
Prof. Dr. S. Sauter
Institut für Mathematik
Universität Zürich

# Finite Elements for Elliptic Eigenvalue Problems

S. Sauter

Lecture Notes for the Zürich Summerschool 2008

25. 08. 2008

# 1 Introduction

The numerical computation of eigenvalues and eigenfunctions of partial differential equations is of utmost importance in practically all fields of physical and engineering applications. In these lecture notes we will introduce the finite element discretization of eigenvalue problems for elliptic partial differential operators and develop its error analysis.

In this first chapter, we will consider as an introductory example the parabolic problem time-dependent heat conduction in a physical body which leads to a sequence of elliptic eigenvalue problems.

In Chapter 2, we will first introduce some basic concepts in functional analysis such as Banach and Hilbert spaces, dual spaces, compact operators. Eigenvalue problems for elliptic partial differential operators typically can be formulated as an operator eigenvalue problem with a compact operator. Hence, we will introduce here the Fredholm-Riesz-Schauder theory for compact operators which states the basis properties of eigenvalue problems for such operators. This material can be found in any book on functional analysis (see, e.g., [30], [10], [7], [2]). Finally, we will introduce the variational formulation of elliptic partial differential equations, the relevant function spaces (Sobolev spaces) and the concept of weak solutions. Further, we will state the existence and uniqueness theorems in the framework of the Lax-Milgram lemma. Also this material is contained in any standard textbook on this topic and we refer, e.g., to [13] or [15].

In Chapter 3, we will introduce elliptic eigenvalue problems and their finite element discretization. For doing so, we will also define finite element spaces and state their approximation properties. Standard references for these topics are [27], [28], [15], [8], [4]. The standard reference for finite element methods for symmetric and non-symmetric elliptic eigenvalue problems is [3].

In Chapter 4 we will develop the error analysis for finite element discretization for elliptic eigenvalue problems. First, the min/max characterization of the eigenvalues via Rayleigh quotients will be introduced and some monotonicity results will be proved. Then, abstract error estimates will be derived which go back to [19], [21], [12], [25], [26], [24], [23]. Finally, these abstract error estimates will be combined with the approximation properties of finite element spaces resulting in estimates of the eigenvalue and -vector errors which are explicit in the size of the eigenvalue, the spectral gap and the order of approximation.

We start here with some introductory model problem: the problem of heat conduction in a physical body $\Omega \subset \mathbb{R}^3$. We assume that the temperature is held at zero on $\partial\Omega$ for all time and that our goal is to determine the temperature distribution $u(x,t)$ at a point $x = (x_1, x_2, x_3)^\intercal \in \Omega$ and at time $t > 0$. The physical law which describes heat conduction leads to the equation

$$r\dot{u} - \operatorname{div}(A \operatorname{grad} u) = 0 \qquad \text{in } \Omega \times ]0,T]\,. \tag{1.1a}$$

Here $\dot{u}$ denotes the partial derivative with respect to time, the *divergence* div is defined for sufficiently smooth vector fields $w : \Omega \to \mathbb{R}^d$ by

$$\operatorname{div} w(x) = \sum_{i=1}^{d} \frac{\partial w(x)}{\partial x_i}$$

and the *gradient* of a sufficiently scalar function $v$ is $\nabla u = (\partial_i u)_{i=1}^d$ (the div, and grad operators are applied only to the spatial variables and not to the variable $t$). The $d \times d$ matrix

function $A : \Omega \to \mathbb{R}^{d\times d}$ is uniformly positive definite and describes the thermal conductivity of the material and the scalar function $r : \Omega \to \mathbb{R}$ is uniformly positive and describes the material density times the specific heat of the material. As boundary conditions we consider homogenous Dirichlet conditions

$$u(x,t) = 0 \qquad \forall x \in \Gamma, \forall t > 0 \tag{1.1b}$$

and initial conditions are prescribed by

$$u(x,0) = f(x) \qquad \forall x \in \Omega \tag{1.1c}$$

for some given initial temperature distribution $f$.

We employ the ansatz

$$u(x,t) = v(x)\, w(t)$$

which separates the spatial variables $x \in \Omega$ from the temporal variable $t$. This leads to the system of differential equations

$$-\operatorname{div}(A \operatorname{grad} v) = \lambda r v \quad \text{in } \Omega \quad \text{and} \quad v|_{\partial\Omega} = 0 \tag{1.2a}$$

and

$$\dot{w}(t) + \lambda w(t) = 0 \qquad \forall t > 0. \tag{1.2b}$$

It is well known (cf. Remark 2.33(2)) that (1.2a) has eigenvalues

$$0 < \lambda_1 \le \lambda_2 \le \ldots \nearrow \infty$$

and corresponding eigenfunctions $v_i$, $i \in \mathbb{N}$, which can be normalized according to

$$\int_\Omega v_i \overline{v_j} r = \delta_{i,j}.$$

Corresponding to each $\lambda_j$ we find a solution of (1.2b) $w(t) := w_j(t) = a_j \mathrm{e}^{-\lambda_j t}$. Thus, the separated solutions are given by the formal sum

$$u(x,t) = \sum_{j=1}^{\infty} a_j v_j(x) \mathrm{e}^{-\lambda_j t}.$$

The coefficients $a_j$ can be determined via the initial conditions (1.1c) by expanding $f$ into the eigenfunctions $v_j$

$$f = \sum_{j=1}^{\infty} f_j v_j(x) \quad \text{with} \quad f_j := \int_\Omega f \overline{v_j} r$$

so that

$$u(x,t) = \sum_{j=1}^{\infty} \left( \int_\Omega f \overline{v_j} r \right) v_j(x) \mathrm{e}^{-\lambda_j t}. \tag{1.3}$$

We note that from (1.3) and the positivity of the eigenvalues, one can show that $\lim_{t\to\infty} u(x,t) = 0$ and that the decay rate for the temperature $u$ is governed by the factor $\mathrm{e}^{-\lambda_1 t}$.

# 2 Some Basic Facts from Functional Analysis

In this chapter, we will present a few fundamental results from the area of functional analysis. It is not intended as an introduction to functional analysis, instead we will refer to other texts or we will give schematic proofs if we think this might help the reader's understanding of the subject.

## 2.1 Normed Spaces

With $X$ we denote a normed, linear space over the coefficient field $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. A norm $\|\cdot\| : X \to [0, \infty)$ is a mapping with the properties

$$\forall x \in X : \|x\| = 0 \Longrightarrow x = 0\,, \tag{2.1a}$$

$$\forall \lambda \in \mathbb{K} : \|\lambda x\| = |\lambda|\, \|x\|\,, \tag{2.1b}$$

$$\forall x, y \in X : \|x + y\| \leq \|x\| + \|y\|\,. \tag{2.1c}$$

We will use the notation $\|\cdot\|_X$ if the space $X$ is not clear from the context. We call the pair $(X, \|\cdot\|)$ a normed space.

We can define several different norms on $X$. Two norms $\|\cdot\|_1$, $\|\cdot\|_2$ on $X$ are **equivalent** if and only if

$$\exists C > 0 : \quad C^{-1} \|x\|_1 \leq \|x\|_2 \leq C \|x\|_1 \qquad \forall x \in X. \tag{2.2}$$

Equivalent norms induce the same topology on $X$.

**Theorem 2.1 (Nearly orthogonal element)** *Let $X$ be a normed space and let $Y \subset X$ be a closed proper subspace (i.e., $Y \neq X$, $Y$ closed in $X$). For any $0 < \theta < 1$ ($\leq 1$ if $X$ is a Hilbert space) there exists some $x_\theta \in X$ with*

$$\|x_\theta\|_X = 1 \quad \textit{and} \quad \theta \leq \operatorname{dist}(x_\theta, Y) \leq 1.$$

A proof can be found, e.g., in [30, Chap. III, Sec. 2].

## 2.2 Linear Operators

Let $X$ and $Y$ be normed spaces with the respective norms $\|\cdot\|_X$ and $\|\cdot\|_Y$. A linear mapping $T : X \to Y$ is called an operator. An operator $T : X \to Y$ is called **bounded** if

$$\|T\|_{Y \leftarrow X} := \sup \left\{ \|Tx\|_Y / \|x\|_X : \ 0 \neq x \in X \right\} < \infty\,. \tag{2.3}$$

Here $\|T\|_{Y \leftarrow X}$ is the **operator norm**. The set of all bounded linear operators $T : X \to Y$ is denoted by $L(X, Y)$ and together with

$$(T_1 + T_2)x := T_1 x + T_2 x, \ (\lambda T_1)x = T_1(\lambda x), \ \lambda \in \mathbb{K} \tag{2.4}$$

constitutes a normed, linear space $(L(X, Y), \|\cdot\|_{Y \leftarrow X})$. If $X = Y$ we write $L(X)$ instead of $L(X, X)$.

