

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, in preparation for later developments, witness 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 $0 \leq x \leq \ell$ 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 $x + u(x)$.

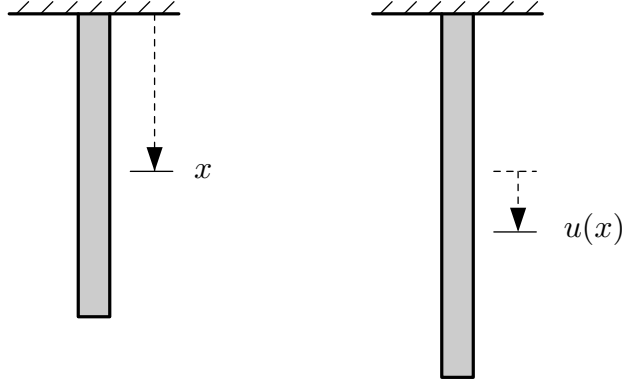


Figure 11.1. Bar with One Fixed Support.

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,

$$u(0) = 0 \quad (11.1)$$

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

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

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 $\Delta x \rightarrow 0$. The result is the strain function

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

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 $n \rightarrow \infty$ and the spring lengths $\Delta x \rightarrow 0$. 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

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

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

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

where

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

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 $n \rightarrow \infty$ and the spring lengths $\Delta x \rightarrow 0$. 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

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

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

$$w(x) = c(x)v(x), \quad (11.5)$$

[†] 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

$$w_k = c_k v_k, \quad \text{or} \quad \mathbf{w} = C \mathbf{v}, \quad (11.6)$$

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,

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

or

$$f = -\frac{dw}{dx}. \quad (11.7)$$

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

$$f_k = \frac{w_{k-1} - w_k}{\Delta x}, \quad w_n = 0, \quad (11.8)$$

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

$$w(\ell) = 0 \quad (11.9)$$

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

$$A^T = \frac{1}{\Delta x} \begin{pmatrix} 1 & -1 & & & & \\ & 1 & -1 & & & \\ & & 1 & -1 & & \\ & & & 1 & -1 & \\ & & & & \ddots & \ddots \\ & & & & & \ddots \end{pmatrix} \longrightarrow -\frac{d}{dx} \quad (11.10)$$

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

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

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

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

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:

$$u(0) = 0, \quad w(\ell) = c(\ell)u'(\ell) = 0. \quad (11.13)$$

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

$$K\mathbf{u} = A^T C A \mathbf{u} = \mathbf{f} \quad (11.14)$$

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

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

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 .

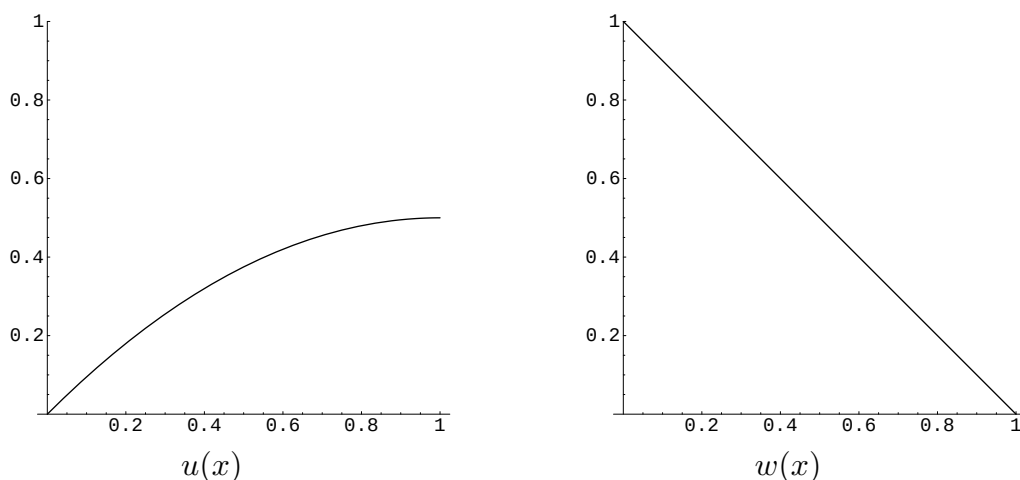


Figure 11.2. Displacement and Stress of a Bar with One Fixed End.

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

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

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

$$u(x) = -\frac{1}{2}\alpha x^2 + ax + b, \quad \text{where} \quad \alpha = \frac{f}{c} \quad (11.16)$$

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

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

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

$$u(x) = \alpha \left(x - \frac{1}{2}x^2 \right), \quad (11.17)$$

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

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

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

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

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

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

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

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

directly for the stresses:

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

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

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

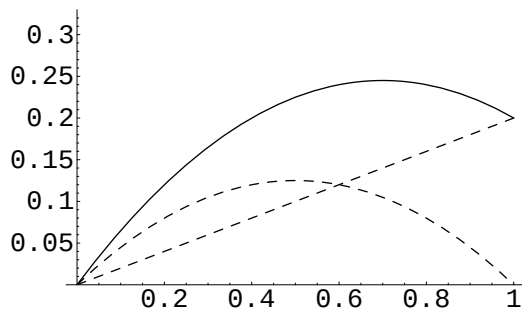


Figure 11.3. Displacements of a Bar with Two Fixed Ends.

The sampled displacement $u(x_k)$ is not exactly equal to u_k , but their difference tends to zero as the number of masses $n \rightarrow \infty$. 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

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

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

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

Thus

$$u(x) = \frac{1}{2}\alpha(x - x^2) + dx \tag{11.19}$$

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 $w(x) = a - fx$ *cannot* be determined without first figuring out the displacement (11.19):

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

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

$$-u'' = f(x), \quad u'(0) = 0, \quad u'(\ell) = 0, \quad (11.20)$$

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

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

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

$$u'(\ell) = - \int_0^\ell f(x) dx = 0, \quad (11.21)$$

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

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

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 (5.80), which requires that $\mathbf{f}$ be orthogonal, with respect to the Euclidean inner product, to all elements of $\text{coker } K$. In physical equilibrium systems K is symmetric, and so $\text{coker } K = \ker K$. 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, $\langle f; 1 \rangle = \int_0^\ell f(x) dx = 0$. 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.

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 $\mathbf{f} = (f_1, f_2, \dots, f_n)^T \in \mathbb{R}^n$ into a linear combination

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

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

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

[†] 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.

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

$$\mathbf{u} = f_1 \mathbf{u}_1 + f_2 \mathbf{u}_2 + \cdots + f_n \mathbf{u}_n. \quad (11.26)$$

Remark: The alert reader will recognize that $\mathbf{u}_1, \dots, \mathbf{u}_n$ 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

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

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 satisfy

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

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 which are the same everywhere except at one single point have exactly the same integral, [53, 158]. 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

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

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, [116, 156]. (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

$$\lim_{n \rightarrow \infty} g_n(x) = 0, \quad x \neq y, \quad (11.29)$$

while the total amount of force remains fixed at

$$\int_0^\ell g_n(x) dx = 1. \quad (11.30)$$

On a formal level, the limit “function”

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

will satisfy the key properties (11.27–28).

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

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

These functions satisfy

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

while[‡]

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

[†] 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.

[‡] 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.

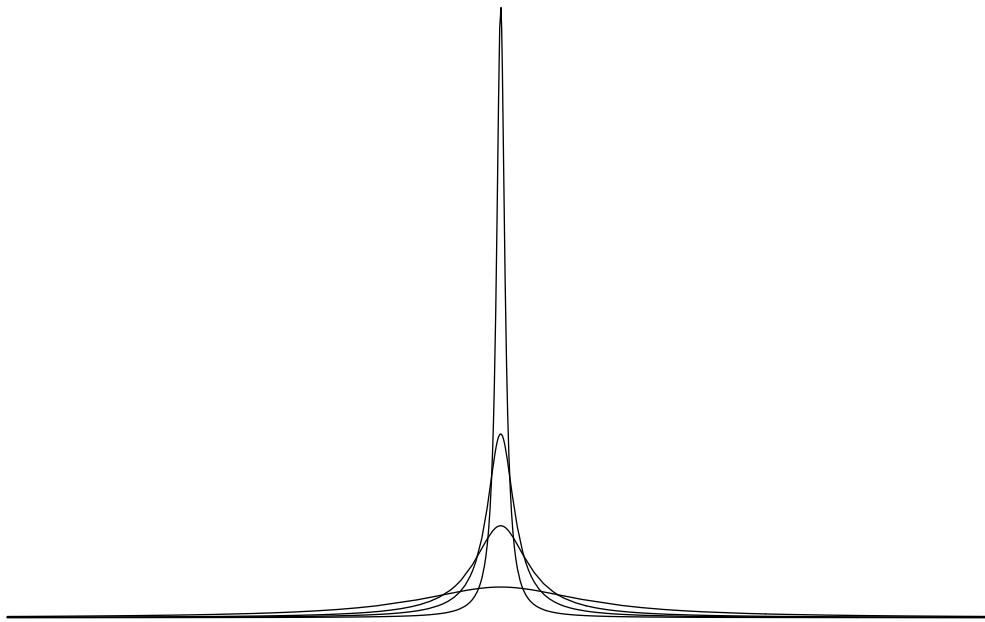


Figure 11.4. Delta Function as Limit.

Therefore, formally, we identify the limiting function

$$\lim_{n \rightarrow \infty} g_n(x) = \delta(x) = \delta_0(x), \quad (11.34)$$

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, 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:

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

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, [53, 158].

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:

$$\delta_y(x) = \delta(x - y). \quad (11.35)$$

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

$$\widehat{g}_n(x) = g_n(x - y) = \frac{n}{\pi (1 + n^2(x - y)^2)}. \quad (11.36)$$

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

$$\int_0^\ell \delta_y(x) u(x) dx = u(y), \quad \text{for} \quad 0 < y < \ell. \quad (11.37)$$

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 (11.37) serves to define a linear functional[†] $L_y: C^0[0, \ell] \rightarrow \mathbb{R}$ that maps a continuous function $u \in C^0[0, \ell]$ 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 .

