Numerics of Partial Differential Equations: Stationary Problems

Lecture Notes

Michael Feischl and Dirk Praetorius

October 1, 2020

Chapter 1

Introduction

1.1 Strong Form and Variational Form

The finite element method is a scheme for the numerical solution of partial differential equations. In this chapter, we introduce the basic concepts for elliptic problems in the frame of the Riesz theorem. To that end, we consider the most standard example, namely the Poisson equation with mixed Dirichlet-Neumann boundary conditions. We aim to solve

$$\begin{aligned} -\Delta u &= f \qquad \text{in } \Omega, \\ u &= 0 \quad \text{on } \Gamma_D, \\ \partial u / \partial n &= \phi \quad \text{on } \Gamma_N, \end{aligned} \tag{1.1}$$

which is said to be the **strong form** of the boundary value problem. Here, Ω denotes a domain in $\mathbb{R}^d$, $d = 2, 3$. The boundary $\Gamma := \partial\Omega$ is split into the Dirichlet boundary Γ_D and the Neumann boundary Γ_N, respectively. To be more precise, we assume that Γ_D and Γ_N are (relatively) open subsets of Γ with $\Gamma_D \cap \Gamma_N = \emptyset$ and $\Gamma = \overline{\Gamma}_D \cup \overline{\Gamma}_N$. The source term $f : \Omega \to \mathbb{R}$ as well as the Neumann data $\phi : \Gamma_N \to \mathbb{R}$ are given, and $u : \Omega \to \mathbb{R}$ is the unknown solution. Moreover,

$$\Delta u(x) := \sum_{j=1}^{d} \frac{\partial^2 u}{\partial x_j^2}(x) \tag{1.2}$$

denotes the Laplace operator, which is defined in the classical sense for a function $u \in C^2(\overline{\Omega})$, where $C^k(\overline{\Omega}) := \big\{ w|_{\overline{\Omega}} \, \big| \, w \in C^k(\mathbb{R}^d) \big\}$. If $u \in C^2(\overline{\Omega})$ solves (1.1), u is said to be a **strong solution** of the mixed boundary value problem.

Throughout the lecture, we shall assume that Ω is a **Lipschitz domain** in $\mathbb{R}^d$, i.e.,

- Ω is a bounded, open, and connected subset of $\mathbb{R}^d$,
- Ω is locally on one side of Γ,
- Γ can locally be parametrized by Lipschitz continuous functions.

An important consequence of this assumption is the validity of the **integration by parts formula**

$$\int_\Omega \frac{\partial u}{\partial x_j} \, v \, dx + \int_\Omega u \, \frac{\partial v}{\partial x_j} \, dx = \int_\Gamma u v n_j \, ds \quad \text{for all } u, v \in C^1(\overline{\Omega}), \tag{1.3}$$

where n_j denotes the j-th component of the outer normal vector of Ω on Γ and where ds denotes the surface measure on Γ. For a precise definition and details, we refer, e.g., to [McL].

Let $u \in C^2(\overline{\Omega})$ be a strong solution of (1.1) and $v \in C^1_D(\overline{\Omega}) := \big\{w \in C^1(\overline{\Omega}) \,\big|\, w|_{\Gamma_D} = 0\big\}$. Multiplication of $-\Delta u = f$ by v, integration over Ω, and integration by parts yield that

$$\int_\Omega fv\,dx = -\int_\Omega (\Delta u)v\,dx = -\sum_{j=1}^d \int_\Omega \frac{\partial^2 u}{\partial x_j^2}\,v\,dx = \sum_{j=1}^d \Big[\int_\Omega \frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_j}\,dx - \int_\Gamma \frac{\partial u}{\partial x_j}\,vn_j\,ds\Big].$$

With $x \cdot y = \sum_{j=1}^d x_j y_j$ the usual scalar product in $\mathbb{R}^d$, we obtain the **first Green formula**

$$\int_\Omega fv\,dx = \int_\Omega \nabla u \cdot \nabla v\,dx - \int_\Gamma \frac{\partial u}{\partial n}\,v\,ds, \tag{1.4}$$

where we have used $\nabla u \cdot n = \partial u/\partial n$. Together with $v|_{\Gamma_D} = 0$ and $\Gamma_N = \Gamma\backslash\overline{\Gamma}_D$, we may plug-in the Neumann data to see that

$$\int_\Omega fv\,dx = \int_\Omega \nabla u \cdot \nabla v\,dx - \int_{\Gamma_N} \frac{\partial u}{\partial n}\,v\,ds = \int_\Omega \nabla u \cdot \nabla v\,dx - \int_{\Gamma_N} \phi v\,ds.$$

Altogether we thus have proven the following proposition:

Proposition 1.1. *Let $u \in C^2(\overline{\Omega})$ solve the strong form* (1.1)*. Then, it holds that*

$$\int_\Omega \nabla u \cdot \nabla v\,dx = \int_\Omega fv\,dx + \int_{\Gamma_N} \phi v\,ds \quad \textit{for all } v \in C^1_D(\overline{\Omega}), \tag{1.5}$$

which is the ***variational form*** *of the boundary value problem* (1.1). ■

This proposition gives a necessary condition for a function u to solve the strong form (1.1). We stress that any strong solution belongs to $C^1_D(\overline{\Omega})$ and that the variational form (1.5) can be understood for $u \in C^1_D(\overline{\Omega})$. This leads to a symmetric variational formulation: Find $u \in C^1_D(\overline{\Omega})$ such that (1.5) holds.

