated physically by flipping a solid rectangular object, e.g., this book, up into the air. It is easy to arrange it to spin around its long axis or its short axis in a stable manner, but it will balk at attempts to make it rotate around its middle axis!

Lyapunov's Method

Systems that incorporate damping, viscosity and/or frictional effects do not typically possess non-constant first integrals. From a physical standpoint, the damping will cause the total energy of the system to be a decreasing function of time. Asymptotically, the system returns to a (stable) equilibrium, and the extra energy has been dissipated away. However, this physical principle captures important mathematical implications for the behavior of solutions. It leads to a useful alternative method for establishing stability of equilibria, even in cases when the linearization stability test is inconclusive. The nineteenth century Russian mathematician Alexander Lyapunov was the first to pinpoint the importance of such functions in dynamics.

Definition 20.32. A *Lyapunov function* for the first order autonomous system $\dot{\mathbf{u}} = \mathbf{F}(\mathbf{u})$ is a continuous real-valued function $L(\mathbf{u})$ that is non-increasing on all solutions $\mathbf{u}(t)$, meaning that

$$L(\mathbf{u}(t)) \;\le\; L(\mathbf{u}(t_0)) \qquad \text{for all} \qquad t > t_0. \tag{20.69}$$

A *strict Lyapunov function* satisfies the strict inequality

$$L(\mathbf{u}(t)) \;<\; L(\mathbf{u}(t_0)) \qquad \text{for all} \qquad t > t_0, \tag{20.70}$$

whenever $\mathbf{u}(t)$ is a *non-equilibrium* solution to the system. (Clearly, the Lyapunov function must be constant on an equilibrium solution.)

We can characterize continuously differentiable Lyapunov functions by using the elementary calculus results that a scalar function is non-increasing if and only if its derivative is non-negative, and is strictly decreasing if its derivative is strictly less than 0.

Lemma 20.33. *A continuously differentiable function $L(\mathbf{u})$ is a Lyapunov function for the system $\dot{\mathbf{u}} = \mathbf{F}(\mathbf{u})$ if and only if it satisfies the* Lyapunov inequality

$$\frac{d}{dt}\, L(\mathbf{u}(t)) = \nabla L(\mathbf{u}) \cdot \mathbf{F}(\mathbf{u}) \;\le\; 0 \qquad \text{for all solutions} \qquad \mathbf{u}(t). \tag{20.71}$$

If

$$\frac{d}{dt}\, L(\mathbf{u}(t)) = \nabla L(\mathbf{u}) \cdot \mathbf{F}(\mathbf{u}) \;<\; 0 \qquad \text{whenever} \qquad \mathbf{F}(\mathbf{u}) \neq \mathbf{0}, \tag{20.72}$$

then $L(\mathbf{u})$ is a strict Lyapunov function.

Indeed, the formula for the derivative of $L(\mathbf{u}(t))$ follows from the same chain rule computation used to establish (20.56)

The proof of Theorem 20.28 can be readily adapted to prove stability of equilibria of a system which has a Lyapunov function; details can be found in [**89**, **99**].

Theorem 20.34. *Let $L(\mathbf{u})$ be a Lyapunov function for the autonomous system $\dot{\mathbf{u}} = \mathbf{F}(\mathbf{u})$. If $\mathbf{u}^\star$ is a strict local minimum of L, then $\mathbf{u}^\star$ is a stable equilibrium point. If $L(\mathbf{u})$ is a strict Lyapunov function, then $\mathbf{u}^\star$ is an asymptotically stable equilibrium.*

Proof: ∎

Warning: Maxima of Lyapunov functions are *not* stable equilibria.

Unlike first integrals, which can, at least in principle, be systematically constructed by solving a first order partial differential equation, finding Lyapunov functions is much more of an art form, usually requiring some combination of physical intuition and inspired guesswork.

Example 20.35. Return to the planar system

$$\frac{du}{dt} = v, \qquad \frac{dv}{dt} = -\alpha \sin u - \beta\, v,$$

describing the damped oscillations of a pendulum, as in (20.52). Physically, we expect that the damping will cause a continual decrease in the total energy in the system, which, by (20.66), is

$$E(u, v) = \tfrac{1}{2}\, m\, v^2 + \kappa\,(1 - \cos u).$$

Let us prove that E is, indeed, a Lyapunov function. We compute its time derivative, when $u(t), v(t)$ is a solution to the damped system. Recalling that $\alpha = \kappa/m$, $\beta = \mu/m$, we find

$$\frac{dE}{dt} = \frac{\partial E}{\partial u}\,\frac{du}{dt} + \frac{\partial E}{\partial v}\,\frac{dv}{dt} = (\kappa \sin u)\, v + (m\, v)\,(-\alpha \sin u - \beta\, v) = -\mu\, v^2 \leq 0,$$

since we are assuming that the frictional coefficient $\mu > 0$. Therefore, the energy satisfies the Lyapunov stability criterion (20.71). Consequently, Theorem 20.34 re-establishes the stability of the energy minima $u^\star_{2\,k} = 2\,k\,\pi$, $v^\star_{2\,k} = 0$, where the damped pendulum is at the bottom of the arc. In fact, since $dE/dt < 0$ except when $v = 0$, with a little more work, the Lyapunov criterion can be used to establish their asymptotic stability.

20.4. Numerical Methods.

Since we have no hope of solving the vast majority of differential equations in explicit, analytic form, the design of suitable numerical algorithms for accurately approximating solutions is essential. The ubiquity of differential equations throughout mathematics and its applications has driven the tremendous research effort devoted to numerical solution schemes, some dating back to the beginnings of the calculus. Nowadays, one has the luxury of choosing from a wide range of excellent software packages that provide reliable and accurate results for a broad range of systems, at least for solutions over moderately

long time periods. However, all of these packages, and the underlying methods, have their limitations, and it is essential that one be able to to recognize when the software is working as advertised, and when it produces spurious results! Here is where the theory, particularly the classification of equilibria and their stability properties, as well as first integrals and Lyapunov functions, can play an essential role. Explicit solutions, when known, can also be used as test cases for tracking the reliability and accuracy of a chosen numerical scheme.

In this section, we survey the most basic numerical methods for solving initial value problems. For brevity, we shall only consider so-called single step schemes, culminating in the very popular and versatile fourth order Runge–Kutta Method. This should only serve as a extremely basic introduction to the subject, and many other important and useful methods can be found in more specialized texts, [**88**, **107**]. It goes without saying that some equations are more difficult to accurately approximate than others, and a variety of more specialized techniques are employed when confronted with a recalcitrant system. But all of the more advanced developments build on the basic schemes and ideas laid out in this section.

Euler's Method

The key issues confronting the numerical analyst of ordinary differential equations already appear in the simplest first order ordinary differential equation. Our goal is to calculate a decent approxiomation to the (unique) solution to the initial value problem

$$\frac{du}{dt} = F(t,u), \qquad\qquad u(t_0) = u_0. \tag{20.73}$$

To keep matters simple, we will focus our attention on the scalar case; however, all formulas and results written in a manner that can be readily adapted to first order systems — just replace the scalar functions $u(t)$ and $F(t,u)$ by vector-valued functions $\mathbf{u}$ and $\mathbf{F}(t,\mathbf{u})$ throughout. (The time t, of course, remains a scalar.) Higher order ordinary differential equations are inevitably handled by first converting them into an equivalent first order system, as discussed in Section 20.1, and then applying the numerical scheme.

The very simplest numerical solution method is named after Leonhard Euler — although Newton and his contemporaries were well aware of such a simple technique. Euler's Method is rarely used in practice because much more efficient and accurate techniques can be implemented with minimal additional work. Nevertheless, the method lies at the core of the entire subject, and must be thoroughly understood before progressing on to the more sophisticated algorithms that arise in real-world computations.

Starting at the initial time t_0, we introduce successive *mesh points* (or sample times)

$$t_0 < t_1 < t_2 < t_3 < \ \cdots \ ,$$

continuing on until we reach a desired final time $t_n = t^\star$. The mesh points should be fairly closely spaced. To keep the analysis as simple as possible, we will always use a uniform *step size*, and so

$$h = t_{k+1} - t_k > 0, \tag{20.74}$$

does not depend on k and is assumed to be relatively small. This assumption serves to simplify the analysis, and does not significantly affect the underlying ideas. For a uniform

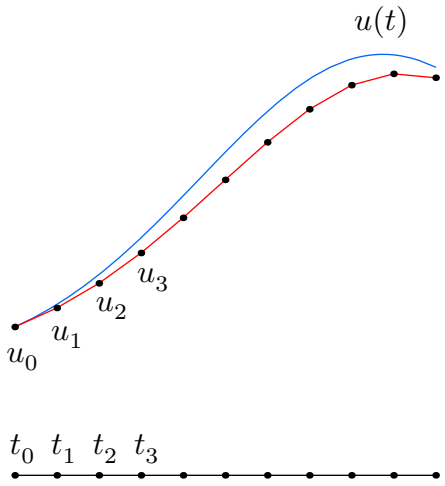


Figure 20.14. Euler's Method.

step size, the k^{th} mesh point is at $t_k = t_0 + k\,h$. More sophisticated *adaptive* methods, in which the step size is adjusted in order to maintain accuracy of the numerical solution, can be found in more specialized texts, e.g., [**88**, **107**]. Our numerical algorithm will recursively compute approximations $u_k \approx u(t_k)$, for $k = 0, 1, 2, 3, \dots$, to the sampled values of the solution $u(t)$ at the chosen mesh points. Our goal is to make the *error* $E_k = u_k - u(t_k)$ in the approximation at each time t_k as small as possible. If required, the values of the solution $u(t)$ between mesh points may be computed by a subsequent interpolation procedure, e.g., the cubic splines of Section 11.4.

As you learned in first year calculus, the simplest approximation to a (continuously differentiable) function $u(t)$ is provided by its tangent line or first order Taylor polynomial. Thus, near the mesh point t_k

$$u(t) \;\approx\; u(t_k) + (t - t_k)\,\frac{du}{dt}(t_k) \;=\; u(t_k) + (t - t_k)\,F(t_k, u(t_k)),$$

in which we replace the derivative du/dt of the solution by the right hand side of the governing differential equation (20.73). In particular, the approximate value of the solution at the subsequent mesh point is

$$u(t_{k+1}) \;\approx\; u(t_k) + (t_{k+1} - t_k)\,F(t_k, u(t_k)). \tag{20.75}$$

This simple idea forms the basis of Euler's Method.

Since in practice we only know the approximation u_k to the value of $u(t_k)$ at the current mesh point, we are forced to replace $u(t_k)$ by its approximation u_k in the preceding formula. We thereby convert (20.75) into the iterative scheme

$$u_{k+1} = u_k + (t_{k+1} - t_k)\,F(t_k, u_k). \tag{20.76}$$

In particular, when based on a uniform step size (20.74), *Euler's Method* takes the simple form

$$u_{k+1} = u_k + h\,F(t_k, u_k). \tag{20.77}$$

As sketched in Figure 20.14, the method starts off approximating the solution reasonably well, but gradually loses accuracy as the errors accumulate.

Euler's Method is the simplest example of a *one-step* numerical scheme for integrating an ordinary differential equation. This refers to the fact that the succeeding approximation, $u_{k+1} \approx u(t_{k+1})$, depends only upon the current value, $u_k \approx u(t_k)$, which is one mesh point or "step" behind.

To begin to understand how Euler's Method works in practice, let us test it on a problem we know how to solve, since this will allow us to precisely monitor the resulting errors in our numerical approximation to the solution.

Example 20.36. The simplest "nontrivial" initial value problem is

$$\frac{du}{dt} = u, \qquad\qquad u(0) = 1,$$

whose solution is, of course, the exponential function $u(t) = e^t$. Since $F(t, u) = u$, Euler's Method (20.77) with a fixed step size $h > 0$ takes the form

$$u_{k+1} = u_k + h\, u_k = (1 + h)\, u_k.$$

This is a linear iterative equation, and hence easy to solve:

$$u_k = (1 + h)^k u_0 = (1 + h)^k$$

is our proposed approximation to the solution $u(t_k) = e^{t_k}$ at the mesh point $t_k = k\, h$. Therefore, the Euler scheme to solve the differential equation, we are effectively approximating the exponential by a power function:

$$e^{t_k} = e^{k\, h} \approx (1 + h)^k$$

When we use simply t to indicate the mesh time $t_k = k\, h$, we recover, in the limit, a well-known calculus formula:

$$e^t = \lim_{h \to 0} (1 + h)^{t/h} = \lim_{k \to \infty} \left(1 + \frac{t}{k} \right)^k. \tag{20.78}$$

A reader familiar with the computation of compound interest will recognize this particular approximation. As the time interval of compounding, h, gets smaller and smaller, the amount in the savings account approaches an exponential.

How good is the resulting approximation? The *error*

$$E(t_k) = E_k = u_k - e^{t_k}$$

measures the difference between the true solution and its numerical approximation at time $t = t_k = k\, h$. Let us tabulate the error at the particular times $t = 1, 2$ and 3 for various values of the step size h. The actual solution values are

$$e^1 = e = 2.718281828\ldots, \qquad e^2 = 7.389056096\ldots, \qquad e^3 = 20.085536912\ldots.$$

In this case, the numerical approximation always underestimates the true solution.

h	$E(1)$	$E(2)$	$E(3)$
.1	$-.125$	$-.662$	-2.636
.01	$-.0134$	$-.0730$	$-\ .297$
.001	$-.00135$	$-.00738$	$-\ .0301$
.0001	$-.000136$	$-.000739$	$-\ .00301$
.00001	$-.0000136$	$-.0000739$	$-\ .000301$

Some key observations:

- For a fixed step size h, the further we go from the initial point $t_0 = 0$, the larger the magnitude of the error.
- On the other hand, the smaller the step size, the smaller the error at a fixed value of t. The trade-off is that more steps, and hence more computational effort[†] is required to produce the numerical approximation. For instance, we need $k = 10$ steps of size $h = .1$, but $k = 1000$ steps of size $h = .001$ to compute an approximation to $u(t)$ at time $t = 1$.
- The error is more or less in proportion to the step size. Decreasing the step size by a factor of $\frac{1}{10}$ decreases the error by a similar amount, but simultaneously increases the amount of computation by a factor of 10.

The final observation indicates that the Euler Method is of *first order*, which means that the error depends *linearly*[‡] on the step size h. More specifically, at a fixed time t, the error is bounded by

$$| E(t) | = | u_k - u(t) | \leq C(t)\, h, \qquad \text{when} \qquad t = t_k = k\, h, \tag{20.79}$$

for some positive $C(t) > 0$ that depends upon the time, and the initial condition, but not on the step size.

Example 20.37. The solution to the initial value problem

$$\frac{du}{dt} = \big(\, 1 - \tfrac{4}{3}\, t \,\big)\, u, \qquad u(0) = 1, \tag{20.80}$$

was found in Example 20.3 by the method of separation of variables:

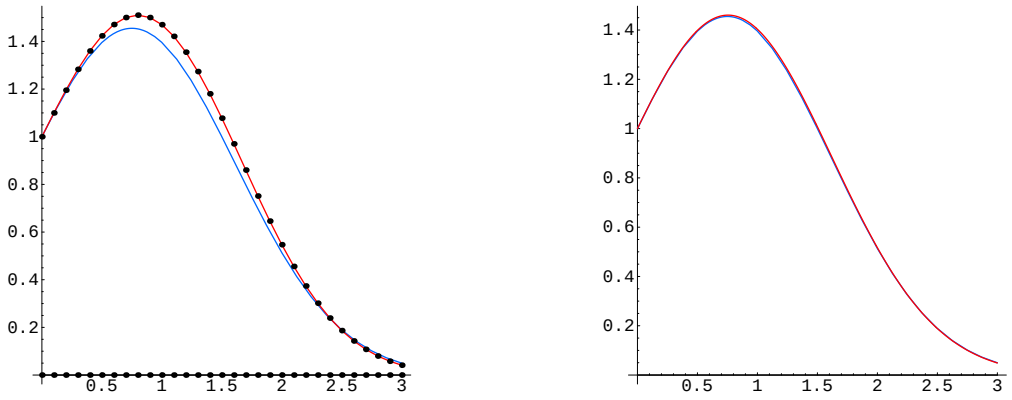
$$u(t) = \exp\big(\, t - \tfrac{2}{3}\, t^2 \,\big). \tag{20.81}$$

Euler's Method leads to the iterative numerical scheme

$$u_{k+1} = u_k + h\, \big(\, 1 - \tfrac{4}{3}\, t_k \,\big)\, u_k, \qquad u_0 = 1.$$

[†] In this case, there happens to be an explicit formula for the numerical solution which can be used to bypass the iterations. However, in almost any other situation, one cannot compute the approximation u_k without having first determined the intermediate values $u_0, \dots, u_{k-1}$.

[‡] See the discussion of the order of iterative methods in Section 19.1 for motivation.



$h = .1$ $\qquad\qquad$ $h = .01$

Figure 20.15. Euler's Method for $\dot{u} = \left(1 - \frac{4}{3}\,t\right) u$.

In Figure 20.15 we compare the graphs of the actual and numerical solutions for step sizes $h = .1$ and .01. In the former plot, we explitiely show the mesh points, but not in the latter, since they are too dense; moreover, the graphs of the numerical and true solutions are almost indistinguishable at this resolution.

The following table lists the numerical errors $E(t_k) = u_k - u(t_k)$ between the computed and actual solution values

$$u(1) = 1.395612425\ldots\,, \qquad u(2) = .513417119\ldots\,, \qquad u(3) = .049787068\ldots\,,$$

for several different step sizes:

h	$E(1)$	$E(2)$	$E(3)$
.1000	.07461761	.03357536	$-.00845267$
.0100	.00749258	.00324416	$-.00075619$
.0010	.00074947	.00032338	$-.00007477$
.0001	.00007495	.00003233	$-.00000747$

As in the previous example, each decrease in step size by a factor of 10 leads to one additional decimal digit of accuracy in the computed solution.

Taylor Methods

In general, the order of a numerical solution method governs both the accuracy of its approximations and the speed at which they converge to the true solution as the step size is decreased. Although the Euler Method is simple and easy to implement, it is only a first order scheme, and therefore of limited utility in serious computations. So, the goal is to devise simple numerical methods that enjoy a much higher order of accuracy.

Our derivation of the Euler Method was based on a first order Taylor approximation to the solution. So, an evident way to design a higher order method is to employ a higher

order Taylor approximation. The Taylor series expansion for the solution $u(t)$ at the succeeding mesh point $t_{k+1} = t_k + h$ has the form

$$u(t_{k+1}) = u(t_k + h) = u(t_k) + h\,\frac{du}{dt}(t_k) + \frac{h^2}{2}\,\frac{d^2u}{dt^2}(t_k) + \frac{h^3}{6}\,\frac{d^3u}{dt^3}(t_k) + \,\cdots\,. \tag{20.82}$$

As we just saw, we can evaluate the first derivative term through use of the underlying differential equation:

$$\frac{du}{dt} = F(t,u). \tag{20.83}$$

The second derivative term can be found by differentiating the equation with respect to t. Invoking the chain rule†,

$$\begin{aligned}\frac{d^2u}{dt^2} = \frac{d}{dt}\,\frac{du}{dt} = \frac{d}{dt}\,F(t,u(t)) &= \frac{\partial F}{\partial t}(t,u) + \frac{\partial F}{\partial u}(t,u)\,\frac{du}{dt}\\ &= \frac{\partial F}{\partial t}(t,u) + \frac{\partial F}{\partial u}(t,u)\,F(t,u) \equiv F^{(2)}(t,u).\end{aligned} \tag{20.84}$$

This operation is known as the *total derivative*, indicating that that we must treat the second variable u as a function of t when differentiating. Substituting (20.83–84) into (20.82) and truncating at order h^2 leads to the *Second Order Taylor Method*

$$\begin{aligned}u_{k+1} &= u_k + h\,F(t_k,u_k) + \frac{h^2}{2}\,F^{(2)}(t_k,u_k)\\ &= u_k + h\,F(t_k,u_k) + \frac{h^2}{2}\,\left(\frac{\partial F}{\partial t}(t_k,u_k) + \frac{\partial F}{\partial u}(t_k,u_k)\,F(t_k,u_k)\right),\end{aligned} \tag{20.85}$$

in which, as before, we replace the solution value $u(t_k)$ by its computed approximation u_k. The resulting method is of second order, meaning that the error function satisfies the quadratic error estimate

$$|\,E(t)\,| = |\,u_k - u(t)\,| \le C(t)\,h^2 \qquad \text{when} \qquad t = t_k = k\,h. \tag{20.86}$$

Example 20.38. Let us explicitly formulate the second order Taylor Method for the initial value problem (20.80). Here

$$\begin{aligned}\frac{du}{dt} &= F(t,u) = \left(1 - \tfrac{4}{3}\,t\right)u,\\ \frac{d^2u}{dt^2} &= \frac{d}{dt}\,F(t,u) = -\tfrac{4}{3}\,u + \left(1 - \tfrac{4}{3}\,t\right)\,\frac{du}{dt} = -\tfrac{4}{3}\,u + \left(1 - \tfrac{4}{3}\,t\right)^2 u,\end{aligned}$$

and so (20.85) becomes

$$u_{k+1} = u_k + h\left(1 - \tfrac{4}{3}\,t_k\right)u_k + \tfrac{1}{2}\,h^2\big[-\tfrac{4}{3}\,u_k + \left(1 - \tfrac{4}{3}\,t_k\right)^2 u_k\big], \qquad u_0 = 1.$$

The following table lists the errors between the values computed by the second order Taylor scheme and the actual solution values, as given in Example 20.37.

† We assume throughout that F has as many continuous derivatives as needed.

h	$E(1)$	$E(2)$	$E(3)$
.100	.00276995	$-.00133328$	.00027753
.010	.00002680	$-.00001216$	.00000252
.001	.00000027	$-.00000012$	.00000002

In accordance with the quadratic error estimate (20.86), a decrease in the step size by a factor of $\frac{1}{10}$ leads in an increase in accuracy of the solution by a factor $\frac{1}{100}$, i.e., an increase in 2 significant decimal places in the numerical approximation of the solution.

Higher order Taylor methods are obtained by including further terms in the expansion (20.82). For example, to derive a third order Taylor method, we include the third order term $(h^3/6)d^3u/dt^3$ in the Taylor expansion, where we evaluate the third derivative by differentiating (20.84), and so

$$\begin{aligned} \frac{d^3u}{dt^3} &= \frac{d}{dt}\,\frac{d^2u}{dt^2} = \frac{d}{dt}\,F^{(2)}(t,u) = \frac{\partial F^{(2)}}{\partial t} \;+\; \frac{\partial F^{(2)}}{\partial u}\,\frac{du}{dt} = \frac{\partial F^{(2)}}{\partial t} \;+\; F\,\frac{\partial F^{(2)}}{\partial u} \\ &= \frac{\partial^2 F}{\partial t^2} + 2\,F\,\frac{\partial^2 F}{\partial t\,\partial u} + F^2\,\frac{\partial^2 F}{\partial u^2} + \frac{\partial F}{\partial t}\,\frac{\partial F}{\partial u} + F\left(\frac{\partial F}{\partial u}\right)^2 \equiv F^{(3)}(t,u), \end{aligned} \tag{20.87}$$

where we continue to make use of the fact that $du/dt = F(t,u)$ is provided by the right hand side of the differential equation. The resulting third order Taylor method is

$$u_{k+1} = u_k + h\,F(t_k,u_k) + \frac{h^2}{2}\,F^{(2)}(t_k,u_k) + \frac{h^3}{6}\,F^{(3)}(t_k,u_k), \tag{20.88}$$

where the last two summand are given by (20.84), (20.87), respectively. The higher order expressions are even worse, and a good symbolic manipulation system is almost essential for accurate computation.

Whereas higher order Taylor methods are easy to motivate, they are rarely used in practice. There are two principal difficulties:

- Owing to their dependence upon the partial derivatives of $F(t,u)$, the right hand side of the differential equation needs to be rather smooth.
- Even worse, the explicit formulae become exceedingly complicated, even for relatively simple functions $F(t,u)$. Efficient evaluation of the multiplicity of terms in the Taylor approximation and avoidance of round off errors become significant concerns.

As a result, mathematicians soon abandoned the Taylor series approach, and began to look elsewhere for high order, efficient integration methods.

Error Analysis

Before pressing on, we need to engage in a more serious discussion of the error in a numerical scheme. A general *one-step* numerical method can be written in the form

$$u_{k+1} = G(h,t_k,u_k), \tag{20.89}$$

where G is a prescribed function of the current approximate solution value $u_k \approx u(t_k)$, the time t_k, and the step size $h = t_{k+1} - t_k$, which, for illustrative purposes, we assume to be fixed. (We leave the discussion of *multi-step methods*, in which G could also depend upon the earlier values $u_{k-1}, u_{k-2}, \dots$, to more advanced texts, e.g., [**88**, **107**].)

In any numerical integration scheme there are, in general, three sources of error.

- The first is the *local error* committed in the current step of the algorithm. Even if we have managed to compute a completely accurate value of the solution $u_k = u(t_k)$ at time t_k, the numerical approximation scheme (20.89) is almost certainly not exact, and will therefore introduce an error into the next computed value $u_{k+1} \approx u(t_{k+1})$. Round-off errors, resulting from the finite precision of the computer arithmetic, will also contribute to the local error.
- The second is due to the error that is already present in the current approximation $u_k \approx u(t_k)$. The local errors tend to accumulate as we continue to run the iteration, and the net result is the *global error*, which is what we actually observe when compaing the numerical apporximation with the exact solution.
- Finally, if the initial condition $u_0 \approx u(t_0)$ is not computed accurately, this *initial error* will also make a contribution. For example, if $u(t_0) = \pi$, then we introduce some initial error by using a decimal approximation, say $\pi \approx 3.14159$.

The third error source will, for simplicity, be ignored in our discussion, i.e., we will assume $u_0 = u(t_0)$ is exact. Further, for simplicity we will assume that round-off errors do not play any significant role — although one must always keep them in mind when analyzing the computation. Since the global error is entirely due to the accumulation of successive local errors, we must first understand the local error in detail.

To measure the local error in going from t_k to t_{k+1}, we compare the exact solution value $u(t_{k+1})$ with its numerical approximation (20.89) under the assumption that the current computed value is correct: $u_k = u(t_k)$. Of course, in practice this is never the case, and so the local error is an artificial quantity. Be that as it may, in most circumstances the local error is (a) easy to estimate, and, (b) provides a reliable guide to the global accuracy of the numerical scheme. To estimate the local error, we assume that the step size h is small and approximate the solution $u(t)$ by its Taylor expansion[†]

$$\begin{aligned} u(t_{k+1}) &= u(t_k) + h\,\frac{du}{dt}(t_k) + \frac{h^2}{2}\,\frac{d^2u}{dt^2}(t_k) + \cdots \\ &= u_k + h\,F(t_k, u_k) + \frac{h^2}{2}\,F^{(2)}(t_k, u_k) + \cdots . \end{aligned} \tag{20.90}$$

In the second expression, we have employed (20.84) and its higher order analogs to evaluate the derivative terms, and then invoked our local accuracy assumption to replace $u(t_k)$ by u_k. On the other hand, a direct Taylor expansion, in h, of the numerical scheme produces

$$u_{k+1} = G(h, t_k, u_k) = G(0, t_k, u_k) + h\,\frac{\partial G}{\partial h}\,(0, t_k, u_k) + \frac{h^2}{2}\,\frac{\partial^2 G}{\partial h^2}\,(0, t_k, u_k) + \cdots . \tag{20.91}$$

[†] In our analysis, we assume that the differential equation, and hence the solution, has sufficient smoothness to justify the relevant Taylor approximation.

The local error is obtained by comparing these two Taylor expansions.

Definition 20.39. A numerical integration method is of *order* n if the Taylor expansions (20.90, 91) of the exact and numerical solutions agree up to order h^n.

For example, the Euler Method

$$u_{k+1} = G(h, t_k, u_k) = u_k + h\, F(t_k, u_k),$$

is already in the form of a Taylor expansion — that has no terms involving $h^2, h^3, \dots$. Comparing with the exact expansion (20.90), we see that the constant and order h terms are the same, but the order h^2 terms differ (unless $F^{(2)} \equiv 0$). Thus, according to the definition, the Euler Method is a first order method. Similarly, the Taylor Method (20.85) is a second order method, because it was explicitly designed to match the constant, h and h^2 terms in the Taylor expansion of the solution. The general Taylor Method of order n sets $G(h, t_k, u_k)$ to be exactly the order n Taylor polynomial, differing from the full Taylor expansion at order h^{n+1}.

Under fairly general hypotheses, it can be proved that, if the numerical scheme has order n as measured by the local error, then the *global error* is bounded by a multiple of h^n. In other words, assuming no round-off or initial error, the computed value u_k and the solution at time t_k can be bounded by

$$| u_k - u(t_k) | \le M\, h^n, \tag{20.92}$$

where the constant $M > 0$ may depend on the time t_k and the particular solution $u(t)$. The error bound (20.92) serves to justify our numerical observations. For a method of order n, decreasing the step size by a factor of $\frac{1}{10}$ will decrease the overall error by a factor of about 10^{-n}, and so, roughly speaking, we anticipate gaining an additional n digits of accuracy — at least up until the point that round-off errors begin to play a role. Readers interested in a more complete error analysis of numerical integration schemes should consult a specialized text, e.g., [**88**, **107**].

The bottom line is the higher its order, the more accurate the numerical scheme, and hence the larger the step size that can be used to produce the solution to a desired accuracy. However, this must be balanced with the fact that higher order methods inevitably require more computational effort at each step. If the total amount of computation has also decreased, then the high order method is to be preferred over a simpler, lower order method. Our goal now is to find another route to the design of higher order methods that avoids the complications inherent in a direct Taylor expansion.

An Equivalent Integral Equation

The secret to the design of higher order numerical algorithms is to replace the differential equation by an equivalent integral equation. By way of motivation, recall that, in general, differentiation is a badly behaved process; a reasonable function can have an unreasonable derivative. On the other hand, integration ameliorates; even quite nasty functions have relatively well-behaved integrals. For the same reason, accurate numerical integration is relatively painless, whereas numerical differentiation should be avoided unless necessary. While we have not dealt directly with integral equations in this text, the

subject has been extensively developed by mathematicians, [**46**], and has many important physical applications, [**apl**].

Conversion of an initial value problem (20.73) to an integral equation is straightforward. We integrate both sides of the differential equation from the initial point t_0 to a variable time t. The Fundamental Theorem of Calculus is used to explicitly evaluate the left hand integral:

$$u(t) - u(t_0) = \int_{t_0}^{t} \dot{u}(s)\, ds = \int_{t_0}^{t} F(s, u(s))\, ds.$$

Rearranging terms, we arrive at the key result.

Lemma 20.40. *The solution $u(t)$ to the the integral equation*

$$u(t) = u(t_0) + \int_{t_0}^{t} F(s, u(s))\, ds \tag{20.93}$$

coincides with the solution to the initial value problem $\dfrac{du}{dt} = F(t, u)$, $u(t_0) = u_0$.

Proof: Our derivation already showed that the solution to the initial value problem satisfies the integral equation (20.93). Conversely, suppose that $u(t)$ solves the integral equation. Since $u(t_0) = u_0$ is constant, the Fundamental Theorem of Calculus tells us that the derivative of the right hand side of (20.93) is equal to the integrand, so $\dfrac{du}{dt} = F(t, u(t))$. Moreover, at $t = t_0$, the upper and lower limits of the integral coincide, and so it vanishes, whence $u(t) = u(t_0) = u_0$ has the correct initial conditions. *Q.E.D.*

Observe that, unlike the differential equation, the integral equation (20.93) requires no additional initial condition — it is automatically built into the equation. The proofs of the fundamental existence and uniqueness Theorems 20.7 and 20.9 for ordinary differential equations are, in fact, based on the integral equation reformulation of the initial value problem; see [**89**, **99**] for details.

The integral equation reformulation is equally valid for systems of first order ordinary differential equations. As noted above, $\mathbf{u}(t)$ and $\mathbf{F}(t, \mathbf{u}(t))$ become vector-valued functions. Integrating a vector-valued function is accomplished by integrating its individual components. Complete details are left to the exercises.

Implicit and Predictor–Corrector Methods

From this point onwards, we shall abandon the original initial value problem, and turn our attention to numerically solving the equivalent integral equation (20.93). Let us rewrite the integral equation, starting at the mesh point t_k instead of t_0, and integrating until time $t = t_{k+1}$. The result is the basic integral formula

$$u(t_{k+1}) = u(t_k) + \int_{t_k}^{t_{k+1}} F(s, u(s))\, ds \tag{20.94}$$

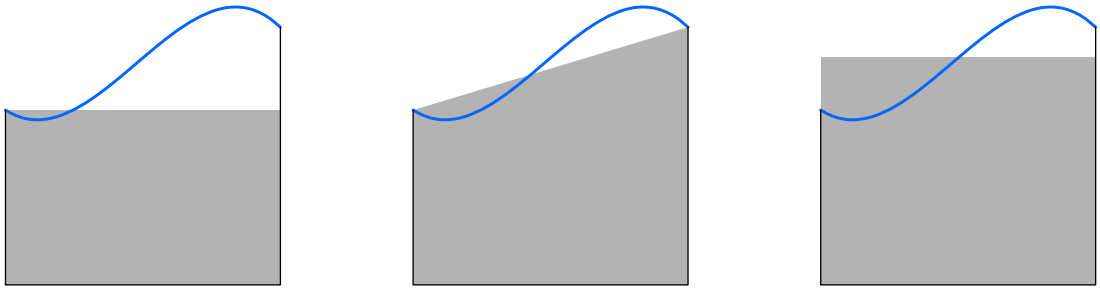


Figure 20.16. Numerical Integration Methods.

that (implicitly) computes the value of the solution at the subsequent mesh point. Comparing this formula with the Euler Method

$$u_{k+1} = u_k + h\, F(t_k, u_k), \qquad \text{where} \qquad h = t_{k+1} - t_k,$$

and assuming for the moment that $u_k = u(t_k)$ is exact, we discover that we are merely approximating the integral by

$$\int_{t_k}^{t_{k+1}} F(s, u(s))\, ds \;\approx\; h\, F(t_k, u(t_k)). \tag{20.95}$$

This is the Left Endpoint Rule for numerical integration — that approximates the area under the curve $g(t) = F(t, u(t))$ between $t_k \le t \le t_{k+1}$ by the area of a rectangle whose height $g(t_k) = F(t_k, u(t_k)) \approx F(t_k, u_k)$ is prescribed by the left-hand endpoint of the graph. As indicated in Figure 20.16, this is a reasonable, but not especially accurate method of numerical integration.