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

$$\int_0^\ell \delta_y(x) u(x) dx = 0, \quad \text{when} \quad y < 0 \quad \text{or} \quad y > \ell, \quad (11.38)$$

because the integrand is identically zero on the entire interval. For technical reasons, we will not attempt to define the integral (11.38) 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 $L: V \rightarrow \mathbb{R}$ on a finite-dimensional inner product space is given by an inner product with a fixed element $\mathbf{a} \in V$, so

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

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

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 $C^0[0, \ell]$, the L^2 inner product

$$L_g[u] = \langle g; u \rangle = \int_0^\ell g(x) u(x) dx \quad (11.39)$$

taken with a fixed function $g \in C^0[0, \ell]$ defines a real-valued linear functional $L_g: C^0[0, \ell] \rightarrow \mathbb{R}$. 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

$$\langle \delta_y; u \rangle = \int_0^\ell \delta_y(x) u(x) dx = u(y) \quad (11.40)$$

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 (11.37) as the limit

$$u(y) = \lim_{n \rightarrow \infty} \int_0^\ell g_n(x) u(x) dx = \int_0^\ell \delta_y(x) u(x) dx \quad (11.41)$$

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 $L_y[u] = u(y)$ represented by the delta function is the limit, $L_y = \lim_{n \rightarrow \infty} L_n$, 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

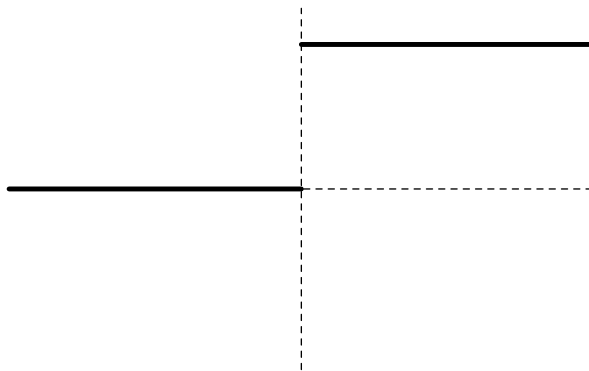


Figure 11.5. The Step Function.

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, [158], 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:

$$g(x)\delta_y(x) = g(y)\delta_y(x), \quad (11.42)$$

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.

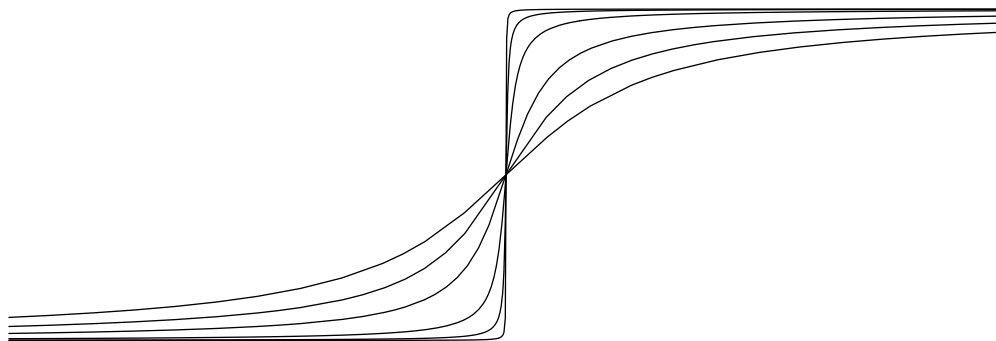


Figure 11.6. Step Function as Limit.

The integral of the delta function is known as a *step function*. More specifically, the basic formulae (11.37, 38) imply that

$$\int_a^x \delta_y(t) dt = \sigma_y(x) = \sigma(x - y) = \begin{cases} 0, & a < x < y, \\ 1, & x > y > a. \end{cases} \quad (11.43)$$

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 (11.43) 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

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

Since

$$\lim_{y \rightarrow \infty} \tan^{-1} y = \frac{1}{2} \pi, \quad \text{while} \quad \lim_{y \rightarrow -\infty} \tan^{-1} y = -\frac{1}{2} \pi,$$

these functions converge to the step function:

$$\lim_{n \rightarrow \infty} f_n(x) = \sigma(x) = \begin{cases} 1, & x > 0, \\ \frac{1}{2}, & x = 0, \\ 0, & x < 0. \end{cases} \quad (11.44)$$

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

The integral of the discontinuous step function (11.43) is the continuous *ramp function*

$$\int_a^x \sigma_y(z) dz = \rho_y(x) = \rho(x - y) = \begin{cases} 0, & a < x < y, \\ x - y, & x > y > a, \end{cases} \quad (11.45)$$

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 $\frac{d\rho}{dx} = \sigma$ has a jump discontinuity, and its second

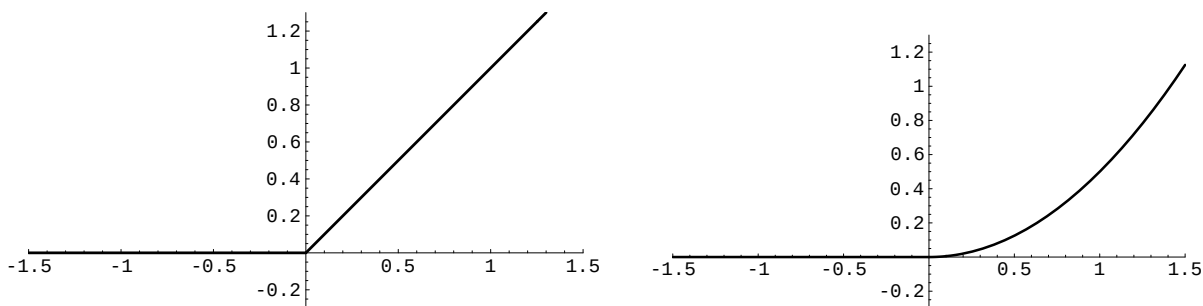


Figure 11.7. First and Second Order Ramp Functions.

derivative $\frac{d^2\rho}{dx^2} = \delta$ 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

$$\rho_n(x-y) = \begin{cases} \frac{(x-y)^n}{n!}, & x > y, \\ 0, & x < y. \end{cases} \quad (11.46)$$

What about differentiation? Motivated by the Fundamental Theorem of Calculus, we shall use formula (11.43) to identify the derivative of the step function with the delta function

$$\frac{d\sigma}{dx} = \delta. \quad (11.47)$$

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

$$f(y^-) = \lim_{x \rightarrow y^-} f(x), \quad f(y^+) = \lim_{x \rightarrow y^+} f(x), \quad (11.48)$$

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.

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

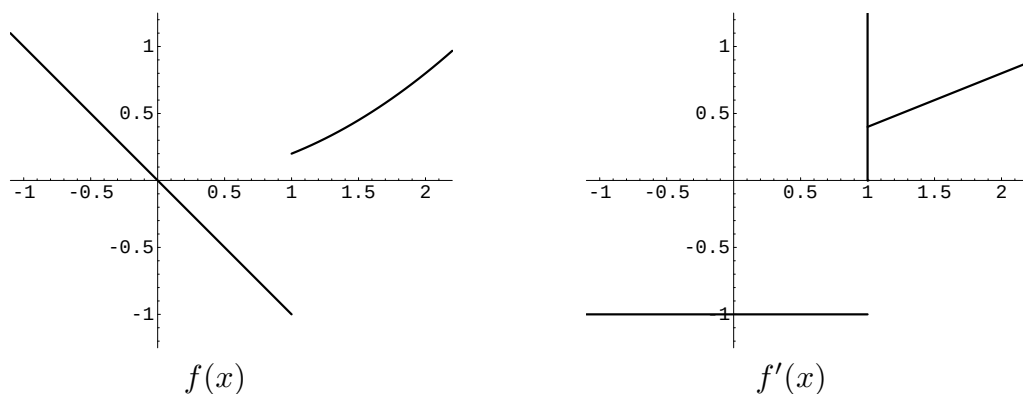


Figure 11.8. The Derivative of a Discontinuous Function.

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

$$\beta = f(y^+) - f(y^-) = \lim_{x \rightarrow y^+} f(x) - \lim_{x \rightarrow y^-} f(x) \quad (11.49)$$

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

$$f(x) = g(x) + \beta \sigma(x - y), \quad (11.50)$$

where $g(x)$ is continuous everywhere, with a removable discontinuity at $x = y$, and differentiable except possibly at the jump. Differentiating (11.50), we find that

$$f'(x) = g'(x) + \beta \delta(x - y), \quad (11.51)$$

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

$$f(x) = \begin{cases} -x, & x < 1, \\ \frac{1}{5}x^2, & x > 1, \end{cases} \quad (11.52)$$

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

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

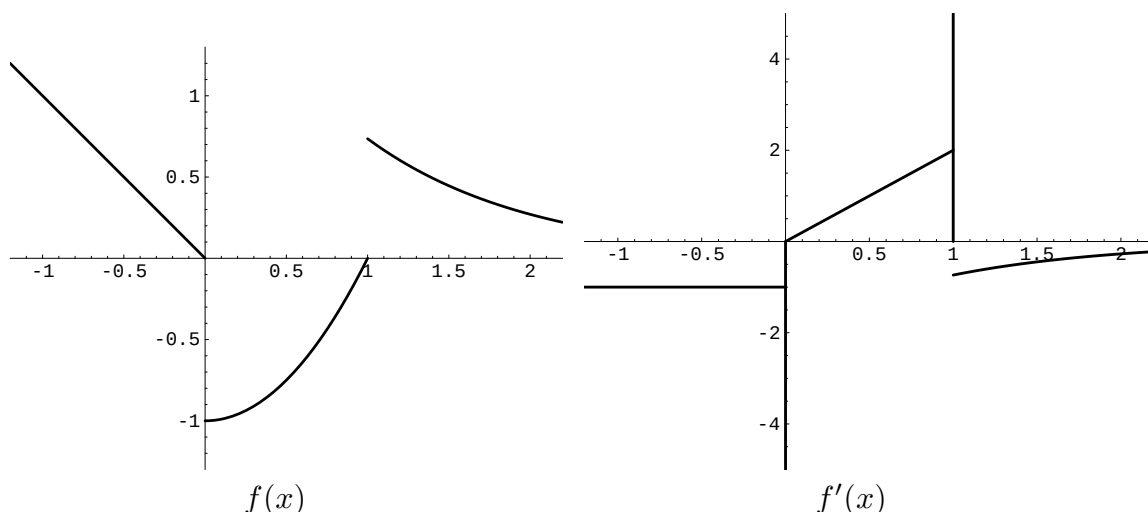


Figure 11.9. The Derivative of a Discontinuous Function.

