Chapter 4

The Ritz Method and the Galerkin Method

Remark 4.1 *Contents.* This chapter studies variational or weak formulations of boundary value problems of partial differential equations in Hilbert spaces. The existence and uniqueness of an appropriately defined weak solution will be discussed. The approximation of this solution with the help of finite-dimensional spaces is called Ritz method or Galerkin method. Some basic properties of this method will be proved.

In this chapter, a Hilbert space V will be considered with inner product $a(\cdot,\cdot)$: $V \times V \to \mathbb{R}$ and norm $\|v\|_V = a(v,v)^{1/2}$. □

4.1 The Theorems of Riesz and Lax–Milgram

Theorem 4.2 Representation theorem of Riesz. *Let $f \in V'$ be a continuous and linear functional, then there is a uniquely determined $u \in V$ with*

$$a(u,v) = f(v) \quad \forall\, v \in V. \tag{4.1}$$

In addition, u is the unique solution of the variational problem

$$F(v) = \frac{1}{2}a(v,v) - f(v) \to \min \ \forall\, v \in V. \tag{4.2}$$

Proof: First, the existence of a solution u of the variational problem will be proved. Since f is continuous, it holds

$$|f(v)| \le c\,\|v\|_V \quad \forall\, v \in V,$$

from what follows that

$$F(v) \ge \frac{1}{2}\,\|v\|_V^2 - c\,\|v\|_V \ge -\frac{1}{2}c^2,$$

where in the last estimate the necessary criterion for a local minimum of the expression of the first estimate is used. Hence, the function $F(\cdot)$ is bounded from below and

$$d = \inf_{v\in V} F(v)$$

exists.

Let $\{v_k\}_{k\in\mathbb{N}}$ be a sequence with $F(v_k) \to d$ for $k \to \infty$. A straightforward calculation (parallelogram identity in Hilbert spaces) gives

$$\|v_k - v_l\|_V^2 + \|v_k + v_l\|_V^2 = 2\,\|v_k\|_V^2 + 2\,\|v_l\|_V^2\,.$$

Using the linearity of $f(\cdot)$ and $d \le F(v)$ for all $v \in V$, one obtains

$$\begin{aligned}
&\|v_k - v_l\|_V^2 \\
&= 2\|v_k\|_V^2 + 2\|v_l\|_V^2 - 4\left\|\frac{v_k+v_l}{2}\right\|_V^2 - 4f(v_k) - 4f(v_l) + 8f\left(\frac{v_k+v_l}{2}\right) \\
&= 4F(v_k) + 4F(v_l) - 8F\left(\frac{v_k+v_l}{2}\right) \\
&\le 4F(v_k) + 4F(v_l) - 8d \to 0
\end{aligned}$$

for $k, l \to \infty$. Hence $\{v_k\}_{k\in\mathbb{N}}$ is a Cauchy sequence. Because V is a complete space, there exists a limit u of this sequence with $u \in V$. Because $F(\cdot)$ is continuous, it is $F(u) = d$ and u is a solution of the variational problem.

In the next step, it will be shown that each solution of the variational problem (4.2) is also a solution of (4.1). It is

$$\begin{aligned}
\Phi(\varepsilon) &= F(u+\varepsilon v) = \frac{1}{2}a(u+\varepsilon v, u+\varepsilon v) - f(u+\varepsilon v) \\
&= \frac{1}{2}a(u,u) + \varepsilon a(u,v) + \frac{\varepsilon^2}{2}a(v,v) - f(u) - \varepsilon f(v).
\end{aligned}$$

If u is a minimum of the variational problem, then the function $\Phi(\varepsilon)$ has a local minimum at $\varepsilon = 0$. The necessary condition for a local minimum leads to

$$0 = \Phi'(0) = a(u,v) - f(v) \quad \text{for all } v \in V.$$

Finally, the uniqueness of the solution will be proved. It is sufficient to prove the uniqueness of the solution of the equation (4.1). If the solution of (4.1) is unique, then the existence of two solutions of the variational problem (4.2) would be a contradiction to the fact proved in the previous step. Let u_1 and u_2 be two solutions of the equation (4.1). Computing the difference of both equations gives

$$a(u_1 - u_2, v) = 0 \quad \text{for all } v \in V.$$

This equation holds, in particular, for $v = u_1 - u_2$. Hence, $\|u_1 - u_2\|_V = 0$, such that $u_1 = u_2$. ■