Exercise 1. Prove the following well-known integral formulae:

- For $f \in C^1(\Omega)^d$, let $\operatorname{div} f := \sum_{j=1}^d \frac{\partial f_j}{\partial x_j}$ denote the divergence operators. Then, there holds the **Gauss divergence theorem**

$$\int_\Omega \operatorname{div} f\,dx = \int_\Gamma f \cdot n\,ds \quad \text{for all } f \in C^1(\overline{\Omega})^d. \tag{1.6}$$

- Besides the first Green formula, there holds the **second Green formula**

$$\int_\Omega (-\Delta u)v\,dx + \int_\Gamma \frac{\partial u}{\partial n}\,v\,ds = \int_\Omega u(-\Delta v)\,dx + \int_\Gamma u\,\frac{\partial v}{\partial n}\,ds \quad \text{for all } u, v \in C^2(\overline{\Omega}). \tag{1.7}$$

Both are easily obtained from the integration by parts formula. □

1.2 Solvability of Variational Form

To look for solutions of the weak form (1.5), we will employ the following Riesz theorem.

Theorem 1.2 (Riesz). *For a Hilbert space H (over $\mathbb{R}$), the mapping*

$$I_H : H \to H^*, \quad I_H(u) := (u\,;\,\cdot)_H \tag{1.8}$$

is linear, isometric, and bijective, i.e., for any $F \in H^*$ there is a unique $u \in H$ such that

$$(u\,;\,v)_H = F(v) \quad \text{for all } v \in H. \tag{1.9}$$

Moreover, it holds that $\|u\|_H = \|F\|_{H^*}$. ■

The proof of this theorem can be found in each textbook of functional analysis.

First, we observe that the left-hand side

$$(u\,;\,v) := \int_\Omega \nabla u \cdot \nabla v \, dx$$

of the variational form (1.5) defines a scalar product on $C^1_D(\overline{\Omega})$, provided the Dirichlet boundary Γ_D is nontrivial: Clearly, $(u\,;\,v)$ is a symmetric bilinear form on $C^1_D(\overline{\Omega})$. It thus only remains to prove definiteness. Note that $0 = (u\,;\,u) = \|\nabla u\|^2_{L^2(\Omega)}$ implies $\nabla u = 0$, whence u is constant in Ω. Together with $u|_{\Gamma_D} = 0$, this proves $u = 0$. Moreover, the right-hand side

$$F(v) := \int_\Omega f v \, dx + \int_{\Gamma_N} \phi v \, ds$$

defines a linear functional on $C^1_D(\overline{\Omega})$ which is continuous with respect to the induced norm $\|v\| := (v\,;\,v)^{1/2}$. We prove this claim only in the special situation $\Gamma = \Gamma_D$ and postpone the abstract proof to a subsequent section.

Lemma 1.3 (Friedrichs' inequality). *Suppose that $\Omega = [a,b] \times [c,d] \subset \mathbb{R}^2$ and $\Gamma_D = \partial\Omega$. Then, it holds that $\|v\|_{L^2(\Omega)} \le \operatorname{diam}(\Omega)\, \|\nabla v\|_{L^2(\Omega)}$ for all $v \in C^1_D(\overline{\Omega})$.*

Proof. For $x = (x_1, x_2) \in \Omega$, it holds that $v(x_1, c) = 0$. Therefore, the fundamental theorem of calculus yields that

$$v(x) = \int_c^{x_2} \partial_2 v(x_1, t)\, dt.$$

The Hölder inequality yields that

$$|v(x)| \le |d-c|^{1/2} \Big(\int_c^{x_2} |\partial_2 v(x_1,t)|^2 \, dt \Big)^{1/2}.$$

Integration over Ω gives

$$\begin{aligned}\|v\|_{L^2(\Omega)}^2 = \int_\Omega |v(x)|^2\,dx \le |d-c| \int_\Omega \int_c^{x_2} |\partial_2 v(x_1,t)|^2\,dt\,dx \\ = |d-c| \int_c^d \int_a^b \int_c^{x_2} |\partial_2 v(x_1,t)|^2\,dt\,dx_1\,dx_2 \\ \le |d-c| \int_c^d \|\partial_2 v\|_{L^2(\Omega)}^2\,dx_2 \\ = |d-c|^2 \|\partial_2 v\|_{L^2(\Omega)}^2.\end{aligned}$$