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

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

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

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

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,

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

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

Example 11.6. The derivative of the absolute value function

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

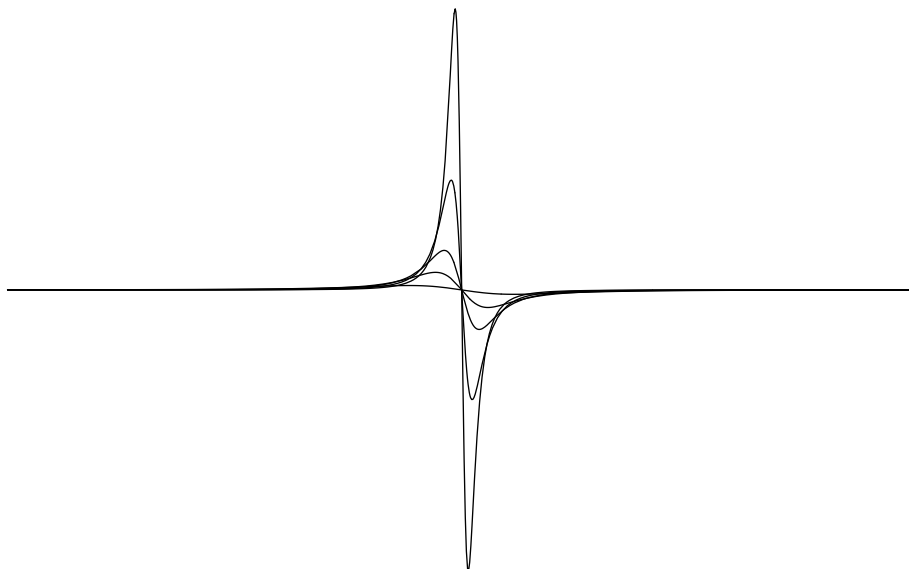


Figure 11.10. Derivative of Delta Function as Limit of Doublets.

is the *sign function*

$$s(x) = a'(x) = \begin{cases} +1, & x > 0, \\ -1, & x < 0. \end{cases} \quad (11.53)$$

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):

$$\frac{d\delta}{dx} = \lim_{n \rightarrow \infty} \frac{dg_n}{dx} = \lim_{n \rightarrow \infty} \frac{-2n^3 x}{\pi(1 + n^2 x^2)^2}. \quad (11.54)$$

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

$$\begin{aligned} \langle \delta'; u \rangle &= \int_{-\infty}^{\infty} \delta'(x) u(x) dx = \lim_{n \rightarrow \infty} \int_{-\infty}^{\infty} g'_n(x) u(x) dx \\ &= - \lim_{n \rightarrow \infty} \int_{-\infty}^{\infty} g_n(x) u'(x) dx = - \int_{-\infty}^{\infty} \delta(x) u'(x) dx = -u'(0). \end{aligned} \quad (11.55)$$

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 $|x| \rightarrow \infty$. 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

$$L'_y[u] = -u'(y) = \langle \delta'_y; u \rangle = \int_0^\ell \delta'_y(x) u(x) dx, \quad \text{where} \quad 0 < y < \ell, \quad (11.56)$$

that maps a continuously differentiable function $u(x)$ to *minus* its derivative at the point y . We note that (11.56) 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 $\tilde{g}_n(x) = g_n(x) + g'_n(x)$ satisfy $\lim_{n \rightarrow \infty} \tilde{g}_n(x) = 0$ for all $x \neq y$, while $\int_{-\infty}^\infty \tilde{g}_n(x) dx = 1$. However, $\lim_{n \rightarrow \infty} \tilde{g}_n = \lim_{n \rightarrow \infty} g_n + \lim_{n \rightarrow \infty} g'_n = \delta + \delta'$. 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 (11.41).

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 differential equation (11.12) takes the special form

$$-\frac{d}{dx} \left(c(x) \frac{du}{dx} \right) = \delta(x-y), \quad 0 < x < \ell, \quad (11.57)$$

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

$$-cu'' = \delta(x-y), \quad u(0) = 0 = u(1), \quad (11.58)$$

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 (11.43),

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

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

$$u(x) = -\frac{\rho(x-y)}{c} + ax + b, \quad (11.59)$$

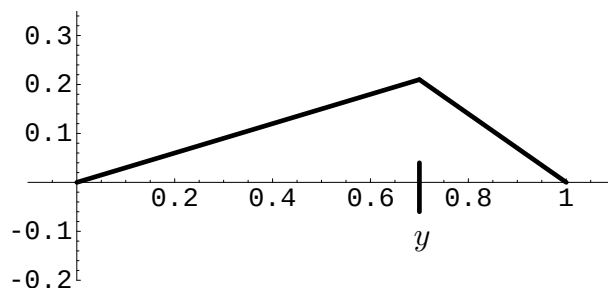


Figure 11.11. Green's function for a Bar with Fixed Ends.

where ρ is the ramp function (11.45). The integration constants a, b are fixed by the boundary conditions; since $0 < y < 1$, we have

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

Therefore, the Green's function for the problem is

$$G(x, y) = -\rho(x - y) + (1 - y)x = \begin{cases} x(1 - y)/c, & x \leq y, \\ y(1 - x)/c, & x \geq y, \end{cases} \quad (11.60)$$

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

$$-cu'' = f(x), \quad u(0) = 0 = u(1), \quad (11.61)$$

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 $0 \leq y \leq 1$ at which impulse forces may be applied, we will use an integral to sum them up, thereby writing the external force as

$$f(x) = \int_0^1 f(y) \delta(x - y) dy. \quad (11.62)$$

In the continuous context, sums are replaced by integrals, and we will interpret (11.62) 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

$$u(x) = \int_0^1 f(y) G(x, y) dy \quad (11.63)$$

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

For the particular boundary value problem (11.61), we use the explicit formula (11.60) for the Green's function. Breaking the integral (11.63) into two parts, for $y < x$ and $y > x$, we arrive at the explicit solution formula

$$u(x) = \frac{1}{c} \int_0^x (1-x)y f(y) dy + \frac{1}{c} \int_x^1 x(1-y) f(y) dy. \quad (11.64)$$

For example, under a constant unit force f , (11.64) reduces to

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

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, [116, 169, 187].

Let us, finally, convince ourselves that the superposition formula (11.64) does indeed give the correct answer. First,

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

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

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

$$\frac{d}{dx} \int_{\alpha(x)}^{\beta(x)} F(x, y) dy = F(x, \beta(x)) \frac{d\beta}{dx} - F(x, \alpha(x)) \frac{d\alpha}{dx} + \int_{\alpha(x)}^{\beta(x)} \frac{\partial F}{\partial x}(x, y) dy \quad (11.65)$$

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, 168]. 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[†]

$$\frac{\partial^2 G}{\partial x^2}(x, y) = 0 \quad \text{for all} \quad x \neq y. \quad (11.66)$$

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

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 (11.58).

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 6.1.6, 6.2.18 and 6.3.17.) 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 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.

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:

$$-\frac{\partial}{\partial x} \left(c(x) \frac{\partial}{\partial x} G(x, y) \right) = 0, \quad \text{for all } x \neq y. \quad (11.67)$$

(ii) Satisfies the homogeneous boundary conditions.

(iii) Is a continuous function of its arguments.

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

(v) Is a symmetric function of its arguments:

$$G(x, y) = G(y, x). \quad (11.68)$$

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

$$u(x) = \int_0^\ell G(x, y) f(y) dy. \quad (11.69)$$

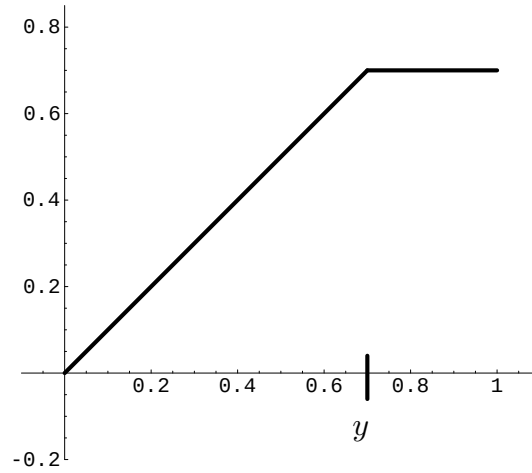


Figure 11.12. Green's Function for Bar with One Fixed and One Free End.

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

$$-cu'' = f(x), \quad u(0) = 0, \quad u'(1) = 0, \quad (11.70)$$

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 $-cu'' = 0$ for all $x \neq y$, the Green's function must be of the form

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

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

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

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

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

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

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

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

$$G(x, y) = \begin{cases} x/c, & x \leq y, \\ y/c, & x \geq y, \end{cases} \quad (11.71)$$

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 (11.69) implies that the solution to the boundary value problem (11.70) can be written as a single integral, namely

$$u(x) = \int_0^1 G(x, y) f(y) dy = \frac{1}{c} \int_0^x y f(y) dy + \frac{1}{c} \int_x^1 x f(y) dy. \quad (11.72)$$

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

11.3. Adjoint 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.

Adjoint 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 $D: U \rightarrow V$ 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

$$\langle u; \tilde{u} \rangle = \int_0^\ell u(x) \tilde{u}(x) dx, \quad \langle\langle v; \tilde{v} \rangle\rangle = \int_0^\ell v(x) \tilde{v}(x) dx, \quad (11.73)$$

[†] 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.