In first year calculus, you no doubt encountered much better methods of approximating the integral of a function. One of these is the *Trapezoid Rule*, which approximates the integral of the function $g(t)$ by the area of a trapezoid obtained by connecting the two points $(t_k, g(t_k))$ and $(t_{k+1}, g(t_{k+1}))$ on the graph of g by a straight line, as in the second Figure 20.16. Let us therefore try replacing (20.95) by the more accurate trapezoidal approximation

$$\int_{t_k}^{t_{k+1}} F(s, u(s))\, ds \;\approx\; \tfrac{1}{2}\, h \,\big[\, F(t_k, u(t_k)) + F(t_{k+1}, u(t_{k+1})) \,\big]. \tag{20.96}$$

Substituting this approximation into the integral formula (20.94), and replacing the solution values $u(t_k), u(t_{k+1})$ by their numerical approximations, leads to the (hopefully) more accurate numerical scheme

$$u_{k+1} = u_k + \tfrac{1}{2}\, h \,\big[\, F(t_k, u_k) + F(t_{k+1}, u_{k+1}) \,\big], \tag{20.97}$$

known as the *Trapezoid Method*. It is an *implicit scheme*, since the updated value u_{k+1} appears on both sides of the equation, and hence is only defined implicitly.

Example 20.41. Consider the differential equation $\dot{u} = \left(1 - \frac{4}{3}t\right) u$ studied in Examples 20.37 and 20.38. The Trapezoid Method with a fixed step size h takes the form

$$u_{k+1} = u_k + \tfrac{1}{2} h \left[\left(1 - \tfrac{4}{3} t_k\right) u_k + \left(1 - \tfrac{4}{3} t_{k+1}\right) u_{k+1} \right].$$

In this case, we can explicit solve for the updated solution value, leading to the recursive formula

$$u_{k+1} = \frac{1 + \frac{1}{2} h \left(1 - \frac{4}{3} t_k\right)}{1 - \frac{1}{2} h \left(1 - \frac{4}{3} t_{k+1}\right)} \, u_k = \frac{1 + \frac{1}{2} h - \frac{2}{3} h t_k}{1 - \frac{1}{2} h + \frac{2}{3} h \, (t_k + h)} \, u_k. \tag{20.98}$$

Implementing this scheme for three different step sizes gives the following errors between the computed and true solutions at times $t = 1, 2, 3$.

h	$E(1)$	$E(2)$	$E(3)$
.100	$-.00133315$	.00060372	$-.00012486$
.010	$-.00001335$	.00000602	$-.00000124$
.001	$-.00000013$	.00000006	$-.00000001$

The numerical data indicates that the Trapezoid Method is of second order. For each reduction in step size by $\frac{1}{10}$, the accuracy in the solution increases by, roughly, a factor of $\frac{1}{100} = \frac{1}{10^2}$; that is, the numerical solution acquires two additional accurate decimal digits. You are asked to formally prove this in Exercise ■.

The main difficulty with the Trapezoid Method (and any other implicit scheme) is immediately apparent. The updated approximate value for the solution u_{k+1} appears on both sides of the equation (20.97). Only for very simple functions $F(t, u)$ can one expect to solve (20.97) explicitly for u_{k+1} in terms of the known quantities t_k, u_k and $t_{k+1} = t_k + h$. The alternative is to employ a numerical equation solver, such as the bisection algorithm or Newton's Method, to compute u_{k+1}. In the case of Newton's Method, one would use the current approximation u_k as a first guess for the new approximation u_{k+1} — as in the continuation method discussed in Example 19.25. The resulting scheme requires some work to program, but can be effective in certain situations.

An alternative, less involved strategy is based on the following far-reaching idea. We already know a half-way decent approximation to the solution value u_{k+1} — namely that provided by the more primitive Euler scheme

$$\widetilde{u}_{k+1} = u_k + h \, F(t_k, u_k). \tag{20.99}$$

Let's use this estimated value in place of u_{k+1} on the right hand side of the implicit equation (20.97). The result

$$\begin{aligned} u_{k+1} &= u_k + \tfrac{1}{2} h \left[F(t_k, u_k) + F(t_k + h, \widetilde{u}_{k+1}) \right] \\ &= u_k + \tfrac{1}{2} h \left[F(t_k, u_k) + F\big(t_k + h, u_k + h \, F(t_k, u_k)\big) \right]. \end{aligned} \tag{20.100}$$

is known as the *Improved Euler Method*. It is a completely explicit scheme since there is no need to solve any equation to find the updated value u_{k+1}.

Example 20.42. For our favorite equation $\dot{u} = \left(1 - \frac{4}{3}t\right)u$, the Improved Euler Method begins with the Euler approximation

$$\widetilde{u}_{k+1} = u_k + h\left(1 - \tfrac{4}{3}t_k\right)u_k,$$

and then replaces it by the improved value

$$\begin{aligned} u_{k+1} &= u_k + \tfrac{1}{2}h\left[\left(1 - \tfrac{4}{3}t_k\right)u_k + \left(1 - \tfrac{4}{3}t_{k+1}\right)\widetilde{u}_{k+1}\right] \\ &= u_k + \tfrac{1}{2}h\left[\left(1 - \tfrac{4}{3}t_k\right)u_k + \left(1 - \tfrac{4}{3}(t_k + h)\right)\left(u_k + h\left(1 - \tfrac{4}{3}t_k\right)u_k\right)\right] \\ &= \left[\left(1 - \tfrac{2}{3}h^2\right)\left[1 + h\left(1 - \tfrac{4}{3}t_k\right)\right] + \tfrac{1}{2}h^2\left(1 - \tfrac{4}{3}t_k\right)^2\right]u_k. \end{aligned}$$

Implementing this scheme leads to the following errors in the numerical solution at the indicated times. The Improved Euler Method performs comparably to the fully implicit scheme (20.98), and significantly better than the original Euler Method.

h	$E(1)$	$E(2)$	$E(3)$
.100	−.00070230	.00097842	.00147748
.010	−.00000459	.00001068	.00001264
.001	−.00000004	.00000011	.00000012

The Improved Euler Method is the simplest of a large family of so-called *predictor–corrector algorithms*. In general, one begins a relatively crude method — in this case the Euler Method — to *predict* a first approximation $\widetilde{u}_{k+1}$ to the desired solution value u_{k+1}. One then employs a more sophisticated, typically implicit, method to *correct* the original prediction, by replacing the required update u_{k+1} on the right hand side of the implicit scheme by the less accurate prediction $\widetilde{u}_{k+1}$. The resulting explicit, corrected value u_{k+1} will, provided the method has been designed with due care, be an improved approximation to the true solution.

The numerical data in Example 20.42 indicates that the Improved Euler Method is of second order since each reduction in step size by $\frac{1}{10}$ improves the solution accuracy by, roughly, a factor of $\frac{1}{100}$. To verify this prediction, we expand the right hand side of (20.100) in a Taylor series in h, and then compare, term by term, with the solution expansion (20.90). First[†],

$$F\big(t_k + h, u_k + h\,F(t_k, u_k)\big) = F + h\big(F_t + F\,F_u\big) + \tfrac{1}{2}h^2\big(F_{tt} + 2F\,F_{tu} + F^2\,F_{uu}\big) + \cdots,$$

where all the terms involving F and its partial derivatives on the right hand side are evaluated at t_k, u_k. Substituting into (20.100), we find

$$u_{k+1} = u_k + h\,F + \tfrac{1}{2}h^2\big(F_t + F\,F_u\big) + \tfrac{1}{4}h^2\big(F_{tt} + 2F\,F_{tu} + F^2\,F_{uu}\big) + \cdots. \quad (20.101)$$

[†] We use subscripts to indicate partial derivatives to save space.

The two Taylor expansions (20.90) and (20.101) agree in their order $1, h$ and h^2 terms, but differ at order h^3. This confirms our experimental observation that the Improved Euler Method is of second order.

We can design a range of numerical solution schemes by implementing alternative numerical approximations to the basic integral equation (20.94). For example, the Midpoint Rule approximates the integral of the function $g(t)$ by the area of the rectangle whose height is the value of the function at the midpoint:

$$\int_{t_k}^{t_{k+1}} g(s)\,ds \;\approx\; h\,g\left(t_k + \tfrac{1}{2}\,h\right), \qquad \text{where} \qquad h = t_{k+1} - t_k. \tag{20.102}$$

See Figure 20.16 for an illustration. The Midpoint Rule is known to have the same order of accuracy as the Trapezoid Rule, [**9**, **34**]. Substituting into (20.94) leads to the approximation

$$u_{k+1} = u_k + \int_{t_k}^{t_{k+1}} F(s,u(s))\,ds \;\approx\; u_k + h\,F\left(t_k + \tfrac{1}{2}\,h, u\left(t_k + \tfrac{1}{2}\,h\right)\right).$$

Of course, we don't know the value of the solution $u\left(t_k + \frac{1}{2}\,h\right)$ at the midpoint, but can predict it through a straightforward adaptation of the basic Euler approximation:

$$u\left(t_k + \tfrac{1}{2}\,h\right) \;\approx\; u_k + \tfrac{1}{2}\,h\,F(t_k,u_k).$$

The result is the *Midpoint Method*

$$u_{k+1} = u_k + h\,F\left(t_k + \tfrac{1}{2}\,h, u_k + \tfrac{1}{2}\,h\,F(t_k,u_k)\right). \tag{20.103}$$

A comparison of the terms in the Taylor expansions of (20.90) and (20.103) reveals that the Midpoint Method is also of second order; see Exercise ■.

Runge–Kutta Methods

The Improved Euler and Midpoint Methods are the most elementary incarnations of a general class of numerical schemes for ordinary differential equations that were first systematically studied by the German mathematicians Carle Runge and Martin Kutta in the late nineteenth century. Runge–Kutta Methods are by far the most popular and powerful general-purpose numerical methods for integrating ordinary differential equations. While not appropriate in all possible situations, Runge–Kutta schemes are surprisingly robust, performing efficiently and accurately in a wide variety of problems. Barring significant complications, they are the method of choice in most basic applications. They comprise the engine that powers most computer software for solving general initial value problems for systems of ordinary differential equations.

The most general *Runge–Kutta Method* takes the form

$$u_{k+1} = u_k + h \sum_{i=1}^{m} c_i\,F(t_{i,k}, u_{i,k}), \tag{20.104}$$

where m counts the number of *terms* in the method. Each $t_{i,k}$ denotes a point in the k^{th} mesh interval, so $t_k \le t_{i,k} \le t_{k+1}$. The second argument $u_{i,k} \approx u(t_{i,k})$ should be viewed as

an approximation to the solution at the point $t_{i,k}$, and so is computed by a simpler Runge–Kutta scheme of the same general format. There is a lot of flexibility in the design of the method, through choosing the coefficients c_i, the times $t_{i,k}$, as well as the scheme (and all parameters therein) used to compute each of the intermediate approximations $u_{i,k}$. As always, the *order* of the method is fixed by the power of h to which the Taylor expansions of the numerical method (20.104) and the actual solution (20.90) agree. Clearly, the more terms we include in the Runge–Kutta formula (20.104), the more free parameters available to match terms in the solution's Taylor series, and so the higher the potential order of the method. Thus, the goal is to arrange the parameters so that the method has a high order of accuracy, while, simultaneously, avoiding unduly complicated, and hence computationally costly, formulae.

Both the Improved Euler and Midpoint Methods are instances of a family of two term Runge–Kutta Methods

$$\begin{aligned} u_{k+1} &= u_k + h\,\big[\,a\,F(t_k,u_k) + b\,F\big(\,t_{k,2},u_{k,2}\,\big)\,\big] \\ &= u_k + h\,\big[\,a\,F(t_k,u_k) + b\,F\big(\,t_k+\lambda\,h,u_k+\lambda\,h\,F(t_k,u_k)\,\big)\,\big], \end{aligned} \tag{20.105}$$

based on the current mesh point, so $t_{k,1} = t_k$, and one intermediate point $t_{k,2} = t_k + \lambda h$ with $0 \le \lambda \le 1$. The basic Euler Method is used to approximate the solution value

$$u_{k,2} = u_k + \lambda\,h\,F(t_k,u_k)$$

at $t_{k,2}$. The Improved Euler Method sets $a = b = \frac{1}{2}$ and $\lambda = 1$, while the Midpoint Method corresponds to $a = 0$, $b = 1$, $\lambda = \frac{1}{2}$. The range of possible values for a, b and λ is found by matching the Taylor expansion

$$\begin{aligned} u_{k+1} &= u_k + h\,\big[\,a\,F(t_k,u_k) + b\,F\big(\,t_k+\lambda\,h,u_k+\lambda\,h\,F(t_k,u_k)\,\big)\,\big] \\ &= u_k + h\,(a+b)\,F(t_k,u_k) + h^2\,b\,\lambda\,\left[\,\frac{\partial F}{\partial t}\,(t_k,u_k) + F(t_k,u_k)\,\frac{\partial F}{\partial u}\,(t_k,u_k)\right] + \,\cdots\,. \end{aligned}$$

(in powers of h) of the right hand side of (20.105) with the Taylor expansion (20.90) of the solution, namely

$$u(t_{k+1}) = u_k + h\,F(t_k,u_k) + \frac{h^2}{2}\,\left[\,F_t(t_k,u_k) + F(t_k,u_k)\,F_u(t_k,u_k)\,\right] + \,\cdots\,,$$

to as high an order as possible. First, the constant terms, u_k, are the same. For the order h and order h^2 terms to agree, we must have, respectively,

$$a + b = 1, \qquad\qquad b\,\lambda = \tfrac{1}{2}.$$

Therefore, setting

$$a = 1-\mu, \qquad b = \mu, \qquad \text{and} \qquad \lambda = \frac{1}{2\,\mu}\,, \qquad \text{where } \mu \text{ is arbitrary}^{\dagger},$$

leads to the following family of two term, second order Runge–Kutta Methods:

$$u_{k+1} = u_k + h \left[(1-\mu)\, F(t_k, u_k) + \mu\, F\left(t_k + \frac{h}{2\,\mu}, u_k + \frac{h}{2\,\mu} F(t_k, u_k) \right) \right]. \qquad (20.106)$$

The case $\mu = \frac{1}{2}$ corresponds to the Improved Euler Method (20.100), while $\mu = 1$ yields the Midpoint Method (20.103). Unfortunately, none of these methods are able to match all of the third order terms in the Taylor expansion for the solution, and so we are left with a one-parameter family of two step Runge–Kutta Methods, all of second order, that include the Improved Euler and Midpoint schemes as particular instances. The methods with $\frac{1}{2} \leq \mu \leq 1$ all perform more or less comparably, and there is no special reason to prefer one over the other.

To construct a third order Runge–Kutta Method, we need to take at least $m \geq 3$ terms in (20.104). A rather intricate computation (best done with the aid of computer algebra) will produce a range of valid schemes; the results can be found in [**88**, **107**]. The algebraic manipulations are rather tedious, and we leave a complete discussion of the available options to a more advanced treatment. In practical applications, a particularly simple fourth order, four term formula has become the most used. The method, often abbreviated as RK4, takes the form

$$u_{k+1} = u_k + \frac{h}{6} \left[F(t_k, u_k) + 2\, F(t_{2,k}, u_{2,k}) + 2\, F(t_{3,k}, u_{3,k}) + F(t_{4,k}, u_{4,k}) \right], \qquad (20.107)$$

where the times and function values are successively computed according to the following procedure:

$$\begin{aligned} t_{2,k} &= t_k + \tfrac{1}{2}\, h, & u_{2,k} &= u_k + \tfrac{1}{2}\, h\, F(t_k, u_k), \\ t_{3,k} &= t_k + \tfrac{1}{2}\, h, & u_{3,k} &= u_k + \tfrac{1}{2}\, h\, F(t_{2,k}, u_{2,k}), \\ t_{4,k} &= t_k + h, & u_{4,k} &= u_k + h\, F(t_{3,k}, u_{3,k}). \end{aligned} \qquad (20.108)$$

The four term RK4 scheme (20.107–108) is, in fact, a fourth order method. This is confirmed by demonstrating that the Taylor expansion of the right hand side of (20.107) in powers of h matches all of the terms in the Taylor series for the solution (20.90) up to and including those of order h^4, and hence the local error is of order h^5. This is not a computation for the faint-hearted — bring lots of paper and erasers, or, better yet, a good computer algebra package! The RK4 scheme is but one instance of a large family of fourth order, four term Runge–Kutta Methods, and by far the most popular owing to its relative simplicity.

Example 20.43. Application of the RK4 Method (20.107–108) to our favorite initial value problem (20.80) leads to the following errors at the indicated times:

---

† Although we should restrict $\mu \geq \frac{1}{2}$ in order that $0 \leq \lambda \leq 1$.

h	$E(1)$	$E(2)$	$E(3)$
.100	-1.944×10^{-7}	1.086×10^{-6}	4.592×10^{-6}
.010	-1.508×10^{-11}	1.093×10^{-10}	3.851×10^{-10}
.001	-1.332×10^{-15}	-4.741×10^{-14}	1.932×10^{-14}

The accuracy is phenomenally good — much better than any of our earlier numerical schemes. Each decrease in the step size by a factor of $\frac{1}{10}$ results in about 4 additional decimal digits of accuracy in the computed solution, in complete accordance with its status as a fourth order method.

Actually, it is not entirely fair to compare the accuracy of the methods using the same step size. Each iteration of the RK4 Method requires four evaluations of the function $F(t, u)$, and hence takes the same computational effort as four Euler iterations, or, equivalently, two Improved Euler iterations. Thus, the more revealing comparison would be between RK4 at step size h, Euler at step size $\frac{1}{4}\,h$, and Improved Euler at step size $\frac{1}{2}\,h$, as these involve roughly the same amount of computational effort. The resulting errors $E(1)$ at time $t = 1$ are listed in the following table.

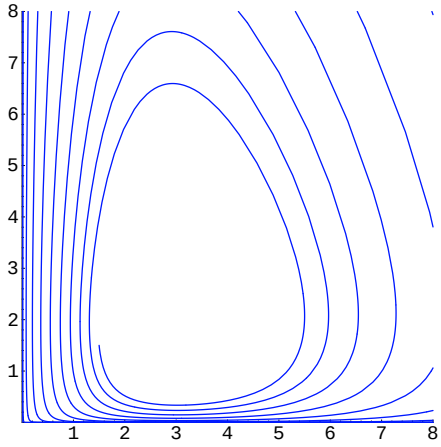
Thus, even taking computational effort into account, the Runge–Kutta Method continues to outperform its rivals. At a step size of .1, it is almost as accurate as the Improved Euler Method with step size .0005, and hence 200 times as much computation, while the Euler Method would require a step size of approximately $.24 \times 10^{-6}$, and would be $4,000,000$ times as slow as Runge–Kutta! With a step size of .001, RK4 computes a solution value that is near the limits imposed by machine accuracy (in single precision arithmetic). The superb performance level and accuracy of the RK4 Method immediately explains its popularity for a broad range of applications.

h	Euler	Improved Euler	Runge–Kutta 4
.1	1.872×10^{-2}	-1.424×10^{-4}	-1.944×10^{-7}
.01	1.874×10^{-3}	-1.112×10^{-6}	-1.508×10^{-11}
.001	1.870×10^{-4}	-1.080×10^{-8}	-1.332×10^{-15}

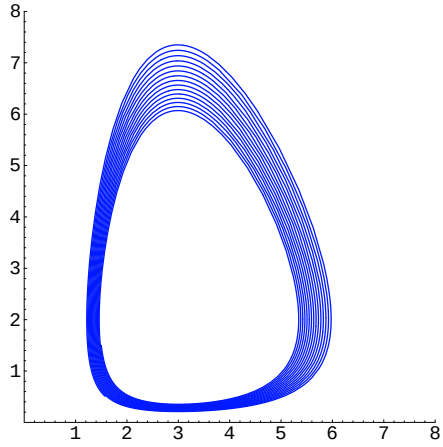
Example 20.44. As noted earlier, by writing the function values as vectors $\mathbf{u}_k \approx \mathbf{u}(t_k)$, one can immediately use all of the preceding methods to integrate initial value problems for first order systems of ordinary differential equations $\dot{\mathbf{u}} = \mathbf{F}(\mathbf{u})$. Consider, by way of example, the Lotka–Volterra system

$$\frac{du}{dt} = 2\,u - u\,v, \qquad \frac{dv}{dt} = -9\,v + 3\,u\,v, \tag{20.109}$$

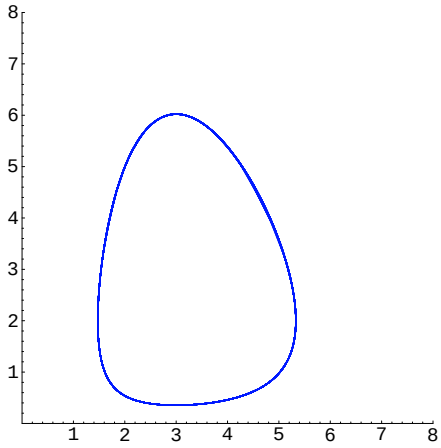
analyzed in Example 20.29. To find a numerical solution, we write $\mathbf{u} = (\,u, v\,)^T$ for the



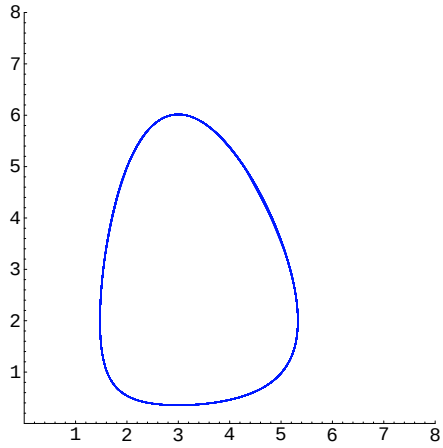
Euler Method, $h = .01$



Euler Method, $h = .001$



Improved Euler Method, $h = .01$



RK4 Method, $h = .01$

Figure 20.17. Numerical Solutions of Predator–Prey Model.

solution vector, while $\mathbf{F}(\mathbf{u}) = (\,2u - uv, -9v + 3uv\,)^T$ is the right hand side of the system. The Euler Method with step size h is given by

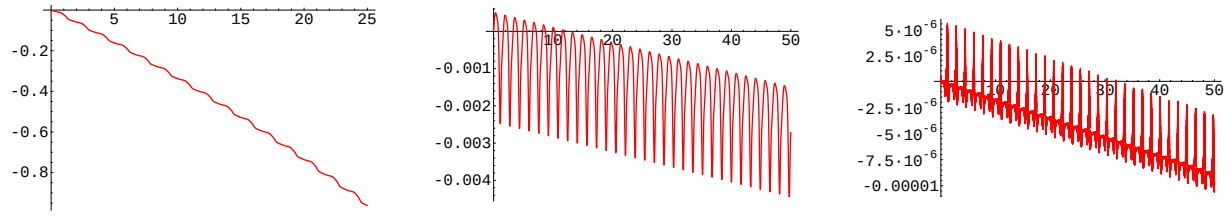
$$\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + h\,\mathbf{F}(\mathbf{u}^{(k)}),$$

or, explicitly, as a first order nonlinear iterative system

$$u^{(k+1)} = u^{(k)} + h\,(2u^{(k)} - u^{(k)}\,v^{(k)}), \qquad v^{(k+1)} = v^{(k)} + h\,(-9v^{(k)} + 3u^{(k)}\,v^{(k)}).$$

The Improved Euler and Runge–Kutta schemes are implemented in a similar fashion. Phase portraits of the three numerical algorithms starting with initial conditions $u^{(0)} = v^{(0)} = 1.5$, and up to time $t = 25$ in the case of the Euler Method, and $t = 50$ for the other two, appear in Figure 20.17. Recall that the solution is supposed to travel periodically around a closed curve, which is the level set

$$I(u, v) = 9\,\log u - 3u + 2\,\log v - v = I(1.5, 1.5) = -1.53988$$



Euler Method
$h = .001$

Improved Euler Method
$h = .01$

RK4 Method
$h = .01$

Figure 20.18. Numerical Evaluation of Lotka–Volterra First Integral.

of the first integral. The Euler Method spirals away from the exact periodic solution, whereas the Improved Euler and RK4 Methods perform rather well. Since we do not have an analytic formula for the solution, we are not able to measure the error exactly. However, the known first integral is supposed to remain constant on the solution trajectories, and so one means of monitoring the accuracy of the solution is to track the variation in the numerical values of $I(u^{(k)}, v^{(k)})$. These are graphed in Figure 20.18; the Improved Euler keeps the value within .0005, while in the RK4 solution, the first integral only experiences change in its the fifth decimal place over the indicated time period. Of course, the ononger one continues to integrate, the more error will gradually creep into the numerical solution. Still, for most practical purposes, the RK4 solution is indistinguishable from the exact solution.

In practical implementations, it is important to monitor the accuracy of the numerical solution, so to gauge when to abandon an insufficiently precise computation. Since accuracy is dependent upon the step size h, one may try adjusting h so as stay within a preassigned error. *Adaptive methods*, allow one to change the step size during the course of the computation, in response to some estimation of the overall error. Insufficiently accurate numerical solutions would necessitate a suitable reduction in step size (or increase in the order of the scheme). On the other hand, if the solution is more accurate than the application requires, one could increase the step size so as to reduce the total amount of computational effort.

How might one decide when a method is giving inaccurate results, since one presumably does not know the true solution and so has nothing to directly test the numerical approximation against? One useful idea is to integrate the differential equation using two different numerical schemes, usually of different orders of accuracy, and then compare the results. If the two solution values are reasonably close, then one is usually safe in assuming that the methods are both giving accurate results, while in the event that they differ beyond some preassigned tolerance, then one needs to re-evaluate the step size. The required adjustment to the step size relies on a more detailed analysis of the error terms. Several well-studied methods are employed in practical situations; the most popular is the Runge–Kutta–Fehlberg Method, which combines a fourth and a fifth order Runge–Kutta scheme for error control. Details can be found in more advanced treatments of the subject, e.g., [**88**, **107**].

Stiff Differential Equations

While the fourth order Runge–Kutta Method with a sufficiently small step size will successfully integrate a broad range of differential equations — at least over not unduly long time intervals — it does occasionally experience unexpected difficulties. While we have not developed sufficiently sophisticated analytical tools to conduct a thorough analysis, it will be instructive to look at why a breakdown might occur in a simpler context.

Example 20.45. The elementary linear initial value problem

$$\frac{du}{dt} = -250\,u, \qquad u(0) = 1, \tag{20.110}$$

is an instructive and sobering example. The explicit solution is easily found; it is a very rapidly decreasing exponential: $u(t) = e^{-250\,t}$.

$$u(t) = e^{-250\,t} \qquad \text{with} \qquad u(1) \approx 2.69 \times 10^{-109}.$$

The following table gives the result of approximating the solution value $u(1) \approx 2.69 \times 10^{-109}$ at time $t = 1$ using three of our numerical integration schemes for various step sizes:

h	Euler	Improved Euler	RK4
.1	6.34×10^{13}	3.99×10^{24}	2.81×10^{41}
.01	4.07×10^{17}	1.22×10^{21}	1.53×10^{-19}
.001	1.15×10^{-125}	6.17×10^{-108}	2.69×10^{-109}

The results are not misprints! When the step size is .1, the computed solution values are perplexingly large, and appear to represent an exponentially growing solution — the complete opposite of the rapidly decaying true solution. Reducing the step size beyond a critical threshold suddenly transforms the numerical solution to an exponentially decaying function. Only the fourth order RK4 Method with step size $h = .001$ — and hence a total of $1,000$ steps — does a reasonable job at approximating the correct value of the solution at $t = 1$.

You may well ask, what on earth is going on? The solution couldn't be simpler — why is it so difficult to compute it? To understand the basic issue, let us analyze how the Euler Method handles such simple differential equations. Consider the initial value problem

$$\frac{du}{dt} = \lambda\,u, \qquad u(0) = 1, \tag{20.111}$$

which has an exponential solution

$$u(t) = e^{\lambda\,t}. \tag{20.112}$$

As in Example 20.36, the Euler Method with step size h relies on the iterative scheme

$$u_{k+1} = (1 + \lambda h)\,u_k, \qquad u_0 = 1,$$

with solution

$$u_k = (1 + \lambda h)^k. \tag{20.113}$$

If $\lambda > 0$, the exact solution (20.112) is exponentially growing. Since $1 + \lambda h > 1$, the numerical iterates are also growing, albeit at a somewhat slower rate. In this case, there is no inherent surprise with the numerical approximation procedure — in the short run it gives fairly accurate results, but eventually trails behind the exponentially growing solution.

On the other hand, if $\lambda < 0$, then the exact solution is exponentially decaying and positive. But now, if $\lambda h < -2$, then $1 + \lambda h < -1$, and the iterates (20.113) grow exponentially fast in magnitude, with alternating signs. In this case, the numerical solution is nowhere close to the true solution; this explains the previously observed pathological behavior. If $-1 < 1 + \lambda h < 0$, the numerical solutions decay in magnitude, but continue to alternate between positive and negative values. Thus, to correctly model the qualitative features of the solution and obtain a numerically respectable approximation, we need to choose the step size h so as to ensure that $0 < 1 + \lambda h$, and hence $h < -1/\lambda$ when $\lambda < 0$. For the value $\lambda = -250$ in the example, then, we must choose $h < \frac{1}{250} = .004$ in order that the Euler Method give a reasonable numerical answer. A similar, but more complicated analysis applies to any of the Runge–Kutta schemes.

Thus, the numerical methods for ordinary differential equations exhibit a form of conditional stability, cf. Section 14.6. Paradoxically, the larger negative λ is — and hence the faster the solution tends to a trivial zero equilibrium — the *more* difficult and expensive the numerical integration.

The system (20.110) is the simplest example of what is known as a *stiff differential equation*. In general, an equation or system is stiff if it has one or more very rapidly decaying solutions. In the case of autonomous (constant coefficient) linear systems $\dot{\mathbf{u}} = A\,\mathbf{u}$, stiffness occurs whenever the coefficient matrix A has an eigenvalue with a large negative real part: $\operatorname{Re} \lambda \ll 0$, resulting in a very rapidly decaying eigensolution. It only takes one such eigensolution to render the equation stiff, and ruin the numerical computation of even the well behaved solutions! Curiously, the component of the actual solution corresponding to such large negative eigenvalues is almost irrelevant, as it becomes almost instanteously tiny. However, the presence of such an eigenvalue continues to render the numerical solution to the system very difficult, even to the point of exhausting any available computing resources. Stiff equations require more sophisticated numerical procedures to integrate, and we refer the reader to [**88**, **107**] for details.

Most of the other methods derived above also suffer from instability due to stiffness of the ordinary differential equation for sufficiently large negative λ. Interestingly, stability for solving the trivial test scalar ordinary differential equation (20.111) suffices to characterize acceptable step sizes h, depending on the size of λ, which, in the case of systems, becomes the eigenvalue. The analysis is not so difficult, owing to the innate simplicity of the test ordinary differential equation (20.111). A significant exception, which also illustrates the test for behavior under rapidly decaying solutions, is the Trapezoid Method (20.97). Let us analyze the behavio of the resulting numerical solution to (20.111). Substituting

$f(t,u) = \lambda\, u$ into the Trapezoid iterative equation (20.97), we find

$$u_{k+1} = u_k + \tfrac{1}{2}\, h\, \big[\, \lambda\, u_k + \lambda\, u_{k+1}\,\big],$$

which we solve for

$$u_{k+1} = \frac{1 + \frac{1}{2}\, h\, \lambda}{1 - \frac{1}{2}\, h\, \lambda}\, u_k \equiv \mu\, u_k.$$

Thus, the behavior of the solution is entirely determined by the size of the coefficient μ. If $\lambda > 0$, then $\mu > 1$ and the numerical solution is exponentially growing, as long as the denominator is positive, which requires $h < 2/\lambda$ to be sufficiently small. In other words, rapidly growing exponential solutions require reasonably small step sizes to accurately compute, which is not surprising. On the other hand, if $\lambda < 0$, then $|\,\mu\,| < 1$, *no matter how large negative* λ *gets*! (But we should also restrict $h < -2/\lambda$ to be sufficiently small, as otherwise $\mu < 0$ and the numerical solution would have oscillating signs, even though it is decaying, and hence vanishing small. If this were part of a larger system, such minor oscillations would not worry us because they would be unnoticeable in the long run.) Thus, the Trapezoid Method is *not* affected by very large negative exponents, and hence not subject to the effects of stiffness. The Trapezoid Method is the simplest example of an A stable method. More precisely, a numerical solution method is called A *stable* if the zero solution is asymptotically stable for the iterative equation resulting from the numerical solution to the ordinary differential equation $\dot u = \lambda\, u$ for all $\lambda < 0$. The big advantage of A stable methods is that they are not affected by stiffness. Unfortunately, A stable methods are few and far between. In fact, they are all implicit one-step methods! *No explicit Runge–Kutta Method is A stable*; see [**107**] for a proof of this disappointing result. Moreover, multistep method, as discussed in Exercise ■, also suffer from the lack of A stability and so are all prone to the effects of stiffnness. Still, when confronted with a very stiff equation, one must discard the sophisticated Runge–Kutta and multi-step methods are resort to a low order, but A stable scheme like the Trapezoid Method.

