Lecture 9
Approximation Techniques

1 The Residual

Consider an equation of the following generic form:

$$A\ f(\overrightarrow{x}) = g(\overrightarrow{x})\,, \tag{1}$$

where A is a linear operator, $f(\overrightarrow{x})$ is the solution, and $g(x)$ is the driving or source function. Let us assume an approximate trial-space expansion for $f(\overrightarrow{x})$ as follows:

$$f(\overrightarrow{x}) \approx \tilde{f}(\overrightarrow{x}) = \sum_{i=1}^{N} f_i B_i(\overrightarrow{x})\,. \tag{2}$$

We further assume that the trial-space functions are chosen so that $f(\overrightarrow{x})$ naturally meets any required boundary conditions. Next we define the residual function as follows:

$$R(\overrightarrow{x}) = g(\overrightarrow{x}) - A\,\tilde{f}(\overrightarrow{x})\,. \tag{3}$$

We would like to choose the expansion coefficients $\{f_i\}_{i=1}^{N}$ so as to make $R(\overrightarrow{x})$ identically zero because then $\tilde{f}(\overrightarrow{x})$ would exactly solve Eq. (1). However, this is obviously almost never possible. The error in the solution is proportional to the size of the residual. To show this, we apply A^{-1} to Eq. (3), and slightly manipulate the resulting equation to obtain

$$f(\overrightarrow{x}) - \tilde{f}(\overrightarrow{x}) = A^{-1}R(\overrightarrow{x})\,, \tag{4}$$

where $f(\overrightarrow{x}) = A^{-1}g(\overrightarrow{x})$. Thus if we can't make the residual identically zero, we should try to make it "small" in some sense.

2 Three Basic Methods

There are several different ways to make the residual "small":

- The Least-Squares Method: choose the expansion coefficients so that the integral of the square of the residual is minimized, i.e. minimize

$$\Gamma = \sum_{\overrightarrow{x}} R^2(\overrightarrow{x})\, d\overrightarrow{x}\,. \tag{5}$$

- The Weighted Residual Method: choose the expansion coefficients so that the residual is orthogonal over the problem domain to a set of N linearly-independent functions $\left\{W_i(\overrightarrow{x})\right\}_{i=1}^{N}$, i.e.,

$$\int_{\overrightarrow{x}} R(\overrightarrow{x})W_i(\overrightarrow{x})\, d\overrightarrow{x} = 0\,, \quad i = 1, N. \tag{6}$$

  The weighting functions form an N-dimensional space of functions called the weighting space. The trial space can be identical to the weighting space, i.e.,

$$B_i(\overrightarrow{x}) = W_i(\overrightarrow{x})\,, \tag{7}$$

  or it can be a different space. In the former case, one obtains the Galerkin method, and in the latter case one obtains the Petrov-Galerkin method. A particularly useful method for transport calculations is the *discontinuous* Galerkin method, which is characterized by a trial space that is discontinuous at the cell interfaces.

- The Collocation Method: choose the expansion coefficients so that the residual is zero at N distinct points $\left\{\overrightarrow{x}_i\right\}_{i=1}^{N}$:

$$R(\overrightarrow{x}_i) = 0\,, \quad i = 1, N. \tag{8}$$

It is clear as to why the residual is made "small" using the least-squares method and the collocation method, but it is less clear for the weighted residual method. The explanation is as follows. If the residual is orthogonal to the weighting functions, the least-squares fit to the residual using the weighting functions is identically zero. Thus the residual must be "near zero". For instance, we represent the least-squares fit as

$$\tilde{R}(\overrightarrow{x}) = \sum_{i=1}^{N} R_i W_i(\overrightarrow{x})\,. \tag{9}$$

The expansion coefficients are chosen to minimize the following functional:

$$\Gamma = \int_X \,[R(\overrightarrow{x}) - \sum_{i=1}^{N} R_i W_i(\overrightarrow{x})]^2 \; d\overrightarrow{x}\;. \tag{10}$$