Definition 4.3 Bounded bilinear form, coercive bilinear form, V-elliptic bilinear form. Let $b(\cdot,\cdot) : V \times V \to \mathbb{R}$ be a bilinear form on the Banach space V. Then it is bounded if

$$|b(u,v)| \le M \|u\|_V \|v\|_V \quad \forall\, u, v \in V, M > 0, \tag{4.3}$$

where the constant M is independent of u and v. The bilinear form is coercive or V-elliptic if

$$b(u,u) \ge m \|u\|_V^2 \quad \forall\, u \in V, m > 0, \tag{4.4}$$

where the constant m is independent of u. □

Remark 4.4 *Application to an inner product.* Let V be a Hilbert space. Then the inner product $a(\cdot,\cdot)$ is a bounded and coercive bilinear form, since by the Cauchy–Schwarz inequality

$$|a(u,v)| \le \|u\|_V \|v\|_V \quad \forall\, u, v \in V,$$

and obviously $a(u,u) = \|u\|_V^2$. Hence, the constants can be chosen to be $M = 1$ and $m = 1$.

Next, the representation theorem of Riesz will be generalized to the case of coercive and bounded bilinear forms. □

Theorem 4.5 Theorem of Lax–Milgram. *Let $b(\cdot,\cdot) : V \times V \to \mathbb{R}$ be a bounded and coercive bilinear form on the Hilbert space V. Then, for each bounded linear functional $f \in V'$ there is exactly one $u \in V$ with*

$$b(u,v) = f(v) \quad \forall\, v \in V. \tag{4.5}$$

Proof: One defines linear operators $T, T' : V \to V$ by

$$a(Tu, v) = b(u, v) \ \forall\, v \in V, \quad a(T'u, v) = b(v, u) \ \forall\, v \in V. \tag{4.6}$$

Since $b(u, \cdot)$ and $b(\cdot, u)$ are continuous linear functionals on V, it follows from Theorem 4.2 that the elements Tu and $T'u$ exist and they are defined uniquely. Because the operators satisfy the relation

$$a(Tu, v) = b(u, v) = a(T'v, u) = a(u, T'v), \tag{4.7}$$

T' is called adjoint operator of T. Setting $v = Tu$ in (4.6) and using the boundedness of $b(\cdot, \cdot)$ yields

$$\|Tu\|_V^2 = a(Tu, Tu) = b(u, Tu) \le M \|u\|_V \|Tu\|_V \implies \|Tu\|_V \le M \|u\|_V$$

for all $u \in V$. Hence, T is bounded. Since T is linear, it follows that T is continuous. Using the same argument, one shows that T' is also bounded and continuous.

Define the bilinear form

$$d(u, v) := a(TT'u, v) = a(T'u, T'v) \quad \forall\, u, v \in V,$$

where (4.7) was used. Hence, this bilinear form is symmetric. Using the coercivity of $b(\cdot, \cdot)$ and the Cauchy–Schwarz inequality gives

$$m^2 \|v\|_V^4 \le b(v, v)^2 = a(T'v, v)^2 \le \|v\|_V^2 \left\|T'v\right\|_V^2 = \|v\|_V^2 a(T'v, T'v) = \|v\|_V^2 d(v, v).$$

Applying now the boundedness of $a(\cdot, \cdot)$ and of T' yields

$$m^2 \|v\|_V^2 \le d(v, v) = a(T'v, T'v) = \left\|T'v\right\|_V^2 \le M \|v\|_V^2 . \tag{4.8}$$

Hence, $d(\cdot, \cdot)$ is also coercive and, since it is symmetric, it defines an inner product on V. From (4.8) one has that the norm induced by $d(v, v)^{1/2}$ is equivalent to the norm $\|v\|_V$. From Theorem 4.2 it follows that there is a exactly one $w \in V$ with

$$d(w, v) = f(v) \quad \forall\, v \in V.$$

Inserting $u = T'w$ into (4.5) gives with (4.6)

$$b(T'w, v) = a(TT'w, v) = d(w, v) = f(v) \quad \forall\, v \in V,$$

hence $u = T'w$ is a solution of (4.5).

