Chapter 14

Variational Approximation of Boundary-Value Problems; Introduction to the Finite Elements Method

14.1 A One-Dimensional Problem: Bending of a Beam

Consider a beam of unit length supported at its ends in 0 and 1, stretched along its axis by a force P, and subjected to a transverse load $f(x)dx$ per element dx, as illustrated in Figure 14.1.

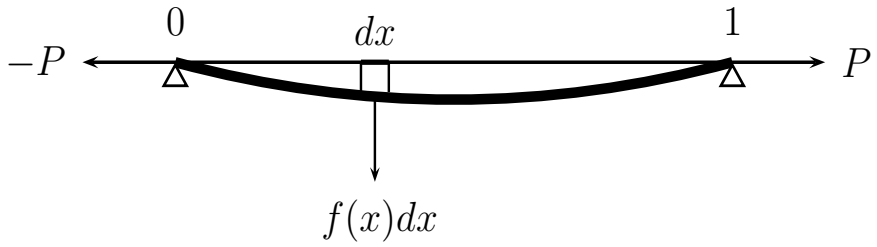

Figure 14.1: Vertical deflection of a beam

The bending moment $u(x)$ at the absissa x is the solution of a boundary problem (BP) of the form

$$\begin{aligned} -u''(x) + c(x)u(x) &= f(x), \quad 0 < x < 1 \\ u(0) &= \alpha \\ u(1) &= \beta, \end{aligned}$$

where $c(x) = P/(EI(x))$, where E is the Young's modulus of the material of which the beam is made and $I(x)$ is the principal moment of inertia of the cross-section of the beam at the abcissa x, and with $\alpha = \beta = 0$.

Remark: The vertical deflection $w(x)$ of the beam and the bending moment $u(x)$ are related by the equation

$$u(x) = -EI\frac{d^2w}{dx^2}.$$

For this problem, we may assume that $c(x) \geq 0$ for all $x \in [0, 1]$.

If we seek a solution $u \in C^2([0, 1])$, that is, a function whose first and second derivatives exist and are continuous, then it can be shown that the problem has a unique solution (assuming c and f to be continuous functions on $[0, 1]$).

Except in very rare situations, this problem has no closed-form solution, so we are led to seek approximations of the solutions.

One one way to proceed is to use the *finite difference method*, where we discretize the problem and replace derivatives by differences.

Another way is to use a *variational approach.*

In this approach, we follow a somewhat surprising path in which we come up with a so-called "weak formulation" of the problem, by using a trick based on integrating by parts!

First, let us observe that we can always assume that $\alpha = \beta = 0$, by looking for a solution of the form $u(x) - (\alpha(1-x) + \beta x)$.

This turns out to be crucial when we integrate by parts.

There are a lot of subtle mathematical details involved to make what follows rigorous, but here, we will take a "relaxed" approach.

First, we need to specify the space of "weak solutions."

This will be the vector space V of continuous functions f on $[0,1]$, with $f(0) = f(1) = 0$, and which are piecewise continuously differentiable on $[0,1]$.

This means that there is a finite number of points $x_0, \ldots,$ x_{N+1} with $x_0 = 0$ and $x_{N+1} = 1$, such that $f'(x_i)$ is undefined for $i = 1, \ldots, N$, but otherwise f' is defined and continuous on each interval (x_i, x_{i+1}) for $i = 0, \ldots, N$.

The space V becomes a Euclidean vector space under the inner product

$$\langle f, g \rangle_V = \int_0^1 (f(x)g(x) + f'(x)g'(x))dx,$$

for all $f, g \in V$. The associated norm is

$$\|f\|_V = \left(\int_0^1 (f(x)^2 + f'(x)^2)dx \right)^{1/2}.$$

Assume that u is a solution of our original boundary problem (BP), so that

$$\begin{aligned} -u''(x) + c(x)u(x) &= f(x), \quad 0 < x < 1 \\ u(0) &= 0 \\ u(1) &= 0. \end{aligned}$$

Multiply the differential equation by any arbitrary *test function* $v \in V$, obtaining

$$-u''(x)v(x) + c(x)u(x)v(x) = f(x)v(x), \qquad (*)$$

and integrate this equation! We get

$$-\int_0^1 u''(x)v(x)dx + \int_0^1 c(x)u(x)v(x)dx = \int_0^1 f(x)v(x)dx. \quad (\dagger)$$

Now, the trick is to use integration by parts on the first term.

Recall that

$$(u'v)' = u''v + u'v',$$

and to be careful about discontinuities, write

$$\int_0^1 u''(x)v(x)dx = \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} u''(x)v(x)dx.$$

Using integration by parts, we have

$$\begin{aligned}
&\int_{x_i}^{x_{i+1}} u''(x)v(x)dx \\
&\qquad = \int_{x_i}^{x_{i+1}} (u'(x)v(x))'dx - \int_{x_i}^{x_{i+1}} u'(x)v'(x)dx \\
&\qquad\quad = [u'(x)v(x)]_{x=x_i}^{x=x_{i+1}} - \int_{x_i}^{x_{i+1}} u'(x)v'(x)dx \\
&= u'(x_{i+1})v(x_{i+1}) - u'(x_i)v(x_i) - \int_{x_i}^{x_{i+1}} u'(x)v'(x)dx.
\end{aligned}$$

It follows that

$$\int_0^1 u''(x)v(x)dx$$

$$= \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} u''(x)v(x)dx =$$

$$\sum_{i=0}^{N} \left(u'(x_{i+1})v(x_{i+1}) - u'(x_i)v(x_i) - \int_{x_i}^{x_{i+1}} u'(x)v'(x)dx \right)$$

$$= u'(1)v(1) - u'(0)v(0) - \int_0^1 u'(x)v'(x)dx.$$

However, the test function v satisfies the boundary conditions $v(0) = v(1) = 0$ (recall that $v \in V$), so we get

$$\int_0^1 u''(x)v(x)dx = -\int_0^1 u'(x)v'(x)dx.$$

Consequently, the equation (†) becomes

$$\int_0^1 (u'v' + cuv)dx = \int_0^1 fvdx, \quad \text{for all } v \in V. \quad (**)$$

