12 Galerkin and Ritz Methods for Elliptic PDEs

12.1 Galerkin Method

We begin by introducing a generalization of the collocation method we saw earlier for two-point boundary value problems. Consider the elliptic PDE

$$Lu(\boldsymbol{x}) = f(\boldsymbol{x}), \tag{110}$$

where L is a linear elliptic partial differential operator such as the Laplacian

$$L = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}, \quad \boldsymbol{x} = (x, y, z) \in \mathbb{R}^3.$$

At this point we will not worry about the boundary conditions that should be posed with (110).

As with the collocation method discussed earlier, we will obtain the approximate solution in the form of a function (instead of as a collection of discrete values). Therefore, we need an approximation space $\mathcal{U} = \text{span}\{u_1, \ldots, u_n\}$, so that we are able to represent the approximate solution as

$$u = \sum_{j=1}^{n} c_j u_j, \quad u_j \in \mathcal{U}. \tag{111}$$

Using the linearity of L we have

$$Lu = \sum_{j=1}^{n} c_j L u_j.$$

We now need to come up with n (linearly independent) conditions to determine the n unknown coefficients c_j in (111). If $\{\Phi_1, \ldots, \Phi_n\}$ is a linearly independent set of linear functionals, then

$$\Phi_i \left[\sum_{j=1}^{n} c_j L u_j - f \right] = 0, \quad i = 1, \ldots, n, \tag{112}$$

is an appropriate set of conditions. In fact, this leads to a system of linear equations

$$Ac = b$$

with matrix

$$A = \begin{bmatrix} \Phi_1 L u_1 & \Phi_1 L u_2 & \ldots & \Phi_1 L u_n \\ \Phi_2 L u_1 & \Phi_2 L u_2 & \ldots & \Phi_2 L u_n \\ \vdots & \vdots & & \vdots \\ \Phi_n L u_1 & \Phi_n L u_2 & \ldots & \Phi_n L u_n \end{bmatrix},$$

coefficient vector $c = [c_1, \ldots, c_n]^T$, and right-hand side vector

$$b = \begin{bmatrix} \Phi_1 f \\ \Phi_2 f \\ \vdots \\ \Phi_n f \end{bmatrix}.$$

Two popular choices are

1. *Point evaluation functionals*, i.e., $\Phi_i(u) = u(\boldsymbol{x}_i)$, where $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ is a set of points chosen such that the resulting conditions are linearly independent, and u is some function with appropriate smoothness. With this choice (112) becomes

$$\sum_{j=1}^{n} c_j L u_j(\boldsymbol{x}_i) = f(\boldsymbol{x}_i), \quad i = 1, \ldots, n,$$

and we now have an extension of the *collocation method* discussed in Chapter 9 to elliptic PDEs is the multi-dimensional setting.

2. If we let $\Phi_i(u) = \langle u, v_i \rangle$, an inner product of the function u with an appropriate *test function* v_i, then (112) becomes

$$\sum_{j=1}^{n} c_j \langle L u_j, v_i \rangle = \langle f, v_i \rangle, \quad i = 1, \ldots, n.$$

If $v_i \in \mathcal{U}$ then this is the classical *Galerkin method*, otherwise it is known as the *Petrov-Galerkin method*.

12.2 Ritz-Galerkin Method

For the following discussion we pick as a model problem a multi-dimensional Poisson equation with homogeneous boundary conditions, i.e.,

$$\begin{aligned} -\nabla^2 u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{113}$$

with domain $\Omega \subset \mathbb{R}^d$. This problem describes, e.g., the steady-state solution of a vibrating membrane (in the case $d = 2$ with shape Ω) fixed at the boundary, and subjected to a vertical force f.

The first step for the Ritz-Galerkin method is to obtain the *weak form* of (113). This is accomplished by choosing a function v from a space $\mathcal{U}$ of smooth functions, and then forming the inner product of both sides of (113) with v, i.e.,

$$-\langle \nabla^2 u, v \rangle = \langle f, v \rangle. \tag{114}$$

To be more specific, we let $d = 2$ and take the inner product

$$\langle u, v \rangle = \iint_{\Omega} u(x, y) v(x, y) dx dy.$$

Then (114) becomes

$$-\iint_{\Omega} (u_{xx}(x, y) + u_{yy}(x, y)) v(x, y) dx dy = \iint_{\Omega} f(x, y) v(x, y) dx dy. \tag{115}$$

In order to be able to complete the derivation of the weak form we now assume that the space $\mathcal{U}$ of test functions is of the form

$$\mathcal{U} = \{v \ : \ v \in C^2(\Omega), \ v = 0 \text{ on } \partial\Omega\},$$

i.e., besides having the necessary smoothness to be a solution of (113), the functions also satisfy the boundary conditions.