This results in $\|v\|_{L^2(\Omega)} \le |d-c|\,\|\partial_2 v\|_{L^2(\Omega)} \le \mathrm{diam}(\Omega)\,\|\nabla v\|_{L^2(\Omega)}$. ■

According to the Hölder and the Friedrichs inequality, we obtain that

$$|F(v)| \le \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} \le \mathrm{diam}(\Omega)\|f\|_{L^2(\Omega)}\|\nabla v\|_{L^2(\Omega)} = \mathrm{diam}(\Omega)\|f\|_{L^2(\Omega)}\|v\|.$$

Therefore, the linear functional F is continuous with respect to $\|\cdot\| := \|\nabla(\cdot)\|_{L^2(\Omega)}$ with operator norm $\|F\|_* \le \mathrm{diam}(\Omega)\|f\|_{L^2(\Omega)}$. If $C_D^1(\overline{\Omega})$ associated with the norm $\|\cdot\|$ *were* a Hilbert space, the Riesz theorem *would* therefore imply the unique solvability of the variational form (1.5). However, $C_D^1(\overline{\Omega})$ is *not* complete and therefore the Riesz theorem does *not* apply.

The remedy is to consider the (unique) completion of $C_D^1(\overline{\Omega})$ with respect to $\|\cdot\|$. This leads to a so-called **Sobolev space** $H_D^1(\Omega)$, which is —by definition— complete and hence a Hilbert space. Density arguments then lead to an extended variational form: Find $u \in H_D^1(\Omega)$ such that

$$\int_\Omega \nabla u \cdot \nabla v\,dx = \int_\Omega fv\,dx + \int_{\Gamma_N} \phi v\,ds \quad \text{for all } v \in H_D^1(\Omega), \tag{1.10}$$

which is the **weak form** of the boundary value problem (1.1). Now, the Riesz theorem applies and proves the unique existence of a **weak solution** $u \in H_D^1(\Omega)$ of (1.10). Later on, we are going to show that

- each strong solution $u \in C^2(\overline{\Omega})$ of (1.1) belongs to $H_D^1(\Omega)$ and is also the unique weak solution of (1.10).
- provided the weak solution $u \in H_D^1(\Omega)$ is smooth, i.e., $u \in C^2(\overline{\Omega})$, the weak solution also solves the strong form (1.1).

In this sense, the strong form (1.1) and the weak form (1.10) are equivalent.

1.3 Finite Element Method

The finite element method for (1.10) essentially consists of replacing the (infinite dimensional) Sobolev space $H_D^1(\Omega)$ by a finite dimensional subspace $X_h \subset H_D^1(\Omega)$: Find $u_h \in X_h$ such that

$$\int_\Omega \nabla u_h \cdot \nabla v_h\,dx = \int_\Omega fv_h\,dx + \int_{\Gamma_N} \phi v_h\,ds \quad \text{for all } v_h \in X_h. \tag{1.11}$$

This problem is equivalent to the solution of a system of linear equations $\mathbf{A}\mathbf{x} = \mathbf{b}$, where the system matrix $\mathbf{A}$ is symmetric and positive definite. Of course, the question of convergence depends on the choice of X_h. Thus, there remain some topics for mathematical discussions later on.

The finite element method is a special **Galerkin scheme**. In this section, we collect the most simple properties of Galerkin schemes. Throughout, H is a (real) Hilbert space, and $\langle\!\langle \cdot\,;\cdot \rangle\!\rangle$ is an equivalent scalar product on H, i.e., there are constants $\alpha, \beta > 0$ such that

$$\alpha \|v\|_H \le |\!|\!| v |\!|\!| \le \beta \|v\|_H \quad \text{for all } v \in H, \tag{1.12}$$

where $|\!|\!| v |\!|\!| := \langle\!\langle v\,;v \rangle\!\rangle^{1/2}$ denotes the induced norm. We stress that $\langle\!\langle \cdot\,;\cdot \rangle\!\rangle$ and $|\!|\!| \cdot |\!|\!|$ are often called **energy scalar product** and **energy norm**, respectively (see also Exercise 5).