Thus, it is natural to introduce the bilinear form $a\colon V \times V \to \mathbb{R}$ given by

$$a(u, v) = \int_0^1 (u'v' + cuv)dx, \quad \text{for all } u, v \in V,$$

and the linear form $\widetilde{f}\colon V \to \mathbb{R}$ given by

$$\widetilde{f}(v) = \int_0^1 f(x)v(x)dx, \quad \text{for all } v \in V.$$

Then, (∗∗) becomes

$$a(u, v) = \widetilde{f}(v), \quad \text{for all } v \in V.$$

We also introduce the *energy function* J given by

$$J(v) = \frac{1}{2}a(v,v) - \widetilde{f}(v) \quad v \in V.$$

Theorem 14.1. *Let u be any solution of the boundary problem (BP).*

(1) Then we have

$$a(u,v) = \widetilde{f}(v), \quad \text{for all } v \in V, \qquad \text{(WF)}$$

where

$$a(u,v) = \int_0^1 (u'v' + cuv)dx, \quad \textit{for all } u, v \in V,$$

and

$$\widetilde{f}(v) = \int_0^1 f(x)v(x)dx, \quad \textit{for all } v \in V.$$

(2) If $c(x) \geq 0$ for all $x \in [0,1]$, then a function $u \in V$ is a solution of (WF) iff u minimizes $J(v)$, that is,

$$J(u) = \inf_{v \in V} J(v),$$

with

$$J(v) = \frac{1}{2}a(v,v) - \widetilde{f}(v) \quad v \in V.$$

Furthermore, u is unique.

Theorem 14.1 shows that every solution u of our boundary problem (BP) is a solution (in fact, unique) of the equation (WF).

The equation (WF) is called the *weak form* or *variational equation* associated with the boundary problem.

This idea to derive these equations is due to *Ritz and Galerkin*.

Now, the natural question is whether the variational equation (WF) has a solution, and whether this solution, if it exists, is also a solution of the boundary problem (it must belong to $C^2([0,1])$, which is far from obvious). Then, (BP) and (WF) would be equivalent.

Some fancy tools of analysis can be used to prove these assertions.

The first difficulty is that the vector space V is not the right space of solutions, because in order for the variational problem to have a solution, it must be complete.

So, we must construct a completion of the vector space V.

This can be done and we get the *Sobolev space* $H_0^1(0,1)$. Then, the question of the regularity of the "weak solution" can also be tackled.

We will not worry about all this. Instead, let us find *approximations* of the problem (WF).

Instead of using the infinite-dimensional vector space V, we consider *finite-dimensional* subspaces V_a (with $\dim(V_a) = n$) of V, and we consider the *discrete problem*:

Find a function $u^{(a)} \in V_a$, such that

$$a(u^{(a)}, v) = \widetilde{f}(v), \quad \text{for all } v \in V_a. \tag{DWF}$$

Since V_a is finite dimensional (of dimension n), let us pick a *basis of functions* $(w_1, \ldots, w_n)$ in V_a, so that every function $u \in V_a$ can we written as

$$u = u_1 w_1 + \cdots + u_n w_n.$$

Then, the equation (DWF) holds iff

$$a(u, w_j) = \widetilde{f}(w_j), \quad j = 1, \ldots, n,$$

and by plugging $u_1 w_1 + \cdots + u_n w_n$ for u, we get a system of k linear equations

$$\sum_{i=1}^{n} a(w_i, w_j) u_i = \widetilde{f}(w_j), \quad 1 \leq j \leq n.$$

Because $a(v,v) \geq \frac{1}{2}\|v\|_{V_a}$, the bilinear form a is symmetric positive definite, and thus the *matrix* $(a(w_i, w_j))$ *is symmetric positive definite, and thus invertible*.

Therefore, (DWF) has a solution given by a *linear system*!

From a practical point of view, we have to compute the integrals

$$a_{ij} = a(w_i, w_j) = \int_0^1 (w_i' w_j' + c w_i w_j) dx,$$

and

$$b_j = \widetilde{f}(w_j) = \int_0^1 f(x) w_j(x) dx.$$

However, if the basis functions are simple enough, this can be done "by hand." Otherwise, numerical integration methods must be used, but there are some good ones.

Let us also remark that the proof of Theorem 14.1 also shows that the unique solution of (DWF) is the unique minimizer of J over all functions in V_a.

It is also possible to compare the approximate solution $u^{(a)} \in V_a$ with the exact solution $u \in V$.

Theorem 14.2. *Suppose $c(x) \geq 0$ for all $x \in [0,1]$. For every finite-dimensional subspace V_a ($\dim(V_a) = n$) of V, for every basis $(w_1, \ldots, w_n)$ of V_a, the following properties hold:*

(1) There is a unique function $u^{(a)} \in V_a$ such that

$$a(u^{(a)}, v) = \widetilde{f}(v), \quad \text{for all } v \in V_a, \qquad \text{(DWF)}$$

and if $u^{(a)} = u_1 w_1 + \cdots + u_n w_n$, then $\mathbf{u} = (u_1, \ldots, u_n)$ is the solution of the linear system

$$A\mathbf{u} = b, \qquad (*)$$

with $A = (a_{ij}) = (a(w_i, w_j))$ and $b_j = \widetilde{f}(w_j)$, $1 \leq i, j \leq n$. Furthermore, the matrix $A = (a_{ij})$ is symmetric positive definite.

(2) The unique solution $u^{(a)} \in V_a$ of (DWF) is the unique minimizer of J over V_a, that is,

$$J(u^{(a)}) = \inf_{v \in V_a} J(v),$$

(3) There is a constant C independent of V_a and of the unique solution $u \in V$ of (WF), such that

$$\left\| u - u^{(a)} \right\|_V \leq C \inf_{v \in V_a} \|u - v\|_V .$$

Let us now give examples of the subspaces V_a used in practice. They usually consist of piecewise polynomial functions.

Pick an integer $N \geq 1$ and subdivide $[0, 1]$ into $N + 1$ intervals $[x_i, x_{i+1}]$, where

$$x_i = hi, \quad h = \frac{1}{N+1}, \quad i = 0, \ldots, N+1.$$

We will use the following fact: every polynomial $P(x)$ of degree $2m+1$ ($m \geq 0$) is completely determined by its values as well as the values of its first m derivatives at two distinct points $\alpha, \beta \in \mathbb{R}$.

There are various ways to prove this.

One way is to use the Bernstein basis, because the kth derivative of a polynomial is given by a formula in terms of its control points.