Chapter 21

The Calculus of Variations

We have already had ample opportunity to exploit Nature's propensity to minimize. Minimization principles form one of the most wide-ranging means of formulating mathematical models governing the equilibrium configurations of physical systems. Moreover, many popular numerical integration schemes such as the powerful finite element method are also founded upon a minimization paradigm. In this chapter, we will develop the basic mathematical analysis of nonlinear minimization principles on infinite-dimensional function spaces — a subject known as the "calculus of variations", for reasons that will be explained as soon as we present the basic ideas. Classical solutions to minimization problems in the calculus of variations are prescribed by boundary value problems involving certain types of differential equations, known as the associated Euler–Lagrange equations. The mathematical techniques that have been developed to handle such optimization problems are fundamental in many areas of mathematics, physics, engineering, and other applications. In this chapter, we will only have room to scratch the surface of this vast and lively area of classical and contemporary research.

The history of the calculus of variations is tightly interwoven with the history of mathematics. The field has drawn the attention of a remarkable range of mathematical luminaries, beginning with Newton, then initiated as a subject in its own right by the Bernoulli family. The first major developments appeared in the work of Euler, Lagrange and Laplace. In the nineteenth century, Hamilton, Dirichlet and Hilbert are but a few of the outstanding contributors. In modern times, the calculus of variations has continued to occupy center stage, witnessing major theoretical advances, along with wide-ranging applications in physics, engineering and all branches of mathematics.

Minimization problems amenable to the methods of the calculus of variations serve to characterize the equilibrium configurations of almost all continuous physical systems, ranging through elasticity, solid and fluid mechanics, electro-magnetism, gravitation, quantum mechanics, string theory, and many, many others. Many geometrical configurations, such as minimal surfaces, can be conveniently formulated as optimization problems. Moreover, numerical approximations to the equilibrium solutions of such boundary value problems are based on a nonlinear finite element approach that reduced the infinite-dimensional minimization problem to a finite-dimensional problem, to which we can apply the optimization techniques learned in Section 19.3; however, we will not pursue this direction here.

We have, in fact, already treated the simplest problems in the calculus of variations. As we learned in Chapters 11 and 15, minimization of a quadratic functional requires solving an associated linear boundary value problem. Just as the vanishing of the gradient of a function of several variables singles out the critical points, among which are the minima,

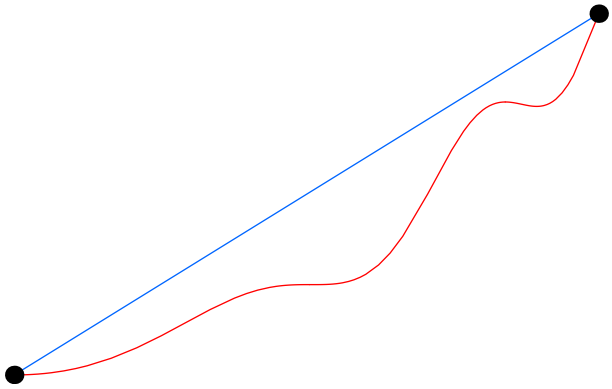


Figure 21.1. The Shortest Path is a Straight Line.

both local and global, so a similar "functional gradient" will distinguish the candidate functions that might be minimizers of the functional. The finite-dimensional calculus leads to a system of algebraic equations for the critical points; the infinite-dimensional functional analog results a boundary value problem for a nonlinear ordinary or partial differential equation whose solutions are the critical functions for the variational problem. So, the passage from finite to infinite dimensional nonlinear systems mirrors the transition from linear algebraic systems to boundary value problems.

21.1. Examples of Variational Problems.

The best way to appreciate the calculus of variations is by introducing a few concrete examples of both mathematical and practical importance. Some of these minimization problems played a key role in the historical development of the subject. And they still serve as an excellent means of learning its basic constructions.

Minimal Curves, Optics, and Geodesics

The *minimal curve problem* is to find the shortest path between two specified locations. In its simplest manifestation, we are given two distinct points

$$\mathbf{a} = (a, \alpha) \qquad \text{and} \qquad \mathbf{b} = (b, \beta) \qquad \text{in the plane} \quad \mathbb{R}^2, \tag{21.1}$$

and our task is to find the curve of shortest length connecting them. "Obviously", as you learn in childhood, the shortest route between two points is a straight line; see Figure 21.1. Mathematically, then, the minimizing curve should be the graph of the particular affine function[†]

$$y = c\,x + d = \frac{\beta - \alpha}{b - a}\,(x - a) + \alpha \tag{21.2}$$

that passes through or interpolates the two points. However, this commonly accepted "fact" — that (21.2) is the solution to the minimization problem — is, upon closer inspection, perhaps not so immediately obvious from a rigorous mathematical standpoint.

[†] We assume that $a \neq b$, i.e., the points $\mathbf{a}, \mathbf{b}$ do not lie on a common vertical line.

Let us see how we might formulate the minimal curve problem in a mathematically precise way. For simplicity, we assume that the minimal curve is given as the graph of a smooth function $y = u(x)$. Then, according to (A.27), the length of the curve is given by the standard arc length integral

$$J[u] = \int_a^b \sqrt{1 + u'(x)^2}\, dx, \tag{21.3}$$

where we abbreviate $u' = du/dx$. The function $u(x)$ is required to satisfy the boundary conditions

$$u(a) = \alpha, \qquad u(b) = \beta, \tag{21.4}$$

in order that its graph pass through the two prescribed points (21.1). The minimal curve problem asks us to find the function $y = u(x)$ that minimizes the arc length functional (21.3) among all "reasonable" functions satisfying the prescribed boundary conditions. The reader might pause to meditate on whether it is analytically obvious that the affine function (21.2) is the one that minimizes the arc length integral (21.3) subject to the given boundary conditions. One of the motivating tasks of the calculus of variations, then, is to rigorously prove that our everyday intuition is indeed correct.

Indeed, the word "reasonable" *is* important. For the arc length functional (21.3) to be defined, the function $u(x)$ should be at least piecewise C^1, i.e., continuous with a piecewise continuous derivative. Indeed, if we were to allow discontinuous functions, then the straight line (21.2) does not, in most cases, give the minimizer; see Exercise ■. Moreover, continuous functions which are not piecewise C^1 need not have a well-defined arc length. The more seriously one thinks about these issues, the less evident the "obvious" solution becomes. But before you get too worried, rest assured that the straight line (21.2) is indeed the true minimizer. However, a fully rigorous proof of this fact requires a careful development of the mathematical machinery of the calculus of variations.

A closely related problem arises in geometrical optics. The underlying physical principle, first formulated by the seventeenth century French mathematician Pierre de Fermat, is that, when a light ray moves through an optical medium, it travels along a path that minimizes the travel time. As always, Nature seeks the most economical[†] solution. In an inhomogeneous planar[‡] optical medium, the speed of light, $c(x, y)$, varies from point to point, depending on the optical properties. Speed equals the time derivative of distance traveled, namely, the arc length of the curve $y = u(x)$ traced by the light ray. Thus,

$$c(x, u(x)) = \frac{ds}{dt} = \sqrt{1 + u'(x)^2}\ \frac{dx}{dt}\ .$$

Integrating from start to finish, we conclude that the total travel time along the curve is equal to

$$T[u] = \int_0^T dt = \int_a^b \frac{dt}{dx}\, dx = \int_a^b \frac{\sqrt{1 + u'(x)^2}}{c(x, u(x))}\, dx. \tag{21.5}$$

[†] Assuming time = money!

[‡] For simplicity, we only treat the two-dimensional case in the text. See Exercise ■ for the three-dimensional version.

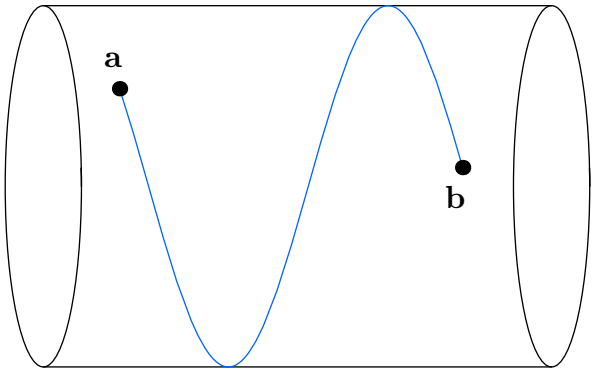


Figure 21.2. Geodesics on a Cylinder.

Fermat's Principle states that, to get from point $\mathbf{a} = (a.\alpha)$ to point $\mathbf{b} = (b, \beta)$, the light ray follows the curve $y = u(x)$ that minimizes this functional subject to the boundary conditions

$$u(a) = \alpha, \qquad u(b) = \beta,$$

If the medium is homogeneous, e.g., a vacuum[†], then $c(x, y) \equiv c$ is constant, and $T[u]$ is a multiple of the arc length functional (21.3), whose minimizers are the "obvious" straight lines traced by the light rays. In an inhomogeneous medium, the path taken by the light ray is no longer evident, and we are in need of a systematic method for solving the minimization problem. Indeed, all of the known laws of geometric optics, lens design, focusing, refraction, aberrations, etc., will be consequences of the geometric and analytic properties of solutions to Fermat's minimization principle, [**optics**].

Another minimization problem of a similar ilk is to construct the *geodesics* on a curved surface, meaning the curves of minimal length. Given two points $\mathbf{a}, \mathbf{b}$ lying on a surface $S \subset \mathbb{R}^3$, we seek the curve $C \subset S$ that joins them and has the minimal possible length. For example, if S is a circular cylinder, then there are three possible types of geodesic curves: straight line segments parallel to the center line; arcs of circles orthogonal to the center line; and spiral helices, the latter illustrated in Figure 21.2. Similarly, the geodesics on a sphere are arcs of great circles. In aeronautics, to minimize distance flown, airplanes follow geodesic circumpolar paths around the globe. However, both of these claims are in need of mathematical justification.

In order to mathematically formulate the geodesic minimization problem, we suppose, for simplicity, that our surface $S \subset \mathbb{R}^3$ is realized as the graph[‡] of a function $z = F(x, y)$. We seek the geodesic curve $C \subset S$ that joins the given points

$$\mathbf{a} = (a, \alpha, F(a, \alpha)), \qquad \text{and} \qquad \mathbf{b} = (b, \beta, F(b, \beta)), \quad \text{lying on the surface} \quad S.$$

[†] In the absence of gravitational effects due to general relativity.

[‡] Cylinders are not graphs, but can be placed within this framework by passing to cylindrical coordinates. Similarly, spherical surfaces are best treated in spherical coordinates. In differential geometry, [**59**], one extends these constructions to arbitrary parametrized surfaces and higher dimensional manifolds.

Let us assume that C can be parametrized by the x coordinate, in the form

$$y = u(x), \qquad z = v(x) = F(x, u(x)),$$

where the last equation ensures that it lies in the surface S. In particular, this requires $a \neq b$. The length of the curve is supplied by the standard three-dimensional arc length integral (B.17). Thus, to find the geodesics, we must minimize the functional

$$\begin{aligned} J[u] &= \int_a^b \sqrt{1 + \left(\frac{dy}{dx}\right)^2 + \left(\frac{dz}{dx}\right)^2}\, dx \\ &= \int_a^b \sqrt{1 + \left(\frac{du}{dx}\right)^2 + \left(\frac{\partial F}{\partial x}\,(x, u(x)) + \frac{\partial F}{\partial u}\,(x, u(x))\,\frac{du}{dx}\right)^2}\; dx, \end{aligned} \tag{21.6}$$

subject to the boundary conditions $u(a) = \alpha$, $u(b) = \beta$. For example, geodesics on the paraboloid

$$z = \tfrac{1}{2}\,x^2 + \tfrac{1}{2}\,y^2 \tag{21.7}$$

can be found by minimizing the functional

$$J[u] = \int_a^b \sqrt{1 + (u')^2 + (x + u\,u')^2}\, dx. \tag{21.8}$$

Minimal Surfaces

The minimal surface problem is a natural generalization of the minimal curve or geodesic problem. In its simplest manifestation, we are given a simple closed curve $C \subset \mathbb{R}^3$. The problem is to find the surface of least total area among all those whose boundary is the curve C. Thus, we seek to minimize the surface area integral

$$\text{area } S = \iint_S dS$$

over all possible surfaces $S \subset \mathbb{R}^3$ with the prescribed boundary curve $\partial S = C$. Such an area–minimizing surface is known as a *minimal surface* for short. For example, if C is a closed plane curve, e.g., a circle, then the minimal surface will just be the planar region it encloses. But, if the curve C twists into the third dimension, then the shape of the minimizing surface is by no means evident.

Physically, if we bend a wire in the shape of the curve C and then dip it into soapy water, the surface tension forces in the resulting soap film will cause it to minimize surface area, and hence be a minimal surface†. Soap films and bubbles have been the source of much fascination, physical, æsthetical and mathematical, over the centuries. The minimal surface problem is also known as *Plateau's Problem*, named after the nineteenth century

† More accurately, the soap film will realize a local but not necessarily global minimum for the surface area functional. Nonuniqueness of local minimizers can be realized in the physical experiment — the same wire may support more than one stable soap film.

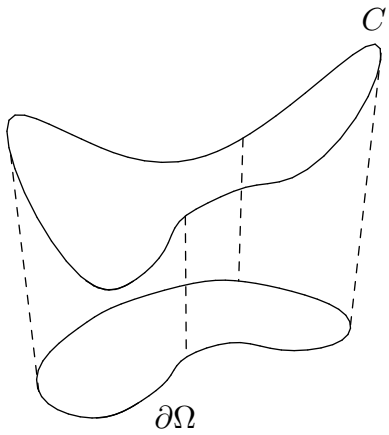


Figure 21.3. Minimal Surface.

French physicist Joseph Plateau who conducted systematic experiments on such soap films. A satisfactory mathematical solution to even the simplest version of the minimal surface problem was only achieved in the mid twentieth century, [**132**, **136**]. Minimal surfaces and related variational problems remain an active area of contemporary research, and are of importance in engineering design, architecture, and biology, including foams, domes, cell membranes, and so on.

Let us mathematically formulate the search for a minimal surface as a problem in the calculus of variations. For simplicity, we shall assume that the bounding curve C projects down to a simple closed curve $\Gamma = \partial\Omega$ that bounds an open domain $\Omega \subset \mathbb{R}^2$ in the (x, y) plane, as in Figure 21.3. The space curve $C \subset \mathbb{R}^3$ is then given by $z = g(x, y)$ for $(x, y) \in \Gamma = \partial\Omega$. For "reasonable" boundary curves C, we expect that the minimal surface S will be described as the graph of a function $z = u(x, y)$ parametrized by $(x, y) \in \Omega$. According to the basic calculus formula (B.39), the surface area of such a graph is given by the double integral

$$J[u] = \iint_\Omega \sqrt{1 + \left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2}\, dx\, dy. \tag{21.9}$$

To find the minimal surface, then, we seek the function $z = u(x, y)$ that minimizes the surface area integral (21.9) when subject to the Dirichlet boundary conditions

$$u(x, y) = g(x, y) \qquad \text{for} \qquad (x, y) \in \partial\Omega. \tag{21.10}$$

As we will see, (21.57), the solutions to this minimization problem satisfy a complicated nonlinear second order partial differential equation.

A simple version of the minimal surface problem, that still contains some interesting features, is to find minimal surfaces with rotational symmetry. A *surface of revolution* is obtained by revolving a plane curve about an axis, which, for definiteness, we take to be the x axis. Thus, given two points $\mathbf{a} = (a, \alpha)$, $\mathbf{b} = (b, \beta) \in \mathbb{R}^2$, the goal is to find the curve $y = u(x)$ joining them such that the surface of revolution obtained by revolving the curve around the x-axis has the least surface area. Each cross-section of the resulting surface is a circle centered on the x axis; see Figure srev■. According to Exercise ■, the area of such

a surface of revolution is given by

$$J[u] = \int_a^b 2\pi\,|\,u\,|\,\sqrt{1+(u')^2}\;dx. \tag{21.11}$$

We seek a minimizer of this integral among all functions $u(x)$ that satisfy the fixed boundary conditions $u(a) = \alpha$, $u(b) = \beta$. The minimal surface of revolution can be physically realized by stretching a soap film between two circular wires, of respective radius α and β, that are held a distance $b-a$ apart. Symmetry considerations will require the minimizing surface to be rotationally symmetric. Interestingly, the revolutionary surface area functional (21.11) is exactly the same as the optical functional (21.5) when the light speed at a point is inversely proportional to its distance from the horizontal axis: $c(x,y) = 1/(2\,\pi\,|\,y\,|)$.

Isoperimetric Problems and Constraints

The simplest *isoperimetric problem* is to construct the simple closed plane curve of a fixed length ℓ that encloses the domain of largest area. In other words, we seek to maximize

$$\text{area}\,\Omega = \iint_\Omega \qquad \text{subject to the constraint} \qquad \text{length}\,\partial\Omega = \oint_{\partial\Omega} = \ell,$$

over all possible domains $\Omega \subset \mathbb{R}^2$. Of course, the "obvious" solution to this problem is that the curve must be a circle whose perimeter is ℓ, whence the name "isoperimetric". Note that the problem, as stated, does not have a unique solution, since if Ω is a maximing domain, any translated or rotated version of Ω will also maximize area subject to the length constraint.

To make progress on the isoperimetric problem, let us assume that the boundary curve is parametrized by its arc length, so $\mathbf{x}(s) = (\,x(s), y(s)\,)^T$ with $0 \le s \le \ell$, subject to the requirement that

$$\left(\frac{dx}{ds}\right)^2 + \left(\frac{dy}{ds}\right)^2 = 1. \tag{21.12}$$

According to (A.56), we can compute the area of the domain by a line integral around its boundary,

$$\text{area}\,\Omega = \oint_{\partial\Omega} x\,dy = \int_0^\ell x\,\frac{dy}{ds}\,ds, \tag{21.13}$$

and thus we seek to maximize the latter integral subject to the arc length constraint (21.12). We also impose periodic boundary conditions

$$x(0) = x(\ell), \qquad\qquad y(0) = y(\ell), \tag{21.14}$$

that guarantee that the curve $\mathbf{x}(s)$ closes up. (Technically, we should also make sure that $\mathbf{x}(s) \neq \mathbf{x}(s')$ for any $0 \le s < s' < \ell$, ensuring that the curve does not cross itself.)

A simpler isoperimetric problem, but one with a less evident solution, is the following. Among all curves of length ℓ in the upper half plane that connect two points $(-a, 0)$ and $(a, 0)$, find the one that, along with the interval $[-a, a]$, encloses the region having the largest area. Of course, we must take $\ell \ge 2a$, as otherwise the curve will be too short

to connect the points. In this case, we assume the curve is represented by the graph of a non-ngebtaive function $y = u(x)$, and we seek to maximize the functional

$$\int_{-a}^{a} u\,dx \qquad \text{subject to the constraint} \qquad \int_{-a}^{a} \sqrt{1+u'^2}\,dx = \ell. \tag{21.15}$$

In the previous formulation (21.12), the arc length constraint was imposed at every point, whereas here it is manifested as an integral constraint. Both types of contraints, pointwise and integral, appear in a wide range of applied and geometrical problems. Such constrained variational problems can profitably be viewed as function space versions of constrained optimization problems. Thus, not surprisingly, their analytical solution will require the introduction of suitable Lagrange multipliers.

21.2. The Euler–Lagrange Equation.

Even the preceding limited collection of examples of variational problems should already convince the reader of the tremendous practical utility of the calculus of variations. Let us now discuss the most basic analytical techniques for solving such minimization problems. We will exclusively deal with classical techniques, leaving more modern direct methods — the function space equivalent of gradient descent and related methods — to a more in–depth treatment of the subject, [**cvar**].

Let us concentrate on the simplest class of variational problems, in which the unknown is a continuously differentiable scalar function, and the functional to be minimized depends upon at most its first derivative. The basic minimization problem, then, is to determine a suitable function $y = u(x) \in \mathrm{C}^1[\,a,b\,]$ that minimizes the *objective functional*

$$J[\,u\,] = \int_a^b L(x,u,u')\,dx. \tag{21.16}$$

The integrand is known as the *Lagrangian* for the variational problem, in honor of Lagrange, one of the main founders of the subject. We usually assume that the Lagrangian $L(x,u,p)$ is a reasonably smooth function of all three of its (scalar) arguments x, u, and p, which represents the derivative u'. For example, the arc length functional (21.3) has Lagrangian function $L(x,u,p) = \sqrt{1+p^2}$, whereas in the surface of revolution problem (21.11), we have $L(x,u,p) = 2\,\pi\,|\,u\,|\sqrt{1+p^2}$. (In the latter case, the points where $u = 0$ are slightly problematic, since L is not continuously differentiable there.)

In order to uniquely specify a minimizing function, we must impose suitable boundary conditions. All of the usual suspects — Dirichlet (fixed), Neumann (free), as well as mixed and periodic boundary conditions — that arose in Chapter 11 are also relevant here. In the interests of brevity, we shall concentrate on the Dirichlet boundary conditions

$$u(a) = \alpha, \qquad u(b) = \beta, \tag{21.17}$$

although some of the exercises will investigate other types of boundary conditions.

The First Variation

According to Section 19.3, the (local) minimizers of a (sufficiently nice) objective function defined on a finite-dimensional vector space are initially characterized as critical

points, where the objective function's gradient vanishes. An analogous construction applies in the infinite-dimensional context treated by the calculus of variations. Every minimizer of a sufficiently nice functional $J[u]$ is a "critical function", meaning that its functional gradient vanishes: $\nabla J[u] = 0$. Indeed, the mathematical justification of this fact outlined in Section 19.3 continues to apply here; see, in particular, the proof of Theorem 19.39. Of course, not every critical point turns out to be a minimum — maxima, saddles, and many degenerate points are also critical. The characterization of In nondegenerate critical points as local minima or maxima relies on the second derivative test, whose functional version, known as the second variation, will be is the topic of the following Section 21.3.

But we are getting ahead of ourselves. The first order of business is to learn how to compute the gradient of a functional defined on an infinite-dimensional function space. As noted in the general Definition 19.36 of the gradient, we must first impose an inner product on the underlying space. The gradient $\nabla J[u]$ of the functional (21.16) will be defined by the same basic directional derivative formula:

$$\langle \nabla J[u], v \rangle = \frac{d}{dt} J[u + t\, v] \Big|_{t=0}. \tag{21.18}$$

Here $v(x)$ is a function that prescribes the "direction" in which the derivative is computed. Classically, v is known as a *variation* in the function u, sometimes written $v = \delta u$, whence the term "calculus of variations". Similarly, the gradient operator on functionals is often referred to as the *variational derivative*, and often written δJ. The inner product used in (21.18) is usually taken (again for simplicity) to be the standard L^2 inner product

$$\langle f, g \rangle = \int_a^b f(x)\, g(x)\, dx \tag{21.19}$$

on function space. Indeed, while the formula for the graident will depend upon the underlying inner product, cf. Exercise ■, the characterization of critical points does not, and so the choice of inner product is primarily dictated by simplicity.

Now, starting with (21.16), for each fixed u and v, we must compute the derivative of the function

$$h(t) = J[u + t\, v] = \int_a^b L(x, u + t\, v, u' + t\, v')\, dx. \tag{21.20}$$

Assuming sufficient smoothness of the integrand allows us to bring the derivative inside the integral and so, by the chain rule,

$$\begin{aligned} h'(t) = \frac{d}{dt} J[u + t\, v] &= \int_a^b \frac{d}{dt} L(x, u + t\, v, u' + t\, v')\, dx \\ &= \int_a^b \left[v\, \frac{\partial L}{\partial u}(x, u + t\, v, u' + t\, v') + v'\, \frac{\partial L}{\partial p}(x, u + t\, v, u' + t\, v') \right] dx. \end{aligned}$$

Therefore, setting $t = 0$ in order to evaluate (21.18), we find

$$\langle \nabla J[u], v \rangle = \int_a^b \left[v\, \frac{\partial L}{\partial u}(x, u, u') + v'\, \frac{\partial L}{\partial p}(x, u, u') \right] dx. \tag{21.21}$$

The resulting integral often referred to as the *first variation* of the functional $J[\,u\,]$. The condition

$$\langle\, \nabla J[\,u\,]\,, v\,\rangle = 0$$

for a minimizer is known as the *weak form* of the variational principle.

To obtain an explicit formula for $\nabla J[\,u\,]$, the right hand side of (21.21) needs to be written as an inner product,

$$\langle\, \nabla J[\,u\,]\,, v\,\rangle = \int_a^b \nabla J[\,u\,]\, v\; dx = \int_a^b h\, v\, dx$$

between some function $h(x) = \nabla J[\,u\,]$ and the variation v. The first summand has this form, but the derivative v' appearing in the second summand is problematic. However, as the reader of Chapter 11 already knows, the secret to moving around derivatives inside an integral is integration by parts. If we set

$$r(x) \equiv \frac{\partial L}{\partial p}\,(x, u(x), u'(x)),$$

we can rewrite the offending term as

$$\int_a^b r(x)\, v'(x)\, dx = \big[\, r(b)\, v(b) - r(a)\, v(a)\,\big] - \int_a^b r'(x)\, v(x)\, dx, \tag{21.22}$$

where, again by the chain rule,

$$r'(x) = \frac{d}{dx}\left(\frac{\partial L}{\partial p}\,(x, u, u')\right) = \frac{\partial^2 L}{\partial x\, \partial p}\,(x, u, u') + u'\,\frac{\partial^2 L}{\partial u\, \partial p}\,(x, u, u') + u''\,\frac{\partial^2 L}{\partial p^2}\,(x, u, u')\,. \tag{21.23}$$

So far we have not imposed any conditions on our variation $v(x)$. We are only comparing the values of $J[\,u\,]$ among functions that satisfy the prescribed boundary conditions, namely

$$u(a) = \alpha, \qquad\qquad u(b) = \beta.$$

Therefore, we must make sure that the varied function $\widehat{u}(x) = u(x) + t\, v(x)$ remains within this set of functions, and so

$$\widehat{u}(a) = u(a) + t\, v(a) = \alpha, \qquad\qquad \widehat{u}(b) = u(b) + t\, v(b) = \beta.$$

For this to hold, the variation $v(x)$ must satisfy the corresponding homogeneous boundary conditions

$$v(a) = 0, \qquad\qquad v(b) = 0. \tag{21.24}$$

As a result, both boundary terms in our integration by parts formula (21.22) vanish, and we can write (21.21) as

$$\langle\, \nabla J[\,u\,]\,, v\,\rangle = \int_a^b \nabla J[\,u\,]\, v\; dx = \int_a^b v\,\left[\, \frac{\partial L}{\partial u}\,(x, u, u') - \frac{d}{dx}\left(\frac{\partial L}{\partial p}\,(x, u, u')\right)\,\right] dx.$$

Since this holds for all variations $v(x)$, we conclude that[†]

$$\nabla J[u] = \frac{\partial L}{\partial u}(x, u, u') - \frac{d}{dx}\left(\frac{\partial L}{\partial p}(x, u, u')\right). \tag{21.25}$$

This is our explicit formula for the functional gradient or variational derivative of the functional (21.16) with Lagrangian $L(x, u, p)$. Observe that the gradient $\nabla J[u]$ of a functional is a *function*.

The *critical functions* $u(x)$ are, by definition, those for which the functional gradient vanishes: satisfy

$$\nabla J[u] = \frac{\partial L}{\partial u}(x, u, u') - \frac{d}{dx}\frac{\partial L}{\partial p}(x, u, u') = 0. \tag{21.26}$$

In view of (21.23), the critical equation (21.26) is, in fact, a second order ordinary differential equation,

$$E(x, u, u', u'') = \frac{\partial L}{\partial u}(x, u, u') - \frac{\partial^2 L}{\partial x\, \partial p}(x, u, u') - u' \frac{\partial^2 L}{\partial u\, \partial p}(x, u, u') - u'' \frac{\partial^2 L}{\partial p^2}(x, u, u') = 0, \tag{21.27}$$

known as the *Euler–Lagrange equation* associated with the variational problem (21.16), in honor of two of the most important contributors to the subject. Any solution to the Euler–Lagrange equation that is subject to the assumed boundary conditions forms a critical point for the functional, and hence is a potential candidate for the desired minimizing function. And, in many cases, the Euler–Lagrange equation suffices to characterize the minimizer without further ado.

Theorem 21.1. *Suppose the Lagrangian function is at least twice continuously differentiable*: $L(x, u, p) \in \mathrm{C}^2$. *Then any* C^2 *minimizer* $u(x)$ *to the corresponding functional* $J[u] = \displaystyle\int_a^b L(x, u, u')\, dx$, *subject to the selected boundary conditions, must satisfy the associated Euler–Lagrange equation* (21.26).

Let us now investigate what the Euler–Lagrange equation tells us about the examples of variational problems presented at the beginning of this section. One word of caution: there do exist seemingly reasonable functionals whose minimizers are not, in fact, C^2, and hence do not solve the Euler–Lagrange equation in the classical sense; see [**13**] for examples. Fortunately, in most variational problems that arise in real-world applications, such pathologies do not appear.

Curves of Shortest Length — Planar Geodesics

Let us return to the most elementary problem in the calculus of variations: finding the curve of shortest length connecting two points $\mathbf{a} = (a, \alpha)$, $\mathbf{b} = (b, \beta) \in \mathbb{R}^2$ in the plane. As we noted in Section 21.2, such planar geodesics minimize the arc length integral

$$J[u] = \int_a^b \sqrt{1 + (u')^2}\, dx \qquad \text{with Lagrangian} \qquad L(x, u, p) = \sqrt{1 + p^2}\,,$$

[†] See Exercise ■ for a complete justification.

subject to the boundary conditions

$$u(a) = \alpha, \qquad u(b) = \beta.$$

Since

$$\frac{\partial L}{\partial u} = 0, \qquad \frac{\partial L}{\partial p} = \frac{p}{\sqrt{1+p^2}},$$

the Euler–Lagrange equation (21.26) in this case takes the form

$$0 \;=\; -\,\frac{d}{dx}\,\frac{u'}{\sqrt{1+(u')^2}} \;=\; -\,\frac{u''}{(1+(u')^2)^{3/2}}.$$

Since the denominator does not vanish, this is the same as the simplest second order ordinary differential equation

$$u'' = 0. \tag{21.28}$$

Therefore, the solutions to the Euler–Lagrange equation are all affine functions, $u = c\,x + d$, whose graphs are straight lines. Since our solution must also satisfy the boundary conditions, the only critical function — and hence the sole candidate for a minimizer — is the straight line

$$y = \frac{\beta - \alpha}{b - a}\,(x - a) + \alpha \tag{21.29}$$

passing through the two points. Thus, the Euler–Lagrange equation helps to reconfirm our intuition that straight lines minimize distance.

Be that as it may, the fact that a function satisfies the Euler–Lagrange equation and the boundary conditions merely confirms its status as a critical function, and does not guarantee that it is the minimizer. Indeed, any critical function is also a candidate for *maximizing* the variational problem, too. The nature of a critical function will be elucidated by the second derivative test, and requires some further work. Of course, for the minimum distance problem, we "know" that a straight line cannot maximize distance, and must be the minimizer. Nevertheless, the reader should have a small nagging doubt that we have completely solved the problem at hand ...

Minimal Surface of Revolution

Consider next the problem of finding the curve connecting two points that generates a surface of revolution of minimal surface area. For simplicity, we assume that the curve is given by the graph of a *non-negative* function $y = u(x) \geq 0$. According to (21.11), the required curve will minimize the functional

$$J[u] = \int_a^b u\,\sqrt{1+(u')^2}\,dx, \qquad \text{with Lagrangian} \qquad L(x,u,p) = u\,\sqrt{1+p^2}\,, \tag{21.30}$$

where we have omitted an irrelevant factor of $2\,\pi$ and used positivity to delete the absolute value on u in the integrand. Since

$$\frac{\partial L}{\partial u} = \sqrt{1+p^2}\,, \qquad \frac{\partial L}{\partial p} = \frac{u\,p}{\sqrt{1+p^2}}\,,$$

the Euler–Lagrange equation (21.26) is

$$\sqrt{1+(u')^2} \;-\; \frac{d}{dx}\,\frac{u\,u'}{\sqrt{1+(u')^2}} \;=\; \frac{1+(u')^2-u\,u''}{(1+(u')^2)^{3/2}} = 0. \tag{21.31}$$

Therefore, to find the critical functions, we need to solve a nonlinear second order ordinary differential equation — and not one in a familiar form.

