Jim Lambers
MAT 461/561
Spring Semester 2009-10
Lecture 27 Notes

These notes correspond to Section 11.5 in the text.

The Rayleigh-Ritz Method

The finite-difference method for boundary value problems, unlike the Shooting Method, is more flexibile in that it can be generalized to boundary value problems in higher space dimensions. However, even then, it is best suited for problems in which the domain is relatively simple, such as a rectangular domain. We now consider an alternative approach that, in higher dimensions, is more readily applied to problems on domains with complicated geometries. This method is known as the *Rayleigh-Ritz Method*.

We consider the linear boundary value problem

$$-\frac{d}{dx}\left(p(x)\frac{dy}{dx}\right)+q(x)y=f(x),\quad 0<x<1,$$

with boundary conditions

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

We assume that $p(x)\geq\delta>0$ for some constant δ, and that $q(x)\geq 0$, on $[0,1]$.

If we multiply both sides of this equation by a *test function* $u(x)$, and then integrate over the domain $[0,1]$, we obtain

$$-\int_0^1 u(x)\frac{d}{dx}\left(p(x)\frac{dy}{dx}(x)\right)+u(x)q(x)y(x)\,dx=\int_0^1 u(x)f(x)\,dx.$$

Applying integration by parts, we obtain

$$\int_0^1 u(x)\frac{d}{dx}\left(p(x)\frac{dy}{dx}\right)\,dx=u(x)p(x)y'(x)\big|_0^1-\int_0^1 p(x)\frac{du}{dx}(x)\frac{dy}{dx}(x)\,dx.$$

Let $C^2[0,1]$ be the space of all functions with two continuous derivatives on $[0,1]$, and let $C_0^2[0,1]$ be the space of all functions in $C^2[0,1]$ that are equal to zero at the endpoints $x=0$ and $x=1$. If we require that our test function $u(x)$ belongs to $C_0^2[0,1]$, then $u(0)=u(1)=0$, and the boundary term in the above application of integration by parts vanishes. We then have

$$\int_0^1 p(x)\frac{du}{dx}(x)\frac{dy}{dx}(x)+q(x)u(x)y(x)\,dx=\int_0^1 u(x)f(x)\,dx.$$

This is called the *weak form* of the boundary value problem, because it only requires that the *first* derivative of $y(x)$ exist, as opposed to the original boundary value problem, that requires the existence of the *second* derivative.

It can be shown that both of these problems have the same solution $y \in C_0^2[0,1]$. Furthermore, y is a solution to the boundary value problem (in either form) if and only if it is the unique function in $C_0^2[0,1]$ that minimizes the functional

$$I[u] = \frac{1}{2}\int_0^1 p(x)[u'(x)]^2 + q(x)[u(x)]^2\,dx - \int_0^1 u(x)f(x)\,dx.$$

The problem of minimizing $I[u]$ is called the *variational form* of the boundary value problem.

There is a close connection between this functional, and the functional on $\mathbb{R}^n$,

$$\phi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{x}^T\mathbf{b}.$$

The function $u(x)$ corresponds to the vector $\mathbf{x}$, the right-hand side function $f(x)$ corresponds to the right-hand side vector $\mathbf{b}$, and the inner product $\mathbf{u}^T\mathbf{v}$ corresponds to the inner product

$$\langle u, v\rangle = \int_0^1 u(x)v(x)\,dx.$$

The matrix A corresponds to the operator

$$Lu = -\frac{d}{dx}\left(p(x)\frac{du}{dx}\right) + q(x)u.$$

Like A, a symmetric positive definite matrix, this operator is *self-adjoint* and positive definite, meaning that

$$\langle u, Lv\rangle = \langle v, Lu\rangle, \quad u, v \in C_0^2[0,1],$$

and, for all nonzero $u \in C_0^2[0,1]$,

$$\langle u, Lu\rangle = -\int_0^1 u(x)\frac{d}{dx}\left(p(x)\frac{du}{dx}(x)\right) + q(x)[u(x)]^2\,dx = \int_0^1 p(x)[u'(x)]^2 + q(x)[u(x)]^2\,dx > 0.$$

To find an approximation of the minimizer of $I[u]$, we restrict ourselves to a subspace of $C_0^2[0,1]$ by requiring that

$$y(x) = \sum_{j=1}^n c_j\phi_j(x),$$

where $\phi_1, \phi_2, \ldots, \phi_n$ form a *basis of trial functions*. For now, we only assume that these trial functions belong to $C_0^2[0,1]$, and are linearly independent. Substituting this form into $I[y]$ yields

$$I[y] = \frac{1}{2}\int_0^1 p(x)\left[\sum_{j=1}^n c_j\phi_j'(x)\right]^2 + q(x)\left[\sum_{j=1}^n c_j\phi_j(x)\right]^2 dx - \int_0^1 f(x)\sum_{j=1}^n c_j\phi_j(x)\,dx.$$