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 (7.74), the adjoint D^* of the derivative operator must satisfy the inner product identity

$$\langle\langle D[u]; v \rangle\rangle = \langle u; D^*[v] \rangle \quad \text{for all} \quad u \in U, \quad v \in V. \quad (11.74)$$

First, we compute the left hand side:

$$\langle\langle D[u]; v \rangle\rangle = \left\langle\left\langle \frac{du}{dx}; v \right\rangle\right\rangle = \int_0^\ell \frac{du}{dx} v \, dx. \quad (11.75)$$

On the other hand, the right hand side should equal

$$\langle u; D^*[v] \rangle = \int_0^\ell u D^*[v] \, dx. \quad (11.76)$$

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 (11.75), 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

$$\int_0^\ell \frac{du}{dx} v \, dx = [u(\ell) v(\ell) - u(0) v(0)] - \int_0^\ell u \frac{dv}{dx} \, dx. \quad (11.77)$$

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

$$- \int_0^\ell u \frac{dv}{dx} \, dx = \int_0^\ell u \left[- \frac{dv}{dx} \right] \, dx = \left\langle u; - \frac{dv}{dx} \right\rangle = \langle u; -D[v] \rangle. \quad (11.78)$$

Equating (11.75) and (11.78), 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 (11.74), we must have

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

and so

$$\left(\frac{d}{dx} \right)^* = D^* = -D = - \frac{d}{dx}. \quad (11.79)$$

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 (11.77) vanish:

$$u(\ell) v(\ell) - u(0) v(0) = 0, \quad (11.80)$$

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

$$u(0) = 0, \quad u(\ell) = 0, \quad (11.81)$$

will ensure that (11.80) holds, and therefore validate (11.79). The homogeneous boundary conditions serve to define the vector space

$$U = \{ u(x) \in C^2[0, \ell] \mid u(0) = u(\ell) = 0 \}$$

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

The fixed boundary conditions (11.81) are not the only possibilities that ensure the vanishing of the boundary terms (11.80). An evident alternative is to require that the strain vanish at both endpoints, $v(0) = v(\ell) = 0$. In this case, the strain space

$$V = \{ v(x) \in C^1[0, \ell] \mid v(0) = v(\ell) = 0 \}$$

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

$$U = \{ u(x) \in C^2[0, \ell] \mid u'(0) = u'(\ell) = 0 \}.$$

Thus, this case corresponds to the free boundary conditions of Example 11.3. Again, restricting $D: U \rightarrow V$ to these particular vector spaces ensures that the boundary terms (11.80) vanish, and so (11.79) holds in this situation too.

Let us list the most important combinations of boundary conditions that imply the vanishing of the boundary terms (11.80), and so ensure the validity of the adjoint equation (11.79).

Self-Adjoint Boundary Conditions for a Bar

- | | |
|---------------------------------|--|
| (a) Both ends fixed: | $u(0) = u(\ell) = 0.$ |
| (b) One free and one fixed end: | $u(0) = u'(\ell) = 0 \quad \text{or} \quad u'(0) = u(\ell) = 0.$ |
| (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 6.3.11.

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

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

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

$$K[u] = f, \quad \text{where} \quad K = D^* \circ D = -D^2. \quad (11.82)$$

We note that

$$K^* = (D^* \circ D)^* = D^* \circ (D^*)^* = D^* \circ D = K, \quad (11.83)$$

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

$$\langle K[u]; \tilde{u} \rangle = \int_0^\ell [-u''(x) \tilde{u}(x)] dx = \int_0^\ell [-u(x) \tilde{u}''(x)] dx = \langle u; K[\tilde{u}] \rangle \quad (11.84)$$

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

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

$$\langle u; \tilde{u} \rangle = \int_0^\ell u(x) \tilde{u}(x) dx, \quad u, \tilde{u} \in U, \quad (11.85)$$

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

$$\langle\langle v; \tilde{v} \rangle\rangle = \int_0^\ell v(x) \tilde{v}(x) c(x) dx, \quad v, \tilde{v} \in V, \quad (11.86)$$

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 (11.86) to define a *bona fide* inner product, is in accordance with the underlying physical assumptions.

Let us recompute the adjoint of the derivative operator $D: U \rightarrow V$, this time with respect to the inner products (11.85–86). Now we need to compare

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

Integrating the first expression by parts, we find

$$\int_0^\ell \frac{du}{dx} cv dx = [u(\ell) c(\ell) v(\ell) - u(0) c(0) v(0)] - \int_0^\ell u \frac{d(cv)}{dx} dx = \int_0^\ell u \left[-\frac{d(cv)}{dx} \right] dx, \quad (11.87)$$

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

provided that we choose our boundary conditions so that

$$u(\ell)c(\ell)v(\ell) - u(0)c(0)v(0) = 0. \quad (11.88)$$

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

$$D^*[v] = -\frac{d(cv)}{dx} = -c\frac{dv}{dx} - c'v. \quad (11.89)$$

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

$$K[u] = -\frac{d}{dx} \left(c(x) \frac{du}{dx} \right), \quad (11.90)$$

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 (11.84),

$$\begin{aligned} G(z, y) &= \int_0^\ell G(x, y) \delta(x - z) dx = \langle G(x, y); \delta(x - z) \rangle = \langle G(x, y); K[G(x, z)] \rangle \quad (11.91) \\ &= \langle K[G(x, y)]; G(x, z) \rangle = \langle \delta(x - y); G(x, z) \rangle = \int_0^\ell G(x, z) \delta(x - y) dx = G(y, z), \end{aligned}$$

for any $0 < y, z < \ell$, which validates[†] the symmetry equation (11.68).

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 $K: U \rightarrow U$ 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 (11.83). 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,

$$\langle K[u]; u \rangle = \langle D^*[D[u]]; u \rangle = \langle\langle D[u]; D[u] \rangle\rangle = \|D[u]\|^2 \geq 0, \quad (11.92)$$

[†] Symmetry at the endpoints follows from continuity.

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., $u \in \ker D$. 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 $C^1[0, \ell]$ 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 $\ker D \subset \tilde{U} = \{u'(0) = u'(\ell) = 0\}$ is a one-dimensional subspace. Therefore, 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

$$\mathcal{P}[u] = \frac{1}{2} \|D[u]\|^2 - \langle u; f \rangle = \int_0^\ell \left[\frac{1}{2} c(x) u'(x)^2 - f(x) u(x) \right] dx. \quad (11.93)$$

A proof of this general fact appears following Theorem 7.61. Pay attention: the norm in (11.93) refers to the strain space V , and so is associated with the weighted inner product (11.86), 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

$$-u'' = f(x), \quad u(0) = 0, \quad u(\ell) = 0. \quad (11.94)$$

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

$$\langle K[u]; u \rangle = \int_0^\ell [-u''(x)u(x)] dx = \int_0^\ell [u'(x)]^2 dx > 0 \quad (11.95)$$

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

$$\mathcal{P}[u] = \frac{1}{2} \|u'\|^2 - \langle u; f \rangle = \int_0^\ell \left[\frac{1}{2} u'(x)^2 - f(x)u(x) \right] dx.$$

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 $-u_\star'' = f$,

$$\begin{aligned} \mathcal{P}[u] &= \int_0^\ell \left[\frac{1}{2} (u')^2 + u_\star'' u \right] dx = u_\star'(\ell) u(\ell) - u_\star'(0) u(0) + \int_0^\ell \left[\frac{1}{2} (u')^2 - u_\star' u' \right] dx \\ &= \int_0^\ell \frac{1}{2} (u' - u_\star')^2 dx - \int_0^\ell \frac{1}{2} (u_\star')^2 dx, \end{aligned} \quad (11.96)$$

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

$$K[u] = -\frac{d}{dx} \left(c(x) \frac{du}{dx} \right) = f(x), \quad u(0) = \alpha, \quad u(\ell) = \beta. \quad (11.97)$$

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

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

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

$$h(x) = \alpha + \frac{\beta - \alpha}{\ell} x. \quad (11.98)$$

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

$$\tilde{u}(x) = u(x) - h(x) \quad (11.99)$$

satisfies the homogeneous Dirichlet boundary conditions

$$\tilde{u}(0) = \tilde{u}(\ell) = 0. \quad (11.100)$$

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

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

or, explicitly,

$$-\frac{d}{dx} \left(c(x) \frac{d\tilde{u}}{dx} \right) = \tilde{f}(x), \quad \text{where} \quad \tilde{f}(x) = f(x) + \frac{d}{dx} \left(c(x) \frac{dh}{dx} \right). \quad (11.101)$$

For the particular choice (11.98),

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

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

$$u(x) = \tilde{u}(x) + h(x). \quad (11.102)$$

We know that the homogeneous Dirichlet boundary value problem (11.100–101) is positive definite, and so we can characterize its solution by a minimum principle, namely as the minimizer of the quadratic energy functional

$$\mathcal{P}[\tilde{u}] = \frac{1}{2} \|\tilde{u}'\|^2 - \langle \tilde{u}; \tilde{f} \rangle = \int_0^\ell \left[\frac{1}{2} c(x) \tilde{u}'(x)^2 - \tilde{f}(x) \tilde{u}(x) \right] dx. \quad (11.103)$$

Let us rewrite the minimization principle in terms of the original displacement function $u(x)$. Replacing $\tilde{u}$ and $\tilde{f}$ by their formulae (11.99, 101) yields

$$\begin{aligned} \mathcal{P}[u] &= \frac{1}{2} \|u' - h'\|^2 - \langle u - h; f - K[h] \rangle \\ &= \left[\frac{1}{2} \|u'\|^2 - \langle u; f \rangle \right] - \left[\langle u'; h' \rangle - \langle u; K[h] \rangle \right] + \left[\frac{1}{2} \|h'\|^2 + \langle h; f - K[h] \rangle \right] \\ &= \mathcal{P}[u] - \left[\langle u'; h' \rangle - \langle u; K[h] \rangle \right] + C_0. \end{aligned} \quad (11.104)$$

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 (11.93) for the actual displacement $u(x)$. The middle terms can be explicitly evaluated:

$$\begin{aligned} \langle u'; h' \rangle - \langle u; K[h] \rangle &= \int_0^\ell \left[c(x) h'(x) u'(x) + (c(x) h'(x))' u(x) \right] dx \\ &= \int_0^\ell \frac{d}{dx} [c(x) h'(x) u(x)] dx = c(\ell) h'(\ell) u(\ell) - c(0) h'(0) u(0). \end{aligned} \quad (11.105)$$

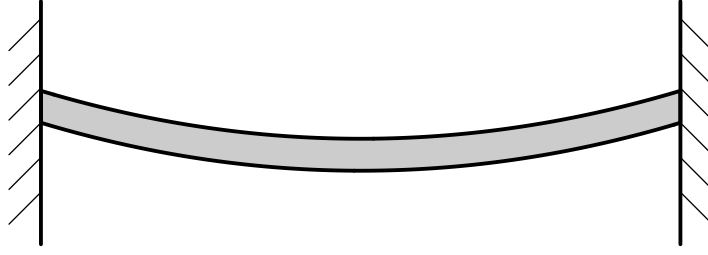


Figure 11.13. Bending of a Beam.