For example, for $m = 1$, every degree 3 polynomial can be written as

$$P(x) = (1-x)^3\, b_0 + 3(1-x)^2 x\, b_1 + 3(1-x)x^2\, b_2 + x^3\, b_3,$$

with $b_0, b_1, b_2, b_3 \in \mathbb{R}$, and we showed that

$$\begin{aligned} P'(0) &= 3(b_1 - b_0) \\ P'(1) &= 3(b_3 - b_2). \end{aligned}$$

Given $P(0)$ and $P(1)$, we determine b_0 and b_3, and from $P'(0)$ and $P'(1)$, we determine b_1 and b_2.

In general, for a polynomial of degree m written as

$$P(x) = \sum_{j=0}^{m} b_j B_j^m(x)$$

in terms of the Bernstein basis $(B_0^m(x), \ldots, B_m^m(x))$ with

$$B_j^m(x) = \binom{m}{j}(1-x)^{m-j}x^j,$$

it can be shown that the kth derivative of P at zero is given by

$$P^{(k)}(0) = m(m-1)\cdots(m-k+1)\left(\sum_{i=0}^{k}\binom{k}{i}(-1)^{k-i}\,b_i\right),$$

and there is a similar formula for $P^{(k)}(1)$.

Actually, we need to use the Bernstein basis of polynomials $B_k^m[r,s]$, where

$$B_j^m[r,s](x) = \binom{m}{j}\left(\frac{s-x}{s-r}\right)^{m-j}\left(\frac{x-r}{s-r}\right)^{j},$$

with $r < s$, in which case

$$\begin{aligned} &P^{(k)}(0)\\ &= \frac{m(m-1)\cdots(m-k+1)}{(s-r)^k}\left(\sum_{i=0}^{k}\binom{k}{i}(-1)^{k-i}\,b_i\right), \end{aligned}$$

with a similar formula for $P^{(k)}(1)$. In our case, we set $r = x_i, s = x_{i+1}$.

Now, if the $2m+2$ values

$$P(0), P^{(1)}(0), \ldots, P^{(m)}(0), P(1), P^{(1)}(1), \ldots, P^{(m)}(1)$$

are given, we obtain a triangular system that determines uniquely the $2m+2$ control points $b_0, \ldots, b_{2m+1}$.

Recall that $C^m([0,1])$ denotes the set of C^m functions f on $[0,1]$, which means that $f, f^{(1)}, \ldots, f^{(m)}$ exist are are continuous on $[0,1]$.

We define the vector space V_N^m as the subspace of $C^m([0,1])$ consisting of all functions f such that

1. $f(0) = f(1) = 0$.
2. The restriction of f to $[x_i, x_{i+1}]$ is a polynomial of degree $2m+1$, for $i = 0, \ldots, N$.

Observe that the functions in V_N^0 are the piecewise affine functions f with $f(0) = f(1) = 0$; an example is shown in Figure 14.2.

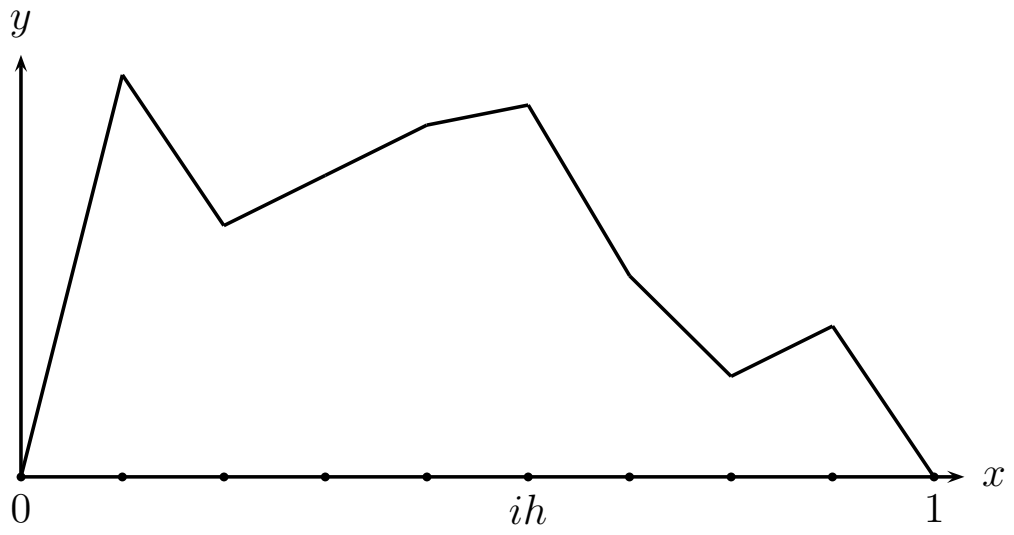

Figure 14.2: A piecewise affine function

This space has dimension N, and a basis consists of the "hat functions" w_i, where the only two nonflat parts of the graph of w_i are the line segments from $(x_{i-1}, 0)$ to $(x_i, 1)$, and from $(x_i, 1)$ to $(x_{i+1}, 0)$, for $i = 1, \ldots, N$, see Figure 14.3.

The basis functions w_i have a small support, which is good because in computing the integrals giving $a(w_i, w_j)$, we find that we get a tridiagonal matrix.

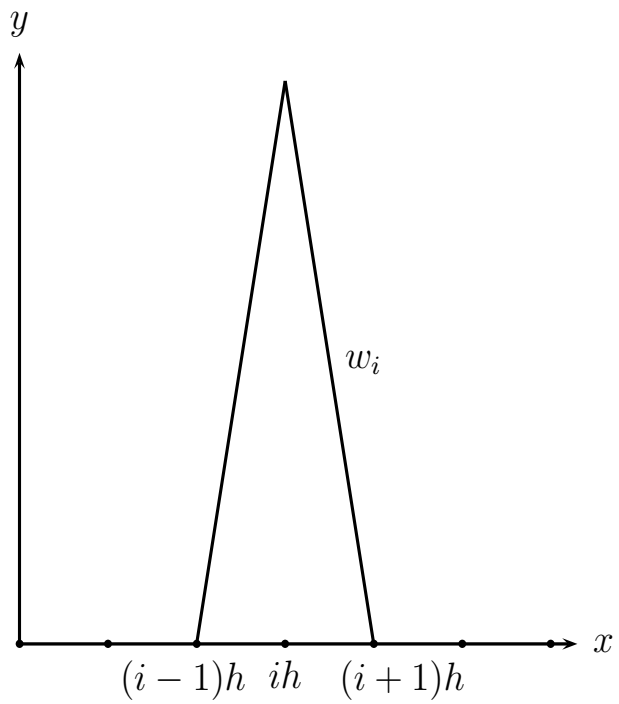

Figure 14.3: A basis "hat function"

They also have the nice property that every function $v \in V_N^0$ has the following expression on the basis (w_i):

$$v(x) = \sum_{i=1}^{N} v(ih) w_i(x), \quad x \in [0,1].$$

In general, it it not hard to see that V_N^m has dimension $mN + 2(m-1)$.