Computing the gradient of $I[y]$ with respect to $c_1, c_2, \ldots, c_j$ and requiring it to vanish yields the system of equations

$$\sum_{j=1}^{n} \left[\int_0^1 p(x)\phi_i'(x)\phi_j'(x) + q(x)\phi_i(x)\phi_j(x)\,dx \right] c_j = \int_0^1 \phi_i(x)f(x)\,dx, \quad i = 1, 2, \ldots, n.$$

This system can be written in matrix-vector form

$$A\mathbf{c} = \mathbf{b}$$

where $\mathbf{c}$ is a vector of the unknown coefficients $c_1, c_2, \ldots, c_n$ and

$$a_{ij} = \int_0^1 p(x)\phi_i'(x)\phi_j'(x) + q(x)\phi_i(x)\phi_j(x)\,dx, \quad i, j = 1, 2, \ldots, n,$$

$$b_i = \int_0^1 \phi_i(x)f(x)\,dx, \quad i = 1, 2, \ldots, n.$$

It is worth noting that this is the same system of equations that is obtained if we substitute our representation of $y(x)$ into the weak form

$$\int_0^1 p(x)\frac{du}{dx}(x)\frac{dy}{dx}(x) + q(x)u(x)y(x)\,dx = \int_0^1 u(x)f(x)\,dx,$$

which we require to hold on the subspace of test functions that is actually equal to the space spanned by the trial functions. This is the approach that serves as the basis for the *Galerkin method*, which is equivalent to the Rayleigh-Ritz method for this particularly boundary value problem, but this equivalence does not hold for more general problems.

It is also worth nothing that substituting this representation into the original, differential form of the boundary value problem, and requiring the resulting system of equations to hold at various points x_i in $[0, 1]$, is the basic idea behind the *collocation method*. There is a strong connection between Galerkin and collocation methods, and, in some cases, they are equivalent. The common thread between all three approaches–Rayleigh-Ritz, Galerkin, and collocation–is that the solution is approximated by a linear combination of trial functions, and the coefficients are obtained by solving a system of equations.

Returning to the Rayleigh-Ritz method, we must choose trial functions $\phi_1, \phi_2, \ldots, \phi_n$. A simple choice is a set of piecewise linear "hat" functions. We divide the interval $[0, 1]$ into $n+1$ subintervals $[x_{i-1}, x_i]$, with spacing $h_i = x_{i+1} - x_i$, for $i = 1, 2, \ldots, n+1$, where $x_0 = 0$ and $x_{n+1} = 1$. Then we define

$$\phi_i(x) = \left\{ \begin{array}{ll} 0 & 0 \leq x \leq x_{i-1} \\ \frac{1}{h_{i-1}}(x - x_{i-1}) & x_{i-1} < x \leq x_i \\ \frac{1}{h_i}(x_{i+1} - x) & x_i < x \leq x_{i+1} \\ 0 & x_{i+1} < x \leq 1 \end{array} \right., \quad i = 1, 2, \ldots, n.$$

These functions automatically satisfy the boundary conditions. Because they are only piecewise linear, their derivatives are discontinuous. They are

$$\phi_i'(x) = \begin{cases} 0 & 0 \le x \le x_{i-1} \\ \frac{1}{h_{i-1}} & x_{i-1} < x \le x_i \\ -\frac{1}{h_i} & x_i < x \le x_{i+1} \\ 0 & x_{i+1} < x \le 1 \end{cases}, \quad i = 1, 2, \ldots, n.$$

It follows from these definitions that $\phi_i(x)$ and $\phi_j(x)$ cannot simultaneously be nonzero at any point in $[0,1]$ unless $|i-j| \le 1$. This yields a symmetric tridiagonal matrix A with entries

$$\begin{aligned} a_{ii} &= \left(\frac{1}{h_{i-1}}\right)^2 \int_{x_{i-1}}^{x_i} p(x)\,dx + \left(-\frac{1}{h_i}\right)^2 \int_{x_i}^{x_{i+1}} p(x)\,dx + \\ &\quad \left(\frac{1}{h_{i-1}}\right)^2 \int_{x_{i-1}}^{x_i} (x - x_{i-1})^2 q(x)\,dx + \left(-\frac{1}{h_i}\right)^2 \int_{x_i}^{x_{i+1}} (x_{i+1} - x)^2 q(x)\,dx, \quad i = 1, 2, \ldots, n, \\ a_{i,i+1} &= -\frac{1}{h_i^2} \int_{x_i}^{x_{i+1}} p(x)\,dx + \frac{1}{h_i^2} \int_{x_i}^{x_{i+1}} (x_{i+1} - x)(x - x_i) q(x)\,dx, \quad i = 1, 2, \ldots, n-1, \\ a_{i+1,i} &= a_{i,i+1}, \quad i = 1, 2, \ldots, n-1. \end{aligned}$$