The uniqueness of the solution is proved analogously as in the symmetric case. ■

4.2 Weak Formulation of Boundary Value Problems

Remark 4.6 *Model problem.* Consider the Poisson equation with homogeneous Dirichlet boundary conditions

$$\begin{array}{rcll} -\Delta u & = & f & \text{in } \Omega \subset \mathbb{R}^d, \\ u & = & 0 & \text{on } \partial\Omega. \end{array} \tag{4.9}$$

□

Definition 4.7 Weak formulation of (4.9)**.** Let $f \in L^2(\Omega)$. A weak formulation of (4.9) consists in finding $u \in V = H_0^1(\Omega)$ such that

$$a(u, v) = (f, v) \quad \forall\, v \in V \tag{4.10}$$

with

$$a(u, v) = (\nabla u, \nabla v) = \int_\Omega \nabla u(\mathbf{x}) \cdot \nabla v(\mathbf{x}) \ d\mathbf{x}$$

and $(\cdot, \cdot)$ is the inner product in $L^2(\Omega)$. □

Remark 4.8 *On the weak formulation.*

- The weak formulation is also called variational formulation.
- As usual in mathematics, 'weak' means that something holds for all appropriately chosen test functions.
- Formally, one obtains the weak formulation by multiplying the strong form of the equation (4.9) with the test function, by integrating the equation on Ω, and applying integration by parts. Because of the Dirichlet boundary condition, on can use as test space $H_0^1(\Omega)$ and therefore the integral on the boundary vanishes.
- The ansatz space for the solution and the test space are defined such that the arising integrals are well defined.
- The weak formulation reduces the necessary regularity assumptions for the solution by the integration and the transfer of derivatives to the test function. Whereas the solution of (4.9) has to be in $C^2(\overline{\Omega})$, the solution of (4.10) has to be only in $H_0^1(\Omega)$. The latter assumption is much more realistic for problems coming from applications.
- The regularity assumption on the right hand side can be relaxed to $f \in H^{-1}(\Omega)$.

□

Theorem 4.9 Existence and uniqueness of the weak solution. *Let $f \in L^2(\Omega)$. There is exactly one solution of* (4.10).

Proof: Because of the Poincaré inequality (3.9), there is a constant c with

$$\|v\|_{L^2(\Omega)} \le c\,\|\nabla v\|_{L^2(\Omega)} \quad \forall\, v \in H_0^1(\Omega).$$

It follows for $v \in H_0^1(\Omega) \subset H^1(\Omega)$ that

$$\begin{aligned} \|v\|_{H^1(\Omega)} &= \left(\|v\|_{L^2(\Omega)}^2 + \|\nabla v\|_{L^2(\Omega)}^2\right)^{1/2} \le \left(c\,\|\nabla v\|_{L^2(\Omega)}^2 + \|\nabla v\|_{L^2(\Omega)}^2\right)^{1/2} \\ &\le C\,\|\nabla v\|_{L^2(\Omega)} \le C\,\|v\|_{H^1(\Omega)}\,. \end{aligned}$$

Hence, $a(\cdot,\cdot)$ is an inner product on $H_0^1(\Omega)$ with the induced norm

$$\|v\|_{H_0^1(\Omega)} = a(v,v)^{1/2},$$

which is equivalent to the norm $\|\cdot\|_{H^1(\Omega)}$.

Define for $f \in L^2(\Omega)$ the linear functional

$$\tilde{f}(v) := \int_\Omega f(\mathbf{x})v(\mathbf{x})\; d\mathbf{x} \quad \forall\, v \in H_0^1(\Omega).$$

Applying the Cauchy–Schwarz inequality (3.5) and the Poincaré inequality (3.9)

$$\left|\tilde{f}(v)\right| = |(f,v)| \le \|f\|_{L^2(\Omega)}\,\|v\|_{L^2(\Omega)} \le c\,\|f\|_{L^2(\Omega)}\,\|\nabla v\|_{L^2(\Omega)} = c\,\|f\|_{L^2(\Omega)}\,\|v\|_{H_0^1(\Omega)}$$

shows that this functional is continuous on $H_0^1(\Omega)$. Applying the representation theorem of Riesz, Theorem 4.2, gives the existence and uniqueness of the weak solution of (4.10). In addition, $u(\mathbf{x})$ solves the variational problem

$$F(v) = \frac{1}{2}\,\|\nabla v\|_2^2 - \int_\Omega f(\mathbf{x})v(\mathbf{x})\; d\mathbf{x} \to \min \quad \text{for all } v \in H_0^1(\Omega).$$