Remark. In the following, we state all results with respect to the norm $\|\cdot\|_H$, which involves the constants $\alpha, \beta > 0$. Analogously, one may state the results with respect to the energy norm $|\!|\!| \cdot |\!|\!| = \|\cdot\|_H$, which corresponds to $\alpha = \beta = 1$. □

For given $F \in H^*$, the Riesz theorem proves the existence and uniqueness of a solution $u \in H$ of

$$\langle\!\langle u\,;v \rangle\!\rangle = F(v) \quad \text{for all } v \in H, \tag{1.13}$$

for what we use the short-hand notation

$$\langle\!\langle u\,;\cdot \rangle\!\rangle = F \in H^* \tag{1.14}$$

to implicitly indicate that this equation holds (pointwise) for all $v \in H$. Now, the Galerkin method simply consists in replacing the continuous space H by some finite dimensional subspace: Let X_h be a finite-dimensional (and hence closed) subspace of H. Since the Riesz theorem applies to the Hilbert space X_h as well, there is a unique **Galerkin solution** $u_h := \mathbb{G}_h u \in X_h$ such that

$$\langle\!\langle \mathbb{G}_h u\,;\cdot \rangle\!\rangle = F \in X_h^*. \tag{1.15}$$

For $u \in H$ and the corresponding functional $\langle\!\langle u\,;\cdot \rangle\!\rangle \in H^*$, this defines the **Galerkin projection**

$$\mathbb{G}_h : H \to X_h \quad \text{where } \mathbb{G}_h u \in X_h \text{ solves} \quad \langle\!\langle \mathbb{G}_h u\,;\cdot \rangle\!\rangle = \langle\!\langle u\,;\cdot \rangle\!\rangle \in X_h^*. \tag{1.16}$$

Note that $\mathbb{G}_h u \in X_h$ is characterized by the **Galerkin orthogonality**

$$\langle\!\langle u - \mathbb{G}_h u\,;v_h \rangle\!\rangle = 0 \quad \text{for all } v_h \in X_h. \tag{1.17}$$

Before we proceed with the theoretical analysis of Galerkin schemes, we treat an implementational issue. The following theorem is the fundamental observation: Usually, only the scalar product $\langle\!\langle \cdot\,;\cdot \rangle\!\rangle$ and the right-hand side $F \in H^*$ are known, while the exact solution $u \in H$ of (1.13) is unknown. Then, the Galerkin solution $\mathbb{G}_h u \in X_h$ can be computed by solving a linear system of equations — without knowledge of u.

Theorem 1.4. *Let $\{\phi_1, \dots, \phi_N\}$ be a basis of X_h. We define the Galerkin matrix $A \in \mathbb{R}^{N\times N}$ and the vector $b \in \mathbb{R}^N$ by*

$$A_{jk} := \langle\!\langle \phi_k\,;\phi_j \rangle\!\rangle \quad \textit{and} \quad b_j := F(\phi_j). \tag{1.18}$$

Then, A is symmetric and positive definite and, in particular, a regular matrix. Moreover, there holds $\mathbb{G}_h u = \sum_{j=1}^N x_j \phi_j$, where the vector $x \in \mathbb{R}^N$ solves $Ax = b$.

Proof. **1. step.** Symmetry of A clearly follows from the symmetry of $\langle\!\langle \cdot\,;\cdot \rangle\!\rangle$.

2. step. For any $x \in \mathbb{R}^N$ and $v_h := \sum_{j=1}^N x_j \phi_j$, it holds that

$$|||v_h|||^2 = \langle\!\langle v_h\,; v_h \rangle\!\rangle = \sum_{j,k=1}^N x_j x_k \langle\!\langle \phi_j\,; \phi_k \rangle\!\rangle = x \cdot Ax.$$

This proves $Ax \cdot x > 0$ for all $x \neq 0$. By definition, A is positive definite and hence regular.

3. step. Determine Galerkin solution: Let $x \in \mathbb{R}^n$ be the unique solution of the linear Galerkin system $Ax = b$. We use the basis representation $\mathbb{G}_h u = \sum_{j=1}^N y_j \phi_j$ of the Galerkin solution with some coefficient vector $y \in \mathbb{R}^n$. By use of the linearity of $\langle\!\langle \cdot\,;\cdot \rangle\!\rangle$, equation (1.15) becomes

$$b_k = F(\phi_k) = \langle\!\langle \mathbb{G}_h u\,; \phi_k \rangle\!\rangle = \sum_{j=1}^N y_j \langle\!\langle \phi_j\,; \phi_k \rangle\!\rangle = (Ay)_k \quad \text{for all } k = 1, \dots, N.$$

Therefore, the coefficient vector $y \in \mathbb{R}^N$ satisfies $Ay = b$. This proves $x = y$, i.e., we obtain $\mathbb{G}_h u$ by solving $Ax = b$. ■

Remark. We just remark that Theorem 1.4 can be applied for *any* orthogonal-type projection, e.g., the L^2-orthogonal projection onto a discrete space. □

We now proceed with the abstract analysis of Galerkin schemes. The following two lemmata provide elementary properties of the Galerkin projection. The first lemma proves stability of the method with respect to changes of the right-hand side F.