This minimization is achieved by requiring that $\frac{\partial \Gamma}{\partial R_i} = 0, \; i = 1, N$. This results in the following matrix equation for the vector of expansion coefficients:

$$\mathcal{W}\;(\overrightarrow{R}) = \overrightarrow{\xi}\;, \tag{11a}$$

where the elements of $\mathcal{W}$ are defined by

$$w_{i,j} = \int_X W_i(\overrightarrow{x})W_j(\overrightarrow{x})\, d\overrightarrow{x}\ , \tag{11b}$$

and

$$\overrightarrow{R} = (R_1, R_2, \cdots R_N)\ , \tag{11c}$$

$$\overrightarrow{\xi} = (\xi_1, \xi_2, \cdots \xi_N)\ , \tag{11d}$$

$$\xi_i = \int_X R(\overrightarrow{x})W_i(\overrightarrow{x})\, d\overrightarrow{x}\ . \tag{11e}$$

Thus the vector of expansion coefficients is given by

$$\overrightarrow{R} = \mathcal{W}^{-1}\,\overrightarrow{\xi}\ . \tag{11f}$$

If the residual is orthogonal to the weighting function, the $\overrightarrow{\xi}$ vector is identically zero, and it follows from Eq. (11f) that the expansion coefficients must also be zero.

3 Examples

Consider the following equation:

$$\frac{df}{dx} + \sigma f = 0\,, \tag{12}$$

which is defined over the interval, $[0, x_0]$, with the boundary condition, $f(0) = 1$. The solution to this equation is

$$f(x) = \exp(-\sigma x)\,. \tag{13}$$

The Taylor-series expansion about $x = 0$ is

$$f(x) = 1 - \tau + \frac{1}{2}\tau^2 - \frac{1}{6}\tau^3 + \frac{1}{24}\tau^4 + O(\tau^5)\,, \tag{14}$$

where $\tau = \sigma x$.

3.1 A Least-Squares Example

We next approximately solve Eq. (12) using the least-squares method in conjunction with a linear trial space that satisfies the boundary condition:

$$\tilde{f}(x) = 1 + ax\,, \tag{15}$$

where a is a constant to be determined. Substituting from Eq. (15) into Eq. (13), and forming the least-squares functional, we get:

$$\Gamma = \int_0^{x_0} (a + \sigma(1 + ax))^2 \ dx\,. \tag{16}$$

To minimize Γ, we set $\frac{\partial \Gamma}{\partial a} = 0$:

$$\int_0^{x_0} 2\left[a + \sigma(1 + ax)\right](1 + \sigma x) \ dx = 0\,. \tag{17}$$

Manipulating Eq. (17), we get

$$\int_0^{x_0} a + \sigma + (2a\sigma + \sigma^2)x + a\sigma^2 x^2 \ dx = 0\,, \tag{18}$$

$$(a + \sigma)x_0 + (2a\sigma + \sigma^2)\frac{x_0^2}{2} + a\sigma^2\frac{x_0^3}{3} = 0\,, \tag{19}$$

$$a = -\frac{\sigma(6 + 3\sigma x_0)}{6 + 6\sigma x_0 + 2\sigma^2 x_0^2}\,. \tag{20}$$

Thus the approximate solution is

$$\tilde{f}(x) = 1 - \frac{\sigma x(6 + 3\sigma x_0)}{6 + 6\sigma x_0 + 2\sigma^2 x_0^2}\,. \tag{21}$$

Evaluating the solution at $x = x_0$ gives

$$\tilde{f}(x_0) = 1 - \frac{\sigma x_0(6 + 3\sigma x_0)}{6 + 6\sigma x_0 + 2\sigma^2 x_0^2}\,. \tag{22}$$

