Finite Element Methods

Solving Operator Equations Via Minimization

We start with several definitions.

Definition 1. Let $\mathcal{V}$ be an inner product space. A linear operator $L\colon D \subset \mathcal{V} \to \mathcal{V}$ is said to be **positive definite** if $\langle v, Lv \rangle > 0$ for every $v \neq 0$ in D. If $\langle v, Lv \rangle \geq 0$ for every $v \in D$, then L is said to be **positive semidefinite**.

The first step is to show that some operator equations may be solved by minimizing a related quadratic functional.

Theorem 2. Suppose $\mathcal{V}$ is a real inner product space, and $K\colon \mathcal{V} \to \mathcal{V}$ is a self-adjoint positive definite linear operator. Suppose the operator equation

$$K[u] = f$$

has a solution. Then this solution, call it $u_\star$, is unique. Moreover, if we define an associated quadratic functional

$$Q[u] \ := \ \frac{1}{2}\langle u, K[u] \rangle - \langle f, u \rangle \tag{1}$$

for all admissable $u \in \mathcal{V}$, then $Q[u_\star] < Q[u]$ for all admissable $u \neq u_\star$.

Proof. To establish uniqueness of solution, suppose $u, v \in \mathcal{V}$ are such that $K[u] = K[v] = f$. Then

$$\langle u - v, K[u-v] \rangle \ = \ \langle u - v, K[u] - K[v] \rangle \ = \ \langle u - v, f - f \rangle \ = \ \langle u - v, 0 \rangle \ = \ 0 \ .$$

By positive definiteness, $u = v$.

Now note that, for any admissible u

$$
\begin{aligned}
Q[u] &= \frac{1}{2}\langle u, K[u]\rangle - \langle u, f\rangle \;=\; \frac{1}{2}\langle u, K[u]\rangle - \langle u, k[u_\star]\rangle \\
&= \frac{1}{2}\langle u, K[u]\rangle - \frac{1}{2}\langle u, K[u_\star]\rangle - \frac{1}{2}\langle u, K[u_\star]\rangle \\
&= \frac{1}{2}\langle u, K[u-u_\star]\rangle - \frac{1}{2}\langle u_\star, K[u]\rangle \\
&= \frac{1}{2}\langle u, K[u-u_\star]\rangle - \frac{1}{2}\Big\{\langle u_\star, K[u-u_\star]\rangle + \langle u_\star, K[u_\star] - K[u]\rangle\Big\} - \frac{1}{2}\langle u_\star, K[u]\rangle \\
&= \frac{1}{2}\langle u-u_\star, K[u-u_\star]\rangle - \frac{1}{2}\langle u_\star, K[u_\star]\rangle \;.
\end{aligned}
$$

But $\langle u-u_\star, K[u-u_\star]\rangle \geq 0$ and achieves its minimum value zero when and only when $u = u_\star$. Thus, Q also is minimized (uniquely) when $u = u_\star$. □

Example 1:

Consider the ODE/BVP

$$-y'' = f(x), \quad 0 < x < \ell, \qquad \text{subject to BCs} \qquad y(0) = 0 = y(\ell). \tag{2}$$

We are working here with the operator $K[y] = -y''$, the one-dimensional negative Laplacian, subject to homogeneous Dirchlet BCs, so it is self-adjoint and positive definite. To see the latter of these assertions, note that for each ϕ that is twice-differentiable in $(0,\ell)$ with $\phi'' \in L^2(0,\ell)$ (the natural inner product space for us to work in), we have

$$\Big\langle \phi, K[\phi]\Big\rangle \;=\; -\int_0^\ell \phi(x)\phi''(x)\,dx \;=\; -\phi(x)\phi'(x)\Big|_0^\ell + \int_0^\ell [\phi'(x)]^2\,dx \;=\; \Big\langle\!\Big\langle \phi, \phi \Big\rangle\!\Big\rangle \;\geq\; 0,$$

where $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ denotes the *1-dimensional Dirichlet inner product*

$$\Big\langle\!\Big\langle \phi, \psi \Big\rangle\!\Big\rangle \;:=\; \int_0^\ell \phi'(x)\psi'(x)\,dx \;. \tag{3}$$

Note that $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ is not truly an inner product in many contexts, as one can have $\Big\langle\!\Big\langle \phi, \phi \Big\rangle\!\Big\rangle = 0$ with $\phi \neq 0$. However, with our BCs, $\Big\langle\!\Big\langle \phi, \phi \Big\rangle\!\Big\rangle = 0$ implies $\phi \equiv 0$.