Going back to our problem (the bending of a beam), assuming that c and f are constant functions, it is not hard to show that the linear system $(*)$ becomes

$$A\mathbf{u} = b,$$

with

$$A = \frac{1}{h}\begin{pmatrix} 2+\frac{2c}{3}h^2 & -1+\frac{c}{6}h^2 & & & \\ -1+\frac{c}{6}h^2 & 2+\frac{2c}{3}h^2 & -1+\frac{c}{6}h^2 & & \\ \ddots & \ddots & \ddots & & \\ & -1+\frac{c}{6}h^2 & 2+\frac{2c}{3}h^2 & -1+\frac{c}{6}h^2 & \\ & & & -1+\frac{c}{6}h^2 & 2+\frac{2c}{3}h^2 \end{pmatrix}$$

and

$$b = h\begin{pmatrix} f \\ f \\ \vdots \\ f \\ f \end{pmatrix}.$$

We can also find a basis of $2N+2$ cubic functions for V_N^1 consisting of functions with small support.

This basis consists of the N functions w_i^0 and of the $N+2$ functions w_i^1 uniquely determined by the following conditions:

$$\begin{aligned} w_i^0(x_j) &= \delta_{ij}, \quad 1 \le j \le N,\ 1 \le i \le N \\ (w_i^0)'(x_j) &= 0, \quad 0 \le j \le N+1,\ 1 \le i \le N \\ w_i^1(x_j) &= 0, \quad 1 \le j \le N,\ 0 \le i \le N+1 \\ (w_i^1)'(x_j) &= \delta_{ij}, \quad 0 \le j \le N+1,\ 0 \le i \le N+1 \end{aligned}$$

with $\delta_{ij} = 1$ iff $i = j$ and $\delta_{ij} = 0$ if $i \neq j$.

Some of these functions are displayed in Figure 14.4.

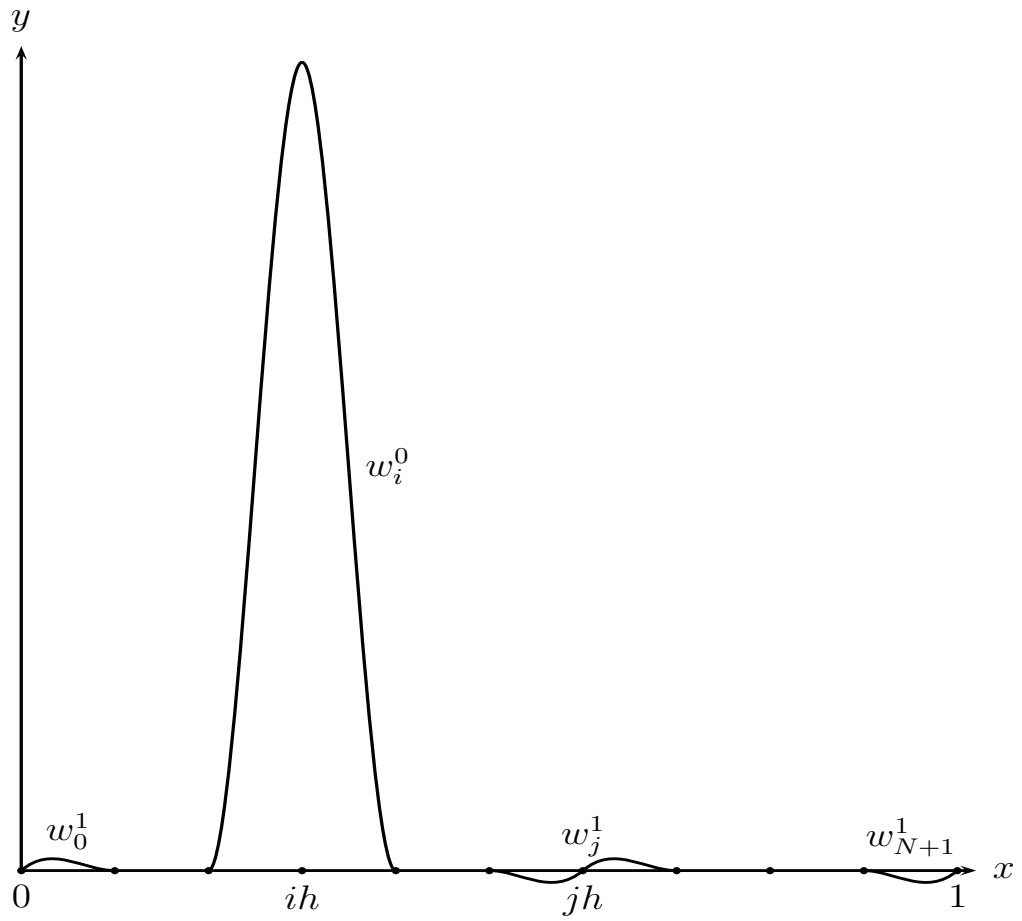

Figure 14.4: The basis functions w_i^0 and w_j^1

For every function $v \in V_N^1$, we have

$$v(x) = \sum_{i=1}^{N} v(ih) w_i^0(x) + \sum_{j=0}^{N+1} v'(jh) w_j^1(x), \quad x \in [0,1].$$

If we order these basis functions as

$$w_0^1, w_1^0, w_1^1, w_2^0, w_2^1, \ldots, w_N^0, w_N^1, w_{N+1}^1,$$

we find that if $c = 0$, the matrix A of the system $(*)$ is tridiagonal by blocks, where the blocks are 2×2, 2×1, or 1×2 matrices, and with single entries in the top left and bottom right corner.