Expanding this solution about $x_0 = 0$ gives

$$\tilde{f}(x_0) = 1 - \tau + \frac{1}{2}\tau^2 - \frac{1}{6}\tau^3 + O(\tau^5)\,, \tag{23}$$

where $\tau = \sigma x_0$. Comparing Eq. (23) with Eq. (14), we find that the $\tilde{f}(x_0)$ is accurate through third order. This is quite good for a linear continuous approximation. On the other hand, for any cell thickness greater than approximately 2.45 mean-free-paths, the solution at x_0 is negative and therefore non-physical. Furthermore, in the limit as $x_0 \to \infty$, we find that $\tilde{f}(x_0) \to -\frac{1}{2}$, which is asymptotically incorrect and non-physical. If we integrate Eq. (12) over the interval, $[0, x_0]$, we obtain the analog of the balance equation:

$$f(x_0) - f(0) + \int_0^{x_0} \sigma f \; dx = 0\,. \tag{24}$$

If we substitute $\tilde{f}$ into Eq. (24), we find that it is not satisfied:

$$\tilde{f}(x_0) - f(0) + \int_0^{x_0} \sigma \tilde{f} \; dx = -\frac{\tau^3}{12 + 12\tau + 4\tau^2}\,. \tag{25}$$

This is in keeping with the fact that least-squares methods are generally not conservative. However, also note that Eq. (25) is nonetheless met through second order in τ, which reflects the fact that it is a convergent method.

3.2 A Weighted Residual Example

We next approximately solve Eq. (12) using the weighted-residual method in conjunction with the linear trial space defined by Eq. (15). We will use the a weight function of unity to ensure a conservative solution. Following Eq. (6), we substitute from Eq. (15) into Eq. (12), and integrate the resulting equation over the interval, $[0, x_0]$, to obtain the equation for a:

$$\int_0^{x_0} \left[a + \sigma(1 + ax)\right]\ dx = 0\,. \tag{26}$$

Manipulating Eq. (26), we get

$$a(x_0 + \frac{1}{2}\sigma x_0^2) + \sigma x_0 = 0\,,$$

$$a = -\frac{2\sigma}{2 + \sigma x_0}\,. \tag{27}$$

Substituting from Eq. (27) into Eq. (15), we obtain the approximate solution:

$$\tilde{f}(x) = 1 - \frac{\sigma x}{1 + \frac{1}{2}\sigma x_0}\,. \tag{28}$$

Evaluating Eq. (28) at $x = x_0$, we get

$$\tilde{f}(x_0) = 1 - \frac{\sigma x_0}{1 + \frac{1}{2}\sigma x_0}\,. \tag{29}$$

Expanding Eq. (29) about $x_0 = 0$, we obtain

$$\tilde{f}(x_0) = 1 - \tau + \frac{1}{2}\tau^2 - \frac{1}{4}\tau^3 + O(\tau^4)\,, \tag{30}$$

where $\tau = \sigma x_0$. Comparing Eq. (29) with Eq. (14), we find that $\tilde{f}(x_0)$ is correct through second order. This is good for a linear approximation. However, for any cell thickness greater than 2 mean-free-paths, the solution at x_0 is negative and therefore non-physical. Furthermore, in the limit as $x_0 \rightarrow \infty$, we find that $\tilde{f}(x_0) \rightarrow -1$, which is asymptotically incorrect and non-physical. Nonetheless if we substitute from Eq. (28) into Eq. (24), we find that the resulting equation is satisfied. This is in keeping with the fact that a weight function of unity should result in a conservative solution.

3.3 A Collocation Example

We next approximately solve Eq. (12) using the weighted-residual method in conjunction with the linear trial space defined by Eq. (15). We choose $x_0/2$ as the collocation point for reasons explained later. Following Eq. (8), we obtain the equation for the constant a:

$$a + \sigma(1 + a x_0/2) = 0\,. \tag{31}$$

Solving for a, we get

$$a = -\frac{2\sigma}{2 + \sigma x_0}\,. \tag{32}$$

Comparing Eq. (32) with Eq. (27), we find that the collocation solution is identical to the weighted-residual solution. To see why this is so, one need simply perform the integral in Eq. (26) using a one-point quadrature with the quadrature point equal to $x_0/2$, and the quadrature weight equal to $\Delta x = x_0$:

