6

Galerkin's Method

It is necessary to solve differential equations. (Newton)

Ideally, I'd like to be the eternal novice, for then only, the surprises would be endless. (Keith Jarret)

In Chapters 3 and 5, we discussed the numerical solution of the simple initial value problem $u'(x) = f(x)$ for $a < x \leq b$ and $u(a) = u_0$, using piecewise polynomial approximation. In this chapter, we introduce *Galerkin's method* for solving a general differential equation, which is based on seeking an (approximate) solution in a (finite-dimensional) space spanned by a set of basis functions which are easy to differentiate and integrate, together with an orthogonality condition determining the coefficients or coordinates in the given basis. With a finite number of basis functions, Galerkin's method leads to a system of equations with finitely many unknowns which may be solved using a computer, and which produces an approximate solution. Increasing the number of basis functions improves the approximation so that in the limit the exact solution may be expressed as an infinite series. In this book, we normally use Galerkin's method in the computational form with a finite number of basis functions. The basis functions may be global polynomials, piecewise polynomials, trigonometric polynomials or other functions. The finite element method in basic form is Galerkin's method with piecewise polynomial approximation. In this chapter, we apply Galerkin's method to two examples with a variety of basis functions. The first example is an initial value problem that models population growth and we use a global polynomial approximation. The second example is a boundary

value problem that models the flow of heat in a wire and we use piecewise polynomial approximation, more precisely piecewise linear approximation. This is a classic example of the *finite element method.* For the second example, we also discuss the *spectral method* which is Galerkin's method with trigonometric polynomials.

The idea of seeking a solution of a differential equation as a linear combination of simpler basis functions, is old. Newton and Lagrange used power series with global polynomials and Fourier and Riemann used Fourier series based on trigonometric polynomials. These approaches work for certain differential equations posed on domains with simple geometry and may give valuable qualitative information, but cannot be used for most of the problems arising in applications. The finite element method based on piecewise polynomials opens the possibility of solving general differential equations in general geometry using a computer. For some problems, combinations of trigonometric and piecewise polynomials may be used.

6.1. Galerkin's method with global polynomials

6.1.1. A population model

In the simplest model for the growth of a population, like the population of rabbits in West Virginia, the rate of growth of the population is proportional to the population itself. In this model we ignore the effects of predators, overcrowding, and migration, for example, which might be okay for a short time provided the population of rabbits is relatively small in the beginning. We assume that the time unit is chosen so that the model is valid on the time interval $[0, 1]$. We will consider more realistic models valid for longer intervals later in the book. If $u(t)$ denotes the population at time t then the differential equation expressing the simple model is $\dot{u}(t) = \lambda u(t)$, where λ is a positive real constant and $\dot{u} = du/dt$. This equation is usually posed together with an initial condition $u(0) = u_0$ at time zero, in the form of an initial value problem:

$$\begin{cases} \dot{u}(t) = \lambda u(t) & \text{for } 0 < t \leq 1, \\ u(0) = u_0. \end{cases} \tag{6.1}$$

The solution of (6.1), $u(t) = u_0 \exp(\lambda t)$, is a smooth increasing function when $\lambda > 0$.

6.1.2. Galerkin's method

We now show how to compute a polynomial approximation U of u in the set of polynomials $V^{(q)} = \mathcal{P}^q(0,1)$ on $[0,1]$ of degree at most q using Galerkin's method. We know there are good approximations of the solution u in this set, for example the Taylor polynomial and interpolating polynomials, but these require knowledge of u or derivatives of u at certain points in $[0,1]$. The goal here is to compute a polynomial approximation of u using only the information that u solves a specified differential equation and has a specified value at one point. We shall see that this is precisely what Galerkin's method achieves. Since we already know the analytic solution in this model case, we can use this knowledge to evaluate the accuracy of the approximations.

Because $\{t^j\}_{j=0}^q$ is a basis for $V^{(q)}$, we can write $U(t) = \sum_{j=0}^q \xi_j t^j$ where the coefficients $\xi_j \in \mathbb{R}$ are to be determined. It is natural to require that $U(0) = u_0$, that is $\xi_0 = u_0$, so we may write

$$U(t) = u_0 + \sum_{j=1}^{q} \xi_j t^j,$$

where the "unknown part" of U, namely $\sum_{j=1}^q \xi_j t^j$, is in the subspace $V_0^{(q)}$ of $V^{(q)}$ consisting of the functions in $V^{(q)}$ that are zero at $t = 0$, i.e. in $V_0^{(q)} = \{v : v \in V^{(q)}, v(0) = 0\}$.

Problem 6.1. Prove that $V_0^{(q)}$ is a subspace of $V^{(q)}$.