Fortunately, there is a little trick[†] we can use to find the solution. If we multiply the equation by u', we can then rewrite the result as an exact derivative

$$u'\left(\frac{1+(u')^2-u\,u''}{(1+(u')^2)^{3/2}}\right) \;=\; \frac{d}{dx}\,\frac{u}{\sqrt{1+(u')^2}} = 0.$$

We conclude that the quantity

$$\frac{u}{\sqrt{1+(u')^2}} = c, \tag{21.32}$$

is constant, and so the left hand side is a *first integral* for the differential equation, as per Definition 20.26. Solving for[‡]

$$\frac{du}{dx} = u' = \frac{\sqrt{u^2-c^2}}{c}$$

results in an autonomous first order ordinary differential equation, which we can immediately solve:

$$\int \frac{c\;du}{\sqrt{u^2-c^2}} = x+\delta,$$

where δ is a constant of integration. The most useful form of the left hand integral is in terms of the inverse to the hyperbolic cosine function $\cosh z = \frac{1}{2}(e^z+e^{-z})$, whereby

$$\cosh^{-1}\frac{u}{c} = x+\delta, \qquad \text{and hence} \qquad u = c\,\cosh\left(\frac{x+\delta}{c}\right). \tag{21.33}$$

In this manner, we have produced the general solution to the Euler–Lagrange equation (21.31). Any solution that also satisfies the boundary conditions provides a critical function for the surface area functional (21.30), and hence is a candidate for the minimizer.

The curve prescribed by the graph of a hyperbolic cosine function (21.33) is known as a *catenary*. It is *not* a parabola, even though to the untrained eye it looks similar. Interestingly, the catenary is the same profile as a hanging chain; see Exercise ■. Owing to their minimizing properties, catenaries are quite common in engineering design — for instance the cables in a suspension bridge such as the Golden Gate Bridge are catenaries, while the arch in St. Louis is an inverted catenary.

[†] Actually, as with many tricks, this is really an indication that something profound is going on. Noether's Theorem, a result of fundamental importance in modern physics that relates symmetries and conservation laws, [**76**, **141**], underlies the integration method; see Exercise ■ for further details.

[‡] The square root is real since, by (21.32), $|\,u\,| \le |\,c\,|$.

So far, we have not taken into account the boundary conditions It turns out that there are three distinct possibilities, depending upon the configuration of the boundary points:

(a) There is precisely one value of the two integration constants c, δ that satisfies the two boundary conditions. In this case, it can be proved that this catenary is the unique curve that minimizes the area of its associated surface of revolution.

(b) There are two different possible values of c, δ that satisfy the boundary conditions. In this case, one of these is the minimizer, and the other is a spurious solution — one that corresponds to a saddle point for the surface area functional.

(c) There are *no* values of c, δ that allow (21.33) to satisfy the two boundary conditions. This occurs when the two boundary points $\mathbf{a}, \mathbf{b}$ are relatively far apart. In this configuration, the physical soap film spanning the two circular wires breaks apart into two circular disks, and this defines the minimizer for the problem, i.e., unlike cases (a) and (b), there is *no* surface of revolution that has a smaller surface area than the two disks. However, the "function"† that minimizes this configuration consists of two vertical lines from the boundary points to the x axis, along with that segment on the axis lying between them. More precisely, we can approximate this function by a sequence of genuine functions that give progressively smaller and smaller values to the surface area functional (21.11), but the actual minimum is not attained among the class of (smooth) functions; this is illustrated in Figure cats■.

Further details are relegated to the exercises.

Thus, even in such a reasonably simple example, a number of the subtle complications can already be seen. Lack of space precludes a more detailed development of the subject, and we refer the interested reader to more specialized books on the calculus of variations, including [**46**, **76**].

The Brachistochrone Problem

The most famous classical variational principle is the so-called *brachistochrone problem*. The Latin word "brachistochrone" means "minimal time". An experimenter lets a bead slide down a wire that connects two fixed points. The goal is to shape the wire in such a way that, starting from rest, the bead slides from one end to the other in minimal time. Naïve guesses for the wire's optimal shape, including a straight line, a parabola, a circular arc, or even a catenary are wrong. One can do better through a careful analysis of the associated variational problem. The brachistochrone problem was originally posed by the Swiss mathematician Johann Bernoulli in 1696, and served as an inspiration for much of the subsequent development of the subject.

We take, without loss of generality, the starting point of the bead to be at the origin: $\mathbf{a} = (0, 0)$. The wire will bend downwards, and so, to avoid distracting minus signs in the subsequent formulae, we take the vertical y axis to point downwards. The shape of the wire will be given by the graph of a function $y = u(x) \geq 0$. The end point $\mathbf{b} = (b, \beta)$ is

† Here "function" must be taken in a *very* broad sense, as this one does not even correspond to a generalized function!

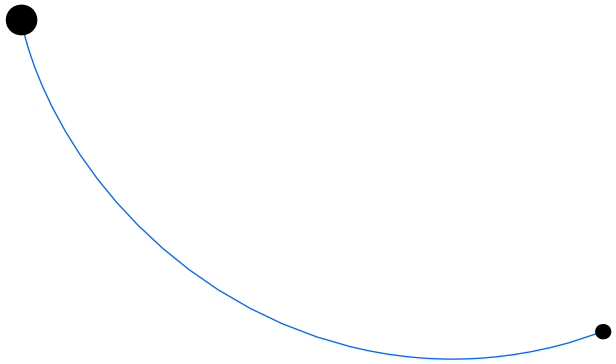


Figure 21.4. The Brachistrochrone Problem.

assumed to lie below and to the right, and so $b > 0$ and $\beta > 0$. The set-up is sketched in Figure 21.4.

To mathematically formulate the problem, the first step is to find the formula for the transit time of the bead sliding along the wire. Arguing as in our derivation of the optics functional (21.5), if $v(x)$ denotes the instantaneous speed of descent of the bead when it reaches position $(x, u(x))$, then the total travel time is

$$T[u] = \int_0^\ell \frac{ds}{v} = \int_0^b \frac{\sqrt{1 + (u')^2}}{v}\, dx, \tag{21.34}$$

where ℓ is the overall length of the wire.

We shall use conservation of energy to determine a formula for the speed v as a function of the position along the wire. The kinetic energy of the bead is $\frac{1}{2} m v^2$, where m is its mass. On the other hand, due to our sign convention, the potential energy of the bead when it is at height $y = u(x)$ is $-m g u(x)$, where g the gravitational force, and we take the initial height as the zero potential energy level. The bead is initially at rest, with 0 kinetic energy and 0 potential energy. Assuming that frictional forces are negligible, conservation of energy implies that the total energy must remain equal to 0, and hence

$$0 = \tfrac{1}{2} m v^2 - m g u.$$

We can solve this equation to determine the bead's speed as a function of its height:

$$v = \sqrt{2 g u}\,. \tag{21.35}$$

Substituting this expression into (21.34), we conclude that the shape $y = u(x)$ of the wire is obtained by minimizing the functional

$$T[u] = \int_0^b \sqrt{\frac{1 + (u')^2}{2 g u}}\; dx, \tag{21.36}$$

subjet to the boundary conditions

$$u(0) = 0, \qquad u(b) = \beta. \tag{21.37}$$

The associated Lagrangian is

$$L(x,u,p) = \sqrt{\frac{1+p^2}{u}}\,,$$

where we omit an irrelevant factor of $\sqrt{2\,g}$ (or adopt physical units in which $g = \frac{1}{2}$). We compute

$$\frac{\partial L}{\partial u} \;=\; -\,\frac{\sqrt{1+p^2}}{2\,u^{3/2}}\,, \qquad\qquad \frac{\partial L}{\partial p} \;=\; \frac{p}{\sqrt{u\,(1+p^2)}}\,.$$

Therefore, the Euler–Lagrange equation for the brachistochrone functional is

$$-\,\frac{\sqrt{1+(u')^2}}{2\,u^{3/2}} \;-\; \frac{d}{dx}\,\frac{u'}{\sqrt{u\,(1+(u')^2)}} \;=\; -\,\frac{2\,u\,u'' + (u')^2 + 1}{2\,\sqrt{u\,(1+(u')^2)}} \;=\; 0. \tag{21.38}$$

Thus, the minimizing functions solve the nonlinear second order ordinary differential equation

$$2\,u\,u'' + (u')^2 + 1 = 0.$$

Rather than try to solve this differential equation directly, we note that the Lagrangian does not depend upon x, and therefore we can use the result of Exercise ■, that states that the Hamiltonian function

$$H(x,u,p) = L - p\,\frac{\partial L}{\partial p} = \frac{1}{\sqrt{u(1+p^2)}}$$

defines a first integral. Thus,

$$H(x,u,u') = \frac{1}{\sqrt{u(1+(u')^2)}} = k, \qquad \text{which we rewrite as} \qquad u(1+(u')^2) = c,$$

where $c = 1/k^2$ is a constant. (This can be checked by directly calculating $dH/dx \equiv 0$.) Solving for the derivative u' results in the first order autonomous ordinary differential equation

$$\frac{du}{dx} = \sqrt{\frac{c-u}{u}}\,.$$

As in (20.7), this equation can be explicitly solved by separation of variables, and so

$$\int \sqrt{\frac{u}{c-u}}\; du = x + k.$$

The left hand integration relies on the trigonometric substitution

$$u = \tfrac{1}{2}\,c(1-\cos\theta),$$

whereby

$$x + k = \tfrac{1}{2}\,c \int \sqrt{\frac{1-\cos\theta}{1+\cos\theta}}\;\sin\theta\,d\theta = \tfrac{1}{2}\,c\int\,(1-\cos\theta)\,d\theta = \tfrac{1}{2}\,c(\theta - \sin\theta).$$

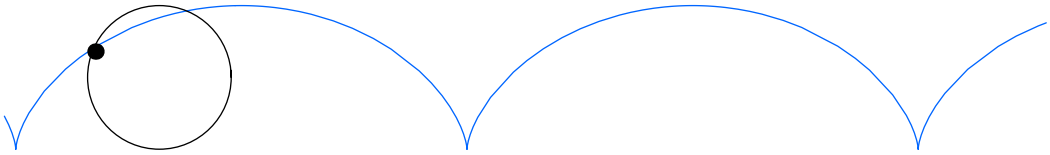


Figure 21.5. A Cycloid.

The left hand boundary condition implies $k = 0$, and so the solution to the Euler–Lagrange equation are curves parametrized by

$$x = r(\theta - \sin\theta), \qquad u = r(1 - \cos\theta). \tag{21.39}$$

With a little more work, it can be proved that the parameter $r = \frac{1}{2}\,c$ is uniquely prescribed by the right hand boundary condition, and moreover, the resulting curve supplies the global minimizer of the brachistochrone functional, [**cvar**]. The miniming curve is known as a *cycloid*, which, according to Exercise ∎, can be visualized as the curve traced by a point sitting on the edge of a rolling wheel of radius r, as plotted in Figure 21.5. Interestingly, in certain configurations, namely if $\beta < 2\,b/\pi$, the cycloid that solves the brachistrochrone problem dips below the right hand endpoint $\mathbf{b} = (b, \beta)$, and so the bead is moving upwards when it reaches the end of the wire.

21.3. The Second Variation.

The solutions to the Euler–Lagrange boundary value problem are the critical functions for the variational principle, meaning that they cause the functional gradient to vanish. For finite-dimensional optimization problems, being a critical point is only a necessary condition for minimality. One must impose additional conditions, based on the second derivative of the objective function at the critical point, in order to guarantee that it is a minimum and not a maximum or saddle point. Similarly, in the calculus of variations, the solutions to the Euler–Lagrange equation may also include (local) maxima, as well as other non-extremal critical functions. To distinguish between the possibilties, we need to formulate a second derivative test for the objective functional. In the calculus of variations, the second derivative of a functional is known as its *second variation*, and the goal of this section is to construct and analyze it in its simplest manifestation.

For a finite-dimensional objective function $F(u_1, \dots, u_n)$, the second derivative test was based on the positive definiteness of its Hessian matrix. The justification was based on the second order Taylor expansion of the objective function at the critical point. In an analogous fashion, we expand an objective functional $J[\,u\,]$ near the critical function. Consider the scalar function

$$h(t) = J[\,u + t\,v\,],$$

where the function $v(x)$ represents a variation. The second order Taylor expansion of $h(t)$ takes the form

$$h(t) = J[\,u + t\,v\,] = J[\,u\,] + t\,K[\,u; v\,] + \tfrac{1}{2}\,t^2\,Q[\,u; v\,] + \,\cdots\,.$$

The first order terms are linear in the variation v, and, according to our earlier calculation, given by an inner product

$$h'(0) = K[\,u; v\,] = \langle\, \nabla J[\,u\,]\,, v\,\rangle$$

between the variation and the functional gradient. In particular, if u is a critical function, then the first order terms vanish,

$$K[\,u; v\,] = \langle\, \nabla J[\,u\,]\,, v\,\rangle = 0$$

for all allowable variations v, meaning those that satisfy the homogeneous boundary conditions. Therefore, the nature of the critical function u — minimum, maximum, or neither — will, in most cases, be determined by the second derivative terms

$$h''(0) = Q[\,u; v\,].$$

As argued in our proof of the finite-dimensional Theorem 19.42, if u is a minimizer, then $Q[\,u; v\,] \geq 0$. Conversely, if $Q[\,u; v\,] > 0$ for all $v \not\equiv 0$, i.e., the second variation is positive definite, then the critical function u will be a strict local minimizer. This forms the crux of the second derivative test.

Let us explicitly evaluate the second variational for the simplest functional (21.16). Consider the scalar function

$$h(t) = J[\,u + t\,v\,] = \int_a^b L(x, u + t\,v, u' + t\,v')\,dx,$$

whose first derivative $h'(0)$ was already determined in (21.21); here we require the second variation

$$Q[\,u; v\,] = h''(0) = \int_a^b \big[\, A\,v^2 + 2\,B\,v\,v' + C\,(v')^2 \,\big]\,dx, \tag{21.40}$$

where the coefficient functions

$$A(x) = \frac{\partial^2 L}{\partial u^2}\,(x, u, u')\,, \qquad B(x) = \frac{\partial^2 L}{\partial u\,\partial p}\,(x, u, u')\,, \qquad C(x) = \frac{\partial^2 L}{\partial p^2}\,(x, u, u')\,, \tag{21.41}$$

are found by evaluating certain second order derivatives of the Lagrangian at the critical function $u(x)$. In contrast to the first variation, integration by parts will not eliminate all of the derivatives on v in the quadratic functional (21.40), which causes significant complications in the ensuing analysis.

The second derivative test for a minimizer relies on the positivity of the second variation. To formulate conditions that the critical function be a minimizer for the functional, So, we need to establish criteria guaranteeing the positive definiteness of such a quadratic functional is positive definite, meaning that $Q[\,u; v\,] > 0$ for all non-zero allowable variations $v(x) \not\equiv 0$. Clearly, if the integrand is positive definite at each point, so

$$A(x)\,v^2 + 2\,B(x)\,v\,v' + C(x)\,(v')^2 > 0 \quad \text{whenever} \quad a < x < b, \quad \text{and} \quad v(x) \not\equiv 0, \tag{21.42}$$

then $Q[\,u; v\,] > 0$ is also positive definite.

Example 21.2. For the arc length minimization functional (21.3), the Lagrangian is $L(x,u,p) = \sqrt{1+p^2}$. To analyze the second variation, we first compute

$$\frac{\partial^2 L}{\partial u^2} = 0, \qquad \frac{\partial^2 L}{\partial u\, \partial p} = 0, \qquad \frac{\partial^2 L}{\partial p^2} = \frac{1}{(1+p^2)^{3/2}}\,.$$

For the critical straight line function

$$u(x) = \frac{\beta - \alpha}{b-a}\,(x-a) + \alpha, \qquad \text{with} \qquad p = u'(x) = \frac{\beta-\alpha}{b-a}\,,$$

we find

$$A(x) = \frac{\partial^2 L}{\partial u^2} = 0, \quad B(x) = \frac{\partial^2 L}{\partial u\, \partial p} = 0, \quad C(x) = \frac{\partial^2 L}{\partial p^2} = \frac{(b-a)^3}{\big[\,(b-a)^2 + (\beta-\alpha)^2\,\big]^{3/2}} \equiv k.$$

Therefore, the second variation functional (21.40) is

$$Q[u;v] = \int_a^b k\,(v')^2\,dx,$$

where $k > 0$ is a positive constant. Thus, $Q[u;v] = 0$ vanishes if and only if $v(x)$ is a constant function. But the variation $v(x)$ is required to satisfy the homogeneous boundary conditions $v(a) = v(b) = 0$, and hence $Q[u;v] > 0$ for all allowable nonzero variations. Therefore, we conclude that the straight line is, indeed, a (local) minimizer for the arc length functional. We have at last justified our intuition that the shortest distance between two points is a straight line!

In general, as the following example demonstrates, the pointwise positivity condition (21.42) is overly restrictive.

Example 21.3. Consider the quadratic functional

$$Q[v] = \int_0^1 \big[\,(v')^2 - v^2\,\big]\,dx. \tag{21.43}$$

We claim that $Q[v] > 0$ for all nonzero $v \not\equiv 0$ subject to homogeneous Dirichlet boundary conditions $v(0) = 0 = v(1)$. This result is not trivial! Indeed, the boundary conditions play an essential role, since choosing $v(x) \equiv c \neq 0$ to be any constant function will produce a negative value for the functional: $Q[v] = -\,c^2$.

To prove the claim, consider the quadratic functional

$$\widetilde{Q}[v] = \int_0^1 (v' + v\tan x)^2\,dx \;\geq\; 0,$$

which is clearly non-negative, since the integrand is everywhere ≥ 0. Moreover, by continuity, the integral vanishes if and only if v satisfies the first order linear ordinary differential equation

$$v' + v\tan x = 0, \qquad \text{for all} \qquad 0 \leq x \leq 1.$$

The only solution that also satisfies boundary condition $v(0) = 0$ is the trivial one $v \equiv 0$. We conclude that $\widetilde{Q}[v] = 0$ if and only if $v \equiv 0$, and hence $\widetilde{Q}[v]$ is a positive definite quadratic functional on the space of allowable variations.

Let us expand the latter functional,

$$\begin{aligned}\widetilde{Q}[v] &= \int_0^1 \big[\, (v')^2 + 2\,v\,v' \tan x + v^2 \tan^2 x \,\big]\, dx \\ &= \int_0^1 \big[\, (v')^2 - v^2\,(\tan x)' + v^2 \tan^2 x \,\big]\, dx = \int_0^1 \big[\, (v')^2 - v^2 \,\big]\, dx = Q[v].\end{aligned}$$

In the second equality, we integrated the middle term by parts, using $(v^2)' = 2\,v\,v'$, and noting that the boundary terms vanish owing to our imposed boundary conditions. Since $\widetilde{Q}[v]$ is positive definite, so is $Q[v]$, justifying the previous claim.

To appreciate how subtle this result is, consider the almost identical quadratic functional

$$\widehat{Q}[v] = \int_0^4 \big[\, (v')^2 - v^2 \,\big]\, dx, \tag{21.44}$$

the only difference being the upper limit of the integral. A quick computation shows that the function $v(x) = x(4 - x)$ satisfies the boundary conditions $v(0) = 0 = v(4)$, but

$$\widehat{Q}[v] = \int_0^4 \big[\, (4 - 2\,x)^2 - x^2(4 - x)^2 \,\big]\, dx = -\frac{64}{5} < 0.$$

Therefore, $\widehat{Q}[v]$ is *not* positive definite. Our preceding analysis does not apply because the function $\tan x$ becomes singular at $x = \frac{1}{2}\pi$, and so the auxiliary integral $\displaystyle\int_0^4 (v' + v \tan x)^2\, dx$ does not converge.

The complete analysis of positive definiteness of quadratic functionals is quite subtle. Indeed, the strange appearance of $\tan x$ in the preceding example turns out to be an important clue! In the interests of brevity, let us just state without proof a fundamental theorem, and refer the interested reader to [**76**] for full details.

Theorem 21.4. *Let $A(x), B(x), C(x) \in \mathrm{C}^0[a, b]$ be continuous functions. The quadratic functional*

$$Q[v] = \int_a^b \big[\, A\,v^2 + 2\,B\,v\,v' + C\,(v')^2 \,\big]\, dx$$

is positive definite, so $Q[v] > 0$ for all $v \not\equiv 0$ satisfying the homogeneous Dirichlet boundary conditions $v(a) = v(b) = 0$, provided

(a) $C(x) > 0$ for all $a \le x \le b$, and

(b) For any $a < c \le b$, the only solution to its linear Euler–Lagrange boundary value problem

$$-(C\,v')' + (A - B')\,v = 0, \qquad\qquad v(a) = 0 = v(c), \tag{21.45}$$

is the trivial function $v(x) \equiv 0$.

Remark: A value c for which (21.45) has a nontrivial solution is known as a *conjugate point* to a. Thus, condition (b) can be restated that the variational problem has no conjugate points in the interval $(a, b]$.

Example 21.5. The quadratic functional

$$Q[v] = \int_0^b \big[\, (v')^2 - v^2 \,\big]\, dx \tag{21.46}$$

has Euler–Lagrange equation

$$-v'' - v = 0.$$

The solutions $v(x) = k \sin x$ satisfy the boundary condition $v(0) = 0$. The first conjugate point occurs at $c = \pi$ where $v(\pi) = 0$. Therefore, Theorem 21.4 implies that the quadratic functional (21.46) is positive definite *provided* the upper integration limit $b < \pi$. This explains why the original quadratic functional (21.43) is positive definite, since there are no conjugate points on the interval $[0, 1]$, while the modified version (21.44) is *not*, because the first conjugate point π lies on the interval $(0, 4]$.

In the case when the quadratic functional arises as the second variation of a functional (21.16), then the coefficient functions A, B, C are given in terms of the Lagrangian $L(x, u, p)$ by formulae (21.41). In this case, the first condition in Theorem 21.4 requires

$$\frac{\partial^2 L}{\partial p^2}\,(x, u, u') > 0 \tag{21.47}$$

for the minimizer $u(x)$. This is known as the *Legendre condition*. The second, *conjugate point condition* requires that the so-called *linear variational equation*

$$-\frac{d}{dx}\left(\frac{\partial^2 L}{\partial p^2}\,(x, u, u')\,\frac{dv}{dx}\right) + \left(\frac{\partial^2 L}{\partial u^2}\,(x, u, u') - \frac{d}{dx}\,\frac{\partial^2 L}{\partial u\,\partial p}\,(x, u, u')\right) v = 0 \tag{21.48}$$

has no nontrivial solutions $v(x) \not\equiv 0$ that satisfy $v(a) = 0$ and $v(c) = 0$ for $a < c \le b$. In this way, we have arrived at a rigorous form of the second derivative test for the simplest functional in the calculus of variations.

Theorem 21.6. *If the function $u(x)$ satisfies the Euler–Lagrange equation* (21.26)*, and, in addition, the Legendre condition* (21.47) *and there are no conjugate points on the interval, then $u(x)$ is a strict local minimum for the functional.*

21.4. Multi-dimensional Variational Problems.

The calculus of variations encompasses a very broad range of mathematical applications. The methods of variational analysis can be applied to an enormous variety of physical systems, whose equilibrium configurations inevitably minimize a suitable functional, which, typically, represents the potential energy of the system. Minimizing configurations appear as critical functions at which the functional gradient vanishes. Following similar computational procedures as in the one-dimensional calculus of variations, we find that the critical functions are characterized as solutions to a system of partial differential equations,

known as the Euler–Lagrange equations associated with the variational principle. Each solution to the boundary value problem specified by the Euler–Lagrange equations subject to appropraite boundary conditions is, thus, a candidate minimizer for the variational problem. In many applications, the Euler–Lagrange boundary value problem suffices to single out the physically relevant solutions, and one need not press on to the considerably more difficult second variation.

Implementation of the variational calculus for functionals in higher dimensions will be illustrated by looking at a specific example — a first order variational problem involving a single scalar function of two variables. Once this is fully understood, generalizations and extensions to higher dimensions and higher order Lagrangians are readily apparent. Thus, we consider an objective functional

$$J[u] = \iint_\Omega L(x, y, u, u_x, u_y)\, dx\, dy, \tag{21.49}$$

having the form of a double integral over a prescribed domain $\Omega \subset \mathbb{R}^2$. The *Lagrangian* $L(x, y, u, p, q)$ is assumed to be a sufficiently smooth function of its five arguments. Our goal is to find the function(s) $u = f(x, y)$ that minimize the value of $J[u]$ when subject to a set of prescribed boundary conditions on $\partial\Omega$, the most important being our usual Dirichlet, Neumann, or mixed boundary conditions. For simplicity, we concentrate on the Dirichlet boundary value problem, and require that the minimizer satisfy

$$u(x, y) = g(x, y) \qquad \text{for} \qquad (x, y) \in \partial\Omega. \tag{21.50}$$

The First Variation

The basic necessary condition for an extremum (minimum or maximum) is obtained in precisely the same manner as in the one-dimensional framework. Consider the scalar function

$$h(t) \equiv J[u + t\, v] = \iint_\Omega L(x, y, u + t\, v, u_x + t\, v_x, u_y + t\, v_y)\, dx\, dy$$

depending on $t \in \mathbb{R}$. The *variation* $v(x, y)$ is assumed to satisfy homogeneous Dirichlet boundary conditions

$$v(x, y) = 0 \qquad \text{for} \qquad (x, y) \in \partial\Omega, \tag{21.51}$$

to ensure that $u + t\, v$ satisfies the same boundary conditions (21.50) as u itself. Under these conditions, if u is a minimizer, then the scalar function $h(t)$ will have a minimum at $t = 0$, and hence

$$h'(0) = 0.$$

When computing $h'(t)$, we assume that the functions involved are sufficiently smooth so as to allow us to bring the derivative inside the integral, and then apply the chain rule. At $t = 0$, the result is

$$h'(0) = \frac{d}{dt}\, J[u + t\, v]\, \Big|_{t=0} = \iint_\Omega \left(v\, \frac{\partial L}{\partial u} + v_x\, \frac{\partial L}{\partial p} + v_y\, \frac{\partial L}{\partial q} \right) dx\, dy, \tag{21.52}$$

where the derivatives of L are all evaluated at x, y, u, u_x, u_y. To identify the functional gradient, we need to rewrite this integral in the form of an inner product:

$$h'(0) = \langle\, \nabla J[u]\,, v\,\rangle = \iint_\Omega h(x,y)\, v(x,y)\, dx\, dy, \qquad \text{where} \qquad h = \nabla J[u].$$

To convert (21.52) into this form, we need to remove the offending derivatives from v. In two dimensions, the requisite integration by parts formula is based on Green's Theorem A.26, and the required formula can be found in (15.88), namely,

$$\iint_\Omega \frac{\partial v}{\partial x}\, w_1 + \frac{\partial v}{\partial y}\, w_2\, dx\, dy = \oint_{\partial\Omega} v\,(-w_2\, dx + w_1\, dy) - \iint_\Omega v\left(\frac{\partial w_1}{\partial x} + \frac{\partial w_2}{\partial y}\right) dx\, dy, \tag{21.53}$$

in which w_1, w_2 are arbitrary smooth functions. Setting $w_1 = \dfrac{\partial L}{\partial p}$, $w_2 = \dfrac{\partial L}{\partial q}$, we find

$$\iint_\Omega \left(v_x \frac{\partial L}{\partial p} + v_y \frac{\partial L}{\partial q}\right) dx\, dy = -\iint_\Omega v\left[\frac{\partial}{\partial x}\left(\frac{\partial L}{\partial p}\right) + \frac{\partial}{\partial y}\left(\frac{\partial L}{\partial q}\right)\right] dx\, dy,$$

where the boundary integral vanishes owing to the boundary conditions (21.51) that we impose on the allowed variations. Substituting this result back into (21.52), we conclude that

$$h'(0) = \iint_\Omega v\left[\frac{\partial L}{\partial u} - \frac{\partial}{\partial x}\left(\frac{\partial L}{\partial p}\right) - \frac{\partial}{\partial y}\left(\frac{\partial L}{\partial q}\right)\right] dx\, dy = \langle\, \nabla J[u]\,, v\,\rangle, \tag{21.54}$$

where

$$\nabla J[u] = \frac{\partial L}{\partial u} - \frac{\partial}{\partial x}\left(\frac{\partial L}{\partial p}\right) - \frac{\partial}{\partial y}\left(\frac{\partial L}{\partial q}\right)$$

is the desired first variation or functional gradient. Since the gradient vanishes at a critical function, we conclude that the minimizer $u(x,y)$ must satisfy the *Euler–Lagrange equation*

$$\frac{\partial L}{\partial u}(x,y,u,u_x,u_y) \;-\; \frac{\partial}{\partial x}\left(\frac{\partial L}{\partial p}(x,y,u,u_x,u_y)\right) \;-\; \frac{\partial}{\partial y}\left(\frac{\partial L}{\partial q}(x,y,u,u_x,u_y)\right) \;=\; 0. \tag{21.55}$$

Once we explicitly evaluate the derivatives, the net result is a second order partial differential equation, which you are asked to write out in full detail in Exercise ∎.

Example 21.7. As a first elementary example, consider the Dirichlet minimization problem

$$J[u] = \iint_\Omega \tfrac{1}{2}\left(u_x^2 + u_y^2\right) dx\, dy \tag{21.56}$$

that we first encountered in our analysis of the Laplace equation (15.100). In this case, the associated Lagrangian is

$$L = \tfrac{1}{2}(p^2 + q^2), \qquad \text{with} \qquad \frac{\partial L}{\partial u} = 0, \qquad \frac{\partial L}{\partial p} = p = u_x, \qquad \frac{\partial L}{\partial q} = q = u_y.$$

Therefore, the Euler–Lagrange equation (21.55) becomes

$$-\frac{\partial}{\partial x}\,(u_x) - \frac{\partial}{\partial y}\,(u_y) = -\,u_{xx} - u_{yy} = -\,\Delta u = 0,$$

which is the two-dimensional Laplace equation. Subject to the selected boundary conditions, the solutions, i.e., the harmonic functions, are critical functions for the Dirichlet variational principle. This reconfirms the Dirichlet characterization of harmonic functions as minimizers of the variational principle, as stated in Theorem 15.14.

However, the calculus of variations approach, as developed so far, leads to a much weaker result since it only singles out the harmonic functions as *candidates* for minimizing the Dirichlet integral; they could just as easily be maximing functions or saddle points. When dealing with a quadratic variational problem, the direct algebraic approach is, when applicable, the more powerful, since it assures us that the solutions to the Laplace equation really do minimize the integral among the space of functions satisfying the appropriate boundary conditions. However, the direct method is restricted to quadratic variational problems, whose Euler–Lagrange equations are linear partial differential equations. In nonlinear cases, one really does need to utilize the full power of the variational machinery.

Example 21.8. Let us derive the Euler–Lagrange equation for the minimal surface problem. From (21.9), the surface area integral

$$J[u] = \iint_\Omega \sqrt{1 + u_x^2 + u_y^2}\; dx\, dy \qquad \text{has Lagrangian} \qquad L = \sqrt{1 + p^2 + q^2}\,.$$

Note that

$$\frac{\partial L}{\partial u} = 0, \qquad \frac{\partial L}{\partial p} = \frac{p}{\sqrt{1 + p^2 + q^2}}\,, \qquad \frac{\partial L}{\partial q} = \frac{q}{\sqrt{1 + p^2 + q^2}}\,.$$

Therefore, replacing $p \to u_x$ and $q \to u_y$ and then evaluating the derivatives, the Euler–Lagrange equation (21.55) becomes

$$-\frac{\partial}{\partial x}\,\frac{u_x}{\sqrt{1 + u_x^2 + u_y^2}} - \frac{\partial}{\partial y}\,\frac{u_y}{\sqrt{1 + u_x^2 + u_y^2}} = \frac{-\,(1 + u_y^2)u_{xx} + 2u_x u_y u_{xy} - (1 + u_x^2)u_{yy}}{(1 + u_x^2 + u_y^2)^{3/2}} = 0.$$

Thus, a surface described by the graph of a function $u = f(x, y)$ is a critical function, and hence a candidate for minimizing surface area, provided it satisfies the *minimal surface equation*

$$(1 + u_y^2)\, u_{xx} - 2\, u_x\, u_y\, u_{xy} + (1 + u_x^2)\, u_{yy} = 0. \tag{21.57}$$

We are confronted with a complicated, nonlinear, second order partial differential equation, which has been the focus of some of the most sophisticated and deep analysis over the past two centuries, with significant progress on understanding its solution only within the past 70 years. In this book, we have not developed the sophisticated analytical, geometrical, and numerical techniques that are required to have anything of substance to say about its solutions. We refer the interested reader to the advanced texts [**132**, **136**] for further developments in this fascinating problem.

Example 21.9. The small deformations of an elastic body $\Omega \subset \mathbb{R}^n$ are described by the *displacement* field, $\mathbf{u}\colon \Omega \to \mathbb{R}^n$. Each material point $\mathbf{x} \in \Omega$ in the undeformed body will move to a new position $\mathbf{y} = \mathbf{x} + \mathbf{u}(\mathbf{x})$ in the deformed body

$$\widetilde{\Omega} = \{\, \mathbf{y} = \mathbf{x} + \mathbf{u}(\mathbf{x}) \mid \ \mathbf{x} \in \Omega \,\}.$$

The one-dimensional case governs bars, beams and rods, two-dimensional bodies include thin plates and shells, while $n = 3$ for fully three-dimensional solid bodies. See [**8**, **85**] for details and physical derivations.

For small deformations, we can use a linear theory to approximate the much more complicated equations of nonlinear elasticity. The simplest case is that of a homogeneous and isotropic planar body $\Omega \subset \mathbb{R}^2$. The equilibrium mechanics are described by the deformation function $\mathbf{u}(\mathbf{x}) = (\, u(x,y), v(x,y)\,)^T$. A detailed physical analysis of the constitutive assumptions leads to a minimization principle based on the following functional:

$$\begin{aligned} J[u,v] &= \iint_\Omega \left[\, \tfrac{1}{2}\,\mu \,\|\, \nabla \mathbf{u} \,\|^2 + \tfrac{1}{2}\,(\lambda+\mu)(\nabla\cdot\mathbf{u})^2 \,\right] dx\, dy \\ &= \iint_\Omega \left[\, \left(\, \tfrac{1}{2}\,\lambda + \mu \,\right)(u_x^2 + v_y^2) + \tfrac{1}{2}\,\mu\,(u_y^2 + v_x^2) + (\lambda+\mu)\,u_x\,v_y \,\right] dx\, dy. \end{aligned} \tag{21.58}$$

The parameters λ, μ are known as the *Lamé moduli* of the material, and govern its intrinsic elastic properties. They are measured by performing suitable experiments on a sample of the material. Physically, (21.58) represents the stored (or potential) energy in the body under the prescribed displacement. Nature, as always, seeks the displacement that will minimize the total energy.