**Exercise 2.2** *(a) Show that for all $x \in X$ and $T \in L(X,Y)$ we have*

$$\|Tx\|_Y \leq \|T\|_{Y \leftarrow X} \|x\|_X \,. \tag{2.5}$$

*(b) Show that for $T_1 \in L(Y,Z)$, $T_2 \in L(X,Y)$ we have $T_1 T_2 \in L(X,Z)$ and*

$$\|T_1 T_2\|_{Z \leftarrow X} \leq \|T_1\|_{Z \leftarrow Y} \, \|T_2\|_{Y \leftarrow X} \,. \tag{2.6}$$

**Definition 2.3** *The sequence $(T_n)_n \subset L(X,Y)$ converges to $T$ if*

$$T_n \to T \Longleftrightarrow \|T - T_n\|_{Y \leftarrow X} \to 0 \;\; \textit{for} \;\; n \to \infty.$$

*It converges pointwise to $T$ if*

$$\forall x \in X : \; \|T_n x - Tx\|_Y \to 0 \;\; \textit{for} \;\; n \to \infty.$$

## 2.3 Banach Spaces

The sequence $\{x_n\} \subset X$ is called **Cauchy convergent** if $\sup\{\|x_n - x_m\|_X : n, m \geq k\} \to 0$ for $k \to \infty$. $X$ is called **complete** if all Cauchy sequences converge to an $x \in X$. A complete, normed, linear space is called a **Banach space**.

**Proposition 2.4** *Let $X$ be a normed space and $Y$ a Banach space. Then $L(X,Y)$ is a Banach space.*

The Banach space $X$ is called **separable** if there exists a countable, dense subset $A = \{a_n : n \in \mathbb{N}\} \subset X$.

## 2.4 Embeddings

Let $X, Y$ be Banach spaces with $X \subset Y$. The injection (or embedding) $I : X \to Y$ is defined by $Ix = x$ for all $x \in X$ and is clearly linear. If $I$ is bounded:

$$\forall x \in X : \; \|x\|_Y \leq C \|x\|_X \,, \tag{2.7}$$

we have $I \in L\,(X,Y)$. If $X$ is also dense in $Y$, we call $X$ densely and continuously embedded in $Y$.

## 2.5 Hilbert Spaces

Let $X$ be a vector space. A mapping $(\cdot,\cdot) : X \times X \to \mathbb{K}$ is called an **inner product** on $X$ if

$$(x,x) > 0 \;\; \forall x \in X \backslash \{0\} \;, \tag{2.8a}$$

$$(\lambda x + y, z) = \lambda (x,z) + (y,z) \quad \forall \lambda \in \mathbb{K},\, x,y,z \in X \,, \tag{2.8b}$$

$$(x,y) = \overline{(y,x)} \quad \forall x,y \in X \,. \tag{2.8c}$$

A Banach space $(X, \|\cdot\|_X)$ is called a **Hilbert space** if there exists an inner product on $X$, such that $\|x\|_X = (x,x)^{1/2}$ for all $x \in X$.

Furthermore, from (2.8) we have the Cauchy-Schwarz inequality

$$|(x,y)| \leq \|x\| \, \|y\| \quad \forall x,y \in X. \tag{2.9}$$

Two vectors $x, y \in X$ are **orthogonal** if $(x,y) = 0$. We denote this by $x \perp y$. For $A \subset X$, $A^\perp := \{x \in X \mid \forall a \in A : (x,a) = 0\}$ is a closed subspace of $X$.

**Proposition 2.5** *Let $X$ be a Hilbert space and $U \subset X$ a closed subspace. Then we have $X = U \oplus U^\perp$, i.e.:*

$$\forall x \in X : x = u + v,\ u \in U,\ v \in U^\perp,\ \|x\|^2 = \|u\|^2 + \|v\|^2 .$$

A system of orthonormal vectors $(v_\iota)_{\iota \in \mathcal{I}}$ in a Hilbert space $X$ is an **orthonormal basis** of $X$ if, for every $x \in X$, the Fourier expansion

$$x = \sum_{\iota \in \mathcal{I}} (x, v_\iota)\, v_\iota$$

converges.

**Theorem 2.6** *For every Hilbert space, there exists an orthonormal basis.*

A proof can be found, e.g., in [16, Theorem 65.1].

## 2.6 Dual Spaces

### 2.6.1 Dual Space of a Normed, Linear Space

Let $X$ be a normed, linear space over $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. The dual space $X'$ of $X$ is the space of all bounded, linear mappings (functionals)

$$X' = L(X, \mathbb{K}) .$$

$X'$ is a Banach space with norm

$$\|x'\|_{X'} := \|x'\|_{\mathbb{K} \leftarrow X} = \sup \{|x'(x)| / \|x\|_X : x \in X \setminus \{0\}\} . \tag{2.10}$$

For $x'(x)$ one can also write

$$\langle x, x' \rangle_{X \times X'} = \langle x', x \rangle_{X' \times X} = x'(x) , \tag{2.11}$$

where $\langle \cdot, \cdot \rangle_{X \times X'}$, $\langle \cdot, \cdot \rangle_{X' \times X}$ are called **dual forms** or **duality pairings**.

**Lemma 2.7** *Let $X \subset Y$ be continuously embedded. Then $Y' \subset X'$ is continuously embedded.*

**Proof.** For $y' \in Y'$, $X \subset Y$ gives us that $y'$ is defined on $X$. We therefore have $Y' \subset X'$. Since $X \subset Y$, we have, due to (2.7), that

$$\|y'\|_{Y'} = \sup_{x \in Y \setminus \{0\}} \{|y'(x)| / \|x\|_Y\} \geq C^{-1} \sup_{x \in X \setminus \{0\}} \{|y'(x)| / \|x\|_X\} = C^{-1} \|y'\|_{X'}$$

and therefore that $\|y'\|_{X'} \leq C \|y'\|_{Y'}$. This proves that the embedding $Y' \subset X'$ is continuous. ■

### 2.6.2 Dual Operator

**Proposition 2.8** *Let $X, Y$ be Banach spaces and let $T \in L(X,Y)$. For $y' \in Y'$,*

$$\langle Tx, y' \rangle_{Y \times Y'} = \langle x, x' \rangle_{X \times X'} \quad \forall x \in X \tag{2.12}$$

*defines a unique $x' \in X'$. The mapping $y' \to x'$ is linear and defines the **dual operator** $T' : Y' \to X'$ as given by $T'y' = x'$. Furthermore, we have $T' \in L(Y', X')$ and*

$$\|T'\|_{X' \leftarrow Y'} = \|T\|_{Y \leftarrow X} . \tag{2.13}$$

One of the most general principles in functional analysis is the extension of continuous linear operators which are defined on some subspace of a Banach space to the whole Banach space. We will state here the version of the Hahn-Banach extension theorem in Banach spaces.

**Theorem 2.9** *Let $X$ be a Banach space, $M$ a subspace of $X$ and $f_0$ a continuous linear functional defined on $M$. Then there exists a continuous linear functional $f$ defined on $X$ such that i) $f$ is an extension of $f_0$ and ii) $\|f_0\|_{\mathbb{C} \leftarrow M} = \|f\|_{\mathbb{C} \leftarrow X}$.*

The proof can be found, e.g., in [30, Chap. IV, Sec. 5].

**Corollary 2.10** *Let $X$ be a Banach space and $x_0 \in X \setminus \{0\}$. Then, there exists a continuous linear functional $f_0$ on $X$ such that*

$$f_0(x_0) = \|x_0\|_X \quad \text{and} \quad \|f_0\|_{X'} = 1.$$

### 2.6.3 Adjoint Operator

Let $X$ be a Hilbert space over $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$. For all $y \in X$,

$$f_y(\cdot) := (\cdot, y)_X : X \to \mathbb{K}$$

is continuous and linear: We have $f_y(\cdot) \in X'$ and $\|f_y\|_{X'} = \|y\|_X$. The converse is a result of Riesz' theorem.

**Theorem 2.11 (Riesz Representation Theorem)** *Let $X$ be a Hilbert space. For all $f \in X'$ there exists a unique $y_f \in X$ such that*

$$\|f\|_{X'} = \|y_f\|_X \quad \text{and} \quad f(x) = (x, y_f)_X \qquad \forall x \in X.$$

**Corollary 2.12** *Let $X$ be a Hilbert space. We use the same notation as in Theorem 2.11.*

*a) There exists a bounded, invertible conjugate linear form $J_X : X \to X'$ with $J_X y = f_y$, $J_X^{-1} f = y_f$. The mapping $J_X$ is an isometry: $\|J_X\|_{X' \leftarrow X} = \|J_X^{-1}\|_{X \leftarrow X'} = 1$.*

*b) $X'$ is a Hilbert space with inner product $(x', y')_{X'} := \overline{(J_X^{-1} x',\, J_X^{-1} y')_X}$.*

*c) $\|x'\|_{X'}$ in (2.10) is equal to $(x', x')_{X'}^{1/2}$.*

*d) $X \cong X''$ with $x(x') := x'(x)$ and we identify $X$ with $X''$. In particular, we have $J_{X'} = J_X^{-1}$, $J_X = (J_X)'$, $T'' = T$ for $T \in L(X,Y)$ if $Y = Y''$ and if both are Hilbert spaces.*

*e) If $\mathbb{K} = \mathbb{R}$, the spaces $X$ and $X'$ can be identified with each other by means of the isomorphism $J_X$. Then we have $X := X' \Longrightarrow J_X = I$.*

**Definition 2.13** *Let $X, Y$ be Hilbert spaces and $T \in L(X, Y)$. The adjoint operator of $T$ is given by $T^* := J_X^{-1} T' J_Y \in L(Y, X)$.*

We have

$$\|T\|_{Y \leftarrow X} = \|T^*\|_{X \leftarrow Y} \quad \text{and} \quad (Tx, y)_Y = (x, T^* y)_X \quad \forall x \in X, y \in Y. \tag{2.14}$$

**Definition 2.14**

*a. $T \in L(X)$ is self adjoint if $T = T^*$.*

*b. $T \in L(X)$ is a projection if $T^2 = T$.*

**Proposition 2.15** *Let $X_0 \subset X$ be a closed subspace of the Hilbert space $X$. For $x \in X$ there exists a unique $x_0(x) \in X_0$ with*

$$\|x - x_0\|_X = \min\{\|x - y\|_X : y \in X_0\}. \tag{2.15}$$

*The mapping $x \to x_0 =: Px$ is an orthogonal projection.*

### 2.6.4 Weak Convergence

The Bolzano-Weierstraß theorem states that in $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ every bounded sequence has at least one accumulation point. This statement only holds in a weaker form when considering infinite-dimensional function spaces. First we will need to define the concept of weak convergence.