We determine the coefficients by requiring U to satisfy the differential equation in (6.1) in a suitable "average" sense. Of course U can't satisfy the differential equation at every point because the exact solution is not a polynomial. In Chapter 4, we gave a concrete meaning to the notion that a function be zero on average by requiring the function to be orthogonal to a chosen subspace of functions. The Galerkin method is based on this idea. We define the *residual error* of a function v for the equation (6.1) by

$$R(v(t)) = \dot{v}(t) - \lambda v(t).$$

The residual error $R(v(t))$ is a function of t once v is specified. $R(v(t))$ measures how well v satisfies the differential equation at time t. If the residual is identically zero, that is $R(v(t)) \equiv 0$ for all $0 \le t \le 1$, then the equation is satisfied and v is the solution. Since the exact solution u is not a polynomial, the residual error of a function in $V^{(q)}$ that satisfies the initial condition is never identically zero, though it can be zero at distinct points.

The Galerkin approximation U is the function in $V^{(q)}$ satisfying $U(0) = u_0$ such that its residual error $R(U(t))$ is orthogonal to all functions in $V_0^{(q)}$, i.e.,

$$\int_0^1 R(U(t))v(t)\,dt = \int_0^1 (\dot{U}(t) - \lambda U(t))v(t)\,dt = 0 \quad \text{for all } v \in V_0^{(q)}. \tag{6.2}$$

This is the *Galerkin orthogonality* property of U, or rather of the residual $R(U(t))$. Since the coefficient of U with respect to the basis function 1 for $V^{(q)}$ is already known ($\xi_0 = u_0$), we require (6.2) to hold only for functions v in $V_0^{(q)}$. By way of comparison, note that the true solution satisfies a stronger orthogonality condition, namely

$$\int_0^1 (\dot{u} - \lambda u)v\,dt = 0 \quad \text{for } \textit{all} \text{ functions } v. \tag{6.3}$$

We refer to the set of functions where we seek the Galerkin solution U, in this case the space $V^{(q)}$ of polynomials w satisfying $w(0) = u_0$, as the *trial space* and the space of the functions used for the orthogonality condition, which is $V_0^{(q)}$, as the *test space*. In this case, the trial and test space are different because of the *non-homogeneous* initial condition $w(0) = u_0$ (assuming $u_0 \neq 0$), satisfied by the trial functions and the homogeneous boundary condition $v(0) = 0$ satisfied by the test functions $v \in V_0^{(q)}$. In general, different methods are obtained choosing the trial and test spaces in different ways.

6.1.3. The discrete system of equations

We now show that (6.2) gives an invertible system of linear algebraic equations for the coefficients of U. Substituting the expansion for U

into (6.2) gives

$$\int_0^1 \left(\sum_{j=1}^{q} j\xi_j t^{j-1} - \lambda u_0 - \lambda \sum_{j=1}^{q} \xi_j t^j \right) v(t)\, dt = 0 \quad \text{for all } v \in V_0^{(q)}.$$

It suffices to insure that this equation holds for every basis function for $V_0^{(q)}$, yielding the set of equations:

$$\sum_{j=1}^{q} j\xi_j \int_0^1 t^{j+i-1}\, dt - \lambda \sum_{j=1}^{q} \xi_j \int_0^1 t^{j+i}\, dt = \lambda u_0 \int_0^1 t^i\, dt, \quad i = 1, ..., q,$$

where we have moved the terms involving the initial data to the right-hand side. Computing the integrals gives

$$\sum_{j=1}^{q} \left(\frac{j}{j+i} - \frac{\lambda}{j+i+1} \right) \xi_j = \frac{\lambda}{i+1} u_0, \quad i = 1, ..., q. \tag{6.4}$$

This is a $q \times q$ system of equations that has a unique solution if the matrix $A = (a_{ij})$ with coefficients

$$a_{ij} = \frac{j}{j+i} - \frac{\lambda}{j+i+1}, \quad i, j = 1, ..., q,$$

is invertible. It is possible to prove that this is the case, though it is rather tedious and we skip the details. In the specific case $u_0 = \lambda = 1$ and $q = 3$, the approximation is

$$U(t) \approx 1 + 1.03448t + .38793t^2 + .301724t^3,$$

which we obtain solving a 3×3 system.

Problem 6.2. Compute the Galerkin approximation for $q = 1, 2, 3$, and 4 assuming that $u_0 = \lambda = 1$.

Plotting the solution and the approximation for $q = 3$ in Fig. 6.1, we see that the two essentially coincide.

Since we know the exact solution u in this case, it is natural to compare the accuracy of U to other approximations of u in $V^{(q)}$. In Fig. 6.2, we plot the errors of U, the third degree polynomial interpolating u at $0, 1/3, 2/3$, and 1, and the third degree Taylor polynomial

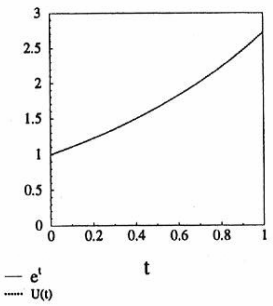

Figure 6.1: The solution of $\dot{u} = u$ and the third degree Galerkin approximation.