Thus, the solution (if one exists) of our problem is the function $y_\star$ minimizing

$$Q[y] \;:=\; \frac{1}{2}\langle y, Ky\rangle - \langle f, y\rangle \;=\; \frac{1}{2}\langle\!\langle y, y\rangle\!\rangle - \langle f, y\rangle \;.$$

■

It is interesting to note that, while the operator K requires its arguments to be (at least piecewise) twice differentiable, Q (in its formulation involving the Dirichlet inner product) only requires arguments (admissible functions) from

$$\mathcal{A} \;=\; \left\{ v\colon [0,\ell] \to \mathbb{R} \,\middle|\, v \text{ is continuous, } v' \text{ is PWC and bdd., and } v(0) = 0 = v(\ell) \right\} \;.$$

As was the case when we solved IBVPs using Fourier series, if we solve (2) by a process which minimizes the associated functional Q, the result may be a *weak solution*.

Example 2: Poisson Problem in the Plane with Dirichlet BCs

Consider the problem

$$-\Delta u = f, \quad (x,y) \in \Omega, \qquad \text{with} \qquad u = 0 \;\text{ for }\; (x,y) \in \partial\Omega.$$

Here we assume that Ω is a bounded, connected region in $\mathbb{R}^n$ (say, $n = 2$ or 3) with piecewise smooth boundary $\partial\Omega$. Working under the inner product of $L^2(\Omega)$, our operator $K = -\Delta$ (with the prescribed BCs) is once again self-adjoint and positive definite. The argument for the latter is similar to the above

$$\begin{aligned}\left\langle \phi, K[\phi] \right\rangle \;&=\; -\int_\Omega \phi(\mathbf{x})\, \Delta\phi(\mathbf{x})\, d\mathbf{x} \;=\; -\int_{\partial\Omega} \phi(\mathbf{x})\, (\nabla\phi \cdot \mathbf{n})(\mathbf{x})\, d\sigma + \int_\Omega \nabla\phi(\mathbf{x}) \cdot \nabla\phi(\mathbf{x})\, d\mathbf{x} \\ &=\; \int_\Omega \|\nabla\phi(\mathbf{x})\|^2\, d\mathbf{x} \;=\; \left\langle\!\left\langle \phi, \phi \right\rangle\!\right\rangle \;\geq\; 0\,,\end{aligned}$$

where the *n-dimensional Dirichlet inner product* (truly an inner product because of the BCs) is given by

$$\left\langle\!\left\langle \phi, \psi \right\rangle\!\right\rangle \;:=\; \int_\Omega \nabla\phi(\mathbf{x}) \cdot \nabla\psi(\mathbf{x})\, d\mathbf{x} \tag{4}$$

Hence, the solution to our problem, when it exists, is the unique minimizer $u_\star$ of the functional

$$Q[u] \;:=\; \frac{1}{2}\,\langle u, Ku\rangle - \langle f, u\rangle \;=\; \frac{1}{2}\,\langle\!\langle u, u\rangle\!\rangle - \langle f, u\rangle \;.$$

■

Note that, in the case $\Omega \subset \mathbb{R}^2$,

$$Q[u] \;=\; \iint_\Omega \left(\frac{1}{2}\, u_x^2 + \frac{1}{2}\, u_y^2 - f u \right) dx\, dy$$

and is defined (at least) for all functions u which are piecewise C^1 in Ω and satisfy the homogeneous Dirichlet BCs.