Now we rewrite the left-hand side of (115):

$$\begin{aligned}\iint_\Omega (u_{xx}+u_{yy})\,v dxdy &= \iint_\Omega \left[(u_x v)_x + (u_y v)_y - u_x v_x - u_y v_y\right] dxdy \\ &= \iint_\Omega \left[(u_x v)_x + (u_y v)_y\right] dxdy - \iint_\Omega \left[u_x v_x - u_y v_y\right] dxdy \quad (116)\end{aligned}$$

By using Green's Theorem (integration by parts)

$$\iint_\Omega (P_x + Q_y) dxdy = \int_{\partial\Omega} (Pdy - Qdx)$$

the first integral on the right-hand side of (116) turns into

$$\iint_\Omega \left[(u_x v)_x + (u_y v)_y\right] dxdy = \int_{\partial\Omega} \left(u_x v dy - u_y v dx\right).$$

Now the special choice of $\mathcal{U}$, i.e., the fact that v satisfies the boundary conditions, ensures that this term vanishes. Therefore, the weak form of (113) is given by

$$\iint_\Omega \left[u_x v_x + u_y v_y\right] dxdy = \iint_\Omega f v dxdy.$$

Another way of writing the previous formula is of course

$$\iint_\Omega \nabla u \cdot \nabla v dxdy = \iint_\Omega f v dxdy. \quad (117)$$

To obtain a numerical method we now need to require $\mathcal{U}$ to be finite-dimensional with basis $\{u_1, \ldots, u_n\}$. Then we can represent the approximate solution u^h of (113) as

$$u^h = \sum_{j=1}^{n} c_j u_j. \quad (118)$$

The superscript h indicates that the approximate solution is obtained on some underlying discretization of Ω with mesh size h.

Remark 1. In practice there are many ways of discretizing Ω and selecting $\mathcal{U}$.

(a) For example, regular (tensor product) grids can be used. Then $\mathcal{U}$ can consist of tensor products of piecewise polynomials or B-spline functions that satisfy the boundary conditions of the PDE.

(b) It is also possible to use irregular (triangulated) meshes, and again define piecewise (total degree) polynomials or splines on triangulations satisfying the boundary conditions.

(c) More recently, meshfree approximation methods have been introduced as possible choices for $\mathcal{U}$.

2. In the literature the piecewise polynomial approach is usually referred to as the *finite element method*.

3. The discretization of Ω will almost always result in a computational domain that has piecewise linear (Lipschitz-continuous) boundary.

We now return to the discussion of the general numerical method. Once we have chosen a basis for the approximation space $\mathcal{U}$, then it becomes our goal to determine the coefficients c_j in (118). By inserting u^h into the weak form (117), and selecting as trial functions v the basis functions of $\mathcal{U}$ we obtain a system of equations

$$\iint_\Omega \nabla u^h \cdot \nabla u_i dxdy = \iint_\Omega f u_i dxdy, \quad i = 1, \ldots, n.$$

Using the representation (118) of u^h we get

$$\iint_\Omega \nabla \left[\sum_{j=1}^{n} c_j u_j \right] \cdot \nabla u_i dxdy = \iint_\Omega f u_i dxdy, \quad i = 1, \ldots, n,$$

or by linearity

$$\sum_{j=1}^{n} c_j \iint_\Omega \nabla u_j \cdot \nabla u_i dxdy = \iint_\Omega f u_i dxdy, \quad i = 1, \ldots, n. \tag{119}$$

This last set of equations is known as the *Ritz-Galerkin method* and can be written in matrix form

$$Ac = b,$$

where the *stiffness matrix* A has entries

$$A_{i,j} = \iint_\Omega \nabla u_j \cdot \nabla u_i dxdy.$$

Remark 1. The stiffness matrix is usually assembled element by element, i.e., the contribution to the integral over Ω is split into contributions for each *element* (e.g., rectangle or triangle) of the underlying mesh.

2. Depending on the choice of the (finite-dimensional) approximation space $\mathcal{U}$ and underlying discretization, the matrix will have a well-defined structure. This is one of the most important applications driving the design of efficient linear system solvers.

Example One of the most popular finite element versions is based on the use of piecewise linear C^0 polynomials (built either on a regular grid, or on a triangular partition of Ω). The basis functions u_i are "hat functions", i.e., functions that are

piecewise linear, have value one at one of the vertices, and zero at all of its neighbors. This choice makes it very easy to satisfy the homogeneous Dirichlet boundary conditions of the model problem exactly (along a polygonal boundary).

Since the gradients of piecewise linear functions are constant, the entries of the stiffness matrix essentially boil down to the areas of the underlying mesh elements.

Therefore, in this case, the Ritz-Galerkin method is very easily implemented. We generate some examples with Matlab's PDE toolbox `pdetool`.