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,

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

differ by a constant. We conclude that, if the function $\tilde{u}$ minimizes $\mathcal{P}[\tilde{u}]$, then $u = \tilde{u} + h$ 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*

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

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

Warning: The inhomogeneous mixed boundary value problem is trickier, since the extra terms (11.105) *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 xy plane, as sketched in Figure 11.13. Let $0 \leq x \leq \ell$ 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

$$\kappa = \frac{u''}{(1 + (u')^2)^{3/2}}. \quad (11.106)$$

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 (11.106) is replaced by its linear approximation

$$\kappa \approx u''. \quad (11.107)$$

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

$$w(x) = c(x) v(x) = c(x) \frac{d^2 u}{dx^2}, \quad (11.108)$$

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

$$L^*[v] = L^* \circ L[u] = f. \quad (11.109)$$

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:

$$\langle u; \tilde{u} \rangle = \int_a^b u(x) \tilde{u}(x) dx, \quad \langle\langle v; \tilde{v} \rangle\rangle = \int_a^b v(x) \tilde{v}(x) c(x) dx. \quad (11.110)$$

According to the general adjoint equation (7.74), we need to equate

$$\int_0^\ell L[u] v c dx = \langle\langle L[u]; v \rangle\rangle = \langle u; L^*[v] \rangle = \int_0^\ell u L^*[v] dx. \quad (11.111)$$

[†] By definition, [9, 168], 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 A.5.11 for details.

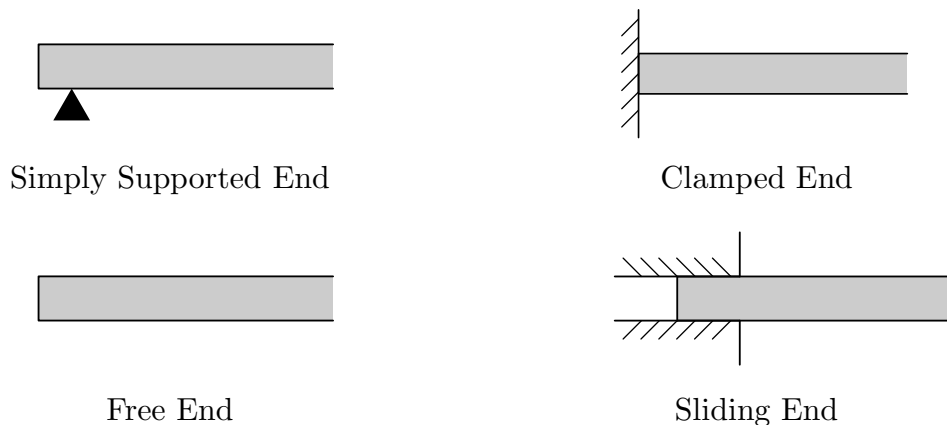


Figure 11.14. Boundary Conditions for a Beam.

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

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

Comparing with (11.111), we conclude that $L^*[v] = D^2(c v)$ provided the boundary terms vanish:

$$\begin{aligned} \left[\frac{du}{dx} c v - u \frac{d(c v)}{dx} \right] \Big|_{x=0}^\ell &= \left[\frac{du}{dx} w - u \frac{dw}{dx} \right] \Big|_{x=0}^\ell \\ &= [u'(\ell) w(\ell) - u(\ell) w'(\ell)] - [u'(0) w(0) - u(0) w'(0)] = 0. \end{aligned} \quad (11.112)$$

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

$$L^*[v] = \frac{d^2(c v)}{dx^2} = f(x). \quad (11.113)$$

A justification of (11.113) based on physical principles can be found in [181]. Combining (11.108, 113), we conclude that the equilibrium configuration of the beam is characterized as a solution to the fourth order ordinary differential equation

$$\frac{d^2}{dx^2} \left(c(x) \frac{d^2 u}{dx^2} \right) = f(x). \quad (11.114)$$

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. (11.112). 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*

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

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 equation (11.114) reduces to the elementary fourth order ordinary differential equation

$$\frac{d^4 u}{dx^4} = 0. \tag{11.115}$$

The general solution is an arbitrary cubic polynomial,

$$u = a x^3 + b x^2 + c x + d. \tag{11.116}$$

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

$$u(0) = 0, \quad u'(0) = \beta, \quad u(1) = 0, \quad u'(1) = 0, \tag{11.117}$$

so that the left end is tilted by a (small) angle $\tan^{-1} \beta$. We substitute the solution formula (11.116) into the boundary conditions (11.117) and solve for

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

The resulting solution

$$u(x) = \beta (x^3 - 2x^2 + x) = \beta x(1 - x)^2 \tag{11.118}$$

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

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



Figure 11.15. Hermite Cubic Spline.

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

$$u(0) = \alpha, \quad u'(0) = 0, \quad u(1) = 0, \quad u'(1) = 0. \quad (11.119)$$

Substituting (11.116) and solving for a, b, c, d , we find that the solution is

$$u(x) = \alpha (1 - x)^2 (2x + 1). \quad (11.120)$$

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 (11.118) and (11.120):

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

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

$$u(0) = 0, \quad u'(0) = 0, \quad u(1) = 0, \quad u'(1) = 0. \quad (11.121)$$

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

$$\frac{d^4 u}{dx^4} = \delta(x - y) \quad (11.122)$$

corresponding to a concentrated unit impulse applied at position y along the beam. Integrating (11.122) four times, using (11.46) with $n = 4$, we produce the general solution

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

to the differential equation (11.122). The boundary conditions (11.121) require

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

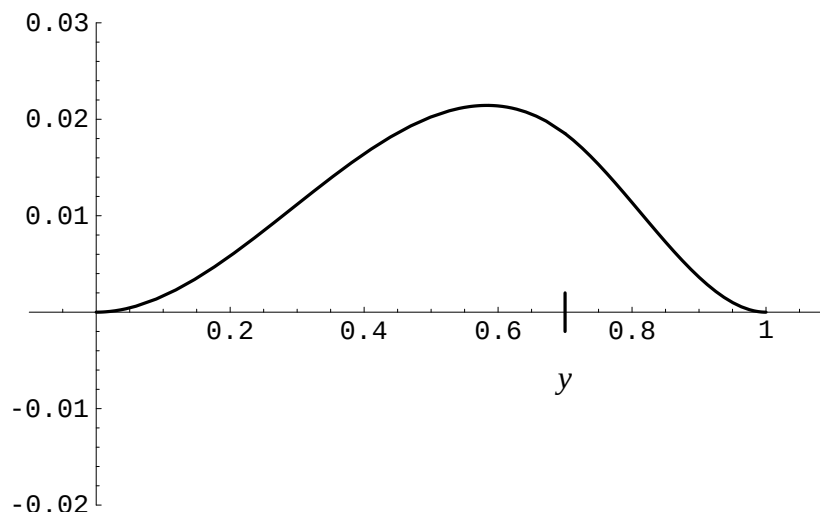


Figure 11.16. Green's Function for a Beam with Two Fixed Ends.

and hence

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

Therefore, the Green's function is

$$G(x, y) = \begin{cases} \frac{1}{6} x^2 (1 - y)^2 (3y - x - 2xy), & x < y, \\ \frac{1}{6} y^2 (1 - x)^2 (3x - y - 2xy), & x > y. \end{cases} \quad (11.123)$$

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. (11.91). 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 (11.115) 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 (11.123) is graphed 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

$$\frac{d^4 u}{dx^4} = f(x), \quad u(0) = u'(0) = u(1) = u'(1) = 0, \quad (11.124)$$

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

$$f(x) = \int_0^1 f(y) \delta(x - y) dy$$

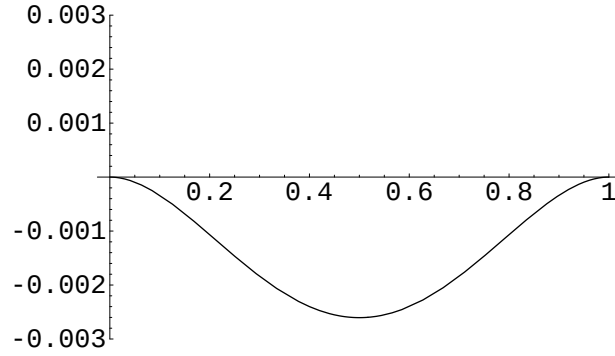


Figure 11.17. Deflection of a Uniform Beam under Gravity.

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

$$\begin{aligned}
 u(x) &= \int_0^1 G(x, y) f(y) dy \\
 &= \frac{1}{6} \int_0^x y^2 (1-x)^2 (3x-y-2xy) f(y) dy + \frac{1}{6} \int_x^1 x^2 (1-y)^2 (3y-x-2xy) f(y) dy.
 \end{aligned} \tag{11.125}$$

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

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

and graphed in Figure 11.17. Although we could, of course, obtain $u(x)$ by integrating the original differential equation (11.124) directly, writing the solution formula (11.125) 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 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[†]

$$\mathcal{P}[u] = \frac{1}{2} \|L[u]\|^2 - \langle u; f \rangle = \int_a^b \left[\frac{1}{2} c(x) u''(x)^2 - f(x) u(x) \right] dx \quad (11.126)$$

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.