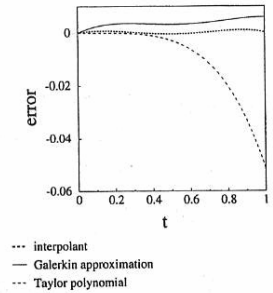

Figure 6.2: The errors of the third degree Galerkin approximation, a third degree interpolant of the solution, and the third degree Taylor polynomial of the solution.

of u computed at $t = 0$. The error of U compares favorably with the error of the interpolant of U and both of these are more accurate than the Taylor polynomial of u in the region near $t = 1$ as we would expect. We emphasize that the Galerkin approximation U attains this accuracy without any specific knowledge of the solution u except the initial data at the expense of solving a linear system of equations.

Problem 6.3. Compute the $L_2(0,1)$ projection into $\mathcal{P}^3(0,1)$ of the exact solution u and compare to U.

Problem 6.4. Determine the discrete equations if the test space is changed to $V^{(q-1)}$.

6.1.4. A surprise: ill-conditioning

Stimulated by the accuracy achieved with $q = 3$, we compute the approximation with $q = 9$. We solve the linear algebraic system in two ways: first exactly using a symbolic manipulation package and then approximately using Gaussian elimination on a computer that uses roughly 16 digits. In general, the systems that come from the discretization of a differential equation are too large to be solved exactly and we are forced to solve them numerically with Gaussian elimination for example.

We obtain the following coefficients ξ_i in the two computations:

exact coefficients	approximate coefficients
.14068...	152.72...
.48864...	−3432.6...
.71125...	32163.2...
.86937...	−157267.8...
.98878...	441485.8...
1.0827...	−737459.5...
1.1588...	723830.3...
1.2219...	−385203.7...
1.2751...	85733.4...

We notice the huge difference, which makes the approximately computed U worthless. We shall now see that the difficulty is related to the fact that the system of equations (6.4) is *ill-conditioned* and this problem is exacerbated by using the standard polynomial basis $\{t^i\}_{i=0}^q$.

Problem 6.5. If access to a symbolic manipulation program and to numerical software for solving linear algebraic systems is handy, then compare the coefficients of U computed exactly and approximately for $q = 1, 2, \ldots$ until significant differences are found.

It is not so surprising that solving a system of equations $A\xi = b$, which is theoretically equivalent to inverting A, is sensitive to errors in the coefficients of A and b. The errors result from the fact that the computer stores only a finite number of digits of real numbers. This sensitivity is easily demonstrated in the solution of the 1×1 "system" of equations $ax = 1$ corresponding to computing the inverse $x = 1/a$ of a given real number $a \neq 0$. In Fig. 6.3, we plot the inverses of two numbers a_1 and a_2 computed from two approximations $\tilde{a}_1$ and $\tilde{a}_2$ of the same accuracy. We see that the corresponding errors in the approximations $\tilde{x} = 1/\tilde{a}_i$ of the exact values $x = 1/a_i$ vary greatly in the two cases, since

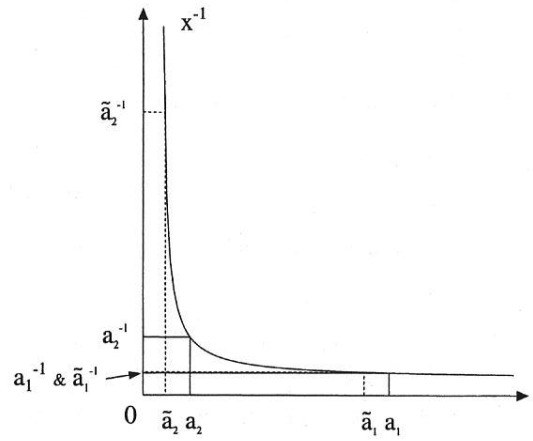

Figure 6.3: The sensitivity of the solution of $ax = 1$ to errors in a.

$1/a_i - 1/\tilde{a}_i = (\tilde{a}_i - a_i)/(a_i\tilde{a}_i)$. The closer a_i is to zero the more sensitive is the solution $1/a_i$ to errors in a_i. This expresses that computing $1/a$ is *ill-conditioned* when a is close to zero.