■

Example 4.10 *A more general elliptic problem.* Consider the problem

$$\begin{aligned} -\nabla\cdot(A(\mathbf{x})\nabla u) + c(\mathbf{x})u &= f \quad \text{in } \Omega \subset \mathbb{R}^d, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{4.11}$$

with $A(\mathbf{x}) \in \mathbb{R}^{d\times d}$ for each point $\mathbf{x} \in \Omega$. It will be assumed that the coefficients $a_{i,j}(\mathbf{x})$ and $c(\mathbf{x}) \geq 0$ are bounded, $f \in L^2(\Omega)$, and that the matrix (tensor) $A(\mathbf{x})$ is for all $\mathbf{x} \in \Omega$ uniformly elliptic, i.e., there are positive constants m and M such that

$$m \left\|\mathbf{y}\right\|_2^2 \leq \mathbf{y}^T A(\mathbf{x})\mathbf{y} \leq M \left\|\mathbf{y}\right\|_2^2 \quad \forall\, \mathbf{y} \in \mathbb{R}^d,\ \forall\, \mathbf{x} \in \Omega.$$

The weak form of (4.11) is obtained in the usual way by multiplying (4.11) with test functions $v \in H_0^1(\Omega)$, integrating on Ω, and applying integration by parts: Find $u \in H_0^1(\Omega)$, such that

$$a(u,v) = f(v) \quad \forall\, v \in H_0^1(\Omega)$$

with

$$a(u,v) = \int_\Omega \left(\nabla u(\mathbf{x})^T A(\mathbf{x}) \nabla v(\mathbf{x}) + c(\mathbf{x})u(\mathbf{x})v(\mathbf{x})\right)\, d\mathbf{x}.$$

This bilinear form is bounded (*exercise*). The coercivity of the bilinear form is proved by using the uniform ellipticity of $A(\mathbf{x})$ and the non-negativity of $c(\mathbf{x})$:

$$\begin{aligned} a(u,u) &= \int_\Omega \nabla u(\mathbf{x})^T A(\mathbf{x}) \nabla u(\mathbf{x}) + c(\mathbf{x})u(\mathbf{x})u(\mathbf{x})\ d\mathbf{x} \\ &\geq \int_\Omega m \nabla u(\mathbf{x})^T \nabla u(\mathbf{x})\ d\mathbf{x} = m \left\|u\right\|_{H_0^1(\Omega)}^2. \end{aligned}$$

Applying the Theorem of Lax–Milgram, Theorem 4.5, gives the existence and uniqueness of a weak solution of (4.11).

If the tensor is not symmetric, $a_{ij}(\mathbf{x}) \neq a_{ji}(\mathbf{x})$ for one pair i, j, then the solution cannot be characterized as the solution of a variational problem. □

4.3 The Ritz Method and the Galerkin Method

Remark 4.11 *Idea of the Ritz method.* Let V be a Hilbert space with the inner product $a(\cdot,\cdot)$. Consider the problem

$$F(v) = \frac{1}{2}a(v,v) - f(v) \to \min, \tag{4.12}$$

where $f : V \to \mathbb{R}$ is a bounded linear functional. As already proved in Theorem 4.2, there is a unique solution $u \in V$ of this variational problem which is also the unique solution of the equation

$$a(u,v) = f(v) \quad \forall\, v \in V. \tag{4.13}$$

For approximating the solution of (4.12) or (4.13) with a numerical method, it will be assumed that V has a countable orthonormal basis (Schauder basis). Then, there are finite-dimensional subspaces $V_1, V_2, \ldots \subset V$ with $\dim V_k = k$, which has the following property: for each $u \in V$ and each $\varepsilon > 0$ there is a $K \in \mathbb{N}$ and a $u_k \in V_k$ with

$$\left\|u - u_k\right\|_V \leq \varepsilon \quad \forall\, k \geq K. \tag{4.14}$$

Note that it is not required that there holds an inclusion of the form $V_k \subset V_{k+1}$.