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

$$u(x_j) = y_j, \quad j = 0, \dots, n. \quad (11.127)$$

The *mesh points* $x_0 < x_1 < x_2 < \dots < x_n$ are distinct and labeled in increasing order. The spline is modeled as an elastic beam, and so satisfies the homogeneous beam equation (11.115). Therefore,

$$u(x) = a_j + b_j(x - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3, \quad \begin{array}{l} x_j \leq x \leq x_{j+1}, \\ j = 0, \dots, n-1, \end{array} \quad (11.128)$$

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 (11.128) in terms of $x - x_j$ is merely for computational convenience.

Our problem is to determine the coefficients

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

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 (11.127). Since it is defined by a different formula on each side of the mesh point, this results in a total of $2n$ conditions:

$$\begin{aligned} u(x_j^+) &= a_j = y_j, \\ u(x_{j+1}^-) &= a_j + b_j h_j + c_j h_j^2 + d_j h_j^3 = y_{j+1}, \end{aligned} \quad j = 0, \dots, n-1, \quad (11.129)$$

[†] 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 (11.110).

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 (11.129) guarantee that $u(x)$ is continuous. The condition $u(x) \in C^1$ 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

$$b_j + 2c_j h_j + 3d_j h_j^2 = u'(x_{j+1}^-) = u'(x_{j+1}^+) = b_{j+1}, \quad j = 0, \dots, n-2. \quad (11.130)$$

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

$$2c_j + 6d_j h_j = u''(x_{j+1}^-) = u''(x_{j+1}^+) = 2c_{j+1}, \quad j = 0, \dots, n-2, \quad (11.131)$$

to ensure that u'' is continuous at the mesh points. We have now imposed a total of $4n - 2$ conditions, namely (11.129–131), 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

$$c_0 = 0, \quad c_{n-1} + 3d_{n-1} h_{n-1} = 0. \quad (11.132)$$

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

$$b_0 = \alpha, \quad b_{n-1} + 2c_{n-1} h_{n-1} + 3d_{n-1} h_{n-1}^2 = \beta. \quad (11.133)$$

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

$$b_0 = b_{n-1} + 2c_{n-1} h_{n-1} + 3d_{n-1} h_{n-1}^2, \quad c_0 = c_{n-1} + 3d_{n-1} h_{n-1}. \quad (11.134)$$

If we also require that the end interpolation values agree,

$$u(x_0) = y_0 = y_n = u(x_n), \quad (11.135)$$

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 $a = x_0 < x_1 < \dots < x_n = b$, and corresponding data values $y_0, y_1, \dots, y_n$, along with one of the three kinds of boundary conditions (11.132), (11.133), or (11.134). Then there exists a unique piecewise cubic spline function $u(x) \in C^2[a, b]$ that interpolates the data, $u(x_0) = y_0, \dots, u(x_n) = y_n$, 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 (11.129) says that

$$a_j = y_j, \quad j = 0, \dots, n-1. \quad (11.136)$$

Next, (11.131–132) imply that

$$d_j = \frac{c_{j+1} - c_j}{3h_j}. \quad (11.137)$$

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

$$c_n = 0.$$

We now substitute (11.136–137) into the second set of equations in (11.129), and then solve the resulting equation for

$$b_j = \frac{y_{j+1} - y_j}{h_j} - \frac{(2c_j + c_{j+1})h_j}{3}. \quad (11.138)$$

Substituting this result and (11.137) back into (11.130), and simplifying, we find

$$h_j c_j + 2(h_j + h_{j+1})c_{j+1} + h_{j+1}c_{j+2} = 3 \left[\frac{y_{j+2} - y_{j+1}}{h_{j+1}} - \frac{y_{j+1} - y_j}{h_j} \right] = z_{j+1}, \quad (11.139)$$

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

$$c_0 = 0, \quad c_n = 0,$$

and so (11.139) constitutes a tridiagonal linear system

$$A \mathbf{c} = \mathbf{z}, \quad (11.140)$$

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

$$A = \begin{pmatrix} 2(h_0 + h_1) & h_1 & & & \\ h_1 & 2(h_1 + h_2) & h_2 & & \\ & h_2 & 2(h_2 + h_3) & h_3 & \\ & & \ddots & \ddots & \ddots \\ & & & h_{n-3} & 2(h_{n-3} + h_{n-2}) & h_{n-2} \\ & & & & h_{n-2} & 2(h_{n-2} + h_{n-1}) \end{pmatrix} \quad (11.141)$$

and right hand side $\mathbf{z} = (z_1, z_2, \dots, z_{n-1})^T$. Once (11.141) has been solved, we will then use (11.136–138) to reconstruct the other spline coefficients a_j, b_j, d_j .

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

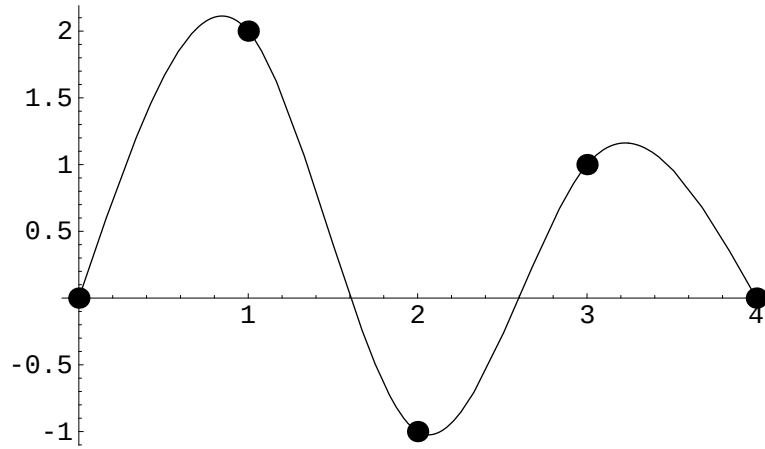


Figure 11.18. A Cubic Spline.

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 (11.140), we can apply our tridiagonal solution algorithm (1.66). Let us specialize to the most important case, when the mesh points are equally spaced in the interval $[a, b]$, so that

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

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

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

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

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

With this convention, the basic equations (11.136–139) are the same. In this case, the coefficient matrix for the linear system

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

is of *circulant tridiagonal* form:

$$A = \begin{pmatrix} 2(h_{n-1} + h_0) & h_0 & & & h_{n-1} \\ h_0 & 2(h_0 + h_1) & h_1 & & \\ & h_1 & 2(h_1 + h_2) & h_2 & \\ & & \ddots & \ddots & \ddots \\ & & & h_{n-3} & 2(h_{n-3} + h_{n-2}) & h_{n-2} \\ h_{n-1} & & & & h_{n-2} & 2(h_{n-2} + h_{n-1}) \end{pmatrix}. \quad (11.142)$$

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 1.7.14.

One immediate application of splines is curve fitting in computer aided design and graphics. The basic problem is to draw a smooth parametrized curve $\mathbf{u}(t) = (u(t), v(t))^T$ 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, [64, 163].

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

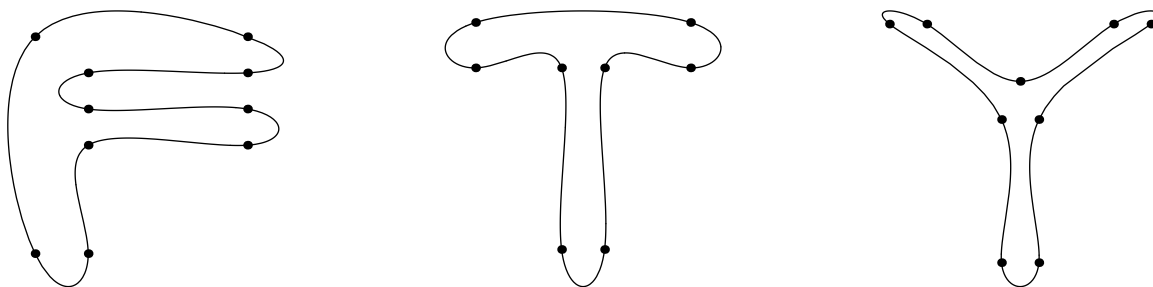


Figure 11.19. Three Sample Spline Letters.

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.

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

$$-\frac{d}{dx} \left(p(x) \frac{du}{dx} \right) + q(x)u = -p(x) \frac{d^2u}{dx^2} - p'(x) \frac{du}{dx} + q(x)u = f(x), \quad (11.143)$$

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

$$u(a) = 0, \quad u(b) = 0. \quad (11.144)$$

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 $a \leq x \leq b$. To ensure positive definiteness of the Sturm–Liouville differential operator, we also assume $q(x) \geq 0$. 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 [117].

Theorem 11.12. *Let $p(x) > 0$ and $q(x) \geq 0$ for $a \leq x \leq b$. Then the Sturm–Liouville boundary value problem (11.143–144) 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, [144].

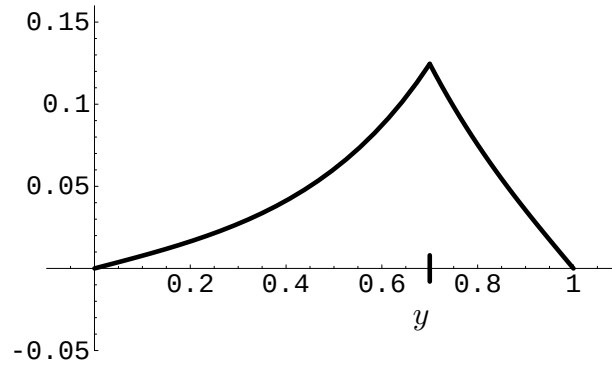


Figure 11.20. Green's Function for the Constant Coefficient Sturm–Liouville Problem.