A different order of the basis vectors would mess up the tridiagonal block structure of A. We leave the details as an exercise.

14.2 A Two-Dimensional Problem: An Elastic Membrane

Consider an elastic membrane attached to a round contour whose projection on the (x_1, x_2)-plane is the boundary Γ of an open, connected, bounded region Ω in the (x_1, x_2)-plane, as illustrated in Figure 14.5.

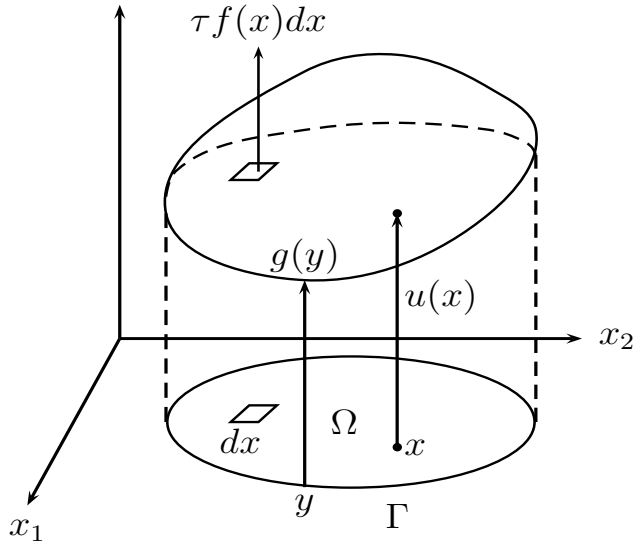

Figure 14.5: An elastic membrane

In other words, we view the membrane as a surface consisting of the set of points (x, z) given by an equation of the form

$$z = u(x),$$

with $x = (x_1, x_2) \in \overline{\Omega}$, where $u\colon \overline{\Omega} \to \mathbb{R}$ is some sufficiently regular function, and we think of $u(x)$ as the vertical displacement of this membrane.

We assume that this membrane is under the action of a vertical force $\tau f(x)dx$ per surface element in the horizontal plane (where τ is the tension of the membrane).

The problem is to find the vertical displacement u as a function of x, for $x \in \overline{\Omega}$.

It can be shown (under some assumptions on Ω, Γ, and f), that $u(x)$ is given by a PDE with boundary condition, of the form

$$\begin{aligned} -\Delta u(x) &= f(x), \quad x \in \Omega \\ u(x) &= g(x), \quad x \in \Gamma, \end{aligned}$$

where $g\colon \Gamma \to \mathbb{R}$ represents the height of the contour of the membrane.

We are looking for a function u in $C^2(\Omega) \cap C^1(\overline{\Omega})$.

The operator Δ is the *Laplacian*, and it is given by

$$\Delta u(x) = \frac{\partial^2 u}{\partial x_1^2}(x) + \frac{\partial^2 u}{\partial x_2^2}(x).$$

This is an example of a *boundary problem*, since the solution u of the PDE must satisfy the condition $u(x) = g(x)$ on the boundary of the domain Ω.

The above equation is known as *Poisson's equation*, and when $f = 0$ as *Laplace's equation*.

It can be proved that if the data f, g and Γ are sufficiently smooth, then the problem has a unique solution.

To get a weak formulation of the problem, first we have to make the boundary condition homogeneous, which means that $g(x) = 0$ on Γ.

It turns out that g can be extended to the whole of $\overline{\Omega}$ as some sufficiently smooth function $\widehat{h}$, so we can look for a solution of the form $u - \widehat{h}$, but for simplicity, let us assume that the contour of Ω lies in a plane parallel to the (x_1, x_2)- plane, so that $g = 0$.

We let V be the subspace of $C^2(\Omega) \cap C^1(\overline{\Omega})$ consisting of functions v such that $v = 0$ on Γ.

As before, we multiply the PDE by a test function $v \in V$, getting

$$-\Delta u(x)v(x) = f(x)v(x),$$

and we "integrate by parts."

In this case, this means that we use a version of Stokes formula known as *Green's first identity*, which says that

$$\int_\Omega -\Delta u\, v\, dx = \int_\Omega (\operatorname{grad} u)\cdot(\operatorname{grad} v)\, dx - \int_\Gamma (\operatorname{grad} u)\cdot n\, v d\sigma$$

(where n denotes the outward pointing unit normal to the surface).

Because $v = 0$ on Γ, the integral $\int_\Gamma$ drops out, and we get an equation of the form

$$a(u, v) = \widetilde{f}(v) \quad \text{for all } v \in V,$$

where a is the bilinear form given by

$$a(u, v) = \int_\Omega \left(\frac{\partial u}{\partial x_1} \frac{\partial v}{\partial x_1} + \frac{\partial u}{\partial x_2} \frac{\partial v}{\partial x_2} \right) dx$$

and $\widetilde{f}$ is the linear form given by

$$\widetilde{f}(v) = \int_\Omega f v dx.$$

We get the same equation as in section 14.2, but over a set of functions defined on a two-dimensional domain.

As before, we can choose a finite-dimensional subspace V_a of V and consider the discrete problem with respect to V_a.

Again, if we pick a basis $(w_1, \ldots, w_n)$ of V_a, a vector $u = u_1 w_1 + \cdots + u_n w_n$ is a solution of the Weak Formulation of our problem iff $\mathbf{u} = (u_1, \ldots, u_n)$ is a solution of the linear system

$$A\mathbf{u} = b,$$

with $A = (a(w_i, w_j))$ and $b = (\widetilde{f}(w_j))$.

However, the integrals that give the entries in A and b are much more complicated.

An approach to deal with this problem is the *method of finite elements*.

The idea is to also discretize the boundary curve Γ.

If we assume that Γ is a *polygonal line*, then we can *triangulate* the domain Ω, and then we consider spaces of functions which are piecewise defined on the triangles of the triangulation of Ω.

The simplest functions are piecewise affine and look like tents erected above groups of triangles.

Again, we can define base functions with small support, so that the matrix A is tridiagonal by blocks.

The finite element method is a vast subject and it is presented in many books of various degrees of difficulty and obscurity.