To compute the Euler–Lagrange equations, we consider the functional variation

$$h(t) = J[u + t\,f, v + t\,g],$$

in which the individual variations f, g are arbitrary functions subject only to the given homogeneous boundary conditions. If u, v minimize J, then $h(t)$ has a minimum at $t = 0$, and so we are led to compute

$$h'(0) = \langle\, \nabla J\,, \mathbf{f}\,\rangle = \iint_\Omega \,(f\,\nabla_u J + g\,\nabla_v J)\,dx\,dy,$$

which we write as an inner product (using the standard L^2 inner product between vector fields) between the variation $\mathbf{f}$ and the functional gradient $\nabla J = (\,\nabla_u J, \nabla_v J\,)^T$. For the particular functional (21.58), we find

$$h'(0) = \iint_\Omega \left[\, \left(\,\lambda + 2\,\mu\,\right)(u_x\,f_x + v_y\,g_y) + \mu\,(u_y\,f_y + v_x\,g_x) + (\lambda+\mu)\,(u_x\,g_y + v_y\,f_x) \,\right] dx\,dy.$$

We use the integration by parts formula (21.53) to remove the derivatives from the variations f, g. Discarding the boundary integrals, which are used to prescribe the allowable boundary conditions, we find

$$h'(0) = -\iint_\Omega \left(\begin{array}{r} \left[\, (\lambda+2\,\mu)\,u_{xx} + \mu\,u_{yy} + (\lambda+\mu)\,v_{xy} \,\right] f + \\ + \left[\, (\lambda+\mu)\,u_{xy} + \mu\,v_{xx} + (\lambda+2\,\mu)\,v_{yy} \,\right] g \end{array} \right) dx\,dy.$$

The two terms in brackets give the two components of the functional gradient. Setting them equal to zero, we derive the second order linear system of Euler–Lagrange equations

$$(\lambda + 2\mu)\, u_{xx} + \mu\, u_{yy} + (\lambda + \mu)\, v_{xy} = 0, \qquad (\lambda + \mu)\, u_{xy} + \mu\, v_{xx} + (\lambda + 2\mu)\, v_{yy} = 0, \tag{21.59}$$

known as *Navier's equations*, which can be compactly written as

$$\mu\, \Delta \mathbf{u} + (\mu + \lambda)\, \nabla(\nabla \cdot \mathbf{u}) = \mathbf{0} \tag{21.60}$$

for the displacement vector $\mathbf{u} = (\, u, v\,)^T$. The solutions to are the critical displacements that, under appropraite boundary conditions, minimize the potential energy functional.

Since we are dealing with a quadratic functional, a more detailed algebraic analysis will demonstrate that the solutions to Navier's equations are the minimizers for the variational principle (21.58). Although only valid in a limited range of physical and kinematical conditions, the solutions to the planar Navier's equations and its three-dimensional counterpart are successfully used to model a wide class of elastic materials.

In general, the solutions to the Euler–Lagrange boundary value problem are critical functions for the variational problem, and hence include all (smooth) local and global minimizers. Determination of which solutions are geniune minima requires a further analysis of the positivity properties of the second variation, which is beyond the scope of our introductory treatment. Indeed, a complete analysis of the positive definiteness of the second variation of multi-dimensional variational problems is quite complicated, and still awaits a completely satisfactory resolution!

Chapter 22

Nonlinear Partial Differential Equations

The ultimate topic to be touched on in this book is the vast and active field of nonlinear partial differential equations. Leaving aside quantum mechanics, which remains to date an inherently linear theory, most real-world physical systems, including gas dynamics, fluid mechanics, elasticity, relativity, ecology, neurology, thermodynamics, and many more, are modeled by *nonlinear* partial differential equations. Attempts to survey, in such a small space, even a tiny fraction of such an all-encompassing range of phenomena, methods, results, and mathematical developments, are doomed to failure. So we will be content to introduce a handful of prototypical, seminal examples that arise in the study of nonlinear waves and that serve to highlight some of the most significant physical and mathematical phenomena not encountered in simpler linear systems. We will only have space to look at simple one-dimensional models; the far more complicated nonlinear systems that govern our three-dimensional dynamical universe quickly lead one to the cutting edge of contemporary research.

Historically, comparatively little was known about the extraordinary range of behavior exhibited by the solutions to nonlinear partial differential equations. Many of the most fundamental phenomena that now drive modern-day research, including solitons, chaos, stability, blow-up and singularity formation, asymptotic properties, etc., remained undetected or at best dimly perceived in the pre-computer era. The last sixty years has witnessed a remarkable blossoming in our understanding, due in large part to the insight offered by the availability of high performance computers coupled with great advances in the understanding and development of suitable numerical approximation schemes. New analytical methods, new mathematical theories, coupled with new computational algorithms have precipitated this revolution in our understanding and study of nonlinear systems, an activity that continues to grow in intensity and breadth. Each leap in computing power coupled with theoretical advances has led to yet deeper understanding of nonlinear phenomena, while simultaneously demonstrating how far we have yet to go. To make sense of this bewildering variety of methods, equations, and results, it is essential build upon a firm foundation on, first of all, linear systems theory, and secondly, nonlinear algebraic equations and nonlinear ordinary differential equations.

Our presentation is arranged according to the order of the underlying differential equation. First order nonlinear partial differential equations model nonlinear waves and arise in gas dynamics, water waves, elastodynamics, chemical reactions, transport of pollutants, flood waves in rivers, chromatography, traffic flow, and a wide range of biological and ecological systems. One of the most important nonlinear phenomena, with no linear counterpart, is the break down of solutions in finite time, resulting in the formation of

discontinuous shock waves. A striking example is the supersonic boom produced by an airplane that breaks the sound barrier. As in the linear wave equation, the signals propagate along the characteristics, but in the nonlinear case the characteristics can cross each other, precipitating the onset of a shock. The characterization of the shock dynamics requires additional physical information, in the form of a conservation law, that supplements the original partial differential equation.

Parabolic second order partial differential equations govern nonlinear diffusion processes, including thermodynamics, chemical reactions, dispersion of pollutants, and population dynamics. The simplest and most well understood is Burgers' equation, which can, surprisingly, be linearized by transforming it to the heat equation. This accident provides an essential glimpse into the world of nonlinear diffusion processes. In the limit, as the diffusion or viscosity tends to zero, the solutions to Burgers' equation tend to the shock wave solutions to the limiting first order dispersionless equation, and thus provides an alternate mechanism for unraveling shock dynamics.

Third order partial differential equations arise in the study of dispersive wave motion, including water waves, plasma waves, waves in elastic media, and elsewhere. We first treat the basic linear dispersive model, comparing and contrasting it with the hyperbolic models we encountered earlier in this text. The distinction between group and wave velocity — observed when, for instance, surface waves propagate over water — is developed. Finally, we introduce the remarkable Korteweg–deVries equation, which serves as a model for waves in shallow water, waves in plasmas, and elsewhere. Despite its intrinsic nonlinearity, it supports stable localized traveling wave solutions, known as *solitons*, that, remarkably, maintain their shape even under collision. The Korteweg–deVries equation is the prototypical example of an integrable system, and this discovery in the mid 1960's inaugurated intense and ongoing research in the remarkable physical models that exhibit integrability, a development that has had many ramifications in both pure and applied mathematics.

22.1. Nonlinear Waves and Shocks.

Before attempting to tackle any nonlinear partial differential equations, we must carefully review the solution to the simplest linear first order partial differential equation.

Linear Transport and Characteristics

The *transport equation*

$$u_t + c\,u_x = 0, \tag{22.1}$$

is so named because it models the transport of, say, a pollutant in a uniform fluid flow. Let us begin by assuming that the *wave speed* c is constant. According to Proposition 14.8, every solution is constant along the characteristic lines of slope

$$\frac{dx}{dt} = c, \qquad \text{namely} \qquad x - ct = \text{constant}. \tag{22.2}$$

As a consequence, the solutions are *traveling waves* of the form

$$u(t,x) = p(x - ct), \tag{22.3}$$

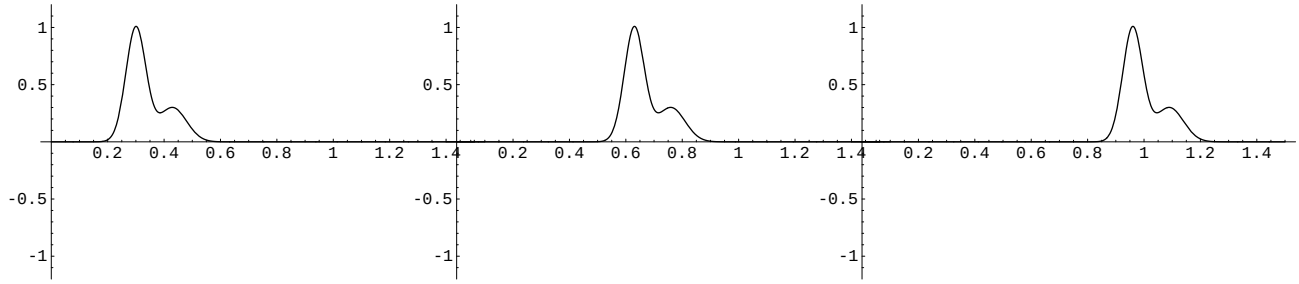


Figure 22.1. Traveling Wave.

where $p(\xi)$ is an arbitrary function of the *characteristic variable* $\xi = x - ct$. To a stationary observer, the solution (22.3) appears as a wave of unchanging form moving at velocity c. When $c > 0$, the wave translates to the right, as illustrated in Figure 22.1. When $c < 0$, the wave moves to the left, while $c = 0$ corresponds to a permanent wave form that remains fixed at its original location.

Slightly more complicated, but still linear, is the non-uniform transport equation

$$u_t + c(x)\, u_x = 0, \tag{22.4}$$

where the wave velocity $c(x)$ depends upon the spatial position. This equation models unidirectional waves propagating through a non-uniform, but static medium. Generalizing the construction (22.2), we define a *characteristic curve* to be a solution to the autonomous ordinary differential equation

$$\frac{dx}{dt} = c(x). \tag{22.5}$$

Thus, unlike the constant velocity version, the characteristics are no longer necessarily straight lines. Nevertheless, the preceding observation remains valid:

Proposition 22.1. *Solutions to the linear transport equation* (22.4) *are constant on its characteristic curves.*

Proof: Let $x(t)$ be a characteristic curve, i.e., a solution to (22.5), parametrized by the time t. Let $h(t) = u(t, x(t))$ be the value of the solution at the point $(t, x(t))$ on the given characteristic curve. Our goal is to prove that $h(t)$ is a constant function of t, and, as usual, this is done by proving that its derivative is identically zero. To differentiate $h(t)$, we invoke the chain rule:

$$\frac{dh}{dt} = \frac{d}{dt}\, u(t, x(t)) = \frac{\partial u}{\partial t}(t, x(t)) + \frac{dx}{dt}\, \frac{\partial u}{\partial x}(t, x(t)) = \frac{\partial u}{\partial t}(t, x(t)) + c(x(t))\, \frac{\partial u}{\partial x}(t, x(t)) = 0.$$

We replaced dx/dt by $c(x)$ since we are assuming that $x(t)$ is a characteristic curve, and hence satisfies (22.5). The final combination of derivatives is zero whenever u solves the transport equation (22.4). *Q.E.D.*

Since the characteristic curve differential equation (22.5) is autonomous, it can be immediately solved:

$$b(x) \equiv \int \frac{dx}{c(x)} = t + k, \tag{22.6}$$

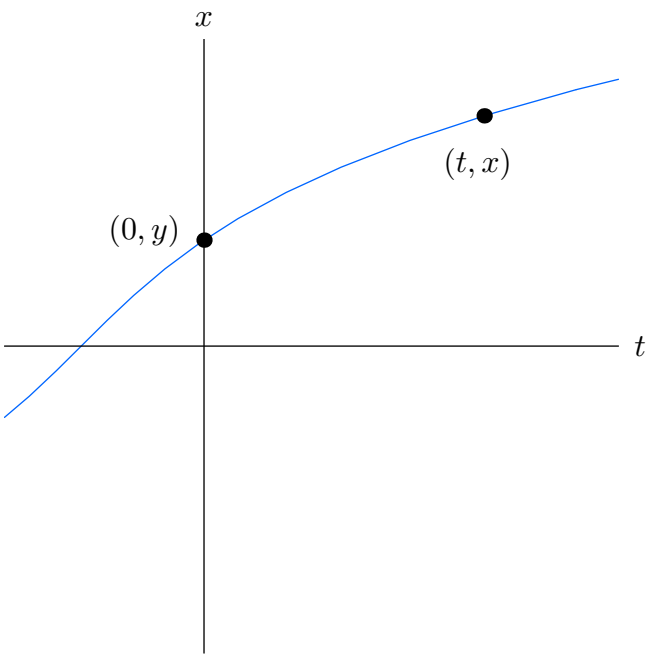


Figure 22.2. Characteristic Curve.

where k is the constant of integration. Thus, the characteristic curves are "parallel", each being a translate of the graph of $t = b(x)$ in the direction of the t axis. The characteristic curves are therefore defined by the formula $x = g(t + k)$, where $g = b^{-1}$ is the inverse function. (See Section 20.1 for full details.)

Observe that the characteristic curves are the level sets of the *characteristic variable* $\xi = b(x) - t$. As a consequence, any function which is constant along the characteristic curves depends only on the value of the characteristic variable at each point, and hence takes the form

$$u(t, x) = p(b(x) - t) \tag{22.7}$$

for some function $p(\xi)$. In other words, the characteristic curves are the common level curves of *all* solutions to the transport equation. It is easy to check directly that, provided $b(x)$ is defined by (22.6), $u(t, x)$ solves the partial differential equation (22.4) for *any* choice of function $p(\xi)$.

To find the solution that satisfies the prescribed initial conditions

$$u(0, x) = f(x) \tag{22.8}$$

we merely substitute the general solution formula (22.7). This leads to the equation

$$p(b(x)) = f(x), \qquad \text{and, therefore,} \qquad p(\xi) = f \circ b^{-1}(\xi) = f\big(g(\xi) \big).$$

The resulting solution formula has a simple graphical interpretation: to find its value $u(t, x)$ at a given point, we look at the characteristic curve passing through (t, x). If this curve intersects the x axis at the point $(0, y)$, then $u(t, x) = u(0, y) = f(y)$. The construction is illustrated in Figure 22.2. Incidentally, if the characteristic curve through (t, x) doesn't intersect the x axis, the solution value $u(t, x)$ is not prescribed by the initial data.

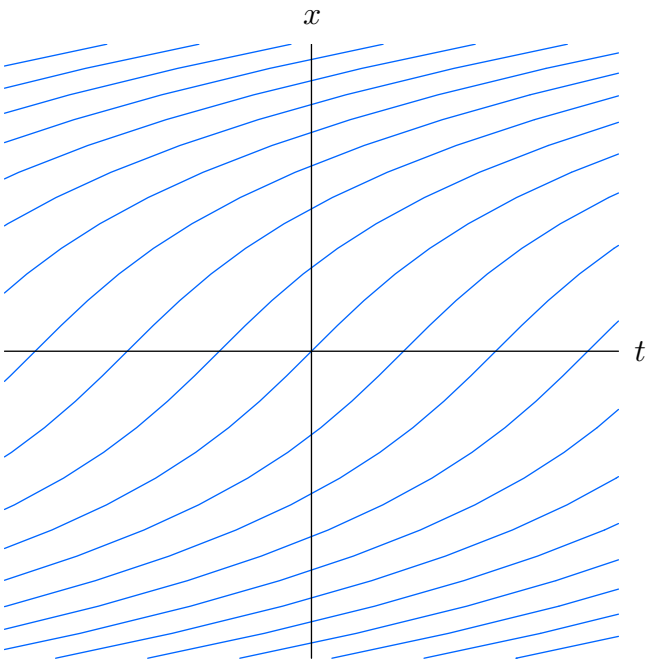


Figure 22.3. Characteristic Curves for $u_t + \dfrac{1}{x^2+1}\, u_x = 0$.

Example 22.2. Let us solve the particular transport equation

$$\frac{\partial u}{\partial t} + \frac{1}{x^2+1}\,\frac{\partial u}{\partial x} = 0 \tag{22.9}$$

by the method of characteristics. According to (22.5), the characteristic curves satisfy the first order ordinary differential equation

$$\frac{dx}{dt} = \frac{1}{x^2+1}\,.$$

Separating variables and integrating, we find

$$\int (x^2+1)\,dx = \tfrac{1}{3}\,x^3 + x = t + k,$$

where k is the integration constant. Some of the resulting characteristic curves are plotted in Figure 22.3.

The characteristic variable is $\xi = \frac{1}{3}\,x^3 + x - t$, and hence the general solution to the equation takes the form

$$u = p\big(\tfrac{1}{3}\,x^3 + x - t\,\big),$$

where $p(\xi)$ is an arbitrary function. A typical solution, corresponding to initial data

$$u(t,0) = \frac{1}{1+(x+2.75)^2}\,,$$

is plotted at times $t = 0, 2, 5, 10, 25, 50$ in Figure 22.4. The fact that the characteristic curves are not straight means that, although the solution remains constant along each

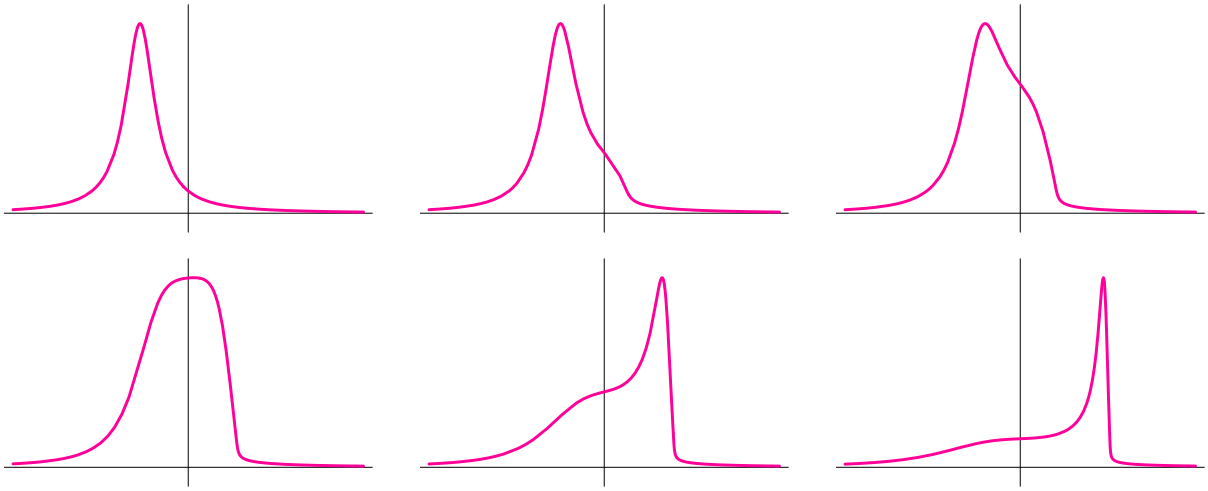


Figure 22.4. Solution to $u_t + \dfrac{1}{x^2+1}\, u_x = 0$.

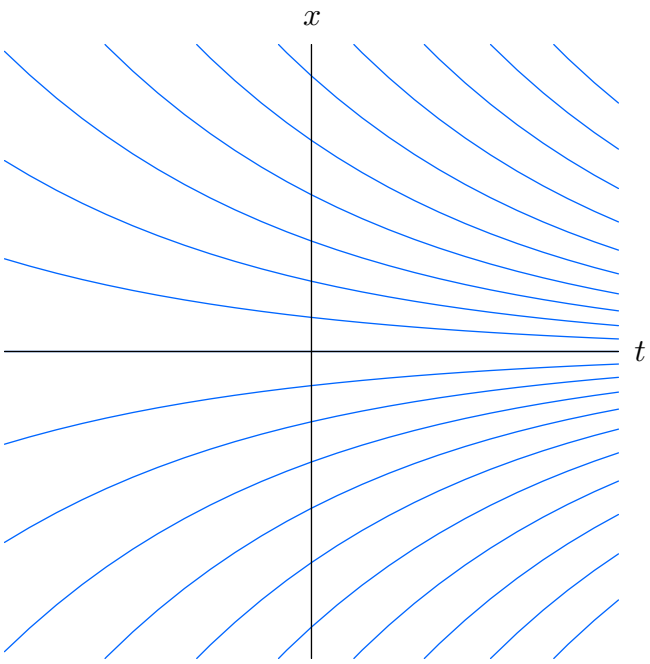


Figure 22.5. Characteristic Curves for $u_t - x\, u_x = 0$.

individual curve, a stationary observer will witness a dynamically changing profile as the wave moves along through the non-uniform medium. The wave speeds up as it aapproaches the origin, and then slows back down once it passes and moves off to the right. As a result, we observe the wave spreading out as it approaches the origin, and then contracting as it moves off to the right.

Example 22.3. Consider the equation

$$u_t - x\, u_x = 0. \tag{22.10}$$

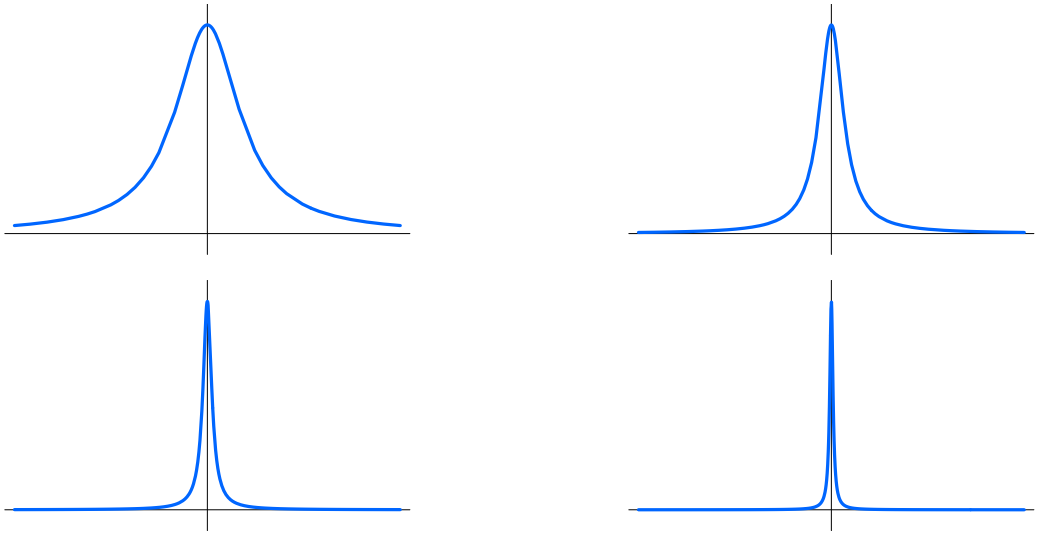


Figure 22.6. Solution to $u_t - x\,u_x = 0$.

In this case, the characteristic curves are the solutions to

$$\frac{dx}{dt} = -x, \qquad \text{and so} \qquad x\,e^t = k, \tag{22.11}$$

where k is the constant of integration; see Figure 22.5. It is easier to adopt $\xi = x\,e^t$ as the characteristic variable here, noting that its level sets are the characteristic curves. The solution therefore takes the form

$$u = p(x\,e^t), \tag{22.12}$$

where $p(\xi)$ is an arbitrary function of $\xi = x\,e^t$. Given the initial data

$$u(0,x) = f(x), \qquad \text{the resulting solution is} \qquad u = f(x\,e^t).$$

For example, the solution

$$u(t,x) = \frac{1}{(x\,e^t)^2 + 1} = \frac{e^{-2\,t}}{x^2 + e^{-2\,t}}$$

corresponding to initial data $u(t,0) = f(x) = (x^2+1)^{-1}$ is plotted in Figure 22.6 at times $t = 0, 1, 2, 3$. Note that since the characteristic curves all converge on the t axis, the solution becomes more and more concentrated at the origin. In the limit, it converges to the function that is zero everywhere except for the value $u(t,0) \equiv 1$ at the origin. *Warning*: The limit is *not* a delta function, since its value at $x = 0$ remains bounded.

A Nonlinear Transport Equation

Perhaps the simplest possible nonlinear partial differential equations is the *nonlinear transport equation*

$$u_t + u\,u_x = 0. \tag{22.13}$$

first systematically studied by Poisson and Riemann in the early nineteenth century. Since it appears in so many applications, this equation appears in the literature under a variety of names, including the Riemann equation, the inviscid Burgers' equation, and the dispersionless Korteweg–deVries equation. It and its multi-dimensional and multi-component generalizations play a crucial role in the modeling of gas dynamics, traffic flow, flood waves in rivers, chromatography, chemical reactions, and other areas; see [**182**].

The first order partial differential equation (22.13) has the form of a transport equation, whose wave velocity $c = u$ depends, not on the position x, but rather on the size of the disturbance. Larger waves move faster, and overtake smaller waves. Waves of elevation, where $u > 0$, move to the right, while waves of depression, where $u < 0$, move to the left.

Fortunately, the method of characteristics that was developed for linear wave equations also works in the present nonlinear situation and leads to a complete solution. Mimicking our previous construction, (22.5), let us define a *characteristic curve* of the nonlinear wave equation (22.13) to be a solution to the ordinary differential equation

$$\frac{dx}{dt} = u(t,x). \tag{22.14}$$

As such, the characteristics depend upon the solution u, which, in turn, is based on the characteristic variable. So we appear to be trapped in a circular argument. The resolution of the apparent conundrum is to observe that, as in the linear case, the solution $u(t,x)$ remains constant along its characteristics, and this fact will allow us to simultaneously specify both.

To prove this claim, suppose that $x = x(t)$ parametrizes a characteristic curve. We need to show that the function $h(t) = u(t,x(t))$, which is obtained by evaluating the solution along the curve, is constant. As usual, constancy is proved by checking that its derivative is identically zero. Invoking the chain rule, and then (22.14), we deduce that

$$\frac{dh}{dt} = \frac{d}{dt}\,u(t,x(t)) = \frac{\partial u}{\partial t}\,(t,x(t)) + \frac{dx}{dt}\,\frac{\partial u}{\partial x}(t,x(t)) = \frac{\partial u}{\partial t}\,(t,x(t)) + u(t,x(t))\,\frac{\partial u}{\partial x}(t,x(t)) = 0.$$

The final expression vanishes because u is assumed to solve the wave equation (22.13) at all values of (t,x), including those on the curve $(t,x(t))$. This verifies our claim that $h(t)$ is constant, and so the solution u is constant on the characteristic curve. This has the implication that the right hand side of equation (22.14) is a constant whenever $x = x(t)$ defines a characteristic curve, and so the derivative dx/dt is a constant — namely the value of u on the curve. In this manner, we arrive at the key deduction that the characteristic curve must be a *straight line*

$$x = u\,t + k, \tag{22.15}$$

whose *characteristic slope* u equals the value assumed by the solution u on it. The larger u is, the steeper the characteristic line, and the faster that part of the wave travels.

The corresponding characteristic variable $\xi = x - t\,u$ depends upon the solution, which can now be written in implicit form

$$u = f(x - t\,u), \tag{22.16}$$

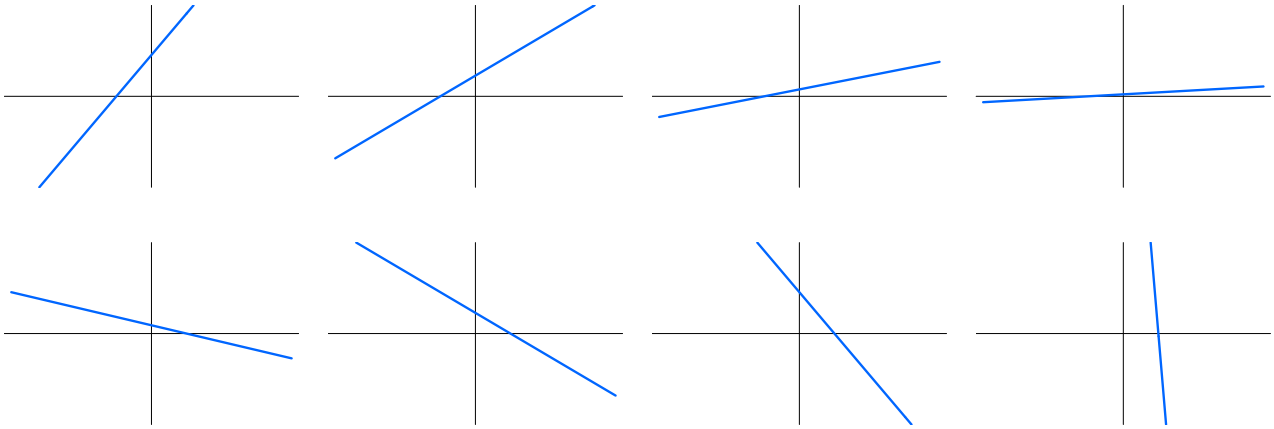


Figure 22.7. Two Solutions to $u_t + u\,u_x = 0$.

where $f(\xi)$ is an arbitrary function of the characteristic variable. The solution $u(t,x)$ can be found by solving the algebraic equation (22.16). For example, if

$$f\xi) = \alpha\,\xi + \beta$$

is an affine function, with α, β constant, then

$$u = \alpha(x - t\,u) + \beta, \qquad \text{and hence} \qquad u(t,x) = \frac{\alpha\,x + \beta}{1 + \alpha\,t} \tag{22.17}$$

is the corresponding solution to the nonlinear transport equation. At each fixed t, the graph of the solution is a straight line. If $\alpha > 0$, the solution flattens out as $t \to \infty$. On the other hand, if $\alpha < 0$, the straight line rapidly steepens to vertical as t approaches the critical time $t_\star = -1/\alpha$, at which point the solution ceases to exist — it is said to "blow up". In Figure 22.7, we graph the solution with $\alpha = 1,\ \beta = .5$, when $t = 0, 1, 5, 20$ on the top row, and $\alpha = -.2,\ \beta = .1$, at times $t = 0, 3, 4, 4.9$ on the bottom row. In the second case, the solution becomes vertical as $t \to 5$ and then ceases to exist.

In general, to construct the solution $u(t,x)$ to the initial value problem

$$u(0,x) = f(x), \tag{22.18}$$

we note that, at $t = 0$, the implicit solution formula (22.16) reduces to $u(0,x) = f(x)$. Thus, the function f coincides with the initial data. However, because (22.16) is an implicit equation for $u(t,x)$, it is not immediately evident

(a) whether it can be solved to give a well-defined value for $u(t,x)$, and,

(b) even granted this, how to describe the solution's qualitative features and dynamical behavior.

A more instructive and revealing strategy is based on the following geometrical construction, inspired by the linear version appearing in Figure 22.2. Through each point $(0,y)$ on the x axis, draw the characteristic line

$$x = t\,f(y) + y \tag{22.19}$$

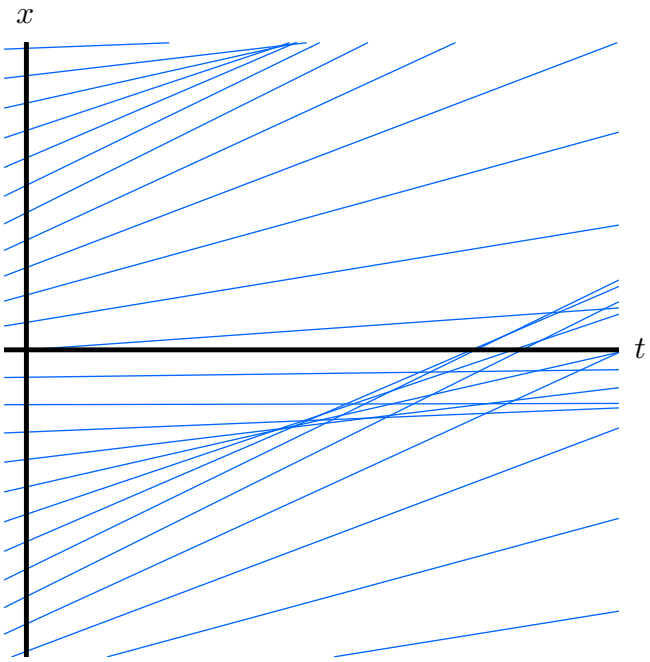


Figure 22.8. Characteristic Lines for $f(x) = \frac{1}{4}\sin(1.8\,x - .8)$.

whose slope, namely $f(y) = u(0, y)$, equals the value of the initial data at that point. According to the preceding argument, the solution will have the same value on the entire characteristic line (22.19), and so

$$u(t, t\,f(y) + y) = f(y).$$

For example, if $f(y) = y$, then $u(t, x) = y$ whenever $x = t\,y + y$; eliminating y, we recover $u(t, x) = x/(t + 1)$, which agrees with one of our straight line solutions (22.17).

Now, the trouble with our construction is immediately apparent from the illustrative Figure 22.8. Any two characteristic lines that are not parallel must cross each other somewhere. The value of the solution must equal to the slope of the characteristic line, and so, at the crossing point, the solution is required to assume two *different* values, one corresponding to each line. Something is clearly amiss, and we need to study the resulting solutions in more depth.