Lemma 1.5. *The Galerkin projection $\mathbb{G}_h$ is a linear and continuous projection onto X_h with*

$$\|\mathbb{G}_h u\|_H \leq \frac{\beta}{\alpha}\, \|u\|_H \quad \textit{for all } u \in H, \tag{1.19}$$

where $\alpha, \beta > 0$ are the norm equivalence constants from (1.12). *Moreover, $\mathbb{G}_h$ is the orthogonal projection onto X_h with respect to the energy scalar product $\langle\!\langle \cdot\,;\cdot \rangle\!\rangle$.*

Proof. For $u_h \in X_h$, the Galerkin orthogonality (1.17) implies $\mathbb{G}_h u_h = u_h$. Therefore $\mathbb{G}_h$ is a projection onto X_h. Also the linearity of $\mathbb{G}_h$ follows from the Galerkin orthogonality (1.17). To see the continuity of $\mathbb{G}_h$, it remains to estimate the operator norm: For $u \in H$ holds

$$|||\mathbb{G}_h u|||^2 = \langle\!\langle \mathbb{G}_h u\,; \mathbb{G}_h u \rangle\!\rangle = \langle\!\langle u\,; \mathbb{G}_h u \rangle\!\rangle \leq |||u|||\,|||\mathbb{G}_h u|||,$$

whence $|||\mathbb{G}_h u||| \leq |||u|||$ and

$$\alpha \|\mathbb{G}_h u\|_H \leq |||\mathbb{G}_h u||| \leq |||u||| \leq \beta \|u\|_H,$$

where we have used the norm equivalence (1.12) on H as well as the Cauchy inequality for the scalar product $\langle\!\langle \cdot\,;\cdot \rangle\!\rangle$. This proves that $\|\mathbb{G}_h u\|_H \leq (\alpha/\beta)\|u\|_H$ and thus continuity of $\mathbb{G}_h$. Finally, we remark that the *unique* orthogonal projection with respect to $\langle\!\langle \cdot\,;\,\cdot \rangle\!\rangle$, is characterized by the orthogonality relation (1.17). ■

The following Céa lemma states that the **Galerkin error** $\|u - \mathbb{G}_h u\|_H$ is quasi-optimal, i.e., it behaves like the best approximation error up to multiplicative constants, which depend only on the continuous setting but not on X_h.

Lemma 1.6 (Céa). *The Galerkin error is quasi-optimal, i.e.,*

$$\|u - \mathbb{G}_h u\|_H \le \frac{\beta}{\alpha} \min_{v_h \in X_h} \|u - v_h\|_H \quad \textit{for all } u \in H, \tag{1.20}$$

where $\alpha, \beta > 0$ are the norm equivalence constants from (1.12)*. With respect to the energy norm, it holds that*

$$|\!|\!| u - \mathbb{G}_h u |\!|\!| = \min_{v_h \in X_h} |\!|\!| u - v_h |\!|\!| \quad \textit{for all } u \in H, \tag{1.21}$$

i.e., the Galerkin solution $\mathbb{G}_h u$ is the best approximation of u with respect to the energy norm.

Proof. For arbitrary $v_h \in X_h$, the Galerkin orthogonality (1.17) proves that

$$|\!|\!| u - \mathbb{G}_h u |\!|\!|^2 = \langle\!\langle u - \mathbb{G}_h u\,;\, u - v_h \rangle\!\rangle \le |\!|\!| u - \mathbb{G}_h u |\!|\!| \, |\!|\!| u - v_h |\!|\!|,$$

which yields (1.21) with an infimum on the right-hand side. Of course, the minimum in (1.21) is attained for $v_h = \mathbb{G}_h u$. With the same arguments as in the proof of the last lemma, we even see that

$$\alpha \|u - \mathbb{G}_h u\|_H \le |\!|\!| u - \mathbb{G}_h u |\!|\!| \le |\!|\!| u - v_h |\!|\!| \le \beta \|u - v_h\|_H,$$

which implies (1.20) with an infimum on the right-hand side. This minimum is attained for $v_h = \Pi_h u$ with $\Pi_h : X \to X_h$ being the orthogonal projection onto X_h with respect to $\|\cdot\|_H$. ■

Exercise 2. Let X be a normed vector space over $\mathbb{R}$ and $X_h \subseteq X$ be a finite dimensional subspace of X. Then, for any $x \in X$, there exists some (not necessarily unique) $x_h \in X_h$ such that

$$\|x - x_h\|_X = \min_{v_h \in X_h} \|x - v_h\|_X,$$

i.e., best approximation errors on finite dimensional spaces as in (1.20) are always attained. Prove that the set of minimizers is convex, closed and bounded (and hence even compact). □

A major advantage of Galerkin methods is that one can prove convergence for any exact solution $u \in H$ if one knows that smooth functions can be approximated well. In the following, think of the subscript $h > 0$ as a mesh-size parameter with corresponding finite dimensional spaces X_h:

Proposition 1.7. *For all $h > 0$, let X_h be a finite-dimensional subspace of H. We assume that there is a dense subspace D of H with approximation property, namely*

$$\lim_{h\to 0} \min_{v_h \in X_h} \|v - v_h\|_H = 0 \quad \textit{for all } v \in D. \tag{1.22}$$

Then, for any $u \in H$, it holds that

$$\lim_{h\to 0} \|u - \mathbb{G}_h u\|_H = 0, \tag{1.23}$$

i.e., the sequence of Galerkin solutions converges to the exact solution u.