Let us simply state three important requirements of the finite element method:

1. "Good" triangulations must be found. This in itself is a vast research topic. Delaunay triangulations are good candidates.
2. "Good" spaces of functions must be found; typically piecewise polynomials and splines.
3. "Good" bases consisting of functions will small support must be found, so that integrals can be easily computed and sparse banded matrices arise.

We now consider boundary problems where the solution varies with time.

14.3 Time-Dependent Boundary Problems: The Wave Equation

Consider a homogeneous string (or rope) of constant cross-section, of length L, and stretched (in a vertical plane) between its two ends which are assumed to be fixed and located along the x-axis at $x = 0$ and at $x = L$.

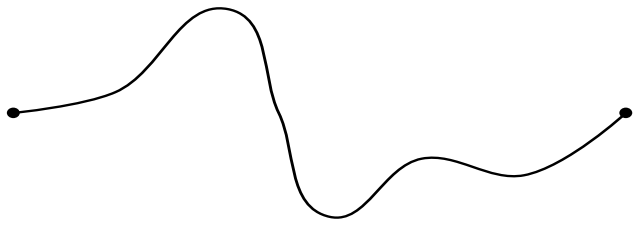

Figure 14.6: A vibrating string

The string is subjected to a transverse force $\tau f(x)dx$ per element of length dx (where τ is the tension of the string).

We would like to investigate the small displacements of the string in the vertical plane, that is, how it vibrates.

Thus, we seek a function $u(x, t)$ defined for $t \geq 0$ and $x \in [0, L]$, such that $u(x, t)$ represents the vertical deformation of the string at the abscissa x and at time t.

It can be shown that u must satisfy the following PDE

$$\frac{1}{c^2}\frac{\partial^2 u}{\partial t^2}(x,t) - \frac{\partial^2 u}{\partial x^2}(x,t) = f(x,t), \quad 0 < x < L,\ t > 0,$$

with $c = \sqrt{\tau/\rho}$, where ρ is the linear density of the string, known as the *one-dimensional wave equation*.

Furthermore, the initial shape of the string is known at $t = 0$, as well as the distribution of the initial velocities along the string;

in other words, there are two functions $u_{i,0}$ and $u_{i,1}$ such that

$$\begin{aligned} u(x,0) &= u_{i,0}(x), \quad 0 \le x \le L, \\ \frac{\partial u}{\partial t}(x,0) &= u_{i,1}(x), \quad 0 \le x \le L. \end{aligned}$$

For example, if the string is simply released from its given starting position, we have $u_{i,1} = 0$.

Lastly, because the ends of the string are fixed, we must have

$$u(0,t) = u(L,t) = 0, \quad t \geq 0.$$

Consequently, we look for a function $u\colon \mathbb{R}_+ \times [0,L] \to \mathbb{R}$ satisfying the following conditions:

$$\begin{aligned}
&\frac{1}{c^2}\frac{\partial^2 u}{\partial t^2}(x,t) - \frac{\partial^2 u}{\partial x^2}(x,t) = f(x,t), \quad 0 < x < L,\ t > 0,\\
&u(0,t) = u(L,t) = 0, \quad t \geq 0 \quad \text{(boundary condition)},\\
&u(x,0) = u_{i,0}(x), \quad 0 \leq x \leq L \quad \text{(intitial condition)},\\
&\frac{\partial u}{\partial t}(x,0) = u_{i,1}(x), \quad 0 \leq x \leq L \quad \text{(intitial condition)}.
\end{aligned}$$

This is an example of a *time-dependent boundary-value problem*, with two *initial conditions*.

To simplify the problem, assume that $f = 0$, which amounts to neglecting the effect of gravity.

In this case, our PDE becomes

$$\frac{1}{c^2}\frac{\partial^2 u}{\partial t^2}(x,t) - \frac{\partial^2 u}{\partial x^2}(x,t) = 0, \quad 0 < x < L,\ t > 0,$$

Let us try our trick of multiplying by a test function v depending only on x, C^1 on $[0, L]$, and such that $v(0) = v(L) = 0$, and integrate by parts.

We get the equation

$$\int_0^L \frac{\partial^2 u}{\partial t^2}(x,t)v(x)dx - c^2 \int_0^L \frac{\partial^2 u}{\partial x^2}(x,t)v(x)dx = 0.$$

For the first term, we get

$$\begin{aligned}\int_0^L \frac{\partial^2 u}{\partial t^2}(x,t)v(x)dx &= \int_0^L \frac{\partial^2}{\partial t^2}[u(x,t)v(x)]dx \\ &= \frac{d^2}{dt^2}\int_0^L u(x,t)v(x)dx \\ &= \frac{d^2}{dt^2}\langle u, v\rangle,\end{aligned}$$

where $\langle u, v\rangle$ is the inner product in $L^2([0, L])$.

The fact that it is legitimate to move $\partial^2/\partial t^2$ outside of the integral needs to be justified rigorously, but we won't do it here.

For the second term, we get

$$-\int_0^L \frac{\partial^2 u}{\partial x^2}(x,t)v(x)dx = -\left[\frac{\partial u}{\partial x}(x,t)v(x)\right]_{x=0}^{x=L} + \int_0^L \frac{\partial u}{\partial x}(x,t)\frac{dv}{dx}(x)dx,$$

and because $v \in V$, we have $v(0) = v(L) = 0$, so we obtain

$$-\int_0^L \frac{\partial^2 u}{\partial x^2}(x,t)v(x)dx = \int_0^L \frac{\partial u}{\partial x}(x,t)\frac{dv}{dx}(x)dx.$$

Our integrated equation becomes

$$\frac{d^2}{dt^2}\langle u, v\rangle + c^2\int_0^L \frac{\partial u}{\partial x}(x,t)\frac{dv}{dx}(x)dx = 0, \quad \text{for all } v \in V \quad \text{and all } t \geq 0.$$

It is natural to introduce the bilinear form $a\colon V \times V \to \mathbb{R}$ given by

$$a(u,v) = \int_0^L \frac{\partial u}{\partial x}(x,t)\frac{\partial v}{\partial x}(x,t)dx,$$

where, for every $t \in \mathbb{R}_+$, the functions $u(x,t)$ and (v,t) belong to V.

Actually, we have to replace V by the subspace of the Sobolev space $H_0^1(0,L)$ consisting of the functions such that $v(0) = v(L) = 0$.