It turns out that there are three basic scenarios. The first, trivial case is when all the characteristic lines are parallel and so the difficulty does not arise. In this case, they all have the same slope, say c, which means that the solution has the same value on each one. Therefore, $u(t, x) \equiv c$ is a trivial constant solution.

The next simplest case occurs when the initial data $f(x)$ is everywhere increasing, so $f(x) \le f(y)$ whenever $x \le y$, which is assured if its derivative is never negative: $f'(x) \ge 0$. In this case, as in sketched in Figure 22.9, the characteristic lines emanating from the x axis fan out into the right half plane, and so never cross each other when $t \ge 0$. Each point (t, x) for $t \ge 0$ lies on a unique characteristic line, and the value of the solution at (t, x) is equal to the slope of the line. Consequently, the solution is well-defined at all future times. Physically, such solutions represent *rarefaction waves*, which gradually spread out as time

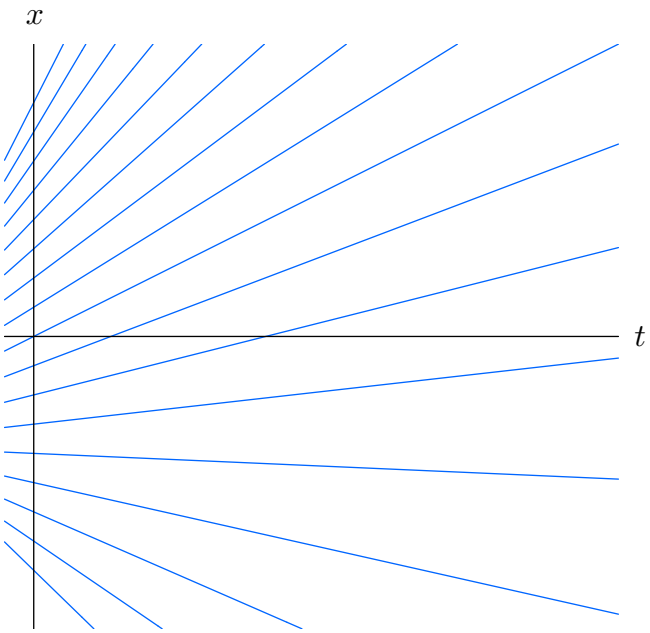


Figure 22.9. Characteristic Lines for a Rarefaction Wave.

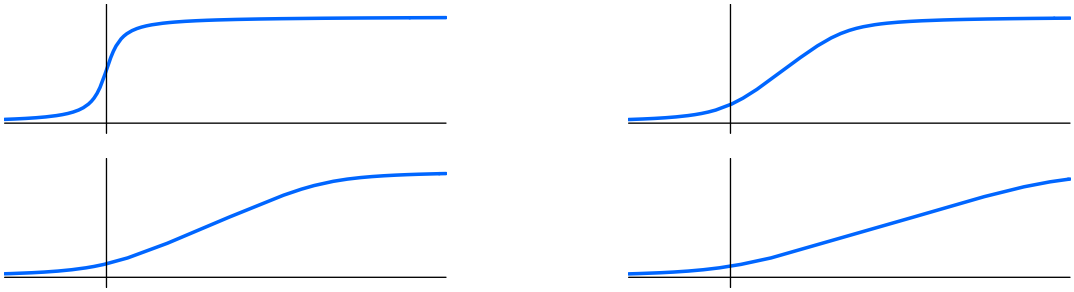


Figure 22.10. Rarefaction Wave.

progresses. A typical example, corresponding to initial data

$$u(0,x) = \tan^{-1} 3x + \frac{\pi}{2},$$

is plotted in Figure 22.10 at successive times $t = 0, 1, 2, 3$. Note how the slope of the solution gradually diminishes as the rarefaction wave spreads out.

The more interesting case is when $f'(x) < 0$. Now some of the characteristic lines starting at $t = 0$ will cross at some point in the future. If (t, x) lies on two or more distinct characteristic lines, the value of the solution $u(t, x)$, which should equal the characteristic slope, is no longer uniquely determined. Although one might be tempted to deal with such multiply-valued solutions in a purely mathematical framework, from a physical standpoint this is unacceptable. The solution $u(t, x)$ is supposed to represent a physical quantity, e.g., density, velocity, pressure, etc., and must therefore assume a unique value at each point. The mathematical model has broken down, and fails to agree with the physical reality.

Before confronting this difficulty, let us first, from a theoretical standpoint, try to understand what happens if we were to continue the solution as a multiply-valued function.

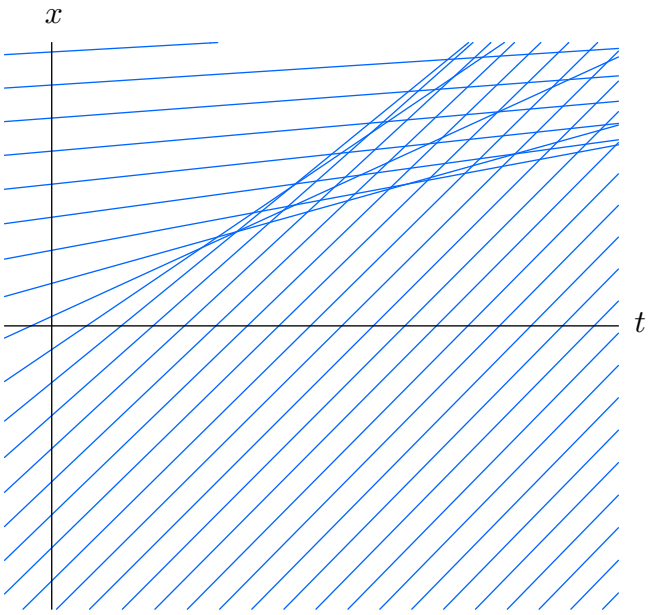


Figure 22.11. Characteristics for a Shock Wave.

To be specific, consider the initial data

$$u(0,x) = \frac{\pi}{6} - \frac{1}{3}\tan^{-1} x, \tag{22.20}$$

appearing in the first plot in Figure 22.12. The corresponding characteristic lines are sketched in Figure 22.11. Initially, they do not cross, and the solution remains a well-defined, single-valued function. However, eventually one reaches a critical time, $t = t_\star > 0$, when the first two characteristic lines cross each other. Subsequently, a wedge-shaped region appears in the (t,x) plane, consisting of points which lie on the intersection of three different characteristic lines with different slopes; at such points, the solution achieves three distinct values. Outside the wedge, the points only belong to a single characteristic line, and the solution remains single-valued. (The boundary of the wedge consists of points where only two characteristic lines cross.)

To fully appreciate what is going on, look now at the sequence of pictures of the multiply-valued solution at successive times in Figure 22.12. Since the initial data is positive, $f(x) > 0$, all the characteristic slopes are positive. As a consequence, all the points on the solution curve will move to the right, at a speed equal to their height. Since the initial data is a decreasing function, points lying to the left will move faster than those on the right, and eventually overtake them. Thus, as time passes, the solution steepens. At the critical time $t_\star$ when the first two characteristic lines cross, say at $x_\star$, the tangent to the solution curve has become vertical:

$$\frac{\partial u}{\partial x}(t, x_\star) \longrightarrow \infty \qquad \text{as} \qquad t \longrightarrow t_\star.$$

Afterwards, the solution graph no longer represents a single-valued function; its overlapping lobes lie over points (t,x) in the aforementioned wedge.

Figure 22.12. Multiply–Valued Solution.

The critical time $t_\star$ can be determined from the implicit solution formula (22.16). Indeed, if we differentiate with respect to x, we find

$$\frac{\partial u}{\partial x} = \frac{\partial}{\partial x} f(x - t\,u) = f'(\xi)\left(1 - t\,\frac{\partial u}{\partial x}\right), \qquad \text{where} \qquad \xi = x - t\,u$$

is the characteristic variable, which is constant along the characteristic lines. Solving,

$$\frac{\partial u}{\partial x} = \frac{f'(\xi)}{1 + t\,f'(\xi)}\,.$$

Therefore, the slope blows up,

$$\frac{\partial u}{\partial x} \longrightarrow \infty, \qquad \text{as} \qquad t \longrightarrow -\frac{1}{f'(\xi)}\,.$$

In other words, if the initial data has negative slope at position x, so $f'(x) < 0$, then the solution along the characteristic line emanating from the point $(0, x)$ will break down at the time $-1/f'(x)$. As a consequence, the earliest critical time is

$$t_\star = \min\left\{ -\frac{1}{f'(x)} \;\middle|\; f'(x) < 0 \right\}. \tag{22.21}$$

For instance, for the particular initial configuration (22.20) represented by the pictures,

$$f'(x) = -\frac{1}{3(1 + x^2)}\,, \qquad \text{and so the critical time is} \qquad t_\star = \min(3(1 + x^2)) = 3.$$

Now, while mathematically plausible, such a multiply-valued solution is physically untenable. So what happens after the critical time $t_\star$? One needs to choose which of the

possible solution values at each point (t, x) contained in the wedge is physically appropriate. Indeed, the mathematics by itself is incapable of specifying how to continue the solution past the critical time at which the characteristics begin to cross. We therefore must return to the underlying physics, and ask what sort of phenomenon are we trying to model. The most instructive is to view the differential equation as a simple model of compressible fluid flow in a single space variable, e.g., motion of gas in a long pipe. If we push a piston down the end of a long pipe then the gas will move ahead of the piston and thereby be compressed. However, if the piston moves too rapidly, the gas piles up on top of itself, and a shock wave forms and propagates down the pipe. Mathematically, the shock is represented by a discontinuity where the solution abruptly changes value.

Conservation Laws and Shocks

One way to resolve our mathematical dilemma relies on the fact that the partial differential equation takes the form of a conservation law, in accordance with the following definition[†].

Definition 22.4. A *conservation law* is an equation of the form

$$\frac{\partial T}{\partial t} + \frac{\partial X}{\partial x} = 0. \tag{22.22}$$

The functions T and X are known, respectively, as the *conserved density* and associated *flux*.

In the simplest situations, the conserved density $T(t, x, u)$ and flux $X(t, x, u)$ depend on the time t, the position x, and the solution $u(t, x)$ to the physical system. (Higher order conservation laws, which also depend upon derivatives of u, will appear in the final section.) We can clearly rewrite the nonlinear transport equation (22.13) in the following conservation law form:

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\tfrac{1}{2} u^2 \right) = 0, \tag{22.23}$$

where the conserved density and flux are, respectively,

$$T = u, \qquad\qquad X = \tfrac{1}{2} u^2.$$

The reason for calling (22.22) a conservation law comes from the following observation.

Proposition 22.5. *Given a conservation law* (22.22),

$$\frac{d}{dt} \int_a^b T \, dx = - \, X \, \Big|_{x=a}^{b}. \tag{22.24}$$

[†] Here we describe the one-dimensional situation. See Exercise ■ for conservation laws for n-dimensional dynamics.

The proof of (22.24) is immediate — assuming sufficient smoothness that allows one to bring the derivative inside the integral sign, and then invoking the Fundamental Theorem of Calculus:

$$\frac{d}{dt}\int_a^b T\,dx = \int_a^b \frac{\partial T}{\partial t}\,dx = -\int_a^b \frac{\partial X}{\partial x}\,dx = -\,X\,\Big|_{x=a}^{b}.$$

Formula (22.24) says that the rate of change of the integrated density over an interval depends only on the flux through its endpoints. In particular, if there is no net flux into or out of the interval, then the integrated density is *conserved*, meaning that it remains constant over time. All physical conservation laws — mass, momentum, energy, and so on — for systems governed by partial differential equations are of this form. (For ordinary differential equations, conservation laws coincide with first integrals, as discussed in Section 20.3.)

For the transport equation (22.23), the integrated conservation law (22.24) takes the specific form

$$\frac{d}{dt}\int_a^b u(t,x)\,dx = \tfrac{1}{2}\big[\,u(t,a)^2 - u(t,b)^2\,\big]. \tag{22.25}$$

Viewing the equation as a model for, say, compressible fluid flow in a pipe, the integral on the left hand side represents the total mass of the fluid contained in the interval $[a,b]$. The right hand side represents the mass flux *into* the interval through its two endpoints, and thus the conservation equation (22.25) is the mathematical formalization of basic mass conservation — mass is neither created nor destroyed, but can only enter a region as a flux through its boundary. In particular, if there is zero mass flux, then we deduce conservation of the total mass.

With this in hand, let us return to the physical context of the nonlinear transport equation. We will assume that mass conservation continues to hold even within a shock, which, from a purely molecular standpoint, makes eminent physical sense. By definition, a *shock* is a jump discontinuity in the solution $u(t,x)$. Suppose that, at time t, a shock occurs at position $x = s(t)$. We require[†] that both the left and right hand limits

$$u_-(t) = u(t,s(t)^-) = \lim_{x\to s(t)^-} u(t,x), \qquad u_+(t) = u(t,s(t)^+) = \lim_{x\to s(t)^+} u(t,x),$$

of the solution on either side of the shock discontinuity are well defined. Let us further assume that, in time, the shock $x = s(t)$ follows a smooth — meaning C^1 — path. Now, referring to Figure 22.13, Consider a small time interval, from t to $t + \Delta t$. During this time, the shock moves from position $a = s(t)$ to position $b = s(t+\Delta t)$. The total mass contained in the interval $[a,b]$ at time t, before the shock has passed through, is

$$m(t) = \int_a^b u(t,x)\,dx \approx \overline{u}_+(t)\,(b-a) = \overline{u}_+(t)\,\big[\,s(t+\Delta t) - s(t)\,\big],$$

[†] With more analytical work, [**182**], the listed assumptions can all be rigorously justified.

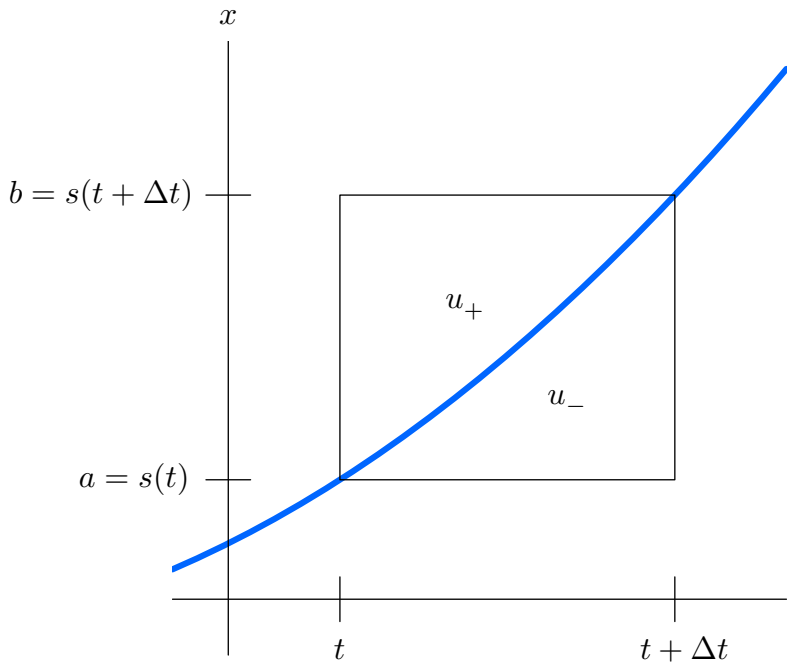


Figure 22.13. Conservation of Mass Near a Shock.

where $\overline{u}_+(t)$ is the average value of $u(t,x)$ over the interval. After the shock has passed, the total mass has become

$$m(t+\Delta t) = \int_a^b u(t+\Delta t, x)\, dx \approx \overline{u}_-(t)\,(b-a) = \overline{u}_-(t)\,\big[\, s(t+\Delta t) - s(t)\,\big],$$

where $\overline{u}_-(t)$ refers to the average value of $u(t+\Delta t, x)$ over the same interval. In the limit as $\Delta t \to 0$, the point $b = s(t+\Delta t) \longrightarrow s(t) = a$, and hence the averages

$$\lim_{\Delta t \to 0} \overline{u}_+(t) = u_+(t), \qquad\qquad \lim_{\Delta t \to 0} \overline{u}_-(t) = u_-(t),$$

tend to the limiting solution values on the right and left hand sides of the shock discontinuity. Thus, the limiting rate of change in mass across the shock at time t is

$$\begin{aligned} \frac{dm}{dt} &= \lim_{\Delta t \to 0} \frac{m(t+\Delta t) - m(t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \big[\, \overline{u}_-(t) - \overline{u}_+(t)\,\big]\, \frac{s(t+\Delta t) - s(t)}{\Delta t} = \big[\, u_-(t) - u_+(t)\,\big]\, \frac{ds}{dt}\,, \end{aligned}$$

which is the product of the shock speed times minus the jump magnitude at the shock discontinuity. On the other hand, at any $t < \tau < t + \Delta t$, the mass flux into the interval $[a,b]$ is, according to the right hand side of (22.25),

$$\tfrac{1}{2}\big[\, u(\tau,a)^2 - u(\tau,b)^2\,\big] \;\longrightarrow\; \tfrac{1}{2}\big[\, u_-(t)^2 - u_+(t)^2\,\big] \qquad \text{as} \qquad \Delta t \longrightarrow 0.$$

For conservation of mass to hold across the shock, the limiting value of the rate of change in mass must equal the limiting mass flux,

$$\big[\, u_-(t) - u_+(t)\,\big]\, \frac{ds}{dt} = \tfrac{1}{2}\big[\, u_-(t)^2 - u_+(t)^2\,\big]\,,$$

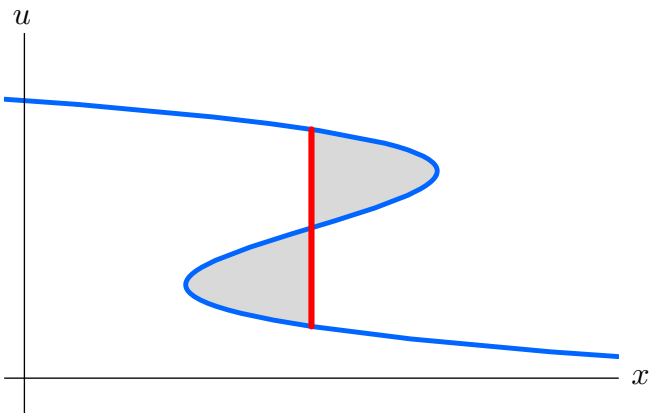


Figure 22.14. Equal Area Rule.

from which we discover the *Rankine–Hugoniot condition*

$$\frac{ds}{dt} = \frac{1}{2}\,\frac{u_-(t)^2 - u_+(t)^2}{u_-(t) - u_+(t)} = \frac{u_-(t) + u_+(t)}{2}\,. \tag{22.26}$$

So, to maintain conservation of mass, the speed of the shock must equal the average of the solution values on either side.

A shock appears when one or more characteristic lines cross. For this to occur, characteristics to the left of the shock must have larger slope (or speed), while those to the right must have smaller slope. Since the shock speed is the average of the two characteristic slopes, this means

$$u_-(t) > \frac{ds}{dt} = \frac{u_-(t) + u_+(t)}{2} > u_+(t). \tag{22.27}$$

While it is theoretically possible to construct a shock solution to (22.13) that maintains the Rankine–Hugoniot constraint (22.26) but violates (22.27), such solutions are excluded on physical grounds, in that they violate causality, [**103**], which requires that characteristics are only allowed to enter shocks, not leave, and, furthermore, are not stable under small perturbations, [**182**]. The dynamics of shock wave solutions is then prescribed by the Rankine–Hugoniot and causality conditions (22.26, 27).

How does one determine the motion of the shock in practice? The answer is beautifully simple. Since the total mass, which at time t is the area under the curve $u(t,x)$, must be conserved, one merely draws the vertical shock line where the areas of the two lobes in the multiply-valued solution are equal, as in Figure 22.14. This *Equal Area Rule* ensures that the total mass of the shock solution matches that of the original (why?), as required by the physical conservation law.

Example 22.6. An illuminating special case is when the initial data has the form of a step function with a single jump discontinuity at the origin:

$$u(0,x) = f(x) = a + b\,\sigma(x) = \begin{cases} a, & x < 0, \\ b, & x > 0. \end{cases} \tag{22.28}$$

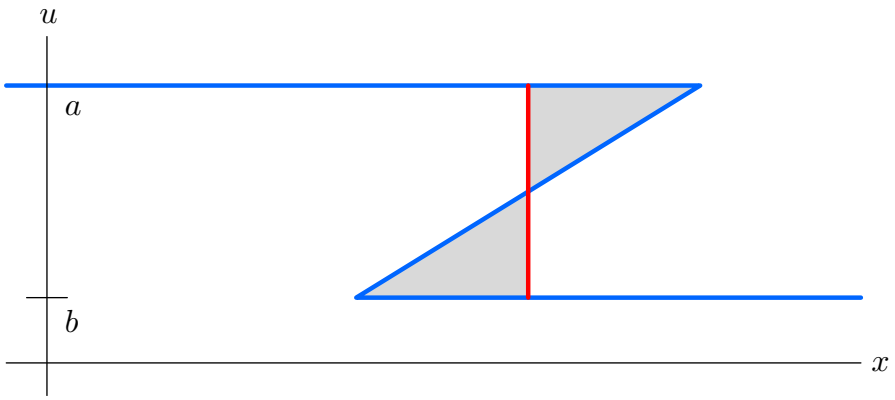


Figure 22.15. Multiply–Valued Step Wave.

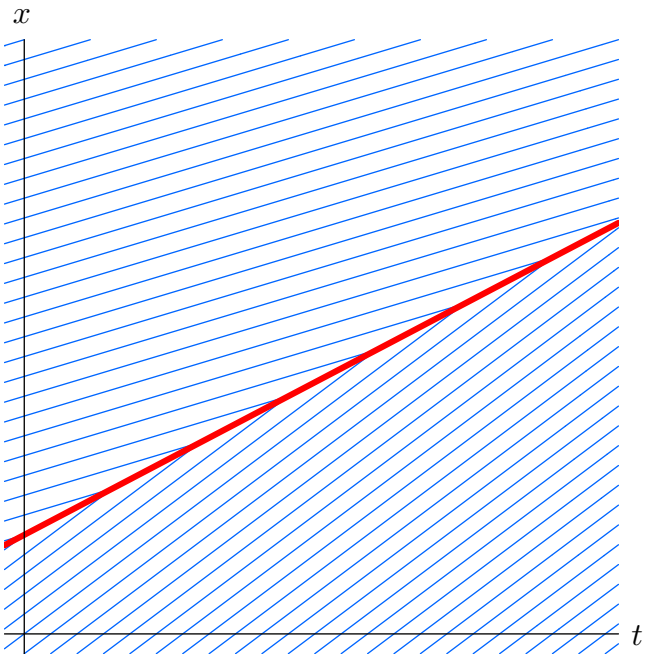


Figure 22.16. Characteristic Lines for the Step Wave Shock.

If† $a > b > 0$, then the initial data is already in the form of a shock wave. For $t > 0$, he mathematical solution constructed by continuing along the characteristic lines is multiply-valued in the region $bt < x < at$, where it assumes both values a and b; see Figure 22.15 . The Equal Area Rule tells us to draw the shock line halfway along, at $x = \frac{1}{2}(a+b)t$, in order that the two triangles have the same area. Therefore, the shock moves with speed $c = \frac{1}{2}(a+b)$ equal to the average of the two speeds at the jump, and so this particular

† Cases where a or b are negative are left to the exercises.

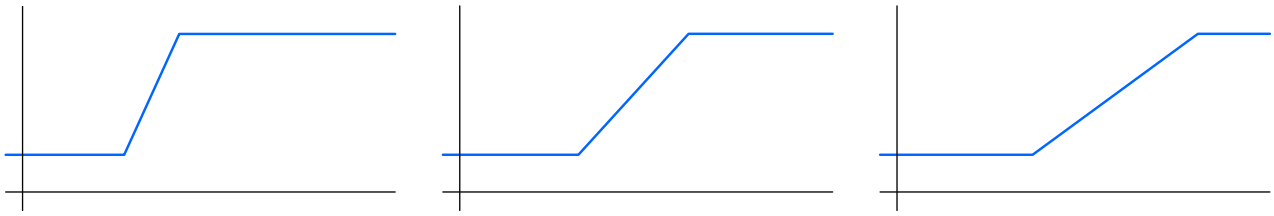


Figure 22.17. Piecewise Affine Rarefaction Wave.

shock wave solution is

$$u(t,x) = a + b\,\sigma(x - ct) = \begin{cases} a, & x < ct, \\ b, & x > ct, \end{cases} \qquad \text{where} \qquad a > c = \frac{a+b}{2} > b. \tag{22.29}$$

A graph of the characteristic lines appears in Figure 22.16.

By way of contrast, suppose $0 < a < b$, so the initial data has a jump upwards. In this case, the characteristic lines diverge from the initial discontinuity, and the mathematical solution is not specified at all in the wedge-shaped region $at < x < bt$. Now our task is to decide how to connect the two regions where the solution is well-defined. The simplest connection is an affine function, i.e., a straight line. Indeed, a simple modification of the rational solution (22.17) produces the function

$$u(t,x) = \frac{x}{t}\,,$$

which not only solves the differential equation, but also has the required values $u(t, at) = a$, and $u(t, bt) = b$ at the two edges of the wedge. The resulting solution is the piecewise affine *rarefaction wave*

$$u(t,x) = \begin{cases} a, & x \le at, \\ x/t, & at \le x \le bt, \\ b, & x \ge bt, \end{cases} \tag{22.30}$$

which is graphed in Figure 22.17. In fact, it can be shown, [**103**], that this is the only solution that preserves the causality condition (22.27).

These two prototypical solutions epitomize the basic phenomenon modeled by the nonlinear transport equation — *rarefaction waves* where the solution spreads out as time progresses, and *compression waves*, that progressively steepen and eventually break into a shock discontinuity. , that correspond to regions where $f'(x) > 0$, and *compression waves*, for $f'(x) < 0$, where the solution contracts and eventually breaks into a shock discontinuity. Anyone caught in a traffic jam recognizes the compression waves, where the vehicles are bunched together and almost stationary, while the interspersed rarefaction waves correspond to freely moving traffic. (An intelligent driver will take advantage of the rarefaction waves moving through the jam to switch lanes!) The familiar, frustrating traffic jam phenomenon is an intrinsic effect of the nonlinear wave model that governs the traffic flow.

Our derivation of the Rankine–Hugoniot condition (22.26) prescribing the shock speed relies on the fact that we can write the original partial differential equation in the form

of a conservation law. But there are other ways to do this; for instance, multiplying the nonlinear transport equation (22.13) by u allows us write it in the alternative conservative form

$$u\,\frac{\partial u}{\partial t} + u^2\,\frac{\partial u}{\partial x} = \frac{\partial}{\partial t}\left(\tfrac{1}{2}\,u^2\right) + \frac{\partial}{\partial x}\left(\tfrac{1}{3}\,u^3\right) = 0. \tag{22.31}$$

Here, the conserved density is $T = \frac{1}{2}\,u^2$, and the associated flux $X = \frac{1}{3}\,u^3$. The integral form equation (22.24) of the conservation law is

$$\frac{d}{dt}\int_a^b \tfrac{1}{2}\,u(t,x)^2\,dx = \tfrac{1}{3}\left[\,u(t,a)^3 - u(t,b)^3\,\right]. \tag{22.32}$$

In some physical models, the integral on the left hand side represents the energy within the interval $[\,a,b\,]$, and the conservation law tells us that energy can only enter the interval as a flux through its ends. If we assume that energy is conserved at a shock, then, repeating our previous argument, we are led to the alternative condition

$$\frac{ds}{dt} = \frac{\frac{1}{3}\,(u_-(t)^3 - u_+(t)^3)}{\frac{1}{2}\,(u_-(t)^2 - u_+(t)^2)} = \frac{2}{3}\,\frac{u_-(t)^2 + u_-(t)\,u_+(t) + u_+(t)^2}{u_-(t) + u_+(t)} \tag{22.33}$$

for the shock speed. Thus, a shock that conserves energy moves at a different speed than one that conserves mass! The evolution of a shock depends not just on the underlying differential equation, but also on the physical assumptions governing the selection of a suitable entropy condition.

The mathematical property that characterizes the shock dynamics is known as an *entropy condition*. Entropy conditions, such as the Rankine–Hugoniot Equal Area Rule (22.26), or the alternative (22.33), allow us to follow the solution beyond the formation of a simple shock. Once a shock forms, it cannot suddenly disappear — the discontinuity remains as the solution propagates. One consequence is the irreversibility of the solutions to the nonlinear transport equation. One cannot simply run time backwards and expect shocks to spontaneously vanish. However, this irreversibility is of a different character than that of the ill-posedness in the backwards heat equation. The nonlinear transport equation can be solved for $t < 0$, but this would result, typically, in the formation of a different collection of shocks, and would not be just the time reversal of the solution.

Continuing past the initial shock formation, as other characteristic lines start to cross, additional shocks appear. The shocks themselves continue propagate, often at different velocities. When a fast moving shock catches up with a slow moving shock, one must then decide how to merge the shocks together to retain a physically consistent solution. The selected entropy condition continues to resolve the ambiguities. However, at this point, the mathematical details have become too complicated for us to pursue in any more detail, and we refer the interested reader to Whitham's book, [**182**], which includes a wide range of applications to equations of gas dynamics, flood waves in rivers, motion of glaciers, chromotography, traffic flow, and many other physical systems.

22.2. Nonlinear Diffusion.

First order partial differential equations, beginning with elementary scalar transport equations, and progressing on to the equations of gas dynamics, the full-blown Euler equa-

tions of fluid mechanics, and yet more complicated systems for plasmas and other complicated physical processes, are used to model conservative wave motion. Such systems fail to account for frictional and/or viscous effects, which are typically modeled by a parabolic diffusion equation such as the heat equation. In this section we investigate the consequences of combining nonlinear wave motion with linear diffusion by analyzing the simplest such model. As we will see, the viscous term helps smooth out abrupt shock discontinuities, and the result is a well-determined and smooth dynamical process. Moreover, in the inviscid limit, as the diffusion term becomes vanishingly small, the smooth viscous solutions converge non-uniformly to the appropriate discontinuous shock wave, leading to an alternative mechanism for analyzing conservative nonlinnear dynamical processes.

Burgers' Equation

The simplest nonlinear diffusion equation is known as[†] *Burgers' equation*

$$u_t + u\,u_x = \gamma\,u_{xx}, \tag{22.34}$$

and is obtained by appending a linear diffusion term to the nonlinear transport equation (22.13). In fluids and gases, one can interpret the right hand side as modeling the effect of viscosity, and so Burgers' equation represents a very simplified version of the equations of viscous fluid mechanics, [**182**]. As with the heat equation, the diffusion coefficient $\gamma > 0$ must be positive in order that initial value problem be well-posed in forwards time.

Since Burgers' equation is first order in t, we expect that its solutions are uniquely prescribed by their initial values, say,

$$u(0,x) = f(x), \qquad -\infty < x < \infty. \tag{22.35}$$

(For simplicity, we will ignore boundary effects here.) Small, slowly varying solutions — more specifically, those for which both $|\,u(t,x)\,|$ and $|\,u_x(t,x)\,|$ are small — tend to act like solutions to the heat equation, smoothing out and decaying to 0 as time progresses. On the other hand, when the solution is large or rapidly varying, the nonlinear term tends to play the dominant role, and we might expect the solution to behave like the nonlinear waves that we analyzed in Section 22.1, perhaps steepening into some sort of shock. But, as we will see, the smoothing effect of the diffusion term, no matter how small, ultimately prevents the appearance of a discontinuous shock. Indeed, it can be proved that, under rather mild assumptions on the initial data, the solution to the initial value problem (22.34, 35) remains smooth and well-defined for all subsequent times, [**182**].