In general, the solution of $Ax = b$ is sensitive to errors in the entries of A when A is "close" to being non-invertible. Recall that a matrix is non-invertible if one row (or column) is a linear combination of the other rows (or columns). In the example of computing the coefficients of the Galerkin approximation with $q = 9$ above, we can see that there might be a problem if we look at the coefficient matrix A:

0.167	0.417	0.550	0.633	0.690	0.732	0.764	0.789	0.809
0.0833	0.300	0.433	0.524	0.589	0.639	0.678	0.709	0.735
0.0500	0.233	0.357	0.446	0.514	0.567	0.609	0.644	0.673
0.0333	0.190	0.304	0.389	0.456	0.509	0.553	0.590	0.621
0.0238	0.161	0.264	0.344	0.409	0.462	0.506	0.544	0.576
0.0179	0.139	0.233	0.309	0.371	0.423	0.467	0.505	0.538
0.0139	0.122	0.209	0.280	0.340	0.390	0.433	0.471	0.504
0.0111	0.109	0.190	0.256	0.313	0.360	0.404	0.441	0.474
0.00909	0.0985	0.173	0.236	0.290	0.338	0.379	0.415	0.447

which is nearly singular since the entries in some rows and columns are quite close. On reflection, this is not surprising because the last two rows are given by $\int_0^1 R(U,t)t^8\,dt$ and $\int_0^1 R(U,t)t^9\,dt$, respectively, and t^8 and t^9 look very similar on $[0,1]$. We plot the two basis functions in Fig. 6.4.

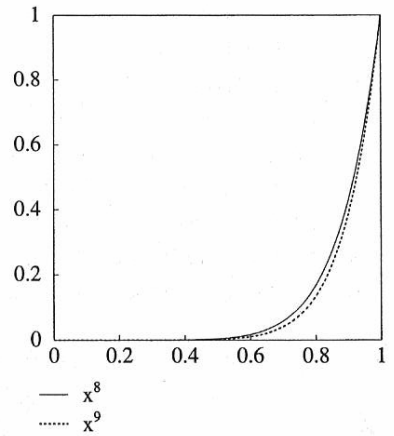

Figure 6.4: The basis functions t^8 and t^9.

In general, linear systems of algebraic equations obtained from the discretization of a differential equation tend to become ill-conditioned as the discretization is refined. This is understandable because refining the discretization and increasing the accuracy of the approximation makes it more likely that computing the residual error is influenced by the finite precision of the computer, for example. However, the degree of ill conditioning is influenced greatly by the differential equation and the choice of trial and test spaces, and even the choice of basis functions for these spaces. The standard monomial basis used above leads to an ill-conditioned system because the different monomials become very similar as the degree increases. This is related to the fact that the monomials are not an orthogonal basis. In general, the best results with respect to reducing the effects of ill-conditioning are obtained by using an orthogonal bases for the trial and test spaces. As an example, the Legendre polynomials, $\{\varphi_i(x)\}$, with $\varphi_0 \equiv 1$ and

$$\varphi_i(x) = (-1)^i \frac{\sqrt{2i+1}}{i!} \frac{d^i}{dx^i}(x^i(1-x)^i), \quad 1 \le i \le q,$$

form an orthonormal basis for $\mathcal{P}^q(0,1)$ with respect to the L_2 inner product. It becomes more complicated to formulate the discrete equations using this basis, but the effects of finite precision are greatly reduced.

Another possibility, which we take up in the second section, is to use piecewise polynomials. In this case, the basis functions are "nearly orthogonal".

Problem 6.6. (a) Show that φ_3 and φ_4 are orthogonal.

6.2. Galerkin's method with piecewise polynomials

We start by deriving the basic model of stationary heat conduction and then formulate a finite element method based on piecewise linear approximation.

6.2.1. A model for stationary heat conduction

We model heat conduction a thin heat-conducting wire occupying the interval $[0,1]$ that is heated by a *heat source* of intensity $f(x)$, see Fig. 6.5. We are interested in the stationary distribution of the temperature $u(x)$

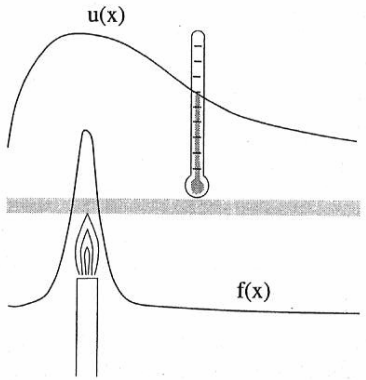

Figure 6.5: A heat conducting wire with a source $f(x)$.

in the wire. We let $q(x)$ denote the heat flux in the direction of the positive x-axis in the wire at $0 < x < 1$. Conservation of energy in a stationary case requires that the net heat flux through the endpoints of an arbitrary sub-interval (x_1, x_2) of $(0,1)$ be equal to the heat produced in (x_1, x_2) per unit time:

$$q(x_2) - q(x_1) = \int_{x_1}^{x_2} f(x)\,dx.$$