Proof. For $v \in D$, the quasi-optimality (1.20) yields that

$$\|u - \mathbb{G}_h u\|_H \le \frac{\beta}{\alpha} \min_{v_h \in X_h} \|u - v_h\|_H \le \frac{\beta}{\alpha} \big(\|u - v\|_H + \min_{v_h \in X_h} \|v - v_h\|_H\big).$$

We have to show that

$$\exists C > 0 \, \forall \varepsilon > 0 \, \exists h_0 > 0 \, \forall h \in (0, h_0) \quad \|u - \mathbb{G}_h u\|_H \le C\,\varepsilon.$$

For $\varepsilon > 0$, let $v \in D$ with $\|u - v\|_H \le \varepsilon$. Choose $h_0 > 0$ according to the approximation assumption (1.22) so that $\min_{v_h \in X_h} \|v - v_h\|_H \le \varepsilon$ for all $h \in (0, h_0)$. We thus finally obtain $\|u - \mathbb{G}_h u\|_H \le 2\beta\varepsilon/\alpha$, which concludes the proof. ■

Although the result of the preceding lemma seems to be very attractive, we stress, however, that the convergence of a Galerkin scheme can be arbitrarily slow. We argue in the abstract setting: If H is a separable Hilbert space, e.g., H is a Sobolev space, there is a countable orthonormal basis $\big\{\phi_j \,\big|\, j \in \mathbb{N}\big\}$. Any $u \in H$ can be written as $u = \sum_{j=1}^{\infty} x_j \phi_j$ with coefficients $(x_n) \in \ell_2$. If we define $X_j := \operatorname{span}\{\phi_1, \dots, \phi_j\}$, it holds that

$$\min_{v_h \in X_h} \|u - v_h\|_H^2 = \sum_{j=k+1}^{\infty} x_j^2.$$

Finally, the decay of the right-hand side can be very slow. One may think of, e.g., $x_j^2 = j^{-(1+\varepsilon)}$ for any $\varepsilon > 0$, so that the series converges but is — in the beginning — almost the divergent harmonic series.

The following exercise shows that the approximation property (1.22) in particular implies that the Hilbert space H has to be separable.

Exercise 3. Suppose that X is a normed space with finite dimensional subspaces $X_\ell \subseteq X_{\ell+1} \subseteq X$ for all $\ell \in \mathbb{N}$. Suppose that $\mathcal{D} \subseteq X$ is a dense subspace such that, for all $x \in X$,

$$\lim_{\ell\to\infty} \min_{x_\ell \in X_\ell} \|x - x_\ell\|_X = 0. \tag{1.24}$$

Then, X is separable, i.e., there is a countable and dense subset $M \subseteq X$. □

Exercise 4. Let $X = \ell_\infty$ and $X_\ell := \big\{(x_n) \in \ell_\infty \,\big|\, x_j = 0 \text{ for all } j \geq \ell\big\}$. Prove that (1.24) fails to hold for any dense subspace $\mathcal{D}$. Note that this also follows if one proves that ℓ_∞ is not separable. □

Remark. All foregoing results of this section hold (in a slightly modified form) in case that $⟪\cdot\,;\cdot⟫$ only is a continuous and elliptic bilinear form on the Hilbert space H, i.e., in all proofs, one can avoid to use the symmetry of $⟪\cdot\,;\cdot⟫$. □

The following exercise explains why $|||\cdot|||$ is called energy norm. In many situations, the function $J(\cdot)$ has the interpretation of a physical energy.

Exercise 5. Let $⟪\cdot\,;\cdot⟫$ be a scalar product on the Hilbert space H such that the norm $|||\cdot|||$ is equivalent to $\|\cdot\|_H$. Let $F \in H^*$ and $u \in H$. Then, the following assertions are equivalent:

- $⟪u\,;\cdot⟫ = F \in H^*$;
- $J(u) = \min\limits_{v\in H} J(v)$, where $J(v) := \frac{1}{2}\,⟪v\,;v⟫ - F(v)$.