**Definition 2.16** *Let $B$ be a Banach space and let $B'$ be its dual space. A sequence $(u_\ell)_{\ell \in \mathbb{N}}$ in $B$ converges* weakly *to an element $u \in B$ if*

$$\lim_{\ell \to \infty} \|f(u) - f(u_\ell)\|_{B'} = 0 \qquad \forall f \in B'.$$

**Theorem 2.17** *Let the Banach space $B$ be reflexive*[1] *and let $(u_\ell)_{\ell \in \mathbb{N}}$ be a bounded sequence in $B$ :*

$$\sup_{\ell \in \mathbb{N}_0} \|u_\ell\|_B \leq C < \infty.$$

*Then there exists a subsequence $\left(u_{\ell_j}\right)_{j \in \mathbb{N}}$ that converges weakly to a $u \in B$.*

The proof can be found in, e.g., [16, Theorem 60.6]. In order to distinguish between the weak convergence of a sequence $(u_\ell)_{\ell \in \mathbb{N}}$ to an element $u$ from the usual (strong) convergence, we use the notation

$$u_\ell \rightharpoonup u.$$

[1] A Banach space $B$ is reflexive if the bi-dual space $B''$ is isomorphic to $B$.

## 2.7 Compact Operators

**Definition 2.18** *The subset $U \subset X$ of the Banach space $X$ is called precompact if every sequence $(x_n)_{n\in\mathbb{N}} \subset U$ has a convergent subsequence $(x_{n_i})_{i\in\mathbb{N}}$. It is compact if, furthermore, $x = \lim_{i\to\infty} x_{n_i} \in U$.*

**Definition 2.19** *Let $X, Y$ be Banach spaces. $T \in L(X,Y)$ is called compact if $\{Tx : x \in X,\ \|x\|_X \leq 1\}$ is precompact in $Y$. The set of all compact linear operators from $X$ into $Y$ is*

$$K(X,Y) := \{T \in L(X,Y) : T \text{ is compact}\}.$$

*If $X = Y$, we simply write $K(X)$ instead of $K(X,X)$.*

We will often consider operators that are composed of several other operators.

**Lemma 2.20** *Let $X, Y, Z$ be Banach spaces, let $T_1 \in L(X,Y)$, $T_2 \in L(Y,Z)$ and let at least one of the operators $T_i$ be compact. Then $T = T_2T_1 \in L(X,Z)$ is also compact.*

**Lemma 2.21** *$T \in L(X,Y)$ compact $\Longrightarrow T' \in L(Y',X')$ compact.*

**Definition 2.22** *Let $Y$ be a Banach space and $X \subset Y$ a subspace that is continuously embedded. The embedding is compact if the injection $I \in L(X,Y)$ is compact. We denote this by $X \subset\subset Y$.*

**Corollary 2.23** *$X \subset\subset Y$ if every sequence $(x_i)_{i\in\mathbb{N}} \subset X$ with $\|x_i\|_X \leq 1$ has a subsequence that converges in $Y$.*

**Remark 2.24** *For $dim(X) < \infty$ or $dim(Y) < \infty$, $T \in L(X,Y)$ is compact.*

**Theorem 2.25 (Heine-Borel)** *Let $X$ be a normed linear space. Then*

$$\overline{B_1(0)} \text{ compact} \iff \dim X < \infty.$$

The following lemma will later be needed for existence theorems when dealing with variational problems.

**Lemma 2.26** *Let $X \subset Y \subset Z$ be Banach spaces with continuous embeddings and let $X \subset\subset Y$. Then for all $\varepsilon > 0$ there exists a constant $C_\varepsilon > 0$ with*

$$\forall x \in X:\ \|x\|_Y \leq \varepsilon \|x\|_X + C_\varepsilon \|x\|_Z.$$

## 2.8 Fredholm-Riesz-Schauder Theory

Throughout this section we assume that $X$ is a Banach space with norm $\|\cdot\|_X$ and that $X \neq \{0\}$ holds.

**Definition 2.27** *Let $X$ be a Banach space and $T \in L(X)$. The* ***resolvent set*** *of $T$ is given by*

$$\rho(T) := \{\lambda \in \mathbb{C} : N(\lambda I - T) = \{0\} \text{ and } R(\lambda I - T) = X\},$$

*where $N(\cdot)$ denotes the null space of an operator and $R(\cdot)$ its range. The spectrum $\sigma(T)$, the point spectrum $\sigma_p(T)$, the continuous spectrum $\sigma_c(T)$, and the residual spectrum $\sigma_r(T)$ are given by*

$$\begin{aligned}
\sigma(T) &:= \mathbb{C}\backslash\rho(T), \\
\sigma_p(T) &:= \{\lambda \in \sigma(T) : N(\lambda I - T) \neq \{0\}\}, \\
\sigma_c(T) &:= \left\{\lambda \in \sigma(T) : N(\lambda I - T) = \{0\} \wedge R(\lambda I - T) \neq X \wedge \overline{R(\lambda I - T)} = X\right\}, \\
\sigma_r(T) &:= \left\{\lambda \in \sigma(T) : N(\lambda I - T) = \{0\} \wedge \overline{R(\lambda I - T)} \neq X\right\}.
\end{aligned}$$

**Remark 2.28**

1. *It holds $\lambda \in \rho(T)$ iff $\lambda I - T : X \to X$ is bijective. From the inverse mapping theorem [30, p.77] we conclude that*

   $$R_\lambda(T) := (\lambda - T)^{-1} \in L(X)$$

   *exists. The function $R_\lambda(T)$ is the* ***resolvent*** *of $T$ and – considered as a function of $\lambda$ – denoted as the* ***resolvent function****.*

2. *$\lambda \in \sigma_p(T)$ is equivalent to*

   $$\exists u \in X\backslash\{0\} : Tu = \lambda u.$$

   *We call $\lambda$ an* ***eigenvalue*** *and $u$ an* ***eigenvector*** *of $T$. The* ***eigenspace*** *of $T$ corresponding to $\lambda$ is $N(\lambda I - T)$. The eigenspace is a $T$-invariant subspace*[2]*.*

**Definition 2.29** *A mapping $A \in L(X, Y)$ is a Fredholm operator if*

1. *$R(A)$ is closed,*
2. *$\dim N(A) < \infty$ and $\operatorname{codim} R(A) < \infty$.*

*The index of a Fredholm operator is*

$$\operatorname{ind}(T) := \dim N(A) - \operatorname{codim} R(A).$$

Note that the finiteness of the co-dimension of $R(A)$ implies that[3] $Y = R(A) \oplus Y_0$ for some finite-dimensional subspace $Y_0 \subset Y$. We have $\operatorname{codim} R(A) := \dim Y_0$ independent of the choice of $Y_0$. The connection between Fredholm operators and compact operators is given by the following theorem.

[2] A subspace $Y \subset X$ is $T$-invariant if $T(Y) \subset Y$.

[3] Recall that, for subspaces $F, G$ of some vector space $E$, the symbol $\oplus$ is used in $F \oplus G$ instead of $+$ if $F \cap G = \{0\}$. $G$ is a complementary space of $F$ if $F \oplus G = E$. In [16, Satz 4.1], it is shown that a complementary space always exists. If $G$ is a complementary space of $F$ in $E$, the co-dimension $\operatorname{codim} F$ is defined by

$$\operatorname{codim} F := \begin{cases} \infty & \text{if } \dim G = \infty, \\ \dim G & \text{if } 1 \leq G < \infty, \\ 0 & \text{if } F = E. \end{cases}$$

In [16, Satz 4.2], it is proved that the definition of the co-dimension of $F$ is independent of the choice of the complementary space $G$.

**Theorem 2.30** *For $T \in K(X)$, the operator $I - T$ is a Fredholm operator of index 0.*

**Proof.** The proof consists of five parts.

1) We prove $\dim N(A) < \infty$ with $A := I - T$.