It is not difficult to verify that the stiffness matrix for our example is symmetric and positive definite. Since the matrix is also very sparse due to the fact that the "hat" basis functions have a very localized support, efficient iterative solvers can be applied. Moreover, it is known that the piecewise linear FEM converges with order $\mathcal{O}(h^2)$.

Remark 1. The Ritz-Galerkin method was independently introduced by Walther Ritz (1908) and Boris Galerkin (1915).

2. The finite element method is one of the most-thoroughly studied numerical methods. Many textbooks on the subject exist, e.g., "The Mathematical Theory of Finite Element Methods" by Brenner and Scott (1994), "An Analysis of the Finite Element Method" by Strang and Fix (1973), or "The Finite Element Method" by Zienkiewicz and Taylor (2000).

12.3 Optimality of the Ritz-Galerkin Method

How does solving the Ritz-Galerkin equations (119) relate to the solution of the strong form (113) of the PDE? First, we remark that the left-hand side of (117) can be interpreted as a new inner product

$$[u, v] = \iint_{\Omega} \nabla u \cdot \nabla v dx dy \tag{120}$$

on the space of functions whose first derivatives are square integrable and that vanish on $\partial\Omega$. This space is a *Sobolev space*, usually denoted by $H_0^1(\Omega)$.

The inner product $[\cdot, \cdot]$ induces a norm $\|v\| = [v, v]^{1/2}$ on $H_0^1(\Omega)$. Now, using this norm, the best approximation to u from $H_0^1(\Omega)$ is given by the function u^h that minimizes $\|u - u^h\|$. Since we define our numerical method via the finite-dimensional subspace $\mathcal{U}$ of $H_0^1(\Omega)$, we need to find u^h such that

$$u - u^h \perp \mathcal{U}$$

or, using the basis of $\mathcal{U}$,

$$\Big[u - u^h, u_i\Big] = 0, \quad i = 1, \ldots, n.$$

Replacing u^h with its expansion in terms of the basis of $\mathcal{U}$ we have

$$\left[u - \sum_{j=1}^{n} c_j u_j, u_i\right] = 0, \quad i = 1, \ldots, n,$$

or

$$\sum_{j=1}^{n} c_j \, [u_j, u_i] = [u, u_i], \quad i = 1, \ldots, n. \tag{121}$$

The right-hand side of this formula contains the exact solution u, and therefore is not useful for a numerical scheme. However, by (120) and the weak form (117) we have

$$\begin{aligned} [u, u_i] &= \iint\limits_{\Omega} \nabla u \cdot \nabla u_i dxdy \\ &= \iint\limits_{\Omega} f u_i dxdy. \end{aligned}$$

Since the last expression corresponds to the inner product $\langle f, u_i \rangle$, (121) can be viewed as

$$\sum_{j=1}^{n} c_j [u_j, u_i] = \langle f, u_i \rangle, \quad i = 1, \ldots, n,$$

which is nothing but the Ritz-Galerkin method (119).

The best approximation property in the Sobolev space $H_0^1(\Omega)$ can also be interpreted as an energy minimization principle. In fact, a smooth solution of the Poisson problem (113) minimizes the energy functional

$$E(u) = \frac{1}{2} \iint\limits_{\Omega} \nabla^2 u dxdy - \iint\limits_{\Omega} f u dxdy$$

over all smooth functions that vanish on the boundary of Ω. By considering the energy of nearby solutions $u + \lambda v$, with arbitrary real λ we see that

$$\begin{aligned} E(u + \lambda v) &= \frac{1}{2} \iint\limits_{\Omega} \nabla(u + \lambda v) \cdot \nabla(u + \lambda v) dxdy - \iint\limits_{\Omega} f(u + \lambda v) dxdy \\ &= \frac{1}{2} \iint\limits_{\Omega} \nabla u \cdot \nabla u dxdy + \lambda \iint\limits_{\Omega} \nabla u \cdot \nabla v dxdy + \frac{\lambda^2}{2} \iint\limits_{\Omega} \nabla v \cdot \nabla v dxdy \\ &\quad - \iint\limits_{\Omega} f u dxdy - \lambda \iint\limits_{\Omega} f v dxdy \\ &= E(u) + \lambda \iint\limits_{\Omega} \left[\nabla u \cdot \nabla v - f v \right] dxdy + \frac{\lambda^2}{2} \iint\limits_{\Omega} \nabla^2 v dxdy \end{aligned}$$

The right-hand side is a quadratic polynomial in λ, so that for a minimum, the term

$$\iint\limits_{\Omega} \left[\nabla u \cdot \nabla v - f v \right] dxdy$$

must vanish for all v. This is again the weak formulation (117).

A discrete "energy norm" is then given by the quadratic form

$$E(u^h) = \frac{1}{2} c^T A c - b c$$

where A is the stiffness matrix, and c is such that the Ritz-Galerkin system (119)

$$Ac = b$$

is satisfied.