In particular, the variational formulation is equivalent to energy minimization, and this result also covers the discrete setting. Derive a formula for the energy error $J(\mathbb{G}_h u) - J(u)$, where $\mathbb{G}_h : H \to X_h$ denotes the Galerkin projection. □

Finally, we comment on an extension of the concept of Galerkin schemes to some nonlinear problems. We note that this framework does, in particular, cover the frame of the Lax–Milgram lemma.

Exercise 6 (Main Theorem on Strongly Monotone Operators (Zarantonello '60)). Let H be a Hilbert space and $A : H \to H^*$ be a Lipschitz continuous and strongly monotone operator, i.e.,

$$\|Au - Av\|_{H^*} \leq L\|u - v\|_H \quad \text{and} \quad \langle Au - Av\,; u - v\rangle_{H^*\times H} \geq M\|u - v\|_H^2 \quad \text{for all } u, v \in H$$

with constants $L, M > 0$ that only depend on A. Then, A is bijective. **Hint:** Injectivity of A follows from the monotonicity of A. To prove surjectivity, we apply a fixed point argument: Let $I_H : H \to H^*$, $I_H(u) := (u\,;\cdot)_H$ denote the Riesz mapping. For given $F \in H^*$ and a certain choice of $C > 0$, the mapping $\Phi(u) := u - CI_H^{-1}(Au - F)$ is a contraction on H. Therefore, the Banach contraction theorem applies and provides a unique $u \in H$ with $u = \Phi(u)$. □

Exercise 7 (Lemma of Lax–Milgram). Use Exercise 6 to derive the Lemma of Lax–Milgram: Let H be a Hilbert space and $a(\cdot,\cdot)$ be a continuous and elliptic bilinear form on H, i.e.,

$$a(u, v) \leq L\,\|u\|_H\|v\|_H \quad \text{and} \quad a(u, u) \geq M\,\|u\|_H^2 \quad \text{for all } u, v \in H,$$

where the constants $L, M > 0$ depend only on $a(\cdot,\cdot)$. Then, given a right-hand side $F \in H^*$, there is a unique $u \in H$ with $a(u, \cdot) = F \in H^*$. □

Bibliography

[Bra] Dietrich Braess: *Finite elements. Theory, fast solvers, and applications in elasticity theory*, Cambridge University Press, Cambridge, 2007.

[McL] William McLean: *Strongly elliptic systems and boundary integral equations*, Cambridge University Press, Cambridge, 2000.

Appendix A

Some Facts from Functional Analysis

I this appendix we collect some results from introductory functional analysis courses which are used throughout. We stick with the case of vector spaces over $\mathbb{R}$.

A.1 Main Theorems from Functional Analysis

Theorem A.1 (Hahn-Banach Extension Theorem). *Let $p : X \to \mathbb{R}$ be a sublinear functional on a linear space X, i.e. $p(x+y) \leq p(x) + p(y)$ and $p(\lambda x) = \lambda p(x)$ for all $x, y \in X$ and $\lambda \geq 0$. If Y is a subspace of X and $f : Y \to \mathbb{R}$ is a linear functional with $f \leq p$ on Y, there is a linear extension $F : X \to \mathbb{R}$ with $F|_Y = f$ and $F \leq p$ on X.* ■

If X is a normed space and $f \in Y^*$, one may choose $p(x) = \|x\|_X \|f\|_{X^*}$ to prove the extension theorem for continuous linear functionals.

Corollary A.2. *If Y is the subspace of a normed space X and $f \in Y^*$, there is an extension $F \in X^*$ with $F|_Y = f$ and $\|F\|_{X^*} = \|f\|_{Y^*}$.* ■

One then considers the subspace $Y := \text{span}\{x\}$ and $f(\lambda x) = \lambda \|x\|_X$ to derive the following corollary:

Corollary A.3. *If X is a normed space and $x \in X$, there is a linear functional $f \in X^*$ with $\|f\|_{X^*} = 1$ and $f(x) = \|x\|_X = \sup\limits_{\|f\|_{X^*}=1} |f(x)|$.* ■

Theorem A.4 (Hahn-Banach Separation Theorem). *Let X be a normed space, and let A and B be convex, nonempty subsets of X with $A \cap B = \emptyset$.*
(i) *If A is open, there is a linear functional $f \in X^*$ and a scalar $\lambda \in \mathbb{R}$ such that $f(x) < \lambda \leq f(y)$ for all $x \in A$ and $y \in B$.*
(ii) *If A is compact and B is closed, there is a linear functional $f \in X^*$ and scalars $\lambda_1, \lambda_2 \in \mathbb{R}$ such that $f(x) \leq \lambda_1 < \lambda_2 \leq f(y)$ for all $x \in A$ and $y \in B$.* ■

If Y is a subspace of X, one can use (ii) to characterize the closure $\overline{Y}$ of Y in X. The proof only needs that each bounded linear functional $f \in Y^*$ is trivial, i.e. $f|_Y = 0$.