Since $Ax = 0$ is equivalent to $x = Tx$ we have

$$B_1(0) \cap N(A) \subset T(B_1(0)),$$

i.e., the unit ball in $N(A)$ is precompact and, hence, $N(A)$ finite dimensional (cf. Theorem 2.25).

2) Next, we prove $R(A)$ is closed.

Let $x \in \overline{R(A)}$ and choose a sequence $(x_n)$ so that $Ax_n \to x$ as $n \to \infty$. W.l.o.g., we may assume that

$$\|x_n\|_X \leq 2d_n \quad \text{with} \quad d_n := \operatorname{dist}(x_n, N(A)),$$

because, otherwise, we choose $a_n \in N(A)$ with $\|x_n - a_n\|_X \leq 2\operatorname{dist}(x_n, N(A))$ and consider $\tilde{x}_n := x_n - a_n$ instead of $x_n$. Note that $\operatorname{dist}(\tilde{x}_n, N(A)) = \operatorname{dist}(x_n, N(A))$.

First, we assume that $d_n \to \infty$ for a subsequence. For $y_n := d_n^{-1} x_n$, we obtain $Ay_n = d_n^{-1} Ax_n \to 0$ as $n \to \infty$. Because $y_n$ is bounded and $T$ is compact, there exists a subsequence which satisfies $Ty_n \to y$ as $n \to \infty$. Hence,

$$y_n = Ay_n + Ty_n \to y$$

and the continuity of $A$ implies

$$Ay = \lim_{n\to\infty} Ay_n = 0,$$

i.e., $y \in N(A)$. This leads to

$$\|y_n - y\|_X \geq \operatorname{dist}(y_n, N(A)) = \operatorname{dist}\left(\frac{x_n}{d_n}, N(A)\right) = \frac{\operatorname{dist}(x_n, N(A))}{d_n} = 1$$

and this is a contradiction. Thus, we have proved that $(d_n)_n$ is bounded and, consequently, $(x_n)_n$ is bounded as well. The compactness of $T$ implies that there exists a subsequence which satisfies $Tx_n \to z$ as $n \to \infty$, i.e.,

$$x \leftarrow Ax_n = A(Ax_n + Tx_n) \to A(x + z).$$

Thus, we have proved $x \in R(A)$.

3) We prove

$$N(A) = \{0\} \implies R(A) = X.$$

Assume that there is some $x \in X \backslash R(A)$. Then

$$A^n x \in R(A^n) \backslash R\left(A^{n+1}\right) \quad \forall n \geq 0, \tag{2.16}$$

because, otherwise, i.e. if $A^n x = A^{n+1} y$ for some $y$ then $A^n(x - Ay) = 0$ and $N(A) = \{0\}$ would imply $x - Ay = 0$ by induction and, in turn, $x \in R(A)$ which is a contradiction.

$R(A^{n+1})$ is closed because

$$A^{n+1} = (I - T)^{n+1} = I + \underbrace{\sum_{k=1}^{n+1} \binom{n+1}{k} (-T)^k}_{\in K(X)}$$

so that, from part 2, we may conclude that $R\left(A^{n+1}\right)$ is closed. Hence, we may choose $a_{n+1} \in R\left(A^{n+1}\right)$ so that

$$\left\|A^n x - a_{n+1}\right\|_X \leq 2 \operatorname{dist}\left(A^n x, R\left(A^{n+1}\right)\right) \overset{(2.16)}{\neq} 0. \tag{2.17}$$

Now, consider

$$x_n := \frac{A^n x - a_{n+1}}{\left\|A^n x - a_{n+1}\right\|_X}.$$

We have $\operatorname{dist}\left(x_n, R\left(A^n\right)\right) \geq 1/2$ because for $y \in R\left(A^{n+1}\right)$ it holds

$$\left\|x_n - y\right\|_X = \frac{\left\|A^n x - \left(a_{n+1} + \left\|A^n x - a_{n+1}\right\|_X y\right)\right\|_X}{\left\|A^n x - a_{n+1}\right\|_X} \geq \frac{\operatorname{dist}\left(A^n x, R\left(A^{n+1}\right)\right)}{\left\|A^n x - a_{n+1}\right\|_X} \overset{(2.17)}{\geq} 1/2.$$

Thus, for all $m > n$, we derive

$$\left\|Tx_n - Tx_m\right\|_X = \left\| x_n - \underbrace{\left(Ax_n + x_m - Ax_m\right)}_{\in R(A^{n+1})} \right\|_X \geq 1/2.$$

Hence, $\left(Tx_n\right)_n$ has no convergent subsequence although $\left(x_n\right)_n$ is a bounded sequence and this is a contradiction to the compactness of $T$.

4) We prove

$$\operatorname{codim} R\left(A\right) \leq \dim N\left(A\right). \tag{2.18}$$

From 1) we obtain that $n := \dim N\left(A\right)$ is finite. Let $x_1, \ldots, x_n$ denote some basis for $N\left(A\right)$. If the assertion (2.18) is wrong, there exist linear independent vectors $y_1, \ldots, y_n$ such that

$$\operatorname{span}\left\{y_1, \ldots, y_n\right\} \oplus R\left(A\right)$$

is a proper subspace of $X$. We choose a dual basis $x_1', \ldots, x_n'$ in $X'$ such that

$$\left\langle x_\ell, x_k'\right\rangle = \delta_{k,\ell} \qquad \forall 1 \leq k, \ell \leq n.$$

We define

$$\tilde{T}x := Tx + \sum_{k=1}^{n} \left\langle x, x_k'\right\rangle y_k$$

and observe that $\tilde{T} \in K\left(X\right)$ because $T$ is compact and $\tilde{T} - T$ has a finite dimensional range. Furthermore, $N\left(\tilde{A}\right) = \{0\}$, where $\tilde{A} := I - \tilde{T}$ because $\tilde{A}x = 0$ implies (due to the choice of $y_k$) $Ax = 0$ and $\left\langle x, x_k'\right\rangle = 0$ for $k = 1, \ldots n$. Hence, $x \in N\left(A\right)$ and there is a representation

$$x = \sum_{k=1}^{n} \alpha_k x_k$$

from which we conclude

$$0 = \left\langle x, x_\ell'\right\rangle = \sum_{k=1}^{n} \alpha_k \left\langle x_k, x_\ell'\right\rangle = \alpha_\ell, \quad \text{i.e., } x = 0.$$

We apply the statement of part 3 to the operator $\tilde{A}$ and derive $R\left(\tilde{A}\right) = X$. Because of

$$\tilde{A}x_\ell = -y_\ell \qquad \forall \ell = 1, \ldots, n$$

and

$$\tilde{A}\left(x - \sum_{\ell=1}^{n} \langle x, x'_\ell \rangle \, x_\ell\right) = Ax \qquad \forall x \in X$$

we obtain

$$X = R\left(\tilde{A}\right) \subset \operatorname{span}\{y_1, \ldots, y_n\} \oplus R(A)$$

and this is the contradiction.

5) It remains to prove

$$n := \dim N(A) \leq \operatorname{codim} R(A) =: m. \tag{2.19}$$

According to part 4 we have $m \leq n$. First, we reduce the assertion to the case $m = 0$. Choose $x_1, \ldots, x_n$ and $x'_1, \ldots, x'_n$ as in part 4 and $y_1, \ldots, y_m$ with

$$X = \operatorname{span}\{y_1, \ldots, y_m\} \oplus R(A).$$

As in part 4 the mapping

$$\tilde{T}x := Tx + \sum_{k=1}^{m} \langle x_k, x'_k \rangle \, y_k$$

is compact and $\tilde{A} := I - \tilde{T}$ is surjective with $N\left(\tilde{A}\right) = \operatorname{span}\{x_i : m < i \leq n\}$. We have to prove that $N\left(\tilde{A}\right) = \{0\}$ which follows from the statement "If $T \in K(X)$ and $\operatorname{codim}(I - T) = 0$, then, $N(I - T) = \{0\}$." by substituting $T \leftarrow \tilde{T}$ therein, i.e., (2.19) for $m = 0$. Thus, we have reduced the assertion to the case $m = 0$.

In the case $m = 0$ it holds $R(A) = X$. We assume that there is some $x_1 \in N(A) \setminus \{0\}$. Because of the surjectivity we may assume (by induction) that there is exist $x_n \in X$, $n \geq 2$, such that $Ax_n = x_{n-1}$. Then, $x_n \in N(A^n) \setminus N(A^{n-1})$. The theorem of the nearly orthogonal element (Theo. 2.1) implies some $y_n \in N(A^n)$ with $\|y_n\|_X = 1$ and $\operatorname{dist}(y_n, N(A^n)) \geq 1/2$. Thus, it follows for all $m < n$

$$\|Ty_n - Ty_m\|_X = \left\| y_n - \underbrace{(Ay_n + y_m - Ay_m)}_{\in N(A^{n-1})} \right\|_X \geq 1/2,$$

i.e., $(Ty_n)_n$ has no convergent subsequence. However, this is a contradiction to the compactness of $T$ in view of the boundedness of $(y_n)_n$. ■

A mapping $F : D \to Y$ from an open subset $D \subset \mathbb{C}$ into a Banach space $Y$ is **complex analytic** if, for any $\lambda_0 \in D$, there exists an open ball $B_{r_0}(\lambda_0) \subset D$ so that $F(\lambda)$ can be represented as a Taylor series about $\lambda_0$ for all $\lambda \in B_{r_0}(\lambda_0)$.

