9 Concepts behind finite-element method

The power of the finite-element method (FEM) becomes evident when handling problems in two or three spatial dimentions. In this lecture (and this course), we will only consider applications of the FEM to problems in one spatial dimention (i.e., BVPs on an interval), so as to demonstrate some of the basic concepts behind this method.

9.1 General idea of FEM

Suppose we are looking for a solution of the BVP

$$y'' + Q(x)y = R(x), \qquad y(0) = y(1) = 0\,. \tag{9.1}$$

This BVP has zero boundary conditions at both end points of the interval; however, this does *not* restrict the generality of the subsequent exposition. Namely, a way to treat nonzero boundary conditions is described in a homework problem.

The idea of the FEM is as follows. Select a set of linearly independent function $\{\phi_j(x)\}_{j=1}^{M}$ such that each ϕ_j satisfies the boundary conditions of the BVP, i.e.

$$\phi_j(0) = \phi_j(1) = 0, \qquad j = 1, \ldots, M\,. \tag{9.2}$$

Then, look for an approximate solution of the BVP in the form

$$Y(x) = \sum_{j=1}^{M} c_j \phi_j(x)\,, \tag{9.3}$$

where coefficients c_j are to be determined. Note that since the ϕ_j's satisfy the boundary conditions of the BVP, so does the solution $Y(x)$. The problem has now become the following: (i) decide which basis functions $\phi_j(x)$ to use and (ii) determine the coefficients c_j so as to make the error between $Y(x)$ and the exact solution $y(x)$ as small as possible.

The term 'basis' describing the set $\{\phi_j\}$ is used here to indicate that functions ϕ_j must be linearly independent and, moreover, their linear superpositions (i.e., the r.h.s. of (9.3)) should be able to approximate functions (i.e., solutions of the BVP) from a sufficiently large class sufficiently closely. A quantitaive characterization of the two 'sufficiently's in the previous sentence is a serious mathematical task, which we will not attempt to undertake. Instead, we will proceed at the intuitive level.

One possible set of basis functions which satisfy boundary conditions (9.2) is

$$\phi_j(x) = \sin(j\pi x), \qquad j = 1, \ldots, M\,. \tag{9.4}$$

Another, more convenient, set will be introduced later on as we proceed. In general, there may be many choices for $\{\phi_j(x)\}$; the decision as to which one to use is usually made on a problem-by-problem basis.

The problem of determining the coefficients c_j can be handled in three different ways, which we will now describe.

9.2 Collocation method

Let us substitute expansion (9.3) into BVP (9.1):

$$\sum_{j=1}^{M} c_j \phi_j''(x) \, + Q(x) \sum_{j=1}^{M} c_j \phi_j(x) \; = \; R(x) \, . \tag{9.5}$$

Recall that due to the choice (9.2), the boundary conditions are satisfied automatically.

Now, ideally, we want Eq. (9.5) to hold identically, i.e. for all $x \in [0,\, 1]$. However, since we only have M free parameters, $\{c_j\}_{j=1}^{M}$, at our disposal, we can only require that Eq. (9.5) be satisfied at M points, called *collocation* points. That is, if $\{x_k\}_{k=1}^{M}$ is a set of such points on $[0,\, 1]$, then we require that the following system of (linear) equations hold:

$$\sum_{j=1}^{M} \left(\phi_j''(x_k) + Q(x_k)\phi_j(x_k) \right) \, c_j \; = \; R(x_k), \qquad k = 1, \ldots, M \, . \tag{9.6}$$

Upon solving this system of M equations for the M unknowns c_j and substituting their values into (9.3), one finds the approximate solution $Y(x)$ of the BVP.