The Ritz approximation of (4.12) and (4.13) is defined by: Find $u_k \in V_k$ with

$$a(u_k, v_k) = f(v_k) \quad \forall\, v_k \in V_k. \tag{4.15}$$

□

Lemma 4.12 Existence and uniqueness of a solution of (4.15)**.** *There exists exactly one solution of* (4.15).

Proof: Finite-dimensional subspaces of Hilbert spaces are Hilbert spaces as well. For this reason, one can apply the representation theorem of Riesz, Theorem 4.2, to (4.15) which gives the statement of the lemma. In addition, the solution of (4.15) solves a minimization problem on V_k. ■

Lemma 4.13 Best approximation property. *The solution of* (4.15) *is the best approximation of u in V_k, i.e., it is*

$$\|u - u_k\|_V = \inf_{v_k \in V_k} \|u - v_k\|_V . \tag{4.16}$$

Proof: Since $V_k \subset V$, one can use the test functions from V_k in the weak equation (4.13). Then, the difference of (4.13) and (4.15) gives the orthogonality, the so-called Galerkin orthogonality,

$$a(u - u_k, v_k) = 0 \quad \forall\, v_k \in V_k. \tag{4.17}$$

Hence, the error $u - u_k$ is orthogonal to the space V_k: $u - u_k \perp V_k$. That means, u_k is the orthogonal projection of u onto V_k with respect of the inner product of V.

Let now $w_k \in V_k$ be an arbitrary element, then it follows with the Galerkin orthogonality (4.17) and the Cauchy–Schwarz inequality that

$$\begin{aligned} \|u - u_k\|_V^2 &= a(u - u_k, u - u_k) = a(u - u_k, u - \underbrace{(u_k - w_k)}_{v_k}) = a(u - u_k, u - v_k) \\ &\le \|u - u_k\|_V \|u - v_k\|_V . \end{aligned}$$

Since $w_k \in V_k$ was arbitrary, also $v_k \in V_k$ is arbitrary. If $\|u - u_k\|_V > 0$, division by $\|u - u_k\|_V$ gives the statement of the lemma. If $\|u - u_k\|_V = 0$, the statement of the lemma is trivially true. ■

Theorem 4.14 Convergence of the Ritz approximation. *The Ritz approximation converges*

$$\lim_{k\to\infty} \|u - u_k\|_V = 0.$$

Proof: The best approximation property (4.16) and property (4.14) give

$$\|u - u_k\|_V = \inf_{v_k \in V_k} \|u - v_k\|_V \le \varepsilon$$

for each $\varepsilon > 0$ and $k \ge K(\varepsilon)$. Hence, the convergence is proved. ■

Remark 4.15 *Formulation of the Ritz method as linear system of equations.* One can use an arbitrary basis $\{\phi_i\}_{i=1}^k$ of V_k for the computation of u_k. First of all, the equation for the Ritz approximation (4.15) is satisfied for all $v_k \in V_k$ if and only if it is satisfied for each basis function ϕ_i. This statement follows from the linearity of both sides of the equation with respect to the test function and from the fact that each function $v_k \in V_k$ can be represented as linear combination of the basis functions. Let $v_k = \sum_{i=i}^k \alpha_i \phi_i$, then from (4.15) it follows that

$$a(u_k, v_k) = \sum_{k=1}^k \alpha_i a(u_k, \phi_i) = \sum_{k=1}^k \alpha_i f(\phi_i) = f(v_k).$$

This equation is satisfied if $a(u_k, \phi_i) = f(\phi_i)$, $i = 1, \ldots, k$. On the other hand, if (4.15) holds then it holds in particular for each basis function ϕ_i.