**Theorem 2.31** *Let $T \in L(X)$. $\rho(T)$ is open and the resolvent function $R_{(\cdot)}(T)$ is a complex analytic mapping from $\rho(T)$ into $L(X)$. It holds*

$$\|R_\lambda(T)\|^{-1}_{X \leftarrow X} \leq \operatorname{dist}(\lambda, \sigma(T)).$$

**Proof.** Let $\lambda \in \rho(T)$. For any $\mu \in \mathbb{C}$, we have

$$(\lambda - \mu) I - T = (\lambda I - T) \underbrace{(I - \mu R_\lambda(T))}_{S(\mu)}.$$

From the theorem on Neumann series (cf. [30, p.69]) it follows that $S(\mu)$ is invertible if

$$|\mu| \, \|R_\lambda(T)\|_{X \leftarrow X} < 1$$

which, in turn, implies that $\lambda - \mu \in \rho(T)$ and

$$R_{\lambda-\mu}(T) = S(\mu)^{-1} R_\lambda(T) = \sum_{k=0}^{\infty} \mu^k R_\lambda(T)^{k+1}.$$

Hence, for $d := \|R_\lambda(T)\|_{X \leftarrow X}^{-1}$, it holds $B_d(\lambda) \subset \rho(T)$. From this we conclude that $\operatorname{dist}(\lambda, \sigma(T)) \geq d$. ■

**Theorem 2.32** *Let $T \in L(X)$. $\sigma(T)$ is compact, non-empty, and it holds*

$$r(T) := \sup_{\lambda \in \sigma(T)} |\lambda| = \lim_{m \to \infty} \|T^m\|_{X \leftarrow X}^{1/m} \leq \|T\|_{X \leftarrow X}$$

*and $r(T)$ is called the spectral radius of $T$.*

**Proof.** Let $\lambda \neq 0$. The theorem on Neumann series implies that $I - \lambda^{-1} T$ is invertible provided $\|\lambda^{-1} T\|_{X \leftarrow X} < 1$, i.e., $|\lambda| > \|T\|_{X \leftarrow X}$ and that

$$R_\lambda(T) = \lambda^{-1} \left(I - \lambda^{-1} T\right)^{-1} = \sum_{k=0}^{\infty} \lambda^{-k-1} T^k. \tag{2.20}$$

Thus, any $\lambda \in \sigma(T)$ must satisfy $|\lambda| \leq \|T\|_{X \leftarrow X}$ and, hence,

$$r(T) \leq \|T\|_{X \leftarrow X}. \tag{2.21}$$

Since

$$\lambda^m I - T^m = (\lambda I - T) p_m(T) = p_m(T) (\lambda I - T)$$

for

$$p_m(T) = \sum_{\ell=0}^{m-1} \lambda^{m-1-\ell} T^\ell$$

we conclude that

$$\begin{aligned} \lambda \in \sigma(T) &\implies \lambda^m \in \sigma(T^m) \\ &\implies |\lambda^m| \leq \|T^m\|_{X \leftarrow X} \ \text{(as a consequence of (2.21))} \\ &\implies |\lambda| \leq \|T^m\|_{X \leftarrow X}^{1/m}. \end{aligned}$$

From this and the definition of the limes superior/inferior it follows that

$$r(T) \leq \liminf_{m \to \infty} \|T^m\|^{1/m}.$$

Next, we will also show

$$r(T) \geq \limsup_{m\to\infty} \|T^m\|_{X\leftarrow X}^{1/m}.$$

According to Theorem 2.31, the resolvent function is a complex analytic function in $\mathbb{C}\backslash\overline{B_r(0)}$ (in $\mathbb{C}$ if $\sigma(T) = \emptyset$) and Cauchy's integral theorem (cf. [30, Chap. V.3]) implies that

$$\frac{1}{2\pi\,\mathrm{i}} \int_{\partial B_s(0)} \lambda^j R_\lambda(T)\, d\lambda$$

is, for any $j \geq 0$ and $s > r$ independent of $s$. By choosing $s > \|T\|_{X\leftarrow X}$ we derive from the representation of $R_\lambda(T)$ as in (2.20) that this integral equals

$$\begin{aligned}\frac{1}{2\pi\,\mathrm{i}} \int_{\partial B_s(0)} \left(\sum_{k=0}^{\infty} \lambda^{j-k-1} T^k\right) d\lambda &= \frac{1}{2\pi} \sum_{k=0}^{\infty} s^{j-k} \left(\int_0^{2\pi} \mathrm{e}^{\mathrm{i}\,\theta(j-k)}\, d\theta\right) T^k \\ &= \sum_{k=0}^{\infty} s^{j-k} \delta_{j,k} T^k = T^j.\end{aligned}$$

Hence, we have for any $j \geq 0$ and $s > r$ the estimate

$$\left\|T^j\right\|_{X\leftarrow X} = \frac{1}{2\pi} \left\| \int_{\partial B_s(0)} \lambda^j R_\lambda(T)\, d\lambda \right\|_{X\leftarrow X} \leq s^{j+1} \sup_{|\lambda|=s} \|R_\lambda(T)\|_{X\leftarrow X}$$

holds. Consequently, we obtain for $s > r$ and any subsequence

$$\left\|T^j\right\|_{X\leftarrow X}^{1/j} \leq s \left(s \sup_{|\lambda|=s} \|R_\lambda(T)\|_{X\leftarrow X}\right)^{1/j} \overset{j\to\infty}{\to} s \text{ or } 0, \tag{2.22}$$

so that

$$\limsup_{j\to\infty} \left\|T^j\right\|_{X\leftarrow X}^{1/j} \leq s.$$

Because this is satisfied for all $s > r$, the assertion on the spectral radius is proved. In the case of $\sigma(T) = \emptyset$ we raise (2.22) to the $j$-th power and get for $j = 0$ and $s \searrow 0$

$$\|I\|_{X\leftarrow X} \leq s \left(\sup_{|\lambda|\leq 1} \|R_\lambda(T)\|_{X\leftarrow X}\right) \to 0,$$

i.e., $I = 0$ and, hence, $X = \{0\}$. ■

In the following, we will investigate the spectrum of compact operators.

**Remark 2.33**

1. *If* $\dim X < \infty$ *holds then* $\sigma(T) = \sigma_p(T)$.
2. *If* $\dim X = \infty$ *and* $T \in K(X)$, *we have* $0 \in \sigma(T)$. *(In general,* 0 *is not an eigenvalue of* $T$*).*

**Proof.** @1: For $\lambda \in \sigma(T)$, the mapping $T$ is not bijective and – because $\dim X < \infty$ – not injective, i.e., $\lambda \in \sigma_p(T)$.

@2: Let $T \in K(X)$ and $0 \in \rho(T)$. Then, (cf. Remark 2.28(1)), $T^{-1} \in L(X)$, and as a consequence of Lemma 2.20 $I = T^{-1}T \in K(X)$. From Theorem 2.25 we conclude that $X$ is finite dimensional and this is a contradiction. ■

The main theorem of this section if the Riesz-Schauder theory.

**Theorem 2.34 (Riesz-Schauder)** *For any operator $T \in K(X)$ it holds*

1. *$\sigma(T) \setminus \{0\}$ consists of countably many (finitely or infinitely many) eigenvalues with 0 as the only possible accumulation point.*
2. *For $\lambda \in \sigma(T) \setminus \{0\}$ we have*

$$1 \leq n_\lambda := \max\left\{n \in \mathbb{N} : N(\lambda I - T)^{n-1} \neq N(\lambda I - T)^n\right\} < \infty.$$

   *$n_\lambda$ is the **index** of $\lambda$ and $\dim N(\lambda I - T)$ is the **multiplicity** of $\lambda$.*
3. *(Riesz decomposition) For $\lambda \in \sigma(T) \setminus \{0\}$ it holds*

$$X = N\left((\lambda I - T)^{n_\lambda}\right) \oplus R\left((\lambda I - T)^{n_\lambda}\right).$$

   *Both subspaces are closed and $T$-invariant and $N(\lambda I - T)^{n_\lambda}$ is finite dimensional.*
4. *For $\lambda \in \sigma(T) \setminus \{0\}$, let $E_\lambda$ be the projection onto $N\left((\lambda I - T)^{n_\lambda}\right)$ according to the decomposition in (3). Then*

$$E_\lambda E_\mu = \delta_{\lambda,\mu} E_\lambda \qquad \forall \lambda, \mu \in \sigma(T) \setminus \{0\}.$$