We also have

$$b_i = \frac{1}{h_{i-1}} \int_{x_{i-1}}^{x_i} (x - x_{i-1}) f(x)\,dx + \frac{1}{h_i} \int_{x_i}^{x_{i+1}} (x_{i+1} - x_i) f(x)\,dx, \quad i = 1, 2, \ldots, n.$$

Treating each distinct type of integral individually, we can express these elements as

$$\begin{aligned} a_{ii} &= Q_{4,i} + Q_{4,i+1} + Q_{2,i} + Q_{3,i}, \quad i = 1, 2, \ldots, n, \\ a_{i,i+1} &= -Q_{4,i+1} + Q_{1,i}, \quad i = 1, 2, \ldots, n-1, \\ b_i &= Q_{5,i} + Q_{6,i}, \quad i = 1, 2, \ldots, n, \end{aligned}$$

where

$$\begin{aligned} Q_{1,i} &= \frac{1}{h_i^2} \int_{x_i}^{x_{i+1}} (x_{i+1} - x)(x - x_i) q(x)\,dx, \quad i = 1, 2, \ldots, n-1, \\ Q_{2,i} &= \left(\frac{1}{h_{i-1}}\right)^2 \int_{x_{i-1}}^{x_i} (x - x_{i-1})^2 q(x)\,dx, \quad i = 1, 2, \ldots, n, \\ Q_{3,i} &= \left(-\frac{1}{h_i}\right)^2 \int_{x_i}^{x_{i+1}} (x_{i+1} - x)^2 q(x)\,dx, \quad i = 1, 2, \ldots, n, \\ Q_{4,i} &= \left(\frac{1}{h_{i-1}}\right)^2 \int_{x_{i-1}}^{x_i} p(x)\,dx, \quad i = 1, 2, \ldots, n, \end{aligned}$$

$$
\begin{aligned}
Q_{5,i} &= \frac{1}{h_{i-1}} \int_{x_{i-1}}^{x_i} (x - x_{i-1}) f(x)\, dx, \quad i = 1, 2, \ldots, n, \\
Q_{6,i} &= \frac{1}{h_i} \int_{x_i}^{x_{i+1}} (x_{i+1} - x_i) f(x)\, dx, \quad i = 1, 2, \ldots, n.
\end{aligned}
$$

These integrals can be expensive to evaluate. For efficiency, we approximate the functions $p(x)$, $q(x)$ and $f(x)$ by piecewise linear polynomial interpolants, and then integrate the resulting integrals exactly. If we define $h = \max_{0 \leq i \leq n} h_i$, then the error in each piecewise linear interpolant is $O(h^2)$, which yields in $O(h^3)$ error in each approximate integral, since the width of each interval of integration is $O(h)$. The resulting approximations are

$$
\begin{aligned}
Q_{1,i} &\approx \frac{h_i}{12}[q(x_i) + q(x_{i+1})], \\
Q_{2,i} &\approx \frac{h_{i-1}}{12}[3q(x_i) + q(x_{i-1})], \\
Q_{3,i} &\approx \frac{h_i}{12}[3q(x_i) + q(x_{i+1})], \\
Q_{4,i} &\approx \frac{1}{2h_{i-1}}[p(x_i) + p(x_{i-1})], \\
Q_{5,i} &\approx \frac{h_{i-1}}{6}[2f(x_i) + f(x_{i-1})], \\
Q_{6,i} &\approx \frac{h_i}{6}[2f(x_i) + f(x_{i+1})].
\end{aligned}
$$

It can be shown that the matrix A with entries defined from these approximate integrals is not only symmetric and tridiagonal, but also positive definite, so it is stable with respect to roundoff error, and can be solved using methods such as the conjugate gradient method that are appropriate for symmetric positive definite systems. It is no surprise that A should be symmetric positive definite, as it arises from a functional involving a self-adjoint, positive definite differential operator.

With h defined as before, it can also be shown that the error in the approximate solution is $O(h^2)$. Higher-order accuracy can be achieved by using higher-degree piecewise polynomials as basis functions, such as cubic splines. Such a choice also helps to ensure that the approximate solution is differentiable, unlike the solution computed using piecewise linear basis functions, which is continuous but not differentiable at the points x_i, $i = 1, 2, \ldots, n$. With cubic splines, the error in the computed solution is $O(h^4)$ as opposed to $O(h^2)$ in the piecewise linear case, due to the two additional degrees of differentiability. However, the drawback is that the matrix arising form the use of higher-degree basis functions is no longer tridiagonal; the upper and lower bandwidth are each equal to the degree of the piecewise polynomial that is used.