FEM: General Rayleigh-Ritz Approach

We have established that a linear operator equation $K[u] = f$ subject to homogeneous Dirichlet BCs, with K self-adjoint and positive definite, may be recast in the **variational form**

$$\min_{u\in D} Q[u], \tag{5}$$

where Q is a **quadratic functional**, and this minimization occurs over some collection D of admissible functions in a larger inner product space $\mathcal{V}$ (probably an L^2 space). Even though the admissible functions D (generally) do not constitute the full space, D is (usually) an infinite-dimensional subspace. The idea behind the Rayleigh-Ritz[1] approach to FEM is to severely restrict the scope of our search for a minimizer. Instead of searching throughout D, we limit our search to admissible functions lying in some *finite-dimensional* subspace $\mathcal{W}$. In particular, we may fix a choice of independent functions $\phi_1, \phi_2, \ldots, \phi_n \in D$ and take $\mathcal{W} = \operatorname{span}(\{\phi_1, \ldots, \phi_n\})$. We then look to solve

$$\min_{u\in\mathcal{W}} Q[u]. \tag{6}$$

Since each $u \in \mathcal{W}$ has the form

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

(6) is really about choosing the best coefficients $\mathbf{c} = (c_1, \ldots, c_n)$. In an abuse of notation, we now write our quadratic functional as if the input is $\mathbf{c}$:

$$Q(\mathbf{c}) \;:=\; \frac{1}{2}\langle u, K[u]\rangle - \langle f, u\rangle \;, \qquad \text{with} \qquad u = u(x;\mathbf{c}) = \sum_{j=1}^{n} c_j\phi_j(x)\,.$$

Plugging the latter expression for u into the functional yields

$$\begin{aligned} Q(\mathbf{c}) \;&=\; \frac{1}{2}\Big\langle \sum_i c_i\phi_i, K\Big[\sum_j c_j\phi_j\Big]\Big\rangle - \Big\langle f, \sum_i c_i\phi_i\Big\rangle \\ &=\; \frac{1}{2}\sum_i\sum_j c_ic_j\,\Big\langle \phi_i, K[\phi_j]\Big\rangle - \sum_i c_i\,\Big\langle f, \phi_i\Big\rangle \\ &=\; \frac{1}{2}\,\mathbf{c}^T\mathbf{M}\mathbf{c} - \mathbf{c}^T\mathbf{b}\,, \end{aligned} \tag{7}$$

where the **stiffness matrix** $\mathbf{M} = (m_{ij})$ and **load vector** $\mathbf{b} = (b_1, \ldots, b_n)$ are given by

$$m_{ij} \;=\; \Big\langle \phi_i, K[\phi_j]\Big\rangle, \qquad b_i \;=\; \Big\langle f, \phi_i\Big\rangle \;, \tag{8}$$

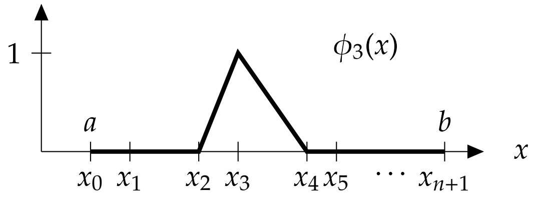
for $i = 1, \ldots, n$, and $j = 1, \ldots, n$. The problem of minimizing a quadratic functional (7) may be a new one to us, but it is a elementary problem in optimization. The **stiffness matrix M**, like the

[1] For short blurbs about Rayleigh and Ritz, see Stanoyevitch, p. 426.

underlying operator K, is symmetric (self-adjoint) and positive definite (so nonsingular), and the minimizer is known to be the unique solution of

$$\mathbf{M}\mathbf{c} = \mathbf{b}, \qquad \text{that is,} \qquad \mathbf{c} = \mathbf{M}^{-1}\mathbf{b}.$$

Example 3: Hat Function Basis for One-Dimensional ODEs/BVPs



It seems one of the most common bases to use in the case of a 1-dimensional ODEs/BVP on $[a, b]$ with homogeneous Dirichlet BCs is one consisting of **hat functions**. Let us take the (possibly non-uniform) partition

$$a = x_0 < x_1 < x_2 < \cdots < x_{n+1} = b,$$

with $h_k = x_{k+1} - x_k$ for $k = 0, \ldots, n$, and for $j = 1, 2, \ldots, n$ let $\phi_j(x)$ be the continuous function which is linear on each subinterval $[x_k, x_{k+1}]$ and whose values at mesh points are given by $\phi_j(x_m) = \delta_{jm}$ (Kronecker delta). A plot of $\phi_4(x)$ appears above at right.