**Proof.** @1: Let $0 \neq \lambda \notin \sigma_p(T)$. Then, $N(I - \lambda^{-1}T) = \{0\}$, i.e., $R(I - \lambda^{-1}T) = X$ (cf. proof, part 3, of Theorem 2.30). This implies $\lambda \in \rho(T)$, i.e., $\sigma(T) \setminus \{0\} \subset \sigma_p(T)$. If $\sigma(T) \setminus \{0\}$ is not finite we choose distinct values $\lambda_n \in \sigma(T) \setminus \{0\}$, $n \in \mathbb{N}$, and eigenvectors $e_n \neq 0$ corresponding to $\lambda_n$ and define

$$X_n := \operatorname{span}\{e_i : 1 \leq i \leq n\}.$$

The eigenvectors are linear independent because, otherwise, if we assume that

$$e_n = \sum_{k=1}^{n-1} \alpha_k e_k$$

with linear independent vectors $e_k$, $1 \leq k \leq n-1$, we would get

$$0 = Te_n - \lambda_n e_n = \sum_{k=1}^{n-1} \alpha_k (Te_k - \lambda_n e_k) = \sum_{k=1}^{n-1} \alpha_k (\lambda_k - \lambda_n) e_k,$$

i.e., $\alpha_k = 0$ for $k = 1, 2, \ldots n-1$, i.e., $e_n = 0$ and this is a contradiction. Hence, $X_{n-1}$ is a proper subspace of $X_n$. Hence, according to the theorem of a nearly orthogonal element (Theorem 2.1), there exists some $x_n \in X_n$ with

$$\|x_n\|_X = 1 \quad \text{and} \quad \operatorname{dist}(x_n, X_{n-1}) \geq 1/2.$$

Since $x_n = \alpha_n e_n + \tilde{x}_n$ for some $\alpha_n \in \mathbb{C}$ and some $\tilde{x}_n \in X_{n-1}$ we conclude from the $T$-invariance of the subspace $X_{n-1}$ that $Tx_n - \lambda_n x_n = T\tilde{x}_n - \lambda_n \tilde{x}_n \in X_{n-1}$. Thus, for all $m < n$ we have

$$\left\| T\left(\frac{x_n}{\lambda_n}\right) - T\left(\frac{x_m}{\lambda_m}\right) \right\|_X = \left\| x_n + \underbrace{\lambda_n^{-1}\left(Tx_n - \lambda_n x_n\right) - \lambda_m^{-1} T x_m}_{\in X_{n-1}} \right\|_X \geq 1/2.$$

Hence, the sequence $\left(T\left(\lambda_n^{-1} x_n\right)\right)_{n\in\mathbb{N}}$ has no accumulation point. The compactness of $T$ implies that $\left(\lambda_n^{-1} x_n\right)_{n\in\mathbb{N}}$ has no bounded subsequence and hence

$$|\lambda_n|^{-1} = \left\|\lambda_n^{-1} x_n\right\|_X \overset{n\to\infty}{\to} \infty,$$

i.e., $\lambda_n \to 0$ as $n \to \infty$. Hence, 0 is the only possible accumulation point of $\sigma(T) \setminus \{0\}$. Thus, $\sigma(T) \setminus B_r(0)$ is finite for any $r > 0$, i.e., $\sigma(T) \setminus \{0\}$ is countable.

@2: Let $A := \lambda I - T$. Then $N\left(A^{n-1}\right) \subset N\left(A^n\right)$ for all $n$. First, we assume that

$$N\left(A^{n-1}\right) \text{ is a proper subset of } N\left(A^n\right) \text{ for all } n \geq 1.$$

Similarly as in the proof of 1) we choose – according to the theorem of the nearly orthogonal element – some $x_n \in N\left(A^n\right)$ such that

$$\|x_n\|_X = 1 \quad \text{and} \quad \operatorname{dist}\left(x_n, N\left(A^{n-1}\right)\right) \geq 1/2.$$

For all $m < n$, we obtain

$$\|Tx_n - Tx_m\|_X = \left\| \lambda x_n - \underbrace{\left(Ax_n + \lambda x_m - Ax_m\right)}_{\in N(A^{n-1})} \right\|_X \geq |\lambda|/2.$$

On the other hand $(x_n)_{n\in\mathbb{N}}$ is a bounded sequence which is a contradiction to the compactness of $T$. Consequently, we find some $n \in \mathbb{N}$ with $N\left(A^{n-1}\right) = N\left(A^n\right)$ and obtain, for $m > n$,

$$\begin{aligned} x \in N\left(A^m\right) &\implies A^{m-n}x \in N\left(A^n\right) = N\left(A^{n-1}\right) \\ &\implies A^{n-1+m-n}x = 0 \implies x \in N\left(A^{m-1}\right), \end{aligned}$$

and hence $N\left(A^m\right) = N\left(A^{m-1}\right)$. Thus, $N\left(A^m\right) = N\left(A^n\right)$ for all $m \geq n$ by induction. Hence, $n_\lambda < \infty$. From $N(A) \neq \{0\}$ we finally obtain $n_\lambda \geq 1$.

@3: Note that

$$N\left(A^{n_\lambda}\right) \oplus R\left(A^{n_\lambda}\right) \subset X$$

because $x \in N\left(A^{n_\lambda}\right) \cap R\left(A^{n_\lambda}\right)$ implies that $A^{n_\lambda}x = 0$ and $x = A^{n_\lambda}y$ for some $y \in X$. In this case, we have $A^{2n_\lambda}y = 0$, i.e., $y \in N\left(A^{2n_\lambda}\right) = N\left(A^{n_\lambda}\right)$ and therefore $x = A^{n_\lambda}y = 0$.

Now, $A^{n_\lambda}$ can be represented as

$$A^{n_\lambda} = \lambda^{n_\lambda} I + \underbrace{\sum_{k=1}^{n_\lambda} \binom{n_\lambda}{k} \lambda^{n_\lambda - k} (-T)^k}_{\in K(X)}.$$

Therefore $\operatorname{codim} R(A^{n_\lambda}) \leq \dim N(A^{n_\lambda}) < \infty$ (cf. proof, part 4, of Theorem 2.30) from which

$$X = N(A^{n_\lambda}) \oplus R(A^{n_\lambda})$$

follows. Since $T$ and $A$ commute the operators $T$ and $A^{n_\lambda}$ commute as well. Hence, both subspaces are $T$-invariant.

Let $T_\lambda$ denote the restriction of $T$ to $R(A^{n_\lambda})$. Note that $T_\lambda \in K(R(A^{n_\lambda}))$, where $R(A^{n_\lambda})$ is a closed subspace (see proof, part 2, of Theorem 2.30) and, hence, $R(A^{n_\lambda})$ is a Banach space. In addition it holds

$$N(\lambda I - T_\lambda) = N(A) \cap R(A^{n_\lambda}) = \{0\}$$

and, hence, $R(\lambda I - T_\lambda) = R(A^{n_\lambda})$ (see proof, part 2, of Theorem 2.30) and we have proved that $\lambda \in \rho(T_\lambda)$.

@4) Let $\mu \in \mathbb{C} \backslash \{\lambda\}$. From the previous reasoning we know that $N(A^{n_\lambda})$ is invariant under $\lambda I - T$.

Auxiliary statement: $(\mu I - T)|_{N(A^{n_\lambda})}$ is injective.

Proof of auxiliary statement: $x \in N(\mu I - T)$ implies $(\lambda - \mu)\, x = Ax$. If, in addition, $A^m x = 0$ for some $m \geq 1$ it follows that

$$(\lambda - \mu)\, A^{m-1} x = A^m x = 0$$

and, because of $\lambda \neq \mu$, $A^{m-1} x = 0$. By induction we derive $x = 0$ and we have proved that

$$N(\mu I - T) \cap N(A^m) = \{0\} \qquad \forall m \geq 1.$$

By setting $m = n_\lambda$ the auxiliary statement follows.

Since $N(A^{n_\lambda})$ is finite dimensional the restriction $\mu I - T : N(A^{n_\lambda}) \to N(A^{n_\lambda})$ is bijective.

Let $\lambda, \mu \in \sigma(T) \setminus \{0\}$ with $\lambda \neq \mu$ and set $A := \lambda I - T$. We just proved that $\mu I - T : N(A^{n_\lambda}) \to N(A^{n_\lambda})$ is bijective. Consequently $(\mu I - T)^{n_\mu} : N(A^{n_\lambda}) \to N(A^{n_\lambda})$ is also bijective, i.e.,

$$N((\lambda I - T)^{n_\lambda}) \subset R((\mu I - T)^{n_\mu}).$$

In other words

$$R(E_\lambda) \subset N(E_\mu).$$

By interchanging $\lambda$ and $\mu$ we obtain $R(E_\mu) \subset N(E_\lambda)$. ■

The property $\sigma(T) \setminus \{0\} \subset \sigma_p(T)$ can be restated as the Fredholm alternative.

**Theorem 2.35 (Fredholm alternative)** *Let $T \in K(X)$ and $\lambda \neq 0$. Then,* ***either***

$$\forall y \in X \quad \exists! x \in X : \quad \lambda x - Tx = y$$

***or***

$$\exists x \in X \backslash \{0\} : \quad \lambda x - Tx = 0.$$

Next, we consider normal operators in Hilbert spaces. In this case, some of the previous assertions can be strengthened.