Then, the weak formulation (variational formulation) of our problem is this:

Find a function $u \in V$ such that

$$\begin{aligned}
&\frac{d^2}{dt^2}\langle u, v\rangle + a(u,v) = 0, \quad \text{for all } v \in V \text{ and all } t \geq 0\\
&u(x,0) = u_{i,0}(x), \quad 0 \leq x \leq L \quad \text{(intitial condition)},\\
&\frac{\partial u}{\partial t}(x,0) = u_{i,1}(x), \quad 0 \leq x \leq L \quad \text{(intitial condition)}.
\end{aligned}$$

It can be shown that there is a positive constant $\alpha > 0$ such that

$$a(u,u) \geq \alpha \|u\|^2_{H^1_0} \quad \text{for all } v \in V$$

(Poincaré's inequality), which shows that a is *positive definite* on V.

The above method is known as the method of *Rayleigh-Ritz*.

A study of the above equation requires some sophisticated tools of analysis which go far beyond the scope of these notes.

Let us just say that there is a countable sequence of solutions with separated variables of the form

$$\begin{aligned} u_k^{(1)} &= \sin\left(\frac{k\pi x}{L}\right)\cos\left(\frac{k\pi ct}{L}\right), \\ u_k^{(2)} &= \sin\left(\frac{k\pi x}{L}\right)\sin\left(\frac{k\pi ct}{L}\right), \quad k \in \mathbb{N}_+, \end{aligned}$$

called *modes* (or *normal modes*).

Complete solutions of the problem are series obtained by combining the normal modes, and they are of the form

$$u(x,t) = \sum_{k=1}^{\infty} \sin\left(\frac{k\pi x}{L}\right)\left(A_k \cos\left(\frac{k\pi ct}{L}\right) + B_k \sin\left(\frac{k\pi ct}{L}\right)\right),$$

where the coefficients A_k, B_k are determined from the Fourier series of $u_{i,0}$ and $u_{i,1}$.

We now consider discrete approximations of our problem.

As before, consider a finite dimensional subspace V_a of V and assume that we have approximations $u_{a,0}$ and $u_{a,1}$ of $u_{i,0}$ and $u_{i,1}$.

If we pick a basis $(w_1, \ldots, w_n)$ of V_a, then we can write our unknown function $u(x,t)$ as

$$u(x,t) = u_1(t)w_1 + \cdots + u_n(t)w_n,$$

where $u_1, \ldots, u_n$ are functions of t.

Then, if we write $\mathbf{u} = (u_1, \ldots, u_n)$, the discrete version of our problem is

$$\begin{aligned} A\frac{d^2\mathbf{u}}{dt^2} + K\mathbf{u} &= 0, \\ u(x,0) &= u_{a,0}(x), \quad 0 \leq x \leq L, \\ \frac{\partial u}{\partial t}(x,0) &= u_{a,1}(x), \quad 0 \leq x \leq L, \end{aligned}$$

where $A = (\langle w_i, w_j \rangle)$ and $K = (a(w_i, w_j))$ are two symmetric matrices, called the *mass matrix* and the *stiffness matrix*, respectively.

In fact, because a and the inner product $\langle -, - \rangle$ are positive definite, these matrices are also positive definite.

We have made some progress since we now have a system of ODE's, and we can solve it by analogy with the scalar case.

So, we look for solutions of the form $\mathbf{U}\cos\omega t$ (or $\mathbf{U}\sin\omega t$), where $\mathbf{U}$ is an n-dimensional vector.

We find that we should have

$$(K - \omega^2 A)\mathbf{U}\cos\omega t = 0,$$

which implies that ω must be a solution of the equation

$$K\mathbf{U} = \omega^2 A\mathbf{U}.$$

Thus, we have to find some λ such that

$$K\mathbf{U} = \lambda A\mathbf{U},$$

a problem known as a *generalized eigenvalue problem*, since the ordinary eigenvalue problem for K is

$$K\mathbf{U} = \lambda\mathbf{U}.$$

Fortunately, because A is SPD, we can reduce this generalized eigenvalue problem to a standard eigenvalue problem.

A good way to do so is to use a *Cholesky decomposition* of A as

$$A = LL^\top,$$

where L is a lower triangular matrix (see Theorem 5.10).

Because A is SPD, it is invertible, so L is also invertible, and

$$K\mathbf{U} = \lambda A\mathbf{U} = \lambda LL^\top\mathbf{U}$$

yields

$$L^{-1}K\mathbf{U} = \lambda L^\top\mathbf{U},$$

which can also be written as

$$L^{-1}K(L^\top)^{-1}L^\top\mathbf{U} = \lambda L^\top\mathbf{U}.$$

Then, if we make the change of variable

$$\mathbf{Y} = L^\top \mathbf{U},$$

using the fact $(L^\top)^{-1} = (L^{-1})^\top$, the equation

$$L^{-1}K(L^\top)^{-1}L^\top \mathbf{U} = \lambda L^\top \mathbf{U}.$$

is equivalent to

$$L^{-1}K(L^{-1})^\top \mathbf{Y} = \lambda \mathbf{Y},$$

a standard eigenvalue problem for the matrix $\widehat{K} = L^{-1}K(L^{-1})^\top$.

Furthermore, we know from Section 5.7 that since K is SPD and L^{-1} is invertible, the matrix $\widehat{K} = L^{-1}K(L^{-1})^\top$ is also SPD.

Consequently, $\widehat{K}$ has positive real eigenvalues $(\omega_1^2, \ldots, \omega_n^2)$ (not necessarily distinct) and it can be diagonalized with respect to an orthonormal basis of eigenvectors, say $\mathbf{Y}^1, \ldots, \mathbf{Y}^n$.

Then, since $\mathbf{Y} = L^\top \mathbf{U}$, the vectors

$$\mathbf{U}^i = (L^\top)^{-1}\mathbf{Y}^i, \quad i = 1, \ldots, n,$$

are linearly independent and are solutions of the generalized eigenvalue problem; that is,

$$K\mathbf{U}^i = \omega_i^2 A\mathbf{U}^i, \quad i = 1, \ldots, n.$$

More is true. Because the vectors $\mathbf{Y}^1, \ldots, \mathbf{Y}^n$ are orthonormal, and because $\mathbf{Y}^i = L^\top \mathbf{U}^i$, from

$$(\mathbf{Y}^i)^\top \mathbf{Y}^j = \delta_{ij},$$

we get

$$(\mathbf{U}^i)^\top A\mathbf{U}^j = \delta_{ij}, \quad 1 \leq i, j \leq n.$$

This suggests defining the functions $U^i \in V_a$ by

$$U^i = \sum_{k=1}^{n} \mathbf{U}^i_k w_k.$$

Then, it immediate to check that

$$a(U^i, U^j) = (\mathbf{U}^i)^\top A \mathbf{U}^j = \delta_{ij},$$

which means that the *functions $(U^1, \ldots, U^n)$ form an orthonormal basis of V_a for the inner product a*.