Now recall that the problem (2)

$$-y'' = f(x), \;\; 0 < x < \ell, \qquad \text{subject to BCs} \qquad y(0) = 0 = y(\ell),$$

may be solved by minimizing the functional

$$Q[v] \; := \; \frac{1}{2}\langle\!\langle v, v\rangle\!\rangle - \langle f, v\rangle \; = \; \frac{1}{2}\int_0^\ell \left[\left(\frac{dv}{dx}\right)^2 - f(x)v(x)\right] dx\,,$$

over the set of admissible functions

$$\mathcal{A} \; = \; \left\{ v\colon [0,\ell] \to \mathbb{R} \,\middle|\, v \text{ is continuous, } v' \text{ is PWC and bdd., and } v(0) = 0 = v(\ell) \right\}.$$

The **piecewise linear Rayleigh-Ritz method** assumes a partition of the interval $[0, \ell]$ and seeks to minimize our functional over the finite-dimensional collection of piecewise linear functions $\mathcal{W} = \text{span}(\{\phi_1, \phi_2, \ldots, \phi_n\})$. That is, we take as our approximate solution

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

where the c_j's are the entries of the vector $\mathbf{c}$ which satisfies $\mathbf{M}\mathbf{c} = \mathbf{b}$, with $\mathbf{M} = (m_{ij})$ and $\mathbf{b}$ having entries given by

$$m_{ij} \;=\; \left\langle \phi_i, \phi_j'' \right\rangle \;=\; \left\langle\!\left\langle \phi_i, \phi_j \right\rangle\!\right\rangle \;=\; \int_0^\ell \phi_i'(x)\phi_j'(x)\,dx \;=\; \cdots \;=\; \begin{cases} \dfrac{1}{h_{i-1}} + \dfrac{1}{h_i}, & i = j, \\ -\dfrac{1}{h_{\min\{i,j\}}}, & |j - i| = 1, \\ 0, & |j - i| > 1, \end{cases} \tag{9}$$

$$b_i \;=\; \left\langle f, \phi_i \right\rangle \;=\; \cdots \;=\; \frac{1}{h_{i-1}}\int_{x_{i-1}}^{x_i} f(x)(x - x_{i-1})\,dx + \frac{1}{h_i}\int_{x_i}^{x_{i+1}} f(x)(x_{i+1} - x)\,dx\,. \tag{10}$$

Some remarks:

- The matrix **M** is tridiagonal (sparse), which is highly desirable, as the number of **elements** (in this case, subintervals of the original domain $[0,\ell]$) may typically be quite large. This sparsity is owing to the fact that the **support** of the basis functions is fairly small and overlaps with the support of just two other basis functions. In settings where more elaborate basis functions are used, this property is still desirable.
- It is generally wise to place more nodes in regions where the (known) coefficient functions of the differential equation undergo more activity.
- When $\max_i h_i$ is small, we might use the *trapezoid rule* to evaluate the integrals in (10) for the b_i's, we get

$$b_i \;\approx\; \frac{1}{2h_{i-1}}[0+f(x_i)h_{i-1}]h_{i-1}+\frac{1}{2h_i}[f(x_i)h_i+0]h_i \;=\; \frac{1}{2}\,f(x_i)(h_{i-1}+h_i)\;.$$

After Stanoyevitch, p. 435ff, apply these ideas to (2) with $\ell = 1$ and

$$f(x) \;=\; 100\sin\left(\operatorname{sign}(x-0.5)\exp\left(\frac{1}{4|x-0.5|^{1.05}+0.3}\right)\right)\exp\left(\frac{1}{4|x-0.5|^{1.2}+0.2}-100(x-0.5)^2\right)\;.$$

The code for doing so is found in the file `stanoP436.m`. Some new Octave/Matlab commands of note:

`diff()`, `sign()`, `end` as an index to a vector

The first few lines of the code contain a choice between two types of meshes, one that is uniform, and one that is *adaptive*, being rather coarse where f is well behaved but much finer in the region $0.35 \le x \le 0.65$ where f is highly oscillatory. There is also a switch controlling whether the elements of the load vector **b** are computed using accurate numerical integration or with the trapezoid rule approximation mentioned above. Results are quite different between the two methods using the uniform mesh, but about identical for the adaptive mesh.

■