**Definition 2.36** *Let $X$ be a Hilbert space over $\mathbb{K}$. Then, $T \in L(X)$ is normal if*

$$T^*T = TT^*,$$

*where $T^*$ is the adjoint of $T$ (cf. Definition 2.13).*

**Proposition 2.37** *Let $X$ be a Hilbert space and let $T \in L(X)$ be a normal operator. Then $\lambda I - T$ is normal for any $\lambda \in \mathbb{K}$ and it holds*

$$T \text{ is normal} \iff \|Tx\|_X = \|T^*x\|_X . \tag{2.23}$$

*Furthermore, for all $\lambda \in \mathbb{C}$ we have*

$$N(\lambda I - T) = N\left(\bar{\lambda} I - T^*\right).$$

**Proof.** The assertion " $\Longrightarrow$ " in (2.23) follows from

$$(Tx, Tx)_X = (x, T^*Tx)_X = (x, TT^*x)_X = (T^*x, T^*x)_X .$$

To prove " $\Longleftarrow$ " in (2.23) we start with the identity

$$\frac{1}{4}\left(\|a+b\|_X^2 - \|a-b\|_X^2\right) = \operatorname{Re}(a,b)_X \qquad \forall a, b \in X.$$

This implies

$$\operatorname{Re}(Tx, Ty)_X = \operatorname{Re}(T^*x, T^*y)_X \qquad \forall x, y \in X.$$

By substituting $\mathrm{i}\, y$ for $y$ in the case $\mathbb{K} = \mathbb{C}$ we get

$$\operatorname{Im}(Tx, Ty)_X = \operatorname{Im}(T^*x, T^*y)_X \qquad \forall x, y \in X.$$

Hence,

$$0 = (Tx, Ty)_X - (T^*x, T^*y)_X = \left((T^*T - TT^*)\, x, y\right)_X \qquad \forall x, y \in X,$$

i.e., $T^*T = TT^*$. ■

**Lemma 2.38** *Let $X$ be a Hilbert space over $\mathbb{K}$ and $X \neq \{0\}$. If $T \in L(X)$ is normal then*

$$r(T) = \|T\|_{X \leftarrow X} .$$

**Proof.** Let $T \neq 0$. By using Theorem 2.32 the statement is proved if we show

$$\|T^m\|_{X \leftarrow X} \geq \|T\|_{X \leftarrow X}^m \qquad \forall m \geq 0.$$

For $m = 0, 1$, this inequality is trivial. For $m \geq 1$ and $x \in X$ it holds

$$\begin{aligned} \|T^m x\|_X^2 &= \left(T^*T^m x, T^{m-1}x\right)_X \leq \|T^*T^m x\|_X \left\|T^{m-1}x\right\|_X \\ &\overset{\text{Theo 2.37}}{=} \left\|T^{m+1}x\right\|_X \left\|T^{m-1}x\right\|_X \leq \left\|T^{m+1}\right\|_{X \leftarrow X} \|T\|_{X \leftarrow X}^{m-1} \|x\|_X^2 , \end{aligned}$$

i.e.,

$$\|T^m\|_{X \leftarrow X}^2 \leq \left\|T^{m+1}\right\|_{X \leftarrow X} \|T\|_{X \leftarrow X}^{m-1} .$$

If we assume by induction that $\|T^m\|_{X \leftarrow X} \geq \|T\|_{X \leftarrow X}^m$, we derive

$$\left\|T^{m+1}\right\|_{X \leftarrow X} \geq \frac{\|T^m\|_{X \leftarrow X}^2}{\|T\|_{X \leftarrow X}^{m-1}} \geq \|T\|_{X \leftarrow X}^{2m-(m-1)} = \|T\|_{X \leftarrow X}^{m+1} .$$

■

**Theorem 2.39** *Let $X$ be a Hilbert space over $\mathbb{C}$ and let $T \in K(X)$ be normal, $T \neq 0$. Then $T$ has the form*

$$Tx = \sum_{k \in N} \lambda_k (x, e_k)_X e_k \tag{2.24}$$

*with $N \subset \mathbb{N}$ and an orthonormal system $(e_k)_{k \in N}$ and $0 \neq \lambda_k \in \mathbb{C}$, where $\lambda_k \to 0$ as $k \to \infty$ if $N$ is infinite. Furthermore, $X$ has the orthogonal decomposition*

$$X = N(T) \oplus \overline{\operatorname{span}\{e_k : k \in N\}}.$$

*The numbers $\lambda_k$ are the eigenvalues of $T$ corresponding to the eigenvectors $e_k$. The values $\lambda_k$ may coincide for different values of $k$. In addition the index satisfies $n_{\lambda_k} = 1$.*

**Proof.** From the spectral theory for compact operators (Theorem 2.34) it follows that $\sigma(T) \setminus \{0\}$ consists of eigenvalues $\lambda_k$, $k \in N \subset \mathbb{N}$ with $\lambda_k \to 0$ as $k \to \infty$ and $N$ is infinite. In **this** ordering (which differs from the numbering in the theorem) we assume that the $\lambda_k$'s are pairwise distinct. The eigenspaces $N_k := N(\lambda_k I - T)$ are finite dimensional. Let $N_0 := N(T)$ and $\lambda_0 := 0$. Proposition 2.37 implies

$$N_k = N\left(\overline{\lambda_k} I - T^*\right) \qquad \forall k \in N \cup \{0\}.$$

First, we prove that the eigenspaces are pairwise orthogonal, i.e.,

$$N_k \perp N_\ell \qquad \forall k, \ell \in N \cup \{0\} \quad \text{with} \quad k \neq \ell. \tag{2.25}$$

Let $x_k \in N_k$ and $x_\ell \in N_\ell$. Then,

$$\lambda_k (x_k, x_\ell)_X = (Tx_k, x_\ell)_X = (x_k, T^* x_\ell)_X = \left(x_k, \overline{\lambda_\ell} x_\ell\right)_X = \lambda_\ell (x_k, x_\ell)_X.$$

Because $\lambda_k \neq \lambda_\ell$, we obtain $(x_k, x_\ell)_X = 0$ and (2.25) is proved. Next, we will show

$$X = \overline{\bigoplus_{k \in N \cup \{0\}} N_k}. \tag{2.26}$$

Choose

$$y \in Y := \left( \bigoplus_{k \in N \cup \{0\}} N_k \right)^\perp.$$

Then, for $x \in N_k$, $k \in N \cup \{0\}$, we get

$$(Ty, x)_X = (y, T^* x)_X = \lambda_k (y, x)_X = 0.$$

Hence, $Ty \in Y$ and $Y$ is a $T$-invariant closed subspace. Consider $T_0 := T|_Y$. Since $T_0$ is normal, there is – provided $Y \neq \{0\}$ – some $\lambda \in \sigma(T_0)$ with $|\lambda| = \|T_0\|_{X \leftarrow X}$ (cf. Lemma 2.38). If $T_0 \neq 0$ then $\lambda$ is an eigenvalue of $T_0$ (according to Theorem 2.34(1)) and, consequently, also an eigenvalue of $T$, i.e., $N_k \cap Y \neq \{0\}$ for some $k \in N$ and this is a contradiction to the definition of $Y$. Hence, $T_0 = 0$, i.e., $Y \subset N(T) = N_0$. But this is also a contradiction and (2.26) is proved.

Let $E_k$, $k \in N \cup \{0\}$ denote the orthogonal projection onto $N_k$. Then,

$$x = \sum_{k \in N \cup \{0\}} E_k x \qquad \forall x \in X$$

and

$$Tx = \sum_{k \in N \cup \{0\}} TE_k x = \sum_{k \in N} \lambda_k E_k x.$$

From this, the representation (2.24) follows if we choose orthonormal basis $(e_{k,\ell})_{\ell=1}^{d_k}$ of $N_k$ with $d_k := \dim N_k$ because

$$E_k x = \sum_{\ell=1}^{d_k} (x, e_{k,\ell})_X \, e_{k,\ell}.$$

The representation (2.24) in particular implies that $N_k = N\left((\lambda_k I - T)^2\right)$ because, for $x \in N\left((\lambda_k I - T)^2\right)$ we have

$$0 = (\lambda_k I - T)^2 x = \sum_{j \in N \cup \{0\}} (\lambda_k - \lambda_j)^2 E_j x,$$

i.e., $E_j x = 0$ for $j \neq k$. Thus, $x = E_k x \in N_k$ and we have proved $n_{\lambda_k} = 1$. ■

**Remark 2.40**

1. *Let $X$ be a Hilbert space and let $T \in L(X)$ be selfadjoint, i.e., $T^* = T$. Then, $\sigma_p(T) \subset \mathbb{R}$ and $\|T\|_{X \leftarrow X}$ or $-\|T\|_{X \leftarrow X}$ is an eigenvalue.*
2. *If $T$ is in addition positive semidefinite, i.e., $(Tx, x)_X \geq 0$ for all $x \in X$, then $\sigma_p(T) \subset [0, \infty[$ and $\|T\|_{X \leftarrow X}$ is an eigenvalue.*