Corollary A.5. *Let Y be a subspace of the normed space X. Then, $x \in X$ satisfies $x \in \overline{Y}$ if and only if $f(x) = 0$ for all $f \in X^*$ with $f|_Y = 0$.*

Proof. For $x \in \overline{Y}$ and $f \in X^*$ with $f|_Y = 0$, continuity yields $f(x) = 0$. The converse implication is proven by contradiction: We assume that $x \notin \overline{Y}$ and choose $f \in X^*$ such that $f(x) < \lambda \leq f(y)$ for all $y \in Y$ and some fixed $\lambda \in \mathbb{R}$. Using that Y is a vector space, we infer that $\lambda \leq f(\pm y) = -f(\mp y) \leq -\lambda$ and thus $f(y) \in [\lambda, -\lambda]$ for all $y \in Y$. As bounded linear functionals are trivial, we obtain $f|_Y = 0$. According to our assumptions, this implies $f(x) = 0$ and thus contradicts $f(x) < \lambda \leq f(0) = 0$. ■

The following corollary is an immediate consequence of the last one.

Corollary A.6. *Let Y be a subspace of the normed space X. Then, Y is dense in X if and only if each functional $f \in X^*$ with $f|_Y = 0$ is trivial, i.e., $f = 0 \in X^*$.* ■

For an operator $T \in L(X;Y)$, one defines $(T^*y^*)(x) := y^*(Tx)$ for all $y^* \in Y^*$ and $x \in X$. From the continuity of T, we see that $T^*y^* \in X^*$, and obviously $T^* : Y^* \to X^*$ is a linear operator. From the corollary of the Hahn-Banach extension theorem, we derive for the operator norm

$$\begin{aligned}\|T^*\| = \sup_{\|y^*\|_{Y^*}=1} \|T^*y^*\|_{X^*} &= \sup_{\|y^*\|_{Y^*}=1} \sup_{\|x\|_X=1} (T^*y^*)(x) \\ &= \sup_{\|x\|_X=1} \sup_{\|y^*\|_{Y^*}=1} (y^*)(Tx) = \sup_{\|x\|_X=1} \|Tx\|_Y = \|T\|,\end{aligned}$$

i.e. there holds $T^* \in L(Y^*;X^*)$ with operator norm $\|T^*\| = \|T\|$. The operator T^* is called the **adjoint operator** of T.

Theorem A.7 (Banach Closed Range Theorem). *For an operator $T \in L(X;Y)$ between Banach spaces X and Y and $T^* \in L(Y^*;X^*)$ its adjoint, the following is pairwise equivalent:*
(i) range(T) *is a closed subspace of Y.*
(ii) range$(T) = (\ker T^*)_\circ := \{y \in Y \,|\, \forall y^* \in \ker(T^*) \quad y^*(y) = 0\}$.
(iii) range(T^*) *is a closed subspace of X^*.*
(iv) range$(T^*) = (\ker T)^\circ := \{x^* \in X^* \,|\, \forall x \in \ker(T) \quad x^*(x) = 0\}$. ■

A.2 Hilbert Spaces

A space X is called **Hilbert space** if it is a Banach space whose norm is induced by a scalar product.

Theorem A.8. *Let Y be the closed subspace of a Hilbert space X and $Y^\perp := \{x \in X \,|\, \forall y \in Y \quad (x\,;\,y)_X = 0\}$ the orthogonal complement. Then, there holds $X = Y \oplus Y^\perp$ in the sense of the linear algebra, i.e. every element $x \in X$ has a unique decomposition $x = y + y^\perp$ with some $y \in Y$ and $y^\perp \in Y^\perp$.* ■

With the orthogonal decomposition $X = Y \oplus Y^{\perp}$, one can define a projection $\pi_Y : X \to Y$ by $x = y + y^{\perp} \mapsto y$.

Corollary A.9. *Let Y be the closed subspace of a Hilbert space X. Then, there is a unique linear operator $\Pi : X \to Y$ with $\Pi|_Y = id$ and $\ker(\Pi) = Y^{\perp}$, which is called* **orthogonal projection** *onto Y. This projection is continuous with operator norm $\|\Pi\| = 1$ and symmetric, i.e. $(x\,;\,y)_X = (\Pi x\,;\,y)_X$ for all $x \in X$ and $y \in Y$. Moreover, the orthogonal projection is the solution operator for the best approximation problem, $\|x - \Pi x\|_X = \min_{y \in Y} \|x - y\|_X$.* ■

The dual space X^* of a Hilbert space X has a straight-forward representation, and one can somehow identify X with X^*.

Theorem A.10 (Riesz). *For a Hilbert space X, the* **Riesz mapping** *$I_X : X \to X^*$, $I_X x := (x\,;\,\cdot)_X \in X^*$, is an isometric isomorphism.* ■