The functions $U^i \in V_a$ are called *modes* (or *modal vectors*).

As a final step, let us look again for a solution of our discrete weak formulation of the problem, this time expressing the unknown solution $u(x,t)$ over the modal basis $(U^1,\dots,U^n)$, say

$$u=\sum_{j=1}^{n}\widetilde{u}_j(t)U^j,$$

where each $\widetilde{u}_j$ is a function of t.

If we write $\mathbf{u}=(u_1,\dots,u_n)$ with $u_k=\sum_{j=1}^{n}\widetilde{u}_j(t)\mathbf{U}_k^j$ for $k=1,\dots,n$, we see that

$$\mathbf{u}=\sum_{j=1}^{n}\widetilde{u}_j\mathbf{U}^j,$$

so using the fact that

$$K\mathbf{U}^j=\omega_j^2A\mathbf{U}^j,\quad j=1,\dots,n,$$

the equation

$$A\frac{d^2\mathbf{u}}{dt^2} + K\mathbf{u} = 0$$

yields

$$\sum_{j=1}^{n}[(\widetilde{u}_j)'' + \omega_j^2\widetilde{u}_j]A\mathbf{U}^j = 0.$$

Since A is invertible and since $(\mathbf{U}^1, \ldots, \mathbf{U}^n)$ are linearly independent, the vectors $(A\mathbf{U}^1, \ldots, A\mathbf{U}^n)$ are linearly independent, and consequently we get the system of n ODEs'

$$(\widetilde{u}_j)'' + \omega_j^2\widetilde{u}_j = 0, \quad 1 \leq j \leq n.$$

Each of these equation has a well-known solution of the form

$$\widetilde{u}_j = A_j \cos \omega_j t + B_j \sin \omega_j t.$$

Therefore, the solution of our approximation problem is given by

$$u = \sum_{j=1}^{n} (A_j \cos \omega_j t + B_j \sin \omega_j t) U^j,$$

and the constants A_j, B_j are obtained from the intial conditions

$$\begin{aligned} u(x,0) &= u_{a,0}(x), \quad 0 \leq x \leq L, \\ \frac{\partial u}{\partial t}(x,0) &= u_{a,1}(x), \quad 0 \leq x \leq L, \end{aligned}$$

by expressing $u_{a,0}$ and $u_{a,1}$ on the modal basis $(U^1, \ldots, U^n)$. Furthermore, the modal functions $(U^1, \ldots, U^n)$ form an orthonormal basis of V_a for the inner product a.

If we use the vector space V_N^0 of piecewise affine functions, we find that the matrices A and K are familar!

Indeed,

$$A = \frac{1}{h}\begin{pmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{pmatrix}$$

and

$$K = \frac{h}{6}\begin{pmatrix} 4 & 1 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 1 & 4 \end{pmatrix}.$$

To conclude this section, let us discuss briefly the wave equation for an elastic membrane, as described in Section 14.2.

This time, we look for a function $u\colon \mathbb{R}_+ \times \Omega \to \mathbb{R}$ satisfying the following conditions:

$$\begin{aligned}
&\frac{1}{c^2}\frac{\partial^2 u}{\partial t^2}(x,t) - \Delta u(x,t) = f(x,t), \quad x \in \Omega,\ t > 0,\\
&u(x,t) = 0, \quad x \in \Gamma, \quad t \geq 0 \quad \text{(boundary condition)},\\
&u(x,0) = u_{i,0}(x), \quad x \in \Omega \quad \text{(intitial condition)},\\
&\frac{\partial u}{\partial t}(x,0) = u_{i,1}(x), \quad x \in \Omega \quad \text{(intitial condition)}.
\end{aligned}$$

Assuming that $f = 0$, we look for solutions in the subspace V of the Sobolev space $H_0^1(\overline{\Omega})$ consisting of functions v such that $v = 0$ on Γ.

Multiplying by a test function $v \in V$ and using Green's first identity, we get the weak formulation of our problem:

Find a function $u \in V$ such that

$$\begin{aligned}
\frac{d^2}{dt^2}\langle u, v\rangle + a(u,v) &= 0, \qquad \text{for all } v \in V \quad \text{and all } t \geq 0 \\
u(x,0) &= u_{i,0}(x), \quad x \in \Omega \quad \text{(intitial condition)}, \\
\frac{\partial u}{\partial t}(x,0) &= u_{i,1}(x), \quad x \in \Omega \quad \text{(intitial condition)},
\end{aligned}$$

where $a\colon V \times V \to \mathbb{R}$ is the bilinear form given by

$$a(u,v) = \int_\Omega \left(\frac{\partial u}{\partial x_1}\frac{\partial v}{\partial x_1} + \frac{\partial u}{\partial x_2}\frac{\partial v}{\partial x_2} \right) dx,$$

and

$$\langle u, v\rangle = \int_\Omega uv dx.$$

As usual, we find approximations of our problem by using finite dimensional subspaces V_a of V.

Picking some basis $(w_1, \ldots, w_n)$ of V_a, and triangulating Ω, as before, we obtain the equation

$$\begin{aligned} A\frac{d^2\mathbf{u}}{dt^2} + K\mathbf{u} &= 0, \\ u(x,0) &= u_{a,0}(x), \quad x \in \Gamma, \\ \frac{\partial u}{\partial t}(x,0) &= u_{a,1}(x), \quad x \in \Gamma, \end{aligned}$$

where $A = (\langle w_i, w_j \rangle)$ and $K = (a(w_i, w_j))$ are two *symmetric positive definite matrices*.

In principle, the problem is solved, but, it may be difficult to find good spaces V_a, good triangulations of Ω, and good bases of V_a, to be able to compute the matrices A and K, and to ensure that they are sparse.