**Proof.** For an eigenpair $(\lambda, x)$ it holds

$$\lambda \|x\|_X^2 = (\lambda x, x)_X = (Tx, x)_X = (x, T^* x)_X = (x, Tx)_X = (x, \lambda x)_X = \bar{\lambda} \|x\|_X^2,$$

i.e., $\lambda = \overline{\lambda}$ because $x \neq 0$. The second statement of 1) follows from Lemma 2.38. The assumption in 2) implies

$$\lambda \|x\|_X^2 = (Tx, x)_X \geq 0, \quad \text{i.e., } \lambda \geq 0.$$

■

## 2.9 Sobolev Spaces

In this subsection, we will generalize the classical notion of derivative for certain subspaces of $L^2$ which are denoted as *Sobolev spaces.*

**Definition 2.41**

1. *For $k \in \mathbb{N}$ and $p \in [1, \infty[$ the **Sobolev space** $W^{k,p}(\Omega)$ and its norm is given by*

$$W^{k,p}(\Omega) := \left\{\varphi \in L^p(\Omega) \mid \forall\, |\alpha| \leq k : D^\alpha \varphi \in L^p(\Omega)\right\},$$

$$\|\varphi\|_{k,p} := \left\{ \sum_{|\alpha| \leq k} \|D^\alpha \varphi\|_{L^p(\Omega)}^p \right\}^{1/p}.$$

*Here $D^\alpha \varphi = \partial_1^{\alpha_1} \partial_2^{\alpha_2} \cdots \partial_d^{\alpha_d} \varphi$ denotes the weak derivative of $\varphi$.*

2. *A seminorm on* $W^{k,p}(\Omega)$ *is given by*

$$|\varphi|_{k,p} := \left\{ \sum_{|\alpha|=k} \|D^\alpha \varphi\|^p_{L^p(\Omega)} \right\}^{1/p}.$$

In the case $p = 2$, $W^{k,2}(\Omega)$ is a Hilbert space and we write short $H^k(\Omega)$ instead of $W^{k,2}(\Omega)$ and skip the index "2" for the corresponding norm and seminorm.

**Theorem 2.42**

1. *The space* $W^{k,p}(\Omega)$ *with norm* $\|\cdot\|_{k,p}$ *is a Banach space.*
2. $C^\infty(\Omega) \cap W^{k,p}(\Omega)$ *is dense in* $W^{k,p}(\Omega)$.
3. $H^k(\Omega)$ *is a Hilbert space with scalar product*

$$(\varphi, \psi)_k := \sum_{|\alpha|\le k} \int_\Omega D^\alpha \varphi D^\alpha \psi.$$

In general, $C_0^\infty(\Omega)$ is not dense in $W^{k,p}(\Omega)$. The closure of $C_0^\infty(\Omega)$ with respect to the norm $\|\cdot\|_{k,p}$ defines the Sobolev space with zero boundary conditions in a "weak" sense.

**Definition 2.43** $W_0^{k,p}(\Omega)$ *is the closure of* $C_0^\infty(\Omega)$ *with respect to the* $W^{k,p}(\Omega)$*-norm. We set* $H_0^k(\Omega) = W_0^{k,2}(\Omega)$.

**Definition 2.44** *The domain* $\Omega$ *has a* ***Lipschitz boundary*** *resp.* $\Omega$ *is a* ***Lipschitz domain****, if there exists* $N \in \mathbb{N}$ *and open sets* $U_1, \ldots, U_N \subset \mathbb{R}^d$ *with the following properties:*

1. $\partial\Omega \subset \bigcup_{i=1}^N U_i$,
2. *For any* $1 \le i \le N$*, the intersection* $\partial\Omega \cap U_i$ *can be represented as the graph of a Lipschitz continuous function.*

**Remark 2.45** *Let* $\Omega$ *be a Lipschitz domain. Then, there exists an exterior normal field almost everywhere on* $\partial\Omega$.

A consequence of the trace theorem is the following alternative characterization of $W_0^{1,p}(\Omega)$.

**Theorem 2.46** $W_0^{1,p}(\Omega) = \{\varphi \in W^{1,p}(\Omega) \mid \varphi|_{\partial\Omega} = 0\}$.

For many applications, the **Friedrichs' inequality** is an essential tool for proving existence and uniqueness.

**Theorem 2.47 (Friedrichs' inequality)** $\|\cdot\|_{k,p}$ *and* $|\cdot|_{k,p}$ *define equivalent norms on* $W_0^{k,p}(\Omega)$.

**Theorem 2.48 (Poincaré inequality)** *Let the space dimension satisfy* $d \ge 2$*. Then,* $|\cdot|_1$ *and* $\|\cdot\|_1$ *are equivalent on* $V := \left\{\varphi \in H^1(\Omega) : \int_\Omega \varphi = 0\right\}$.

**Theorem 2.49 (Sobolev's embedding theorem)** *Let* $\Omega \subset \mathbb{R}^d$ *be a Lipschitz domain. For* $s > d/2$*, the embedding*

$$H^s(\Omega) \subset C^0\left(\overline{\Omega}\right)$$

*is continuous.*

## 2.10 Abstract variational problems

We will transform elliptic boundary value problems – as the starting point for their discretization – into (nearly) equivalent *variational problems.* We begin with the functional analytic prerequisites.

**Theorem 2.50 (Lax-Milgram)** *Let $(X, \|\cdot\|_X)$ be a Banach space, $\ell \in X'$ a continuous linear functional and $a : X \times X \to \mathbb{C}$ a continuos sesquilinear form. Furthermore, we assume that $a$ is hermitian:*

$$a(u,v) = \overline{a(v,u)} \qquad \forall u, v \in X$$

*and coercive: There exists $\alpha > 0$ such that*

$$a(u,u) \geq \alpha \|u\|_X^2 \qquad \forall u \in X.$$

*Then, the functional $J \in C^2(X, \mathbb{R})$,*

$$J(u) := \frac{1}{2} a(u,u) - \operatorname{Re} \ell(u),$$

*has a unique minimizer $u^\star \in X$. This minimizer is the unique solution of*

$$a(u^\star, v) = \ell(v) \qquad \forall v \in X. \tag{2.27}$$

The proof can be found in any textbook on functional analysis and is skipped here.

**Theorem 2.51** *The assumptions and notations are as in Theorem 2.50. Let*

$$A := \|a\|_{\mathbb{C} \leftarrow X \times X} := \sup_{u,v \in X \setminus \{0\}} \frac{|a(u,v)|}{\|u\|_X \|v\|_X}.$$

*Let $S \subset X$ be a finite dimensional subspace of $X$. The unique minimizer of $J$ in $X$ resp. $S$ is denoted by $u \in X$ resp. $u_S \in S$.*

*Then:*

$$\|u - u_S\|_X \leq \frac{A}{\alpha} \inf_{v \in S} \|u - v\|_X. \tag{2.28}$$

*Let, in addition, $H$ be a Hilbert space with scalar product $(\cdot,\cdot)_H$ and norm $\|\cdot\|_H$ so that $X$ is continuously and densely embedded in $H$ with respect to the norm $\|\cdot\|_H$. For $\varphi \in H$, let $u_\varphi \in X$ denote the unique solution of*

$$a(v, u_\varphi) = (\varphi, v)_H \qquad \forall v \in X. \tag{2.29}$$

*Then:*

$$\|u - u_S\|_H \leq A \|u - u_S\|_X \sup_{\varphi \in H \setminus \{0\}} \inf_{v \in S} \frac{\|u_\varphi - v\|_X}{\|\varphi\|_H}.$$

**Remark 2.52** *The first part of Theorem 2.51 is known as "Céa's Lemma", the second part is known as "duality argument of Aubin-Nitsche".*

The assumptions of Theorems 2.50 and 2.51 (essentially) restrict the problem class to positive definite bilinear forms and do not cover non-symmetric problems. We will generalize this theorem by weakening these assumptions.

**Theorem 2.53** *Let $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ be Banach spaces and $X \overset{c}{\hookrightarrow} Y$ and $a_0$, $a_1 : X \times X \to \mathbb{C}$ two continuous sesquilinear forms. We assume that the sesquilinear form $a_0$ is hermitian and coercive. Further, we assume for the sesquilinear form $a_1$ that there is a constant $\bar{A} \in \mathbb{R}_{>0}$ such that*

$$a_1(u,v) \leq \bar{A} \|u\|_X \|v\|_Y \qquad \forall u, v \in X. \tag{2.30}$$

*Let $a := a_0 + a_1$, i.e.,*

$$a(u,v) := a_0(u,v) + a_1(u,v) \qquad \forall u, v \in X.$$

*For all $u \in X \setminus \{0\}$, we assume*

$$a(u,u) \neq 0. \tag{2.31}$$

*Then, the problems*

$$a(u,v) = \ell(v) \qquad \forall v \in X \tag{2.32}$$

*and*

$$a(v,u) = \ell(v) \qquad \forall v \in X$$

*have unique solutions for any continuous, linear functionals $\ell \in X'$.*

**Theorem 2.54** *The assumptions and notations are as in Theorem 2.53. Let $S \subset X$ be a finite-dimensional subspace. Then, the problem*

$$a(u_S, v) = \ell(v) \qquad \forall v \in S \tag{2.33}$$

*has a unique solution $u_S \in S$ for any $\ell \in X'$.*

*In addition, let $a$ be coercive, i.e., there exists $\beta > 0$ with $a(u,u) \geq \beta \|u\|_X^2$ for all $u \in X$. Then, the unique solutions $u$ and $