$$[a + \sigma(1 + ax_0/2)]\,\Delta x = 0\,. \tag{33}$$

Because the quadrature point is equal to $x_0/2$, the quadrature integration is exact (the midpoint rule applied to a linear integrand). Furthermore, if one divides Eq. (33) by Δx, one obtains Eq. (31). Thus in this case, collocation at $x = x_0/2$ is equivalent to an exact quadrature integration of the weighted residual equation, so the collocation method yields the weighted-residual result. This is representative of a general approach often taken with collocation. For instance, let us assume that we have collocated at quadrature points, $\{x_n\}_{n=1}^{N}$ with associated weights $\{w_n\}_{n=1}^{N}$. Then the residual satisfies

$$R(x_n) = 0\,, \quad n = 1, N. \tag{34}$$

Given a set of weight functions, $\{W_i(x)\}_{i=1}^{N}$, we multiply Eq. (34) by $W_i(x_n)w_n$, where i successively takes on each value from 1 to N:

$$W_i(x_n)\,R(x_n)\,w_n = 0 \quad n = 1, N, \quad i = 1, N. \tag{35}$$

Summing Eq. (35) over all n, we get

$$\sum_{n=1}^{N} W_i(x_n)\, R(x_n)\, w_n = 0 \quad i = 1, N. \tag{36}$$

If the quadrature formula is exact for the product of the residual and each weight function, it is clear from Eq. (36) that the weighted residual equations will be satisfied, and therefore that the collocation and weighted residual methods will be equivalent. If not, the collocation method should be "similar" to the weighted residual method.

3.4 Two Discontinuous Galerkin Examples

The discontinuous Galerkin method is based upon a trial space that is continuous within each cell but discontinuous across cell faces. It is critical to uniquely define the approximate solution on each cell face. In particular, the approximate solution must be defined so that it is continuous at each cell face *in the direction of flow.* This "upwinded" face definition is illustrated in Fig. 3.4 for flow in the direction of increasing x. The unknown in Eq. (12) is associated with flow in the direction of increasing x. For simplicity, we first consider a constant-discontinuous approximation for the solution to this equation. The approximate solution takes the following form within the cell $[0, x_0]$:

$$\begin{aligned} \tilde{f} &= 1\,, \quad \text{for } x = 0, \\ &= f_c\,, \quad \text{otherwise.} \end{aligned} \tag{37}$$

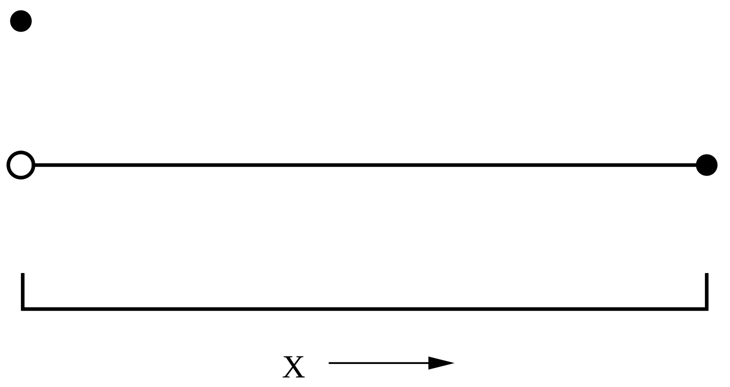

Figure 1: The trial space is discontinuous across cell interfaces, but continuous in the direction of particle flow. The above dependence is appropriate for $\mu > 0$.