Then, one uses as ansatz for the solution also a linear combination of the basis functions

$$u_k = \sum_{j=1}^k u^j \phi_j$$

with unknown coefficients $u^j \in \mathbb{R}$. Using as test functions now the basis functions yields

$$\sum_{j=1}^{k} a(u^j\phi_j, \phi_i) = \sum_{j=1}^{k} a(\phi_j, \phi_i)u^j = f(\phi_i), \quad i = 1, \ldots, k.$$

This equation is equivalent to the linear system of equations $A\mathbf{u} = \mathbf{f}$, where

$$A = (a_{ij})_{i,j=1}^{k} = a(\phi_j, \phi_i)_{i,j=1}^{k}$$

is called stiffness matrix. Note that the order of the indices is different for the entries of the matrix and the arguments of the inner product. The right hand side is a vector of length k with the entries $f_i = f(\phi_i)$, $i = 1, \ldots, k$.

Using the one-to-one mapping between the coefficient vector $(v^1, \ldots, v^k)^T$ and the element $v_k = \sum_{i=1}^{k} v^i\phi_i$, one can show that the matrix A is symmetric and positive definite (*exercise*)

$$\begin{aligned} A = A^T \quad &\iff \quad a(v, w) = a(w, v) \quad \forall\, v, w \in V_k, \\ \mathbf{x}^T A\mathbf{x} > 0 \text{ for } \mathbf{x} \neq \mathbf{0} \quad &\iff \quad a(v, v) > 0 \quad \forall\, v \in V_k, v \neq 0. \end{aligned}$$

□

Remark 4.16 *The case of a bounded and coercive bilinear form.* If $b(\cdot, \cdot)$ is bounded and coercive, but not symmetric, it is possible to approximate the solution of (4.5) with the same idea as for the Ritz method. In this case, it is called Galerkin method. The discrete problem consists in finding $u_k \in V_k$ such that

$$b(u_k, v_k) = f(v_k) \quad \forall\, v_k \in V_k. \tag{4.18}$$

□

Lemma 4.17 Existence and uniqueness of a solution of (4.18)**.** *There is exactly one solution of* (4.18)*.*

Proof: The statement of the lemma follows directly from the Theorem of Lax–Milgram, Theorem 4.5. ■

Remark 4.18 *On the discrete solution.* The discrete solution is not the orthogonal projection into V_k in the case of a bounded and coercive bilinear form, which is not the inner product of V. □

Lemma 4.19 Lemma of Cea, error estimate. *Let $b: V \times V \to \mathbb{R}$ be a bounded and coercive bilinear form on the Hilbert space V and let $f \in V'$ be a bounded linear functional. Let u be the solution of* (4.5) *and u_k be the solution of* (4.18)*, then the following error estimate holds*

$$\|u - u_k\|_V \leq \frac{M}{m} \inf_{v_k \in V_k} \|u - v_k\|_V\,, \tag{4.19}$$

where the constants M and m are given in (4.3) *and* (4.4)*.*

Proof: Considering the difference of the continuous equation (4.5) and the discrete equation (4.18), one obtains the error equation

$$b(u - u_k, v_k) = 0 \quad \forall\, v_k \in V_k,$$

which is also called Galerkin orthogonality. With (4.4), the Galerkin orthogonality, and (4.3) it follows that

$$\begin{aligned} \|u - u_k\|_V^2 \quad &\leq \quad \frac{1}{m} b(u - u_k, u - u_k) = \frac{1}{m} b(u - u_k, u - v_k) \\ &\leq \quad \frac{M}{m} \|u - u_k\|_V \|u - v_k\|_V\,, \quad \forall\, v_k \in V_k, \end{aligned}$$

from what the statement of the lemma follows immediately. ■

Remark 4.20 *On the best approximation error.* It follows from estimate (4.19) that the error is bounded by a multiple of the best approximation error, where the factor depends on properties of the bilinear form $b(\cdot, \cdot)$. Thus, concerning error estimates for concrete finite-dimensional spaces, the study of the best approximation error will be of importance. $\Box$

Remark 4.21 *The corresponding linear system of equations.* The corresponding linear system of equations is derived analogously to the symmetric case. The system matrix is still positive definite but not symmetric. $\Box$

Remark 4.22 *Choice of the basis.* The most important issue of the Ritz and Galerkin method is the choice of the spaces V_k, or more concretely, the choice of an appropriate basis $\{\phi_i\}_{i=1}^k$ that spans the space V_k. From the point of view of numerics, there are the requirements that it should be possible to compute the entries a_{ij} of the stiffness matrix efficiently and that the matrix A should be sparse. $\Box$