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

$$-u'' + \omega^2 u = f(x), \quad u(0) = u(1) = 0. \quad (11.145)$$

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

$$-u'' + \omega^2 u = \delta(x - y), \quad u(0) = u(1) = 0. \quad (11.146)$$

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

$$\cosh \omega x = \frac{e^{\omega x} + e^{-\omega x}}{2}, \quad \sinh \omega x = \frac{e^{\omega x} - e^{-\omega x}}{2}. \quad (11.147)$$

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 (11.146) has the form

$$G(x, y) = \begin{cases} a \sinh \omega x, & x < y, \\ b \sinh \omega(1 - x), & x > y. \end{cases} \quad (11.148)$$

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

$$a \sinh \omega y = b \sinh \omega(1 - y). \quad (11.149)$$

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 (11.146) 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

$$a \omega \cosh \omega y - 1 = -b \omega \cosh \omega(1 - y). \quad (11.150)$$

If we multiply (11.149) by $\omega \cosh \omega (1 - y)$ and (11.150) by $\sinh \omega (1 - y)$ and then add the results together, we find

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

where we used the addition formula for the hyperbolic sine:

$$\sinh(\alpha + \beta) = \sinh \alpha \cosh \beta + \cosh \alpha \sinh \beta. \quad (11.151)$$

Therefore,

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

and the Green's function is

$$G(x, y) = \begin{cases} \frac{\sinh \omega x \sinh \omega (1 - y)}{\omega \sinh \omega}, & x < y, \\ \frac{\sinh \omega (1 - x) \sinh \omega y}{\omega \sinh \omega}, & x > y. \end{cases} \quad (11.152)$$

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 (11.145) is given by the basic superposition formula (11.63), which becomes

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

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

$$\begin{aligned} u(x) &= \int_0^x \frac{\sinh \omega (1 - x) \sinh \omega y}{\omega \sinh \omega} dy + \int_x^1 \frac{\sinh \omega x \sinh \omega (1 - y)}{\omega \sinh \omega} dy \\ &= \frac{\sinh \omega (1 - x) (\cosh \omega x - 1)}{\omega^2 \sinh \omega} + \frac{\sinh \omega x (\cosh \omega (1 - x) - 1)}{\omega^2 \sinh \omega} \\ &= \frac{1}{\omega^2} - \frac{\sinh \omega x + \sinh \omega (1 - x)}{\omega^2 \sinh \omega}. \end{aligned} \quad (11.153)$$

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 (11.144), the domain of L will be the vector space

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

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

To proceed, we must compute the adjoint of $L: U \rightarrow V$. To recover the Sturm–Liouville problem, we use the standard L^2 inner product (11.85) on U , but adopt a weighted inner product

$$\langle\langle \mathbf{v}; \mathbf{w} \rangle\rangle = \int_a^b [p(x)v_1(x)w_1(x) + q(x)v_2(x)w_2(x)] dx, \quad \mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}, \quad (11.154)$$

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 (7.74), the adjoint $L^*: V \rightarrow U$ 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:

$$\begin{aligned} \langle\langle L[u]; \mathbf{v} \rangle\rangle &= \int_a^b [p u' v_1 + q u v_2] dx \\ &= p(b)u(b)v_1(b) - p(a)u(a)v_1(a) + \int_a^b u [-(pv_1)' + qv_2] dx. \end{aligned}$$

The Dirichlet conditions (11.144) ensure that the boundary terms vanish, and therefore,

$$\langle\langle L[u]; \mathbf{v} \rangle\rangle = \int_a^b u [-(pv_1)' + qv_2] dx = \langle u; L^*[\mathbf{v}] \rangle.$$

We conclude that the adjoint operator is given by

$$L^*[\mathbf{v}] = -\frac{d(pv_1)}{dx} + qv_2.$$

The canonical self-adjoint combination

$$K[u] = L^* \circ L[u] = L^* \begin{pmatrix} u' \\ u \end{pmatrix} = -\frac{d}{dx} \left(p \frac{du}{dx} \right) + qu \quad (11.155)$$

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

$$\mathcal{P}[u] = \frac{1}{2} \|L[u]\|^2 - \langle u; f \rangle = \int_a^b \left[\frac{1}{2} p(x) u'(x)^2 + \frac{1}{2} q(x) u(x)^2 - f(x) u(x) \right] dx \quad (11.156)$$

among all C^2 functions satisfying the prescribed boundary conditions. For example, the solution to the constant coefficient Sturm–Liouville problem (11.145) can be characterized as minimizing the quadratic functional

$$\mathcal{P}[u] = \int_0^1 \left[\frac{1}{2} u'^2 + \frac{1}{2} \omega^2 u^2 - f u \right] dx$$

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 [174, 197]. 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

$$\mathcal{P}[u] = \frac{1}{2} \|L[u]\|^2 - \langle f; u \rangle \quad (11.157)$$

for the linear system

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

representing our boundary value problem. The norm in (11.157) is typically based on some form of weighted inner product $\langle\langle v; \tilde{v} \rangle\rangle$ 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

$$\varphi(x) = c_1 \varphi_1(x) + \dots + c_n \varphi_n(x) \quad (11.158)$$

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 (11.158) into (11.157) and expanding we find

$$\mathcal{P}[\varphi] = \frac{1}{2} \sum_{i,j=1}^n m_{ij} c_i c_j - \sum_{i=1}^n b_i c_i = \frac{1}{2} \mathbf{c}^T M \mathbf{c} - \mathbf{c}^T \mathbf{b}, \quad (11.159)$$

where

- (a) $\mathbf{c} = (c_1, c_2, \dots, c_n)^T$ is the vector of unknown coefficients in (11.158),
- (b) $M = (m_{ij})$ is the symmetric $n \times n$ matrix with entries

$$m_{ij} = \langle L[\varphi_i]; L[\varphi_j] \rangle, \quad i, j = 1, \dots, n, \quad (11.160)$$

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

$$b_i = \langle f; \varphi_i \rangle, \quad i = 1, \dots, n. \quad (11.161)$$

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 (11.159) over all possible vectors $\mathbf{c} \in \mathbb{R}^n$. The coefficient matrix M is, in fact, positive definite, since, by the preceding computation,

$$\mathbf{c}^T M \mathbf{c} = \sum_{i,j=1}^n m_{ij} c_i c_j = \|L[c_1 \varphi_1(x) + \dots + c_n \varphi_n]\|^2 = \|L[\varphi]\|^2 > 0 \quad (11.162)$$

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 (11.162) 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 (11.159) is obtained by solving the associated linear system

$$M \mathbf{c} = \mathbf{b}. \quad (11.163)$$

Solving (11.163) 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

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

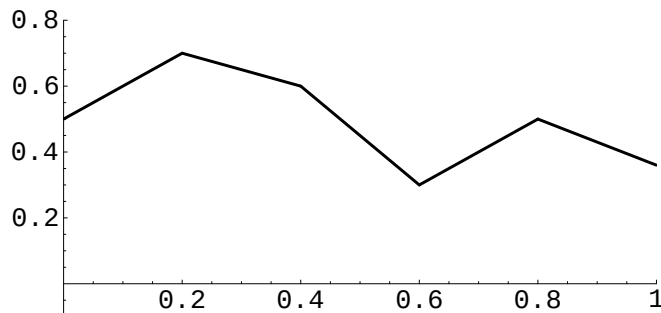


Figure 11.21. A Continuous Piecewise Affine Function.

not be a subspace of U . More importantly, in order to obtain sufficient accuracy, the linear algebraic system (11.163) 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, (11.93), and general Sturm–Liouville problems, (11.156), $\mathcal{P}[u]$ 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 is affine, $f(x) = ax + b$, 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

$$\varphi(a) = \varphi(b) = 0. \quad (11.164)$$

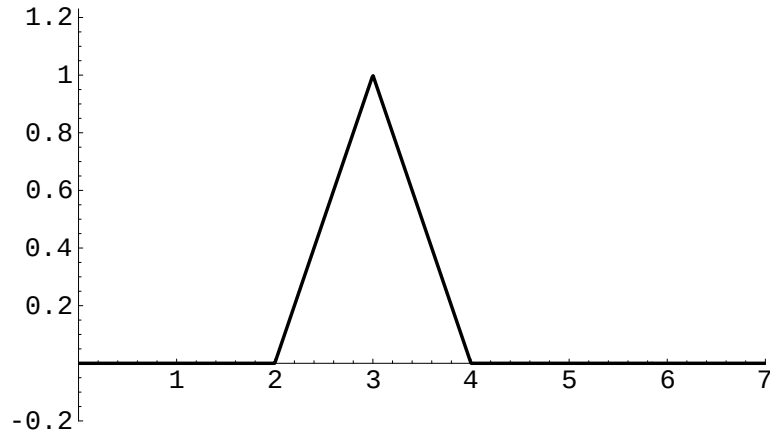


Figure 11.22. A Hat Function.

Thus, on each subinterval

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

Continuity of $\varphi(x)$ requires

$$c_j = \varphi(x_j^+) = \varphi(x_j^-) = c_{j-1} + b_{j-1}h_{j-1}, \quad j = 1, \dots, n-1, \quad (11.165)$$

where $h_{j-1} = x_j - x_{j-1}$ denotes the length of the j^{th} subinterval. The boundary conditions (11.164) require

$$\varphi(a) = c_0 = 0, \quad \varphi(b) = c_{n-1} + h_{n-1}b_{n-1} = 0. \quad (11.166)$$

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 (11.165) and the second boundary condition (11.166) 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 $\varphi''(x)$ 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 $\varphi_*(x)$ will, in practice, provide a reasonable approximation to the actual solution $u_*(x)$.