The simplest explicit solutions are the *traveling waves*, for which

$$u(t,x) = v(\xi) = v(x - ct), \qquad \text{where} \qquad \xi = x - ct,$$

[†] The equation is named after the applied mathematician J.M. Burgers, [**35**], and so the apostrophe goes after the "s". Burgers' equation was apparently first studied as a physical model by Bateman, [**15**], although its solution already appears as an exercise in a nineteenth century ordinary differential equations text, [**71**; vol. 6, p. 102].

indicates a fixed profile, moving to the right with constant speed c. By the chain rule,

$$\frac{\partial u}{\partial t} = -c\,v'(\xi), \qquad \frac{\partial u}{\partial x} = v'(\xi), \qquad \frac{\partial^2 u}{\partial x^2} = v''(\xi).$$

Substituting these expressions into Burgers' equation (22.34), we conclude that $v(\xi)$ must satisfy the nonlinear second order ordinary differential equation

$$-c\,v' + v\,v' = \gamma\,v''.$$

This equation can be solved by first integrating both sides with respect to ξ, and so

$$\gamma\,v' = k - c\,v + \tfrac{1}{2}\,v^2,$$

where k is a constant of integration. As in Section 20.1, the non-constant solutions to such an autonomous first order ordinary differential equation tend to either $\pm\infty$ or to one of the equilibrium points, i.e., the roots of the right hand side, as $t \to \pm\infty$. Thus, to obtain a bounded traveling wave solution $v(\xi)$, the quadratic polynomial on the right hand side must have two real roots, which requires $k < \frac{1}{2}\,c^2$. Assuming this holds, we rewrite the equation in the form

$$2\gamma\,\frac{dv}{d\xi} = (v-a)(v-b), \qquad \text{where} \qquad c = \tfrac{1}{2}\,(a+b). \tag{22.36}$$

To obtain the bounded solutions, we concentrate on the case when $a < v < b$. Integrating (22.36) by the usual method, we find

$$\int \frac{2\gamma\,dv}{(v-a)(v-b)} \;=\; \frac{2\gamma}{b-a}\,\log\left(\frac{b-v}{v-a}\right) \;=\; \xi - \delta,$$

for δ a constant of integration, and hence

$$v(\xi) = \frac{a\,e^{(b-a)(\xi-\delta)/(2\gamma)} + b}{e^{(b-a)(\xi-\delta)/(2\gamma)} + 1}\,.$$

Thus, the bounded traveling wave solutions all have the explicit form

$$u(t,x) = \frac{a\,e^{(b-a)(x-ct-\delta)/(2\gamma)} + b}{e^{(b-a)(x-ct-\delta)/(2\gamma)} + 1}\,.$$

Observe that

$$\lim_{x\to-\infty} u(t,x) = b, \qquad \lim_{x\to\infty} u(t,x) = a,$$

and hence our solution is a monotonically decreasing function going from b to a. The wave travels to the right, unchanged in form, with speed equal to the average of its asymptotic values. In Figure 22.18 we graph sample profiles corresponding to $a = .1$, $b = 1$ for three different values of the diffusion coefficient. Note that the smaller γ is, the sharper the transition layer between the two asymptotic values of the solution. In the inviscid limit

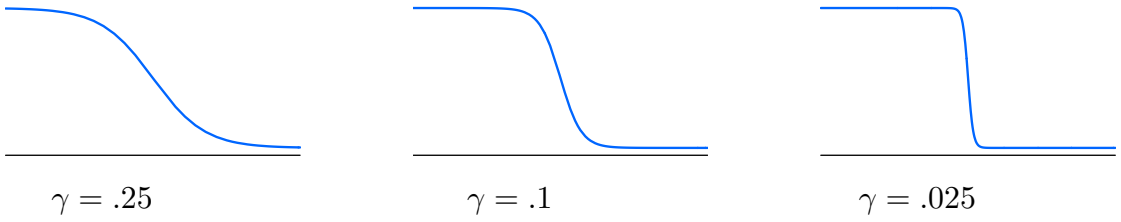


Figure 22.18. Traveling Wave Solutions to Burgers' Equation.

$\gamma \to 0$, the solutions converge to the step shock wave wave solution (22.29) to the nonlinear transport equation, which, as a result, is often referred to as the *inviscid Burgers' equation*.

Indeed, the profound fact is that, in the *inviscid limit* as the diffusion becomes vanishingly small, $\gamma \to 0$, the solutions to Burgers' equation (22.34) converge to the shock wave solution to (22.13) constructed by the Equal Area Rule. This observation is in accordance with our physical intuition, that all physical systems retain a very small dissipative component, that serves to smooth out discontinuities that might appear in a theoretical model that fails to take the dissipation/viscosity/damping/etc. into account. In modern theory, this so-called *viscosity solution method* has been successfully used to characterize the discontinuous solutions to a broad range inviscid nonlinear wave equations is as the limit, as the viscosity goes to zero, of classical solutions to a diffusive version. Thus, the viscosity solutions to the nonlinear transport equation resulting from Burgers' equation are consistent with the Equal Area Rule for drawing the shock discontinuities. More generally, this method allows one to monitor the solutions as they evolve into regimes where multiple shocks merge and interact. We refer the interested reader to [**117**, **182**].

The Hopf–Cole Transformation

By a remarkable stroke of luck, the nonlinear Burgers' equation can be converted into the linear heat equation and thereby explicitly solved. The *linearization* of Burgers' equation first appeared in an obscure exercise in a nineteenth century differential equations textbook, [**71**; vol. 6, p. 102]. Its modern rediscovery by Eberhard Hopf, [**102**], and Julian Cole, [**42**], was a milestone in the modern era of nonlinear partial differential equations, and is named the Hopf–Cole transformation in their honor.

Finding a way to covert a nonlinear differential equation into a linear equation is extremely challenging, and, in, most instances, impossible. On the other hand, the reverse process — "nonlinearizing" a linear equation — is trivial: any nonlinear changes of variables will do the trick! However, the resulting nonlinear equation, while evidently linearizable through the inverse change of variables, is rarely of any independent interest. Sometimes there is a lucky accident, and such "accidental" linearizations can have a profound impact on our understanding of more complicated nonlinear systems.

In the present context, our starting point is the linear heat equation

$$v_t = \gamma\, v_{xx}. \tag{22.37}$$

Among all possible nonlinear changes of dependent variable, one of the simplest that might spring to mind is an exponential function. Let us, therefore, investigate the effect of an

exponential change of variables

$$v(t,x) = e^{\alpha\,\varphi(t,x)}, \qquad \text{so} \qquad \varphi(t,x) = \frac{1}{\alpha}\,\log v(t,x), \tag{22.38}$$

where α is a nonzero constant. The function $\varphi(t,x)$ is real provided $v(t,x) > 0$ is a *positive* solution to the heat equation. Fortunately, this is not hard to arrange: if the initial data $v(0,x) > 0$ is strictly positive, then the resulting solution $v(t,x)$ is positive for all $t > 0$. This follows from the Maximum Principle for the heat equation, cf. Theorem 14.3.

To determine the differential equation satisfied by the function φ, we invoke the chain rule to differentiate (22.38):

$$v_t = \alpha\,\varphi_t\,e^{\alpha\,\varphi}, \qquad v_x = \alpha\,\varphi_x\,e^{\alpha\,\varphi}, \qquad v_{xx} = \big(\,\alpha\,\varphi_{xx} + \alpha^2\,\varphi_x^2\,\big)\,e^{\alpha\,\varphi}.$$

Substituting the first and last formulae into the heat equation (22.37) and canceling a common exponential factor, we conclude that $\varphi(t,x)$ satisfies the nonlinear partial differential equation

$$\varphi_t = \gamma\,\varphi_{xx} + \gamma\,\alpha\,\varphi_x^2, \tag{22.39}$$

known as the *potential Burgers' equation*, for reasons that will soon become apparent.

The second step in the process is to differentiate the potential Burgers' equation with respect to x; the result is

$$\varphi_{tx} = \gamma\,\varphi_{xxx} + 2\gamma\,\alpha\,\varphi_x\,\varphi_{xx}. \tag{22.40}$$

If we now set

$$\frac{\partial \varphi}{\partial x} = u, \tag{22.41}$$

so that φ has the status of a *potential function*, then the resulting partial differential equation

$$u_t = \gamma\,u_{xx} + 2\gamma\,\alpha\,u\,u_x$$

coincides with Burgers' equation (22.34) with $\alpha = -1/(2\gamma)$. In this manner, we have arrived at the famous *Hopf–Cole transformation*.

Theorem 22.7. *If $v(t,x) > 0$ is any positive solution to the linear heat equation $v_t = \gamma\,v_{xx}$, then*

$$u(t,x) = \frac{\partial}{\partial x}\,\big(\,-2\gamma\,\log v(t,x)\,\big) = -2\gamma\,\frac{v_x}{v} \tag{22.42}$$

solves Burgers' equation $u_t + u\,u_x = \gamma\,u_{xx}$.

Do all solutions to Burgers' equation arise in this way? In order to decide, we run the argument in reverse. First, choose a potential function $\widetilde{\varphi}(t,x)$ that satisfies (22.41); for example

$$\widetilde{\varphi}(t,x) = \int_0^x u(t,y)\,dy.$$

If $u(t,x)$ is any solution to Burgers' equation, then $\widetilde{\varphi}(t,x)$ satisfies (22.40). Integrating both sides of the latter equation with respect to x, we conclude that

$$\widetilde{\varphi}_t = \gamma\,\widetilde{\varphi}_{xx} + \gamma\,\alpha\,\widetilde{\varphi}_x^2 + h(t),$$

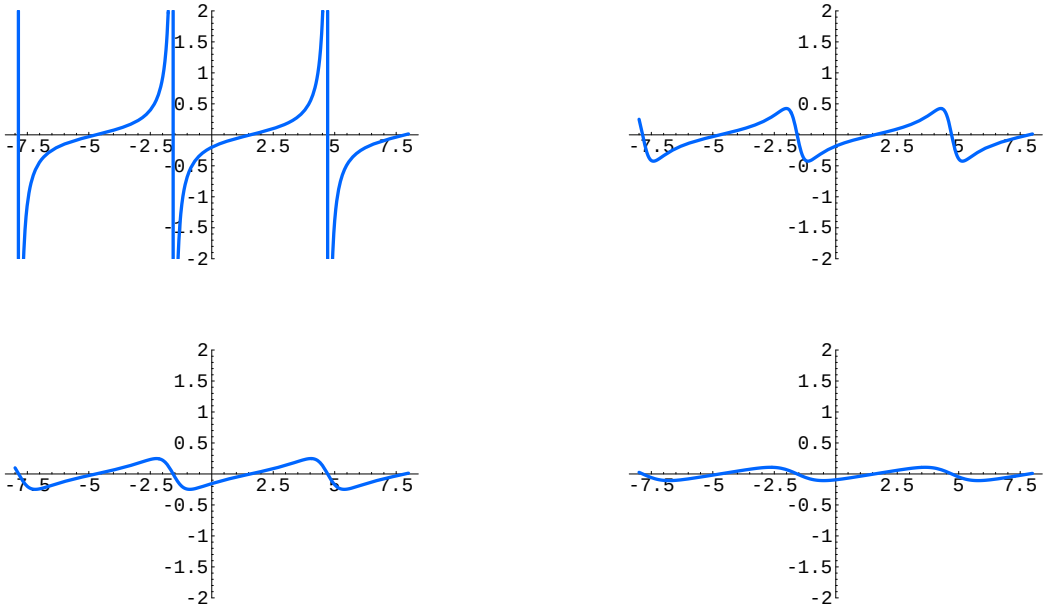


Figure 22.19. Solution to Burgers' Equation.

for some integration "constant" $h(t)$. Thus, unless $h(t) \equiv 0$, our potential function $\widetilde{\varphi}$ doesn't satisfy the potential Burgers' equation (22.39), but that's because we chose the "wrong" potential. Indeed, if we define

$$\varphi(t,x) = \widetilde{\varphi}(t,x) - \eta(t), \qquad \text{where} \qquad \eta'(t) = h(t),$$

then

$$\varphi_t = \widetilde{\varphi}_t - h(t) = \gamma\,\widetilde{\varphi}_{xx} + \gamma\,\alpha\,\widetilde{\varphi}_x^2 = \gamma\,\varphi_{xx} + \gamma\,\alpha\,\varphi_x^2,$$

and hence the modified potential $\varphi(t,x)$ *is* a solution to the potential Burgers' equation (22.39). From this it easily follows that

$$v(t,x) = e^{-\varphi(t,x)/(2\gamma)} \tag{22.43}$$

is a positive solution to the heat equation, from which $u(t,x)$ can be recovered via the Hopf –Cole transformation (22.42). Thus, we have proved that every solution to Burgers' equation comes from a positive solution to the heat equation via the Hopf–Cole transformation.

Example 22.8. As a simple example, the separable solution

$$v(t,x) = a + b\,e^{-\gamma\,\omega^2\,t}\cos\omega\,x$$

to the heat equation leads to the solution

$$u(t,x) = \frac{2\gamma\,b\,\omega\,e^{-\gamma\,\omega^2\,t}\sin\omega\,x}{a + b\,e^{-\gamma\,\omega^2\,t}\cos\omega\,x}$$

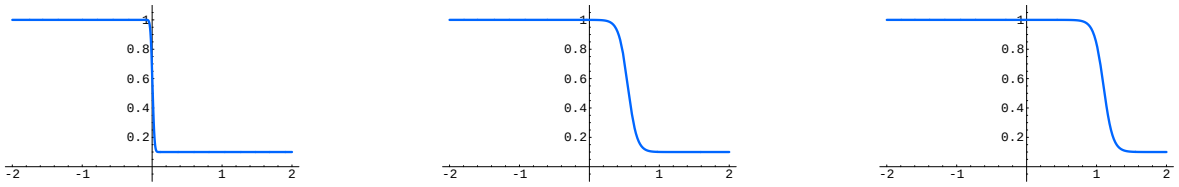


Figure 22.20. Shock Wave Solution to Burgers' Equation.

to Burgers' equation; a typical example is plotted in Figure 22.19. We should require that $a > |\, b\,|$ in order that $v(t,x) > 0$ be a positive solution to the heat equation for $t \geq 0$; otherwise the resulting solution to Burgers' equation will have singularities at the roots of u — see the first graph in Figure 22.19. This particular solution primarily feels the effects of the diffusivity, and rapidly goes to zero.

To solve the initial value problem (22.34–35) for Burgers' equation, we note that, under the Hopf–Cole transformation,

$$v(0,x) = h(x) = \exp\left(-\frac{\varphi(0,x)}{2\gamma}\right) = \exp\left(-\frac{1}{2\gamma}\int_0^x f(x)\,dy\right), \tag{22.44}$$

Remark: The lower limit of the integral can be changed from 0 to any other convenient value without affecting the final form of $u(t,x)$ in (22.42). The only effect is to multiply $v(t,x)$ by an overall constant, which does not change $u(t,x)$.

According to formula (14.60) (adapted to general diffusivity, as in Exercise ■), the solution to the initial value problem (22.37, 44) for the heat equation can be expressed as a convolution integral with the fundamental solution:

$$v(t,x) = \frac{1}{2\sqrt{\pi\,\gamma\,t}}\int_{-\infty}^{\infty} e^{-(x-y)^2/(4\gamma t)}\,h(y)\,dy.$$

Therefore, the solution to the Burgers' initial value problem (22.34, 35) is

$$u(t,x) = \frac{\displaystyle\int_{-\infty}^{\infty} \frac{x-y}{t}\,e^{F(t,x,y)}\,dy}{\displaystyle\int_{-\infty}^{\infty} e^{F(t,x,y)}\,dy} \qquad \text{where} \qquad F(t,x,y) = -\frac{1}{2\gamma}\int_0^y f(z)\,dz - \frac{(x-y)^2}{4\gamma t}\,. \tag{22.45}$$

Example 22.9. To demonstrate the smoothing effect of the diffusion terms, let us see what happens to the initial data

$$u(0,x) = \begin{cases} a, & x < 0, \\ b, & x > 0, \end{cases} \tag{22.46}$$

in the form of a step function. We assume that $a > b$, which would correspond to a shock wave in the inviscid limit $\gamma = 0$. (In Exercise ■, the reader is asked to analyze the case

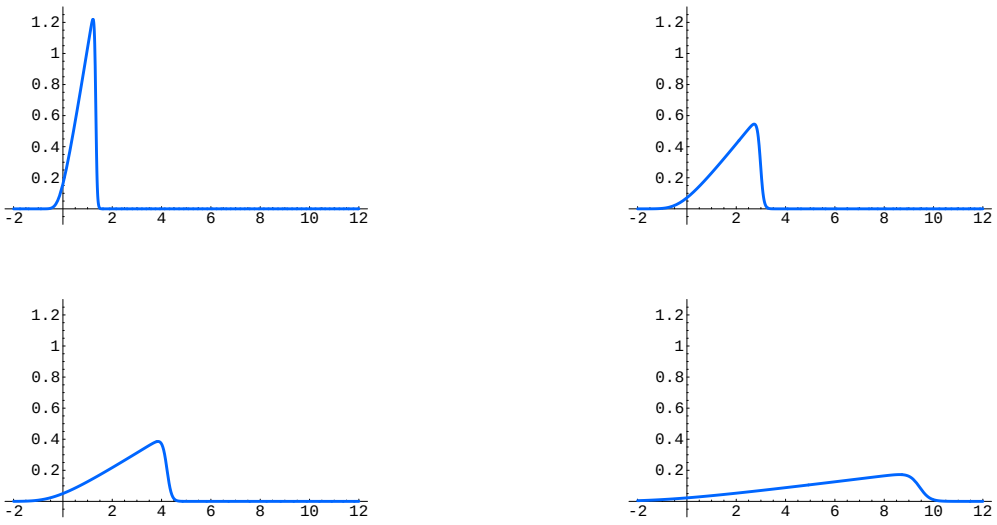


Figure 22.21. Triangular Wave Solution to Burgers' Equation.

$a < b$ which corresponds to a rarefaction wave.) In this case,

$$F(t,x,y) = -\frac{(x-y)^2}{4\gamma t} - \begin{cases} -\dfrac{a\,y}{2\gamma}, & y < 0, \\ -\dfrac{b\,y}{2\gamma}, & y > 0. \end{cases}$$

After some algebraic manipulations, the solution is found to have the explicit form

$$u(t,x) = a \; + \; \frac{b-a}{1 + h(t,x)\, \exp \dfrac{b-a}{2\gamma}\,(x - ct)} \tag{22.47}$$

where

$$c = \frac{a+b}{2}\,, \qquad h(t,x) \;=\; \frac{1 - \operatorname{erf}\left(\dfrac{x-bt}{\sqrt{4\gamma\, t}}\right)}{1 - \operatorname{erf}\left(\dfrac{x-at}{\sqrt{4\gamma\, t}}\right)}\,, \tag{22.48}$$

and $\operatorname{erf} z$ denotes the error function (14.62). The solution, with $a = 1, b = .1$ and $\gamma = .03$ is plotted in Figure 22.20 at times $t = .01, 1.0, 2.0$. Note that the sharp transition region for the shock is immediately smoothed, and the solution rapidly settles into the form of a continuously varying transition layer between the two step heights. The larger the diffusion coefficient in relation to the initial solution heights a, b, the more significant the smoothing effect. Observe that, as $\gamma \to 0$, the function $h(t,x) \to 1$, and hence the solution converges to the shock wave solution (22.29) to the transport equation, in which the speed of the shock is the average of the two initial values.

Example 22.10. Consider the case when the initial data $u(0,x) = \delta(x)$ is a concentrated delta function impulse at the origin. In the solution formula (22.45), starting the

integral for $F(t,x,y)$ at 0 is problematic, but as noted earlier, we are free to select any other starting point, e.g., $-\infty$. Thus, we take

$$F(t,x,y) = -\frac{1}{2\gamma}\int_{-\infty}^{y} \delta(z)\,dz - \frac{(x-y)^2}{4\gamma t} = \begin{cases} -\dfrac{(x-y)^2}{4\gamma t}, & y < 0, \\ -\dfrac{1}{2\gamma} - \dfrac{(x-y)^2}{4\gamma t}, & y > 0. \end{cases}$$

Substituting this into (22.45), we can evaluate the upper integral in elementary terms, while the lower integral involves the error function (14.62); after a little algebra, we find

$$u(t,x) = \sqrt{\frac{4\gamma}{\pi t}}\; \frac{e^{-x^2/(4\gamma t)}}{\coth\left(\dfrac{1}{4\gamma}\right) - \operatorname{erf}\left(\dfrac{x}{\sqrt{4\gamma\, t}}\right)}, \tag{22.49}$$

where

$$\coth z = \frac{\cosh z}{\sinh z} = \frac{e^z + e^{-z}}{e^z - e^{-z}} = \frac{e^{2z}+1}{e^{2z}-1}$$

is the hyperbolic cotangent function. A graph of this solution when $\gamma = .02$ and $a = 1$, at times $t = 1, 5, 10, 50$, appears in Figure 22.21. As you can see, the initial concentration diffuses out, but, unlike the heat equation, the wave does not remain symmetric owing to the advection terms in the equation. The effect is to steepen in front as it propagates. Eventually the triangular wave spreads out as the diffusion progresses.

22.3. Dispersion and Solitons.

Finally, we study a remarkable third order evolution equation that originally arose in the modeling of surface water waves, that serves to introduce yet further phenomena, both linear and nonlinear. The third order derivative models dispersion, in which waves of different frequencies move at different speeds. Coupled with the same nonlinearity as in the inviscid and viscous Burgers' (22.13, 34), the result is one of the most remarkable equations in all of mathematics, with far-reaching implications, not only in fluid mechanics and applications, but even in complex function theory, physics, etc., etc.

Linear Dispersion

So far, in our study of partial differential equations, we have not ventured beyond second order. Higher order equations do occur in applications, particularly in models for wave motion. The simplest linear partial differential equation of a type that we have not yet considered is the third order equation

$$u_t + u_{xxx} = 0 \tag{22.50}$$

It is the third in a hierarchy of simple evolution equations that starts with the simple ordinary differential equation $u_t = u$, then proceeds to the transport equation $u_t = u_x$, and then the heat equation $u_t = u_{xx}$ modeling basic diffusion processes. The third order case (22.50) is a simple model for linear dispersive waves.

To avoid additional complications caused by boundary conditions, we shall only look at the equation on the entire line, so $x \in \mathbb{R}$ The solution to the equation is uniquely specified by initial data

$$u(0,x) = f(x), \qquad -\infty < x < \infty. \tag{22.51}$$

See [**1**] for a proof.

Let us apply the Fourier transform to solve the initial value problem. Let

$$\widehat{u}(t,k) = \frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{\infty} u(t,x)\, e^{-\,\mathrm{i}\,k\,x}\, dx$$

be the spatial Fourier transform of the solution, which is assumed to remain in L^2 at all t, a fact that can be justified a posteriori. In view of the effect of the Fourier transform on derivatives — see Corollary 13.22 — the Fourier transform converts the partial differential equation (22.50) into a first order, linear ordinary differential equation

$$\widehat{u}_t - \,\mathrm{i}\,k^3\,\widehat{u} = 0, \tag{22.52}$$

parametrized by k, with initial conditions

$$\widehat{u}(0,k) = \widehat{f}(k) = \frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{\infty} f(x)\, e^{-\,\mathrm{i}\,k\,x}\, dx \tag{22.53}$$

given by the Fourier transform of (22.51). Solving the initial value problem (22.52–53) by the usual technique, we find

$$\widehat{u}(t,k) = e^{\,\mathrm{i}\,k^3\,t}\,\widehat{f}(k).$$

Inverting the Fourier transform yields the explicit formula for the solution

$$u(t,x) = \frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{\infty} e^{\,\mathrm{i}\,k^3\,t + \mathrm{i}\,k\,x}\,\widehat{f}(k)\, dk \tag{22.54}$$

to the initial value problem for the dispersive wave equation (22.50–51).

Actually, to find the solutions to the differential equation, one does not need the full power of the Fourier transform. Note that (22.54) represents a linear superposition of elementary exponential functions. Let us substitute an exponential ansatz

$$u(t,x) = e^{\,\mathrm{i}\,\omega\,t + \mathrm{i}\,k\,x} \tag{22.55}$$

representing a complex oscillatory wave of *frequency* ω, which indicates the time vibrations, and *wave number* k, which indicates the corresponding oscillations in space. Since

$$\frac{\partial u}{\partial t} = \,\mathrm{i}\,\omega\, e^{\,\mathrm{i}\,\omega\,t + \mathrm{i}\,k\,x}, \qquad \frac{\partial^3 u}{\partial x^3} = -\,\mathrm{i}\,k^3\, e^{\,\mathrm{i}\,\omega\,t + \mathrm{i}\,k\,x},$$

(22.55) satisfies the partial differential equation (22.50) if and only if its frequency and wave number are related by

$$\omega = k^3. \tag{22.56}$$

The result is known as the *dispersion relation* for the partial differential equation. In general, any linear constant coefficient dynamical partial differential equation admits a dispersion relation of the form $\omega = \omega(k)$ which is straightforwardly found by substituting the exponential ansatz (22.55) and canceling the common exponential factors in the resulting equation. In our particular case, the exponential solution of wave number k has the form

$$u_k(t,x) = e^{\,\mathrm{i}\,k^3\,t + \mathrm{i}\,k\,x}.$$

Linear superposition permits us to combine them in integral form, and so, for any (reasonable) function $a(k)$ depending on the wave number,

$$u(t,x) = \frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{\infty} e^{\,\mathrm{i}\,k^3\,t + \mathrm{i}\,k\,x} a(k)\,dk$$

is easily seen to be a solution to the partial differential equation. The Fourier transform solution (22.54) has this form.

Example 22.11. The *fundamental solution* corresponds to a concentrated initial disturbance

$$u(0,x) = \delta(x).$$

since the Fourier transform of the delta function is just $\widehat{\delta}(k) = 1/\sqrt{2\,\pi}$, the resulting solution (22.54) is

$$u(t,x) = \frac{1}{2\,\pi} \int_{-\infty}^{\infty} e^{\,\mathrm{i}\,k^3\,t + \mathrm{i}\,k\,x}\,dk = \frac{1}{\pi} \int_0^{\infty} \cos(k^3\,t + k\,x)\,dk,$$

since the solution is real (or, equivalently, the imaginary part of the integrand is odd) while the real part of the integrand is even. The second integral can be converted into that defining the Airy function,

$$\mathrm{Ai}(z) = \frac{1}{\pi} \int_0^{\infty} \cos\left(\, s\,z + \tfrac{1}{3}\,s^3 \,\right) ds,$$

as in (C.37), by the change of variables

$$s = k\,\sqrt[3]{3\,t}, \qquad\qquad z = \frac{x}{\sqrt[3]{3\,t}}\,,$$

and we conclude that the fundamental solution to the dispersive wave equation (22.50) can be written in terms of the Airy function:

$$u(t,x) = \frac{1}{\sqrt[3]{3\,t}}\ \mathrm{Ai}\left(\ \frac{x}{\sqrt[3]{3\,t}}\ \right).$$

See Figure ee3■ for a graph. Furthermore, writing the general initial data as a superposition of delta functions

$$f(x) = \int_{-\infty}^{\infty} f(\xi)\,d\xi\,\delta(x-\xi),$$

we conclude that the solution has the form

$$u(t,x) = \frac{1}{\sqrt[3]{3\,t}} \int_{-\infty}^{\infty} f(\xi)\ \mathrm{Ai}\left(\frac{x-\xi}{\sqrt[3]{3\,t}} \right) d\xi. \tag{22.57}$$

Although energy is conserved, unlike the heat and diffusion equations, the dispersion of waves means that the solution dies out.

Group velocity and wave velocity.

The Korteweg–deVries Equation

The simplest wave equation that combines dispersion with nonlinearity is the celebrated *Korteweg–deVries equation*

$$u_t + u_{xxx} + u\,u_x = 0. \tag{22.58}$$

The equation was first derived by the French applied mathematician Boussinesq, [**24**; eq. (30)], [**25**; eqs. (283, 291)], in 1872 as a model for surface water waves. It was rediscovered by the Dutch mathematician Korteweg and his student de Vries, [**118**], over two decades later, and, despite Boussinesq's priority, is named after them. In the early 1960's, the American mathematical physicists Martin Kruskal and Norman Zabusky, [**189**], rederived it as a continuum limit of a model of nonlinear mass-spring chains studied by Fermi, Pasta and Ulam, [**66**]. Understanding the puzzling behavior of both systems coming from numerical experiments was the catalyst of one of the most remarkable and far-ranging discoveries of modern mathematics: integrable nonlinear partial differential equations.

The most important special solutions to the Korteweg–deVries equation are the *traveling waves*. We assume that the solution

$$u = v(\xi) = v(x - ct), \qquad \text{where} \qquad \xi = x - ct,$$

is a wave of permanent form, translating to the right with speed c. By the chain rule,

$$\frac{\partial u}{\partial t} = -\,c\,v'(\xi), \qquad \frac{\partial u}{\partial x} = v'(\xi), \qquad \frac{\partial^3 u}{\partial x^3} = v'''(\xi).$$

Substituting these expressions into the Korteweg–deVries equation (22.58), we conclude that $v(\xi)$ must satisfy the nonlinear third order ordinary differential equation

$$v''' + v\,v' - c\,v' = 0. \tag{22.59}$$

Let us further assume that the traveling wave is *localized*, meaning that the solution and its derivatives are small at large distances:

$$\lim_{x \to \pm\infty} u(t,x) = \lim_{x \to \pm\infty} \frac{\partial u}{\partial x}\,(t,x) = \lim_{x \to \pm\infty} \frac{\partial^2 u}{\partial x^2}\,(t,x) = 0.$$

To this end, we impose the boundary conditions

$$\lim_{\xi \to \pm\infty} v(\xi) = \lim_{\xi \to \pm\infty} v'(\xi) = \lim_{\xi \to \pm\infty} v''(\xi) = 0. \tag{22.60}$$

(See Exercise ■ for an analysis of the non-localized traveling wave solutions.)

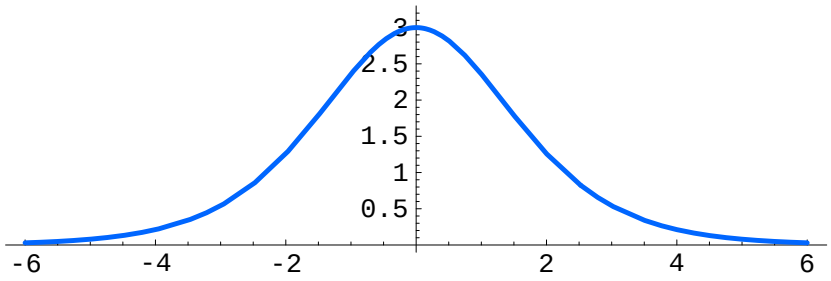


Figure 22.22. Solitary Wave.

The ordinary differential equation (22.59) can, in fact be solved in closed form. First, note that

$$\frac{d}{d\xi}\left[\, v'' + \tfrac{1}{2}\, v^2 - c\, v \,\right] = 0, \qquad \text{and hence} \qquad v'' + \tfrac{1}{2}\, v^2 - c\, v = k,$$

is a first integral, with k indicating the constant of integration. However, the localizing boundary conditions (22.60) imply that $k = 0$. Multiplying the latter equation by v' allows us to integrate a second time

$$\frac{d}{d\xi}\left[\, \tfrac{1}{2}(v')^2 + \tfrac{1}{6}\, v^3 - \tfrac{1}{2}\, c\, v^2 \,\right] = v'\left[\, v'' + \tfrac{1}{2}\, v^2 - c\, v \,\right] = 0.$$

Thus,

$$\tfrac{1}{2}(v')^2 + \tfrac{1}{6}\, v^3 - \tfrac{1}{2}\, c\, v^2 = \ell,$$

where ℓ is a second constant of integration, which, again by the boundary conditions (22.60), is also $\ell = 0$. We conclude that $v(\xi)$ satisfies the first order autonomous ordinary differential equation

$$\frac{dv}{d\xi} = v\sqrt{c - \tfrac{1}{3}\, v}\,.$$

We integrate by the usual method, cf. (20.7):

$$\int \frac{dv}{v\sqrt{c - \frac{1}{3}\, v}} = \xi + \delta.$$

Using a table of integrals, and then solving for v, we conclude that the solution has the form

$$v(\xi) = 3\, c\, \operatorname{sech}^2\left[\, \tfrac{1}{2}\sqrt{c}\, \xi + \delta \,\right],$$

where

$$\operatorname{sech} y = \frac{1}{\cosh y} = \frac{2}{e^y + e^{-y}}\,,$$

is the *hyperbolic secant function*. The solution has the form graphed in Figure 22.22; it has a global maximum at $3\, c \operatorname{sech} 0 = 3\, c$ at $y = 0$, and is an even function, exponentially decay

to 0 as $|\,\xi\,| \to \infty$. The resulting localized traveling wave solutions to the Korteweg–deVries equation are

$$u(t,x) = 3c\,\mathrm{sech}^2\left[\,\tfrac{1}{2}\sqrt{c}\,(x - ct) + \delta\,\right], \tag{22.61}$$

where $c > 0$ and δ are arbitrary constants. The parameter c equals the speed of the wave. It is also equal to one third its amplitude, since the maximum value of $u(t,x)$ is $3c$ at the points $x = ct$, as well as the width, which is on the order of $\sqrt{c}$. The taller and wider the solitary wave, the faster it moves.

The solution (22.61) is known as a *solitary wave solution* since it represents a localized wave that travels unchanged in shape. Such waves were first observed by the British engineer J. Scott Russell, [**159**], who recounts how such a wave was generated by the sudden motion of a barge along an Edinburgh canal and then chasing it on horseback for several miles. Russell's observations were dismissed by his contemporary, the prominent mathematician George Airy, who claimed that such localized disturbances could not exist, basing his analysis upon a linearized theory. Much later, Boussinesq established the proper nonlinear surface wave model (22.58), valid for long waves in shallow water, and also derived the solitary wave solution (22.61), thereby fully exonerating Scott Russell's insight.

These nonlinear traveling wave solutions were discovered by Kruskal and Zabusky, [**189**], to have remarkable properties. For this reason they have been given a special new name — *soliton*. Ordinarily, combining two solutions to a nonlinear equation can be quite unpredictable, and one might expect any number of scenarios to occur. If you start with initial conditions representing a taller wave to the left of a shorter wave, the solution of the Korteweg–deVries equation runs as follows. The taller wave moves faster, and so catches up the shorter wave. They then have a very complicated nonlinear interaction, as expected. But, remarkably, after a while they emerge from the interaction unscathed. The smaller wave is now in back and the larger one in front. After this, they proceed along their way, with the smaller one lagging behind the high speed tall wave. the only effect of their encounter is a phase shift, meaning a change in the value of the phase parameter δ in each wave. See Figure 22.23. After the interaction, the position of the soliton if it had traveled unhindered by the other is shown in a dotted line. Thus, they behave like colliding particles, which is the genesis of the word "soliton".

A similar phenomenon holds for several such soliton solutions. After some time where the various waves interact, they finally emerge with the largest soliton in front, and then in order to the smallest one in back, all progressing at their own speed, and so gradually drawing apart.