The weighting space used in the discontinuous Galerkin method is identical to the trial space except that the weighting space is continuous on $[0, x_0]$. Since we are using a constant trial-space approximation, the weight function is constant over the whole cell:

$$W(x) = 1 \, , \quad \text{for } x \in [0, 1]. \tag{38}$$

Substituting from Eq. (37) into Eq. (12), multiplying by the weight function, and integrating that equation over the cell, we get the equation for f_c, the constant flux value within the cell:

$$\int_0^{x_0} W(x) \left[\frac{\partial \tilde{f}}{\partial x} + \sigma \tilde{f} \right] \, dx = 0 \, . \tag{39}$$

because $\tilde{f}$ jumps at $x = 0$, it has a delta-function derivative at $x = 0$. In particular,

$$\left. \frac{\partial \tilde{f}}{\partial x} \right|_{x=0} = \delta(x) \left(f_c - 1 \right) \, . \tag{40}$$

The validity of the above expression follows from the fact that

$$\tilde{f}(x) = f(0) + \int_0^x \frac{\partial \tilde{f}}{\partial x'} \, dx' \, , \quad \text{for } x \in [0, 1]. \tag{41}$$

One can avoid explicitly defining the derivative of $\tilde{f}$ simply by integrating the derivative term in Eq. (39) by parts:

$$\int_0^{x_0} \left[\frac{\partial \left(W\tilde{f} \right)}{\partial x} - \tilde{f}\frac{\partial W}{\partial x} + \sigma \tilde{f} \right] \, dx = 0 \, . \tag{42}$$

The first term in Eq. (42) is an exact derivative which involves only the evaluation of $\tilde{f}$ on the cell faces, and the second term in Eq. (42) involves only a derivative of the weight function, which never has a delta-function component. Evaluating Eq. (42), we obtain

$$f_c - 1 + \sigma f_c x_0 = 0 \, . \tag{43}$$

Solving Eq. (43) for f_c yields

$$f_c = \frac{1}{1+\sigma x_0} \, , \tag{44}$$

Since the trial space is constant within the cell, it follows that

$$\tilde{f}(x_0) = \frac{1}{1+\sigma x_0} \, , \tag{45}$$

Expanding this expression about $x_0 = 0$, we get

$$\tilde{f}(x_0) = 1 - \tau + \tau^2 \, , \tag{46}$$

where $\tau = \sigma x_0$. Comparing Eq. (46) with Eq. (14), we find that the constant-discontinuous approximation is first-order accurate for $f(x_0)$. This is fairly crude. On the other hand, the solution is monotone decreasing and always positive. Furthermore, in the limit as $x_0 \rightarrow \infty$, we find that $\tilde{f}(x_0) \rightarrow 0$, which is asymptotically correct. Finally, if we substitute from Eq. (45) into Eq. (24), we find that Eq. (24) is satisfied. Thus the constant-discontinuous approximation is conservative, which is consistent with a constant weight function. In general, discontinuous Galerkin methods tend to be much more robust than continuous methods, and are preferred for difficult calculations.

We next consider a linear-discontinuous finite-element approximation to Eq. (12). This the mainstay of modern spatial discretization techniques for the transport equation. The approximate solution takes the following form within the cell $[0, x_0]$:

$$\begin{aligned} \tilde{f} &= 1\,, \quad \text{for } x = 0, \\ &= f_L B_L(x) + f_R B_R(x)\,, \quad \text{otherwise,} \end{aligned} \tag{47}$$

where

$$B_L(x) = \frac{x_0 - x}{x_0}\,, \tag{48}$$

$$B_R(x) = \frac{x}{x_0}\,. \tag{49}$$

The weighting space used in the discontinuous Galerkin method is identical to the trial space except that the weighting space is continuous on $[0, x_0]$. Thus we weight with B_L to obtain the equation for f_L:

$$\frac{1}{2}(f_L + f_R) - 1 + \frac{1}{2}\sigma\left(\frac{2}{3}f_L + \frac{1}{3}f_R\right)x_0 = 0\,, \tag{50}$$

and we weight with B_R to obtain the equation for f_R:

$$f_R - \frac{1}{2}(f_L + f_R) + \frac{1}{2}\sigma\left(\frac{1}{3}f_L + \frac{2}{3}f_R\right)x_0 = 0\,. \tag{51}$$