Linear system (9.6) can be written in the standard form

$$A\vec{\mathbf{c}} = \vec{\mathbf{r}} \, , \tag{9.7}$$

where

$$\begin{gathered} (A)_{kj} = \left(\phi_j''(x_k) + Q(x_k)\phi_j(x_k) \right) , \\ \vec{\mathbf{r}} = \left(R(x_1), \; \ldots \; , \; R(x_M) \right)^T \, . \end{gathered} \tag{9.8}$$

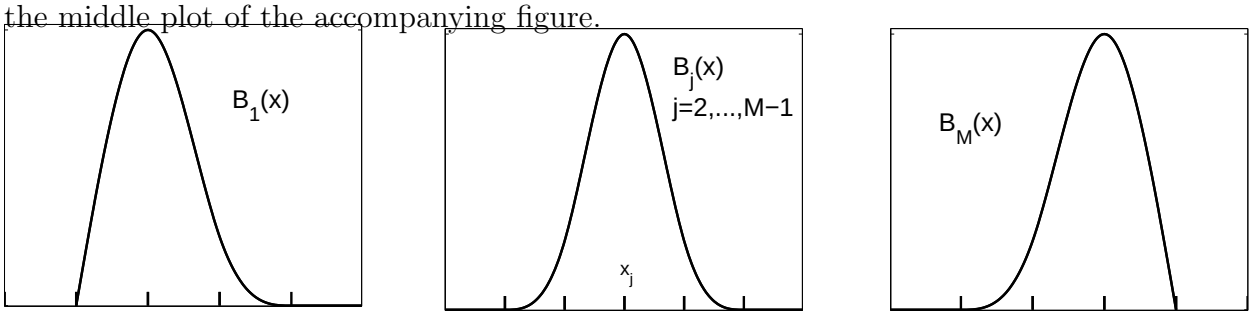
The coefficient matrix A is, in general, *full* (i.e. *not* tri- or pentadiagonal), whereas matrix A that arose in the finite-difference approach of Lecture 8 was tridiagonal. Thus it appears that the collocation methods leads to a system that is more difficult to solve than the system produced by the finite-difference approach. However, one can make matrix A of the collocation method to also be tridiagonal if one chooses the basis functions in a special form. Namely, for $j = 2, \ldots, M-1$, take ϕ_j's to be the following *cubic B-splines*:

$$B_j(x) = \begin{cases} \dfrac{(\Delta x_{j-2})^3}{4h^3}, & x_{j-2} \le x \le x_{j-1} \\ \dfrac{1}{4} + \dfrac{3\Delta x_{j-1}}{4h} \left[1 + \dfrac{\Delta x_{j-1}}{h} - \left(\dfrac{\Delta x_{j-1}}{h} \right)^2 \right], & x_{j-1} \le x \le x_j \\ \dfrac{1}{4} - \dfrac{3\Delta x_{j+1}}{4h} \left[1 - \dfrac{\Delta x_{j+1}}{h} - \left(\dfrac{\Delta x_{j+1}}{h} \right)^2 \right], & x_j \le x \le x_{j+1} \\ -\dfrac{(\Delta x_{j+2})^3}{4h^3}, & x_{j+1} \le x \le x_{j+2} \\ 0, & \text{otherwise} \end{cases} \tag{9.9}$$

$$j = 2, \; \ldots, \; M-1,$$

where $\Delta x_j = x - x_j$, and we have assumed for simplicity that all points are equidistant, so that $h = x_{j+1} - x_j$. Recall that $x_0 = 0$ and $x_{M+1} = 1$ for BVP (9.1). Functions $B_j(x)$ (9.9) have

continuous first and second derivatives everywhere on [0, 1]; a typical such function is shown in the middle plot of the accompanying figure.



When $j = 1$ or $j = M$, these functions have to be slightly modified, so that, say, $B_1(x)$ satisfies the boundary condition at x_0: $B_1(x_0) = 0$, but no condition is placed on its derivative at that point. The plots of B_1 and B_M are shown in the left and right plots of the figure; the analytical expression for, say, B_1 being:

$$B_1(x) = \begin{cases} \dfrac{1}{2(4-3h)}\left[3(6-5h)\dfrac{\Delta x_0}{h} - 9(1-h)\left(\dfrac{\Delta x_0}{h}\right)^2 - \left(\dfrac{\Delta x_0}{h}\right)^3\right], & x_0 \le x \le x_1 \\ \dfrac{1}{4} - \dfrac{3\Delta x_2}{4h}\left[1 - \dfrac{\Delta x_2}{h} - \left(\dfrac{\Delta x_2}{h}\right)^2\right], & x_1 \le x \le x_2 \\ -\dfrac{(\Delta x_3)^3}{4h^3}, & x_2 \le x \le x_3 \\ 0, & \text{otherwise.} \end{cases} \tag{9.10}$$

Now, with $\phi_j(x) = B_j(x)$, system (9.6) has a tridiagonal matrix A because for any x_k, only $B_k(x_k)$ and $B_{k\pm1}(x_k)$ are nonzero, i.e. $B_j(x_k) = 0$ for $|k - j| \ge 1$. Then all one needs in order to write out the explicit form of (9.6) are the values of $B_k(x_k)$ and $B_{k\pm1}(x_k)$. These can be deduced from the entries in the table below.

	x_{k-1}	x_k	x_{k+1}
$B_k(x)$	1/4	1	1/4
$B_k''(x)$	$3/(2h^2)$	$-3/h^2$	$3/(2h^2)$

To conclude this subsection, we point out the advantage of the collocation method over the finite-difference method of Lecture 8: Points x_k do not need to be equidistant. This will have no effect on the form of system (9.6); only the coefficients in (9.9) and (9.10) will be slightly modified. One can use this freedom in distributing the collocation points over the interval so that to place more of them in the region(s) where the solution is expected to change rapidly.

9.3 Galerkin method

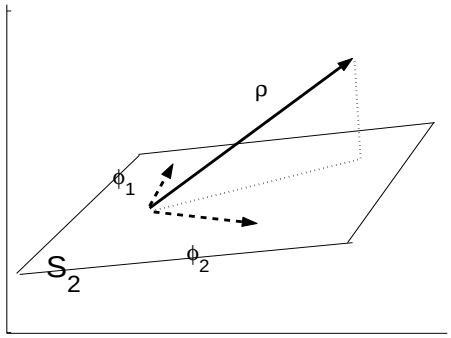
This method allows one to use basis functions that are simpler than the B-splines considered in the previous subsection. The solution obtained, of course, will not be as smooth as that obtained by the collocation method.

As we said above, the approximate solution $Y(x)$ is sought as a linear combination of the basis functions ϕ_j; see (9.3). We can now draw an analogy with Linear Algebra and call the basis functions ϕ_j *vectors* which span (i.e., form) a linear, M-dimensional space S_M. Then, according to (9.3), $Y(x)$ is a vector in that space.

Let us again substitute (9.3) into BVP (9.1) and write the result as

$$\sum_{j=1}^{M} \left(\phi_j''(x) + Q(x)\phi_j(x)\right)\, c_j \;-\; R(x) \;=\; \rho(x)\,. \tag{9.11}$$

(Recall that in the collocation method, we required that the *residual vector* $\rho(x_k) = 0$ at all the collocation points x_k.) Now, the residual vector $\rho(x)$ does *not*, in general, belong to the linear space S_M (in other words, it is not a linear combination of ϕ_j's). Geometrically, we can represent this situation (for $M = 2$) as in the figure on the right. Namely, vector $\rho(x)$ has a component that belongs to S_M and the other component that lies outside of S_M.



The idea of the Galerkin method is to select the coefficients c_j so as to make the residual vector $\rho(x)$ *orthogonal* to all of the basis functions ϕ_j, $j = 1, \ldots, M$. In that case, the projection of $\rho(x)$ on S_M is zero, and hence the "length" of $\rho(x)$ is minimized, since

$$\text{"length" } \rho(x) \;=\; \sqrt{\left(\text{"length" } \rho_{\|\,\text{to}\,S_M}(x)\right)^2 \;+\; \left(\text{"length" } \rho_{\perp\,\text{to}\,S_M}(x)\right)^2}\,.$$

Thus, we need to specify what we mean by 'orthogonal' and 'length' for functions. Two functions $f(x)$ and $g(x)$ are called orthogonal if

$$\int_0^1 f(x)g(x)dx \;=\; 0\,. \tag{9.12}$$

Two remarks are in order. First, note that the integral in (9.12) is over [0, 1]. This is because the BVP we are considering is defined over that interval. If a BVP is defined over $[a, b]$, the corresponding definition of orthogonality would contain $\int_a^b$ instead of $\int_0^1$. Second, (9.12) is *not* the only definition of orthogonality of functions, but just one of those which are used frequently. For different applications, different definitions of function orthogonality may prove to be more convenient.

The definition of 'length' of a function is subordinate to that of orthogonality, namely:

$$||f(x)||_2 = \sqrt{\int_0^1 (f(x))^2\, dx}. \tag{9.13}$$

The subscript '2' of $||...||_2$ is used because the l.h.s. of (9.13) is also known as the L_2-norm of a function, which is different from the ∞-norm that we have considered so far.

Thus, the Galerkin method requires that

$$\int_0^1 \rho(x)\phi_k(x)\, dx \;=\; 0 \qquad \text{for } k = 1, \ldots, M. \tag{9.14}$$

With the account of (9.11), this gives:

$$\sum_{j=1}^{M} c_j \int_0^1 \left(\phi_j''(x) + Q(x)\phi_j(x)\right)\phi_k(x)\,dx \;=\; \int_0^1 R(x)\phi_k(x)\,dx\,; \qquad k = 1,\ldots,M. \tag{9.15}$$

If we now define a matrix A to have the coefficients

$$a_{kj} = \int_0^1 \left(\phi_j''(x) + Q(x)\phi_j(x)\right)\phi_k(x)\,dx \tag{9.16}$$

and the vector $\vec{\mathbf{r}}$ to be

$$\vec{\mathbf{r}} \;=\; \left(\int_0^1 R(x)\phi_1(x)\,dx,\; \ldots,\; \int_0^1 R(x)\phi_M(x)\,dx\right)^T, \tag{9.17}$$

then the system of linear equations (9.15) takes on the familiar form (9.7).

So far, there has been no real advantage of the Galerkin method over the collocation method. Such an advantage arises when we use integration by parts to rewrite (9.16) in the form

$$a_{kj} = -\int_0^1 \phi_j'(x)\phi_k'(x)\,dx \;+\; \int_0^1 Q(x)\phi_j(x)\phi_k(x)\,dx\,. \tag{9.18}$$

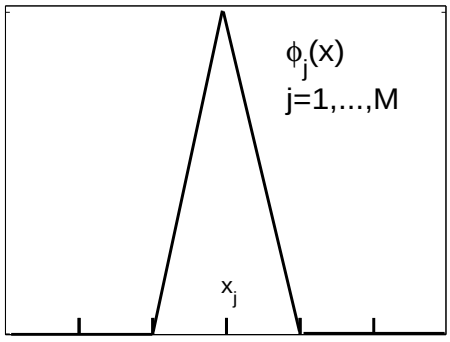
In deriving (9.18), we have used the boundary conditions (9.2). From (9.18), which is equivalent to (9.16), we immediately observe two things, which were not evident from (9.16).

- $a_{kj} = a_{jk}$, i.e. the coefficient matrix in the Galerkin method is *symmetric*.
- To calculate a_{kj}, one only requires ϕ_j', but not ϕ_j'', to exist. Moreover, one does not require ϕ_j' to be continuous; it suffices that it be integrable.

Then, the following simple choice of $\phi_j(x)$ can be made:

$$\phi_j(x) \;=\; \begin{cases} 1 - \dfrac{|\Delta x_j|}{h}, & x_{j-1} \le x \le x_{j+1} \\ 0 & \text{otherwise,} \end{cases} \tag{9.19}$$

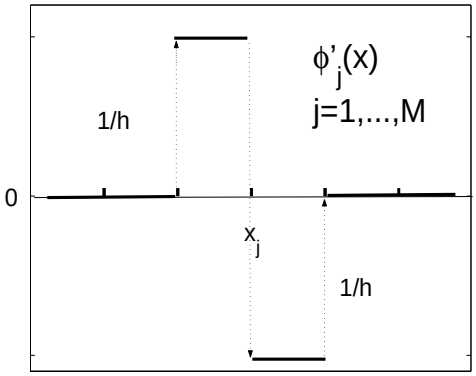
where Δx_j is defined after (9.9). These functions $\phi_j(x)$ are called *hat*-functions, or linear B-splines.



With this choice for ϕ_j's, matrix A is tridiagonal. Indeed, in this case ϕ_j' is as shown on the right, whence one can calculate that

$$\int_0^1 \phi_j'(x)\phi_k'(x)\,dx \;=\; \begin{cases} -\dfrac{1}{h}, & k = j \pm 1 \\ \dfrac{2}{h}, & k = j \\ 0, & \text{otherwise.} \end{cases} \tag{9.20}$$

Quantities $\int_0^1 Q(x)\phi_j(x)\phi_k(x)\,dx$ are nonzero also only for $k = j-1,\, j,\, j+1$, as is evident from the figure above.



To conclude this subsection, let us remark on the issue of the calculation of the second integral in (9.18) and the integrals in (9.17). For brevity, we will only speak here about the former type of integrals; the same will apply to the latter ones. The integrals $\int_0^1 Q(x)\phi_j(x)\phi_k(x)\,dx$ may be evaluated in one of the following ways. First, if their analytical expressions for all required pairs of k and j can be obtained (e.g., with Mathematica or another computer algebra package), the numeric values of these expressions should be used. If such expressions are not available, then the computation may differ depending on whether the ϕ_j's are smooth and non-vanishing over the entire interval $[0,1]$, as, say, functions (9.4), or they are the hat-functions (or any other highly localized functions). In the former case, the integrals in question may be computed by one of the standard methods (say, Simpson's) using the existing subdivision of $[0,1]$, or by Matlab's built-in integrators (`quad` or `quadl`). In the latter case, i.e. for highly localized ϕ_j's, the integrals can be approximated as

$$\int_0^1 Q(x)\phi_j(x)\phi_k(x)\,dx \approx Q(x_{\rm mid})\int_0^1 \phi_j(x)\phi_k(x)\,dx\,, \tag{9.21}$$

where $x_{\rm mid}$ is the middle of the interval over which the product $\phi_j(x)\phi_k(x)$ is nonzero. One can show that the accuracy of approximation (9.21) is $O(h^2)$. In the case when ϕ_j's are the hat-functions (9.19), the integral on the r.h.s. of (9.21) can be explicitly calculated to be:

$$\int_0^1 \phi_j(x)\phi_k(x)\,dx \;=\; \begin{cases} \dfrac{h}{6}, & k=j\pm 1 \\ \dfrac{2h}{3}, & k=j \\ 0, & \text{otherwise.} \end{cases} \tag{9.22}$$

9.4 Rayleigh-Ritz method

This method replaces the problem of solving BVP (9.1) by a problem of finding the minimum of a certain functional. We will not consider this method in more detail, but only mention that: (i) Rayleigh-Ritz method can be shown to be equivalent to Galerkin method, and (ii) the functional mentioned in the previous sentence is the "length" of the residual vector $\rho(x)$, defined according to (9.13).

9.5 Questions for self-assessment

1. Describe the idea behind the collocation method.
2. What condition (or conditions) should the basis functions in the collocation method satisfy?
3. In addition to the above condition(s), what other condition should the basis functions satisfy in order to make the corresponding coefficient matrix tridiagonal?
4. What is the advantage of the collocation method over the finite-diference method?
5. Write down the explicit form of B_M.

6. Verify the entries in the Table in Sec. 9.2.

7. Show how these entries are related to $B_{k\pm1}(x_k)$.

8. Describe the idea behind the Galerkin method.

9. Try to explain the analogy between definition (9.12) of orthogonality of functions and the definition of orthogonality of vectors in R^n. *Hint*: Interpret the integral as (the limit of) a finite, say, Riemann, sum. (The fact that it is the limit is not really important here.)

10. Consequently, explain why (9.13) is analogous to the Euclidean length of a vector.

11. Continuing from the last two questions, try to explain the close analogy between the Galerkin method and the least-squares solution of inconsistent linear systems.

12. What is the advantage of the Galerkin method over the collocation method?