Moreover, starting with an *arbitrary* initial disturbance

$$u(0,x) = f(x)$$

it can be proved that after some time, the solution disintegrates into a finite number of solitons of different heights, moving off to the right, plus a small dispersive tail moving to the left that rapidly disappears. Proving this remarkable result is beyond the scope of this book. It relies on the method of *inverse scattering*, that connects the Korteweg–deVries equation with a linear eigenvalue problem of fundamental importance in one-dimensional quantum mechanics. The solitons correspond to the bound states of a quantum

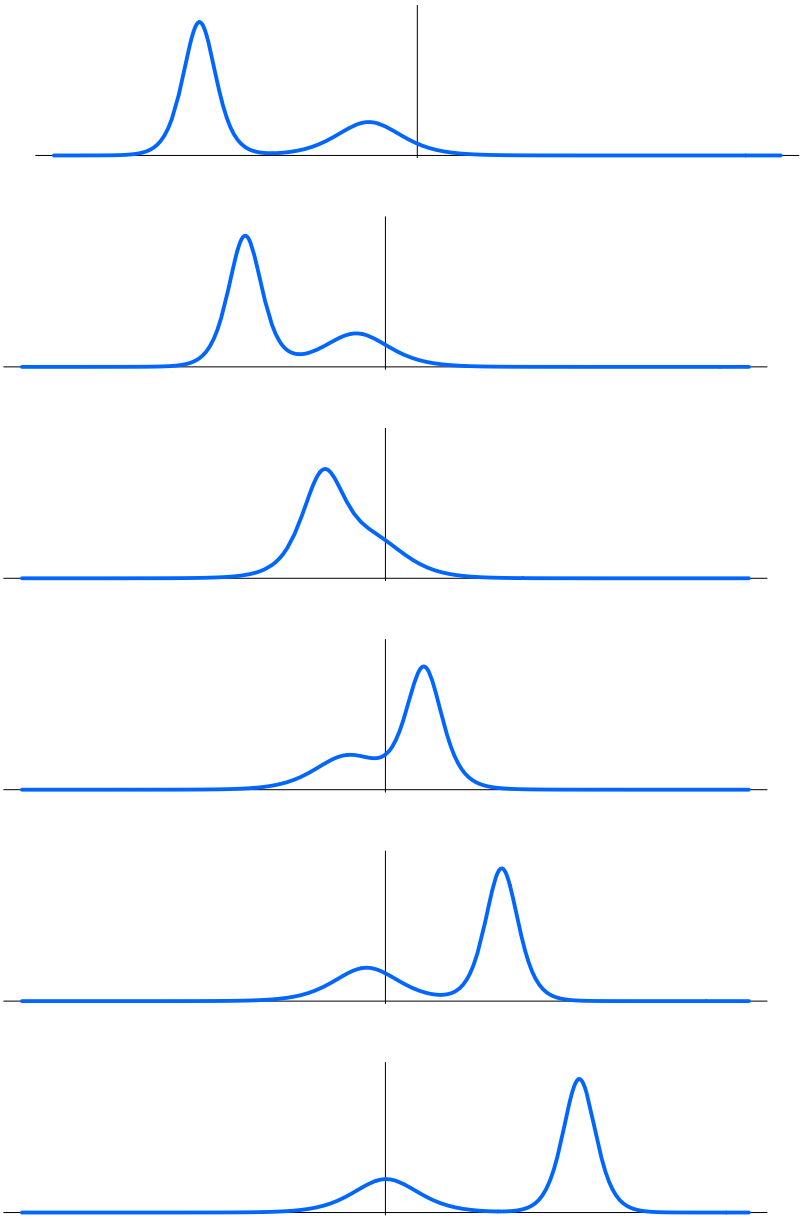


Figure 22.23. Interaction of Two Solitons.

potential. We refer the interested reader to the introductory text [**60**] and the more advanced monograph [**1**] for details.

Like Burgers' equation, the Korteweg–deVries equation can be linearized, but the linearization is considerably more subtle. It relies on the introduction of an auxiliary linear eigenvalue.

Remark: In the Korteweg–deVries equation model, one can find arbitrarily tall soliton

solutions. In physical water waves, if the wave is too tall it will break. Indeed, it can be rigorously proved that the full water wave equations admit solitary wave solutions, but there is a wave of greatest height, beyond which a wave will tend to break. The solitary water waves are not genuine solitons, since there is a small, but measurable, effect when two waves collide.

There is a remarkable transformation, known as the inverse scattering transform, which is a form of nonlinear Fourier transform, that can be used to solve the Korteweg–deVries equation. Its fascinating properties continue to be of great current research interest to this day.

22.4. Important Nonlinear Partial Differential Equations.

We've run out of space, and so cannot do full justice to the final subject. but we decided you should at least meet some of the most important nonlinear partial differential equations arising in modern day physics and engineering. some we've already seen, and we refer back to the minimal surface equation, (21.57).

The cast of major players includes:

Gas dynamics.

The Euler and Navier–Stokes Equations. viscosity. pressure.

The million dollar Cray prize.

The nonlinear Schrödinger equation.

Ginzburg–Landau?

Kuramoto–Sivashinki?

Nonlinear elasticity.

General relativity.

Now we can't even write down the equations without a lot of explanations. A metric on space time is defined by. The riemann curvature tensor. etc.

Free boundary problems.

Many questions remain unanswered and unresolved. Numerical solutions are a challenge. Variational structure, linearization and asymptotics, symmetries and conservation laws can help find explicit solutions, prove existence, etc.

22.5. Conclusion and Bon Voyage.

These are your first wee steps in a vast new realm. We are unable to discuss nonlinear partial differential equations arising in fluid mechanics, in elasticity, in relativity, in differential geometry, in computer vision, in mathematical biology. Chaos and integrability are the two great themes in modern nonlinear applied mathematics, and the student is well-advised to pursue both.

We bid you, dear reader, a fond adieu and wish you unparalleled success in your mathematical endeavors.

References

[**1**] Ablowitz, M.J., and Clarkson, P.A., *Solitons, Nonlinear Evolution Equations and the Inverse Scattering Transform*, L.M.S. Lecture Notes in Math., vol. 149, Cambridge University Press, Cambridge, 1991.

[**2**] Abraham, R., Marsden, J.E., and Ratiu, T., *Manifolds, Tensor Analysis, and Applications*, Springer–Verlag, New York, 1988.

[**3**] Abramowitz, M., and Stegun, I., *Handbook of Mathematical Functions*, National Bureau of Standards Appl. Math. Series, #55, U.S. Govt. Printing Office, Washington, D.C., 1970.

[**4**] Ahlfors, L., *Complex Analysis*, McGraw–Hill, New York, 1966.

[**5**] Airy, G.B., On the intensity of light in the neighborhood of a caustic, *Trans. Cambridge Phil. Soc.* **6** (1838), 379–402.

[**6**] Aki, K., and Richards, P.G., *Quantitative Seismology*, W.H. Freeman, San Francisco, 1980.

[**7**] Alligood, K.T., Sauer, T.D., and Yorke, J.A., *Chaos. An Introduction to Dynamical Systems*, Springer-Verlag, New York, 1997.

[**8**] Antman, S.S., *Nonlinear Problems of Elasticity*, Appl. Math. Sci., vol. 107, Springer–Verlag, New York, 1995.

[**9**] Apostol, T.M., *Calculus*, Blaisdell Publishing Co., Waltham, Mass., 1967–69.

[**10**] Apostol, T.M., *Introduction to Analytic Number Theory*, Springer–Verlag, New York, 1976.

[**11**] Tannenbaum, P., and Arnold, R., *Excursions in Modern Mathematics*, 5th ed., Prentice–Hall, Inc., Englewood Cliffs, N.J., 2004.

[**12**] Baker, G.A., Jr., and Graves–Morris, P., *Padé Approximants*, Encyclopedia of Mathematics and Its Applications, v. 59, Cambridge University Press, Cambridge, 1996.

[**13**] Ball, J.M., and Mizel, V.J., One-dimensional variational problem whose minimizers do not satisfy the Euler-Lagrange equation, *Arch. Rat. Mech. Anal.* **90** (1985), 325–388.

[**14**] Batchelor, G.K., *An Introduction to Fluid Dynamics*, Cambridge University Press, Cambridge, 1967.

[**15**] Bateman, H., Some recent researches on the motion of fluids, *Monthly Weather Rev.* **43** (1915), 63–170.

[**16**] Behrends, E., *Introduction to Markov Chains*, Vieweg, Braunschweig/Wiesbaden, Germany, 2000.

[**17**] Berest, Y., and Winternitz, P., Huygens' principle and separation of variables, *Rev. Math. Phys.* **12** (2000), 159–180.

[**18**] Birkhoff, G., *Hydrodynamics — A Study in Logic, Fact and Similitude*, 1st ed., Princeton University Press, Princeton, 1950.

[**19**] Birkhoff, G., and Rota, G.–C., *Ordinary Differential Equations*, Blaisdell Publ. Co., Waltham, Mass., 1962.

[**20**] Blanchard, P., Devaney, R.L., and Hall, G.R., *Differential Equations*, Brooks–Cole Publ. Co., Pacific Grove, Calif., 1998.

[**21**] Bollobás, B., *Graph Theory*: *an Introductory Course*, Graduate Texts in Mathematics, vol. 63, Springer–Verlag, New York, 1993.

[**22**] Boothby, W.M., *An Introduction to Differentiable Manifolds and Riemannian Geometry*, Academic Press, New York, 1975.

[**23**] Bott, R., and Tu, L.W., *Differential Forms in Algebraic Topology*, Springer–Verlag, New York, 1982.

[**24**] Boussinesq, J., Théorie des ondes et des remous qui se propagent le long d'un canal rectangulaire horizontal, en communiquant au liquide contenu dans ce canal des vitesses sensiblement pareilles de la surface au fond, *J. Math. Pures Appl.* **17** (2) (1872), 55–108.

[**25**] Boussinesq, J., Essai sur la théorie des eaux courants, *Mém. Acad. Sci. Inst. Nat. France* **23** (1) (1877), 1–680.

[**26**] Boyce, W.E., and DiPrima, R.C., *Elementary Differential Equations and Boundary Value Problems*, 7th ed., John Wiley & Sons, Inc., New York, 2001.

[**27**] Bradie, B., *A Friendly Introduction to Numerical Analysis*, Prentice–Hall, Inc., Upper Saddle River, N.J., 2006.

[**28**] Braun, M., *Differential Equations and their Applications: an Introduction to Applied Mathematics*, Springer–Verlag, New York, 1993.

[**29**] Brigham, E.O., *The Fast Fourier Transform*, Prentice–Hall, Inc., Englewood Cliffs, N.J., 1974.

[**30**] Briggs, W.L., *Henson, V.E.*, The DFT. An Owner's Manual for the Discrete Fourier Transform; SIAM, Philadelphia, PA, 1995.

[**31**] Bronstein, M., and Lafaille, S., Solutions of linear ordinary differential equations in terms of special functions, *in*: *Proceedings of the 2002 International Symposium on Symbolic and Algebraic Computation*, T. Mora, ed., ACM, New York, 2002, pp. 23–28.

[**32**] Brown, J.W., and Churchill, R.V., *Fourier Series and Boundary Value Problems*, McGraw-Hill, New York, 1993.

[**33**] Buhmann, M.D., Radial basis functions, *Acta Numer.* **9** (2000), 1–38.

[**34**] Burden, R.L., and Faires, J.D., *Numerical Analysis*, Seventh Edition, Brooks/Cole, Pacific Grove, CA, 2001.

[**35**] Burgers, J.M., A mathematical model illustrating the theory of turbulence, *Adv. Appl. Mech.* **1** (1948), 171–199.

[**36**] Bürgisser, P., Clausen, M., and Shokrollahi, M.A., *Algebraic Complexity Theory*, Springer–Verlag, New York, 1997.

[**37**] Buss, S.A., *3D Computer Graphics*, Cambridge University Press, Cambridge, 2003.

[**38**] Cantwell, B.J., *Introduction to Symmetry Analysis*, Cambridge University Press, Cambridge, 2003.

[**39**] Carmichael, R., *The Theory of Numbers*, Dover Publ., New York, 1959.

[**40**] Clenshaw, C.W., and Olver, F.W.J., Beyond floating point, *J. Assoc. Comput. Mach.* **31** (1984), 319–328.

[**41**] Coddington, E.A., and Levinson, N., *Theory of Ordinary Differential Equations*, McGraw–Hill, New York, 1955.

[**42**] Cole, J.D., On a quasilinear parabolic equation occurring in aerodynamics, *Q. Appl. Math.* **9** (1951), 225–236.

[**43**] Cooley, J.W., and Tukey, J.W., An algorithm for the machine computation of complex Fourier series, *Math. Comp.* **19** (1965), 297–301.

[**44**] Copson, E.T., *Partial Differential Equations*, Cambridge University Press, Cambridge, 1975.

[**45**] Courant, R., *Differential and Integral Calculus*, 2nd ed., Interscience Publ., New York, 1937.

[**46**] Courant, R., and Hilbert, D., *Methods of Mathematical Physics*, vol. I, Interscience Publ., New York, 1953.

[**47**] Courant, R., and Hilbert, D., *Methods of Mathematical Physics*, vol. II, Interscience Publ., New York, 1953.

[**48**] Crowe, M.J., *A History of Vector Analysis*, Dover Publ., New York, 1985.

[**49**] Daubechies, I., Orthonormal bases of compactly supported wavelets, *Commun. Pure Appl. Math.* **41** (1988), 909–996.

[**50**] Daubechies, I., *Ten Lectures on Wavelets*, SIAM, Philadelphia, PA, 1992.

[**51**] Davidson, K.R., and Donsig, A.P., *Real Analysis with Real Applications*, Prentice–Hall, Inc., Upper Saddle River, N.J., 2002.

[**52**] DeGroot, M.H., and Schervish, M.J., *Probability and Statistics*, 3rd ed., Addison–Wesley, Boston, 2002.

[**53**] Demmel, J.W., *Applied Numerical Linear Algebra*, SIAM, Philadelphia, PA, 1997.

[**54**] Devaney, R.L., *An Introduction to Chaotic Dynamical Systems*, Addison–Wesley, Redwood City, Calif., 1989.

[**55**] Dewdney, A.K., *The Planiverse. Computer Contact with a Two-dimensional World*, Copernicus, New York, 2001.

[**56**] Diacu, F., *An Introduction to Differential Equations*, W.H. Freeman and Co., New York, 2000.

[**57**] Dirac, P.A.M., *The Principles of Quantum Mechanics*, Third Edition, Clarendon Press, Oxford, 1947.

[**58**] DLMF Project ■ .

[**59**] do Carmo, M.P., *Differential Geometry of Curves and Surfaces*, Prentice-Hall, Englewood Cliffs, N.J., 1976.

[**60**] Drazin, P.G., and Johnson, R.S., *Solitons: An Introduction*, Cambridge University Press, Cambridge, 1989.

[**61**] Durrett, R., *Essentials of Stochastic Processes*, Springer–Verlag, New York, 1999.

[**62**] Dym, H., and McKean, H.P., *Fourier Series and Integrals*, Academic Press, New York, 1972.

[**63**] Farin, G.E., *Curves and Surfaces for CAGD: A Practical Guide*, Academic Press, London, 2002.

[**64**] Feigenbaum, M.J., Qualitative universality for a class of nonlinear transformations, *J. Stat. Phys.* **19** (1978), 25–52.

[**65**] Feller, W., *An Introduction to Probability Theory and its Applications*, Third Edition., J. Wiley & Sons, New York, 1968.

[**66**] Fermi, E., Pasta, J., and Ulam, S., Studies of nonlinear problems. I., preprint, Los Alamos Report LA 1940, 1955; in: *Nonlinear Wave Motion*, A.C. Newell, ed., Lectures in Applied Math., vol. 15, American Math. Soc., Providence, R.I., 1974, pp. 143–156.

[**67**] Field, J.V., *The Invention of Infinity: Mathematics and Art in the Renaissance*, Oxford University Press, Oxford, 1997.

[**68**] Fletcher, N.H., and Rossing, T.D., *The Physics of Musical Instruments*, Second Edition, Springer–Verlag, New York, 1998.

[**69**] Fine, B., and Rosenberger, G., *The Fundamental Theorem of Algebra*, Undergraduate Texts in Mathematics, Springer–Verlag, New York, 1997.

[**70**] Fleming, W.H., *Functions of Several Variables*, 2d ed., Springer–Verlag, New York, 1977.

[**71**] Forsyth, A.R., *The Theory of Differential Equations*, Cambridge University Press, Cambridge, 1890, 1900, 1902, 1906.

[**72**] Fourier, J., *The Analytical Theory of Heat*, Dover Publ., New York, 1955.

[**73**] Francis, J.G.F., The QR transformation I, II, *Comput. J.* **4** (1961–2), 265–271, 332–345.

[**74**] Gaal, L., *Classical Galois theory*, 4th ed., Chelsea Publ. Co., New York, 1988.

[**75**] Garabedian, P., *Partial Differential Equations*, 2nd ed., Chelsea Publ. Co., New York, 1986.

[**76**] Gel'fand, I.M., and Fomin, S.V., *Calculus of Variations*, Prentice–Hall, Inc., Englewood Cliffs, N.J., 1963.

[**77**] Gohberg, I., and Koltracht, I., Triangular factors of Cauchy and Vandermonde matrices, *Integral Eq. Operator Theory* **26** (1996), 46–59.

[**78**] Goldstein, H., *Classical Mechanics*, Second Edition, Addison–Wesley, Reading, Mass., 1980.

[**79**] Golub, G.H, and Van Loan, C.F., *Matrix Computations*, Johns Hopkins University Press, Baltimore, 1989.

[**80**] Goode, S.W., *Differential Equations and Linear Algebra*, Second Ed., Prentice Hall, Upper Saddle River, NJ, 2000.

[**81**] Gradshteyn, I.S., and Ryzhik, I.W., *Table of Integrals, Series and Products*, Academic Press, New York, 1965.

[**82**] Graver, J.E., *Counting on Frameworks*: *Mathematics to Aid the Design of Rigid Structures*, Dolciani Math. Expo. No. 25, Mathematical Association of America, Washington, DC, 2001.

[**83**] Greene, B., *The Elegant Universe: Superstrings, Hidden Dimensions, and the Quest for the Ultimate Theory*, W. W. Norton, New York, 1999[**84**] Guillemin, V., and Pollack, A., *Differential Topology*, Prentice–Hall, Inc., Englewood Cliffs, N.J., 1974..

[**84**] Guillemin, V., and Pollack, A., *Differential Topology*, Prentice–Hall, Inc., Englewood Cliffs, N.J., 1974.

[**85**] Gurtin, M.E., *An Introduction to Continuum Mechanics*, Academic Press, New York, 1981.

[**86**] Haar, A., Zur Theorie der orthogonalen Funktionensysteme, *Math. Ann.* **69** (1910), 331–371.

[**87**] Haberman, R., *Elementary Applied Partial Differential Equations*, Third Edition, Prentice Hall, Upper Saddle River, NJ, 1998.

[**88**] Hairer, E., Nørsett, S.P., and Wanner, G., *Solving Ordinary Differential Equations*, 2nd ed., Springer–Verlag, New York, 1993–1996.

[**89**] Hale, J.K., *Ordinary Differential Equations*, Second Edition, R. E. Krieger Pub. Co., Huntington, N.Y., 1980.

[**90**] Hall, R.W., and Josić, K., Planetary motion and the duality of force laws, *SIAM Review* **42** (2000), 115–124.

[**91**] Hall, R.W., and Josić, K., The mathematics of musical instruments, *Amer. Math. Monthly* **108** (2001), 347–357.

[**92**] Hamming, R.W., *Numerical Methods for Scientists and Engineers*, McGraw–Hill, New York, 1962.

[**93**] Henrici, P., *Applied and Computational Complex Analysis*, vol. 1, J. Wiley & Sons, New York, 1974.

[**94**] Herrlich, H., and Strecker, G.E., *Category Theory*; *an Introduction*, Allyn and Bacon, Boston, 1973.

[**95**] Herstein, I.N., *Abstract Algebra*, John Wiley & Sons, Inc., New York, 1999.

[**96**] Hestenes, M.R., and Stiefel, E., Methods of conjugate gradients for solving linear systems, *J. Res. Nat. Bur. Standards* **49** (1952), 409–436.

[**97**] Higham, N.J., *Accuracy and Stability of Numerical Algorithms*, Second Edition, SIAM, Philadelphia, 2002.

[**98**] Hille, E., *Ordinary Differential Equations in the Complex Domain*, John Wiley & Sons, New York, 1976.

[**99**] Hirsch, M.W., and Smale, S., *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic Press, New York, 1974.

[**100**] Hobson, E.W., *The Theory of Spherical and Ellipsoidal Harmonics*, Chelsea Publ. Co., New York, 1965.

[**101**] Hoggatt, V.E., Jr., and Lind, D.A., The dying rabbit problem, *Fib. Quart.* **7** (1969), 482–487.

[**102**] Hopf, E., The partial differential equation $u_t + u\,u_x = \mu\,u$, *Commun. Pure Appl. Math.* **3** (1950), 201–230.

[**103**] Howison, S., *Practical Applied Mathematics*, Cambridge University Press, Cambridge, 2005.

[**104**] Hydon, P.E., *Symmetry Methods for Differential Equations*, Cambridge Texts in Appl. Math., Cambridge University Press, Cambridge, 2000.

[**105**] Ince, E.L., *Ordinary Differential Equations*, Dover Publ., New York, 1956.

[**106**] Isaacson, E., and Keller, H.B.., *Analysis of Numerical Methods*, John Wiley & Sons, New York, 1966.

[**107**] Iserles, A., *A First Course in the Numerical Analysis of Differential Equations*, Cambridge University Press, Cambridge, 1996.

[**108**] Jolliffe, I.T., *Principal Component Analysis*, 2nd ed., Springer–Verlag, New York, 2002.

[**109**] Kailath, T., Sayed, A.H., and Hassibi, B., *Linear Estimation*, Prentice–Hall, Inc., Upper Saddle River, N.J., 2000.

[**110**] Kamke, E., *Differentialgleichungen Lösungsmethoden und Lösungen*, vol. 1, Chelsea, New York, 1971.

[**111**] Kammler, D.W., *A First Course in Fourier Analysis*, Prentice Hall, Upper Saddle River, NJ, 2000.

[**112**] Kaplansky, I., *An Introduction to Differential Algebra*, 2nd ed., Hermann, Paris, 1976.

[**113**] Kauffman, L.H., *Knots and Physics*, 2nd ed., World Scientific, Singapore, 1993.

[**114**] Keener, J.P., *Principles of Applied Mathematics. Transformation and Approximation*, Addison–Wesley Publ. Co., New York, 1988.

[**115**] Keller, H.B., *Numerical Methods for Two-Point Boundary-Value Problems*, Blaisdell, Waltham, MA, 1968.

[**116**] Kevorkian, J., *Partial Differential Equations*, Second Edition, Texts in Applied Mathematics, vol. 35, Springer–Verlag, New York, 2000.

[**117**] Knobel, R., *An Introduction to the Mathematical Theory of Waves*, American Mathematical Society, Providence, RI, 2000.

[**118**] Korteweg, D.J., and de Vries, G., On the change of form of long waves advancing in a rectangular channel, and on a new type of long stationary waves, *Phil. Mag.* (5) **39** (1895), 422–443.

[**119**] Krall, A.M., *Applied Analysis*, D. Reidel Publishing Co., Boston, 1986.

[**120**] Kreysig, E., *Advanced Engineering Mathematics*, Eighth Edition, J. Wiley & Sons, New York, 1998.

[**121**] Kublanovskaya, V.N., On some algorithms for the solution of the complete eigenvalue problem, *USSR Comput. Math. Math. Phys.* **3** (1961), 637–657.

[**122**] Landau, L.D., and Lifshitz, E.M., *Quantum Mechanics* (*Non-relativistic Theory*), Course of Theoretical Physics, vol. 3, Pergamon Press, New York, 1977.

[**123**] Lanford, O., A computer-assisted proof of the Feigenbaum conjecture, *Bull. Amer. Math. Soc.* **6** (1982), 427–434.

[**124**] Lighthill, M.J., *Waves in Fluids*, Cambridge University Press, Cambridge, 1978.

[**125**] Mandelbrot, B.B., *The Fractal Geometry of Nature*, W.H. Freeman, New York, 1983.

[**126**] Marsden, J.E., and Tromba, A.J., *Vector Calculus*, 4th ed., W.H. Freeman, New York, 1996.

[**127**] Messiah, A., *Quantum Mechanics*, John Wiley & Sons, New York, 1976.

[**128**] Miller, W., Jr., *Symmetry and Separation of Variables*, Encyclopedia of Mathemtics and Its Applications, vol. 4, Addison–Wesley Publ. Co., Reading, Mass., 1977.

[**129**] Misner, C.W., Thorne, K.S., and Wheeler, J.A., *Gravitation*, W.H. Freeman, San Francisco, 1973.

[**130**] Moon, F.C., *Chaotic Vibrations*, John Wiley & Sons, New York, 1987.

[**131**] Moon, P., and Spencer, D.E., *Field Theory Handbook*, Springer-Verlag, New York, 1971.

[**132**] Morgan, F., *Geometric Measure Theory: a Beginner's Guide*, Academic Press, New York, 2000.

[**133**] Morse, P.M., and Feshbach, H., *Methods of Theoretical Physics*, McGraw–Hill, New York, 1953.

[**134**] Murray, R.N., Li, Z.X., and Sastry, S.S., *A Mathematical Introduction to Robotic Manipulation*, CRC Press, Boca Raton, FL, 1994.

[**135**] Nilsson, J.W., and Riedel, S., *Electric Circuits*, 7th ed., Prentice–Hall, Inc., Upper Saddle River, N.J., 2005.

[**136**] Nitsche, J.C.C., *Lectures on Minimal Surfaces*, Cambridge University Press, Cambridge, 1988.

[**137**] Noether, E., Invariante Variationsprobleme, *Nachr. Konig. Gesell. Wissen. Gottingen, Math.–Phys. Kl.* (1918), 235–257. (See *Transport Theory and Stat. Phys.* **1** (1971), 186–207 for an English translation.)

[**138**] Oberhettinger, F., *Tables of Fourier Transforms and Fourier Transforms of Distributions*, Springer-Verlag, New York, 1990.

[**139**] Oberhettinger, F., and Badii, L., *Tables of Laplace Transforms*, Springer-Verlag, New York, 1973.

[**140**] Olver, F.W.J., *Asymptotics and Special Functions*, Academic Press, New York, 1974.

[**141**] Olver, P.J., *Applications of Lie Groups to Differential Equations*, 2nd ed., Graduate Texts in Mathematics, vol. 107, Springer–Verlag, New York, 1993.

[**142**] Olver, P.J., and Shakiban, C., *Applied Linear Algebra*, Prentice–Hall, Inc., Upper Saddle River, N.J., 2005.

[**143**] O'Neil, P.V., *Advanced Engineering Mathematics*, Fourth Edition, Wadsworth Publ. Co., Belmont, Ca., 1995.

[**144**] Ortega, J.M., *Numerical Analysis; A Second Course*, Academic Press, New York, 1972.

[**145**] Oru**c**c, H., and Phillips, G. M., Explicit factorization of the Vandermonde matrix, *Linear Algebra Appl.* **315** (2000), 113–123.

[**146**] Page, L., Brin, S., Motwani, R., and Winograd, T., The PageRank citation ranking: bringing order to the web, preprint, Stanford University, 1998.

[**147**] Peitgen, H.-O., and Richter, P.H., *The Beauty of Fractals: Images of Complex Dynamical Systems*,Springer–Verlag, New York, 1986.

[**148**] Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P., *Numerical Recipes in C*: *The Art of Scientific Computing*, 2nd ed., Cambridge University Press, Cambridge, 1995.

[**149**] Reed, M., and Simon, B., *Methods of Modern Mathematical Physics*, Academic Press, New York, 1972.

[**150**] Renardy, M., and Rogers, R.C., *An Introduction to Partial Differential Equations*, Academic Press, New York, 1993.

[**151**] Richards, I., and Youn, H., *Theory of Distributions*: *a Non-Technical Introduction*, Cambridge University Press, Cambridge, 1990.

[**152**] Royden, H.L., *Real Analysis*, Macmillan Co., New York, 1988.

[**153**] Rudin, W., *Real and Complex Analysis*, 3rd ed., McGraw–Hill, New York, 1987.

[**154**] Saff, E.B., and Snider, A.D., *Fundamentals of Complex Analysis*, Third Ed., Prentice–Hall, Inc., Upper Saddle River, N.J., 2003.

[**155**] Salomon, D., *Computer Graphics and Geometric Modeling*, Springer–Verlag, New York, 1999.

[**156**] Sapiro, G., *Geometric Partial Differential Equations and Image Analysis*, Cambridge University Press, Cambridge, 2001.

[**157**] Schumaker, L.L., *Spline Functions: Basic Theory*, John Wiley & Sons, New York, 1981.

[**158**] Schwartz, L., *Théorie des distributions*, Hermann, Paris, 1957.

[**159**] Scott Russell, J., On waves, *in*: *Report of the 14*$^{\text{th}}$ *Meeting*, British Assoc. Adv. Sci., 1845, pp. 311–390.

[**160**] Seshadri, R., and Na, T.Y., *Group Invariance in Engineering Boundary Value Problems*, Springer–Verlag, New York, 1985.

[**161**] Sethares, W.A., *Tuning, Timbre, Spectrum, Scale*, Springer–Verlag, New York, 1999.

[**162**] Stewart, J., *Calculus*: *Early Transcendentals*, 5th ed., Thomson Brooks Cole, Belmont, CA, 2003.

[**163**] Strang, G., *Introduction to Applied Mathematics*, Wellesley Cambridge Press, Wellesley, Mass., 1986.

[**164**] Strang, G., *Linear Algebra and its Applications*, Third Ed., Harcourt, Brace, Jovanovich, San Diego, 1988.

[**165**] Strang, G., *Calculus*, Wellesley Cambridge Press, Wellesley, Mass., 1991.

[**166**] Strang, G., Wavelet transforms versus Fourier transforms, *Bull. Amer. Math. Soc.* **28** (1993), 288–305.

[**167**] Strang, G., and Borre, K., *Linear Algebra, Geodesy, and GPS*, Wellesley Cambridge Press, Wellesley, Mass., 1997.

[**168**] Strang, G., and Fix, G.J., *An Analysis of the Finite Element Method*, Prentice–Hall, Inc., Englewood Cliffs, N.J., 1973.

[**169**] Strang, G., and Nguyen, T., *Wavelets and Filter Banks*, Wellesley Cambridge Press, Wellesley, Mass., 1996.

[**170**] Strassen, V., Gaussian elimination is not optimal, *Numer. Math.* **13** (1969), 354–356.

[**171**] Strauss, W.A., *Partial Differential Equations: an Introduction*, John Wiley & Sons, New York, 1992.

[**172**] Tannenbaum, P., *Excursions in Modern Mathematics*, 5th ed., Prentice–Hall, Inc., Upper Saddle River, N.J, 2004.

[**173**] Titchmarsh, E. C., *Theory of Functions*, Oxford University Press, London, 1968.

[**174**] Tychonov, A.N., and Samarski, A.A., *Partial Differential Equations of Mathematical Physics*, Holden–Day, San Francisco, 1964.

[**175**] Ugural, A.C., and Fenster, S.K., *Advanced Strength and Applied Elasticity*, 3rd ed., Prentice–Hall, Inc., Englewood Cliffs, N.J., 1995.

[**176**] Varga, R.S., *Matrix Iterative Analysis*, 2nd ed., Springer–Verlag, New York, 2000.

[**177**] Varga, R.S., *Ger***v***sgorin and His Circles*, Springer–Verlag, New York, 2004.

[**178**] Walter, G.G., and Shen, X., *Wavelets and Other Orthogonal Systems*, 2nd ed., Chapman & Hall/CRC, Boca Raton, Fl, 2001.

[**179**] Watkins, D.S., *Fundamentals of Matrix Computations*, Second Edition,, Wiley–Interscience, New York, 2002.

[**180**] Watson, G.N., *A Treatise on the Theory of Bessel Functions*, Cambridge University Press, Cambridge, 1952.

[**181**] Weinberger, H.F., *A First Course in Partial Differential Equations*, Ginn and Co., Waltham, Mass., 1965.

[**182**] Whitham, G.B., *Linear and Nonlinear Waves*, John Wiley & Sons, New York, 1974.

[**183**] Whittaker, E.T., *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies*, Cambridge University Press, Cambridge, 1937.

[**184**] Whittaker, E.T., and Watson, G.N., *A Course of Modern Analysis*, Cambridge University Press, Cambridge, 1990.

[**185**] Widder, D.V., *The Heat Equation*, Academic Press, New York, 1975.

[**186**] Wolfram, S., *The Mathematica Book*, Third Edition, Cambridge University Press, Cambridge, 1996.

[**187**] Yaglom, I.M., *Felix Klein and Sophus Lie*, Birkhäuser, Boston, 1988.

[**188**] Yale, P.B., *Geometry and Symmetry*, Holden–Day, San Francisco, 1968.

[**189**] Zabusky, N.J., and Kruskal, M.D., Interaction of "solitons" in a collisionless plasma and the recurrence of initial states, *Phys. Rev. Lett.* **15** (1965), 240–243.

[**190**] Zaitsev, V.F., and Polyanin, A.D., *Handbook of Exact Solutions for Ordinary Differential Equations*, CRC Press, Boca Raton, Fl., 1995.

[**191**] Zienkiewicz, O.C., and Taylor, R.L., *The Finite Element Method*, 4th ed., McGraw–Hill, New York, 1989.

[**192**] Zwillinger, D., *Handbook of Differential Equations*, Academic Press, Boston, 1992.

[**193**] Zygmund, A., *Trigonometric Series*, Cambridge University Press, Cambridge, 1968.