The solution for $\tilde{f}(x_0) \equiv f_R$ is

$$\tilde{f}(x_0) = \frac{6 - 2\sigma x_0}{6 + 4\sigma x_0 + \sigma^2 x_0^2}\ . \tag{52}$$

This solution goes to zero for large x_0, and achieves a minimum value of about -0.1 at about 8 mean-free-paths. Thus negative solutions are relatively unlikely and higly damped if they occur. Expanding this expression about $x_0 = 0$ yields

$$\tilde{f}(x_0) = 1 - \tau + \frac{1}{2}\tau^2 + \frac{1}{6}\tau^3 + \frac{1}{36}\tau^4 + O(\tau^5)\,, \tag{53}$$

where $\tau = \sigma x_0$. Comparing Eq. (53) with Eq. (14), we find that this expression is correct through third order. This is a phenomenon known as superconvergence - i.e., convergence at an order that is higher than one would normally expect given the polynomial order of the trial space. The linear-discontinuous method is superconvergent (third-order accurate) for the cell average and outflow edge solutions, but less than second-order accurate for the interior solution associated with the inflow edge (not the inflow edge value itself which meets the exact boundary condition, but rather f_L in this instance). This is excellent behavior and explains the popularity of DFEM methods.

4 Finite-Element Lumping

Finite-element lumping is a process by which the robustness of finite-element schemes is improved at the cost of accuracy. Proper lumping retains the order accuracy of the unlumped scheme unless the latter is superconvergent, in which case, an order of accuracy is generally lost. Robustness relates to resistance to negativities and rapid damping of oscillations. There are many prescriptions for lumping, but all of them have the same result: the span of the discretization stencil is reduced and the diagonal matrix elements increase in size. For example, let us consider the lumped linear-discontinuous method.

The removal and source terms in Eqs. (50 and (51) are lumped via the following replacements: $\left(\frac{2}{3}f_L + \frac{1}{3}f_R\right) \rightarrow f_L$:

$$\frac{1}{2}f_L + f_R - 1 + \frac{1}{2}\sigma_t f_L x_0 = 0\,, \tag{54}$$

and $\left(\frac{2}{3}f_R + \frac{1}{3}f_L\right) \rightarrow f_R$:

$$f_R - \frac{1}{2}(f_L + f_R) + \frac{1}{2}\sigma_t f_R x_0 = 0\,. \tag{55}$$

The gradient terms can be lumped in 1-D curvilinear and multidimensional geometries, but they cannot be lumped in 1-D slabs because the span of the stencil is already minimal. Lumping of gradient terms on non-orthognal meshes is still a very active area of research. When one evaluates the solution for $f(x_0)$ resulting from the lumped equations, one gets

$$f(x_0) = \frac{1}{1 + \sigma x_0 + \frac{1}{2}\sigma^2 x_0^2}\,. \tag{56}$$

This expression goes to zero for large x_0 and is *strictly positive*. Expanding this expression about $x_0 = 0$ yields

$$\tilde{f}(x_0) = 1 - \tau + \frac{1}{2}\tau^2 - \frac{1}{6}\tau^3 + \frac{1}{36}\tau^4 + O(\tau^5)\,, \tag{57}$$

where $\tau = \sigma x_0$. Comparing Eq. (52) with Eq. (14), we find that this expression is correct through third order. This solution is not only second-order accurate, but strictly positive as well. We have lost an order of accuracy, but we still have a second-order method because of the superconvergence of the unlumped scheme.

The lumping process can often be related to the use of quadrature formula to evaluate the finite-element integrals rather than perfoming the exact integrations. For instance, if we use trapzoidal quadrature *after integration by parts* to perform the integrals for the linear-discontinuous method, we obtain the lumped equations. This quadrature rule can be expressed for this purpose as follows:

$$\int_0^{x_0} f(x)\,dx \approx (f(0) + f(x_0))\,\frac{x_0}{2}\,. \tag{58}$$