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

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

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

established:

$$\varphi_j(x) = \begin{cases} \frac{x - x_{j-1}}{x_j - x_{j-1}}, & x_{j-1} \leq x \leq x_j, \\ \frac{x_{j+1} - x}{x_{j+1} - x_j}, & x_j \leq x \leq x_{j+1}, \\ 0, & x \leq x_{j-1} \text{ or } x \geq x_{j+1}, \end{cases} \quad j = 1, \dots, n-1. \quad (11.167)$$

An advantage of using these basis elements is that the resulting coefficient matrix (11.160) turns out to be tridiagonal. Therefore, the tridiagonal Gaussian Elimination algorithm in (1.66) will rapidly produce the solution to the linear system (11.163). 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

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

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 (11.93) based on the quadratic functional

$$\mathcal{P}[u] = \frac{1}{2} \|u'\|^2 - \langle f; u \rangle = \int_0^\ell \left[\frac{1}{2} c(x) u'(x)^2 - f(x) u(x) \right] dx.$$

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

$$x_j = jh = \frac{j\ell}{n}, \quad j = 0, \dots, n.$$

The corresponding finite element basis hat functions are explicitly given by

$$\varphi_j(x) = \begin{cases} (x - x_{j-1})/h, & x_{j-1} \leq x \leq x_j, \\ (x_{j+1} - x)/h, & x_j \leq x \leq x_{j+1}, \\ 0, & \text{otherwise,} \end{cases} \quad j = 1, \dots, n-1. \quad (11.168)$$

The associated linear system (11.163) has coefficient matrix entries

$$m_{ij} = \langle \varphi'_i; \varphi'_j \rangle = \int_0^\ell \varphi'_i(x) \varphi'_j(x) c(x) dx, \quad i, j = 1, \dots, n-1.$$

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,

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

Therefore, the coefficient matrix has the tridiagonal form

$$M = \frac{1}{h^2} \begin{pmatrix} s_0 + s_1 & -s_1 & & & \\ -s_1 & s_1 + s_2 & -s_2 & & \\ & -s_2 & s_2 + s_3 & -s_3 & \\ & & \ddots & \ddots & \ddots \\ & & & -s_{n-3} & s_{n-3} + s_{n-2} & -s_{n-2} \\ & & & & -s_{n-2} & s_{n-2} + s_{n-1} \end{pmatrix}, \quad (11.169)$$

where

$$s_j = \int_{x_j}^{x_{j+1}} c(x) dx \quad (11.170)$$

is the total stiffness of the j^{th} subinterval. For example, in the homogeneous case $c(x) \equiv 1$, the coefficient matrix (11.169) reduces to the very special form

$$M = \frac{1}{h} \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & \ddots & \ddots & \ddots \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}. \quad (11.171)$$

The corresponding right hand side has entries

$$\begin{aligned} b_j &= \langle f; \varphi_j \rangle = \int_0^\ell f(x) \varphi_j(x) dx \\ &= \frac{1}{h} \left[\int_{x_{j-1}}^{x_j} (x - x_{j-1}) f(x) dx + \int_{x_j}^{x_{j+1}} (x_{j+1} - x) f(x) dx \right]. \end{aligned} \quad (11.172)$$

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 (11.170, 172), 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[†], [35, 168], to approximate them:

$$s_j \approx \frac{h}{2} [c(x_j) + c(x_{j+1})], \quad b_j \approx h f(x_j). \quad (11.173)$$

[†] 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.

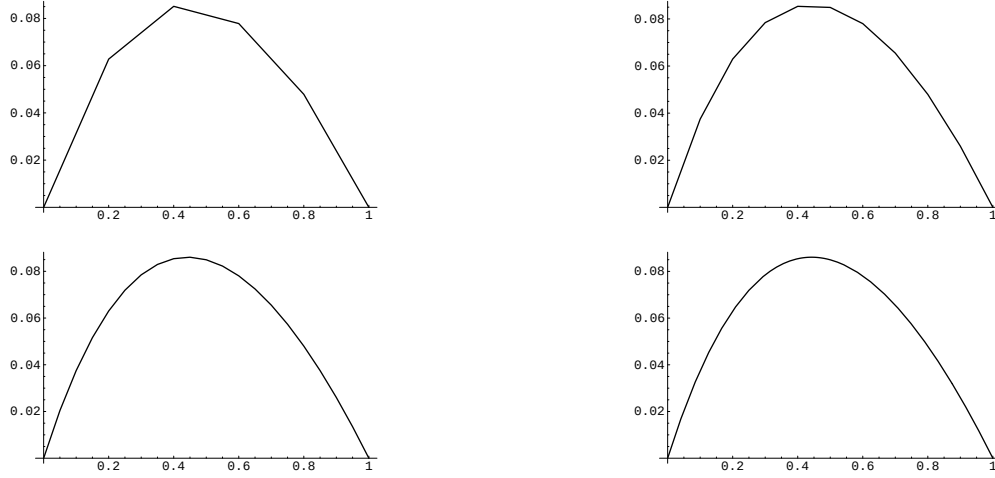


Figure 11.23. Finite Element Solution to (11.175).

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

$$-\frac{c_{j+1} - 2c_j + c_{j-1}}{h^2} = -\frac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1}))}{h^2} = -f(x_j). \quad (11.174)$$

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.

Example 11.15. Consider the boundary value problem

$$-\frac{d}{dx}(x+1)\frac{du}{dx} = 1, \quad u(0) = 0, \quad u(1) = 0. \quad (11.175)$$

The explicit solution is easily found by direct integration:

$$u(x) = -x + \frac{\log(x+1)}{\log 2}. \quad (11.176)$$

It minimizes the associated quadratic functional

$$\mathcal{P}[u] = \int_0^1 \left[\frac{1}{2}(x+1)u'(x)^2 - u(x) \right] dx \quad (11.177)$$

over all possible functions $u \in C^1$ that satisfy the given boundary conditions. The finite element system (11.163) has coefficient matrix given by (11.169) and right hand side (11.172), where

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

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 (11.176). Even when computed on rather coarse meshes, the finite element approximation is quite respectable.

Example 11.16. Consider the Sturm–Liouville boundary value problem

$$-u'' + (x+1)u = xe^x, \quad u(0) = 0, \quad u(1) = 0. \quad (11.178)$$

The solution minimizes the quadratic functional (11.156), which in this particular case is

$$\mathcal{P}[u] = \int_0^1 \left[\frac{1}{2} u'(x)^2 + \frac{1}{2} (x+1) u(x)^2 - e^x u(x) \right] dx, \quad (11.179)$$

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 (11.168). The matrix entries are given by[†]

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

while

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

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. [144].

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 $\tilde{u}(x) = u(x) - h(x)$, where $h(x)$ is any convenient function that satisfies the boundary conditions. For example, for the inhomogeneous Dirichlet conditions

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

we can subtract off the affine function

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

[†] 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 .

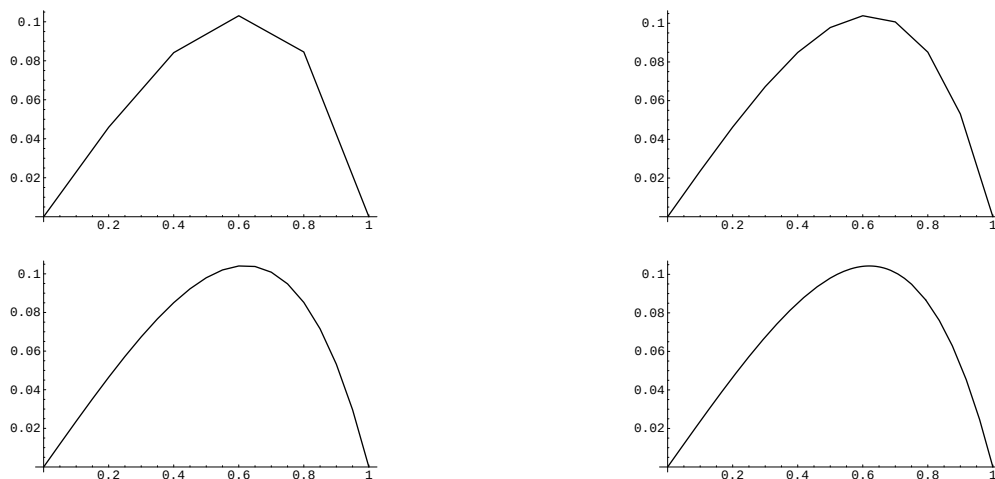


Figure 11.24. Finite Element Solution to (11.178).

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 $\tilde{u}(x) = u(x) - h(x)$ satisfies the amended differential equation

$$K[\tilde{u}] = \tilde{f}, \quad \text{where} \quad \tilde{f} = f - K[h],$$

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 (11.167) 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 (11.160) 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 [188] 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

$$K[\mathbf{u}] = \mathbf{f}, \quad (11.180)$$

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

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

According to the definition (7.74), one can replace K by its adjoint $K^*: W \rightarrow V$, and require

$$\langle \mathbf{u}; K^*[\mathbf{v}] \rangle = \langle \mathbf{f}; \mathbf{v} \rangle \quad \text{for all} \quad \mathbf{v} \in V. \quad (11.181)$$

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 (11.181) 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

$$\langle u; K^*[\varphi] \rangle = \langle f; \varphi \rangle \quad \text{for all} \quad \varphi$$

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

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

[†] The method also straightforwardly extends to nonlinear systems.

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

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

This “semi-weak” formulation is known in mechanics as the *principle of virtual work*, [169]. 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 (11.181) of the differential equation:

$$\langle u; K[\varphi] \rangle = - \int_0^\ell u(x) \frac{d}{dx} \left(c(x) \frac{d\varphi}{dx} \right) dx = \int_0^\ell f(x) \varphi(x) dx = \langle f; \varphi \rangle.\tag{11.183}$$

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

$$\varphi(x) = c_1 \varphi_1(x) + \dots + c_{n-1} \varphi_{n-1}(x)\tag{11.184}$$

satisfy the orthogonality condition (11.181) only for elements $\varphi \in W$ of the subspace. As usual, this only needs to be checked on the basis elements. Substituting (11.184) into the semi-weak form of the system, (11.182), produces a linear system of equations of the form

$$\langle w; K[\varphi_i] \rangle = \sum_{j=1}^n m_{ij} c_j = b_i = \langle f; \varphi_i \rangle, \quad i = 1, \dots, n.\tag{11.185}$$

The reader will recognize this as exactly the same finite element linear system (11.163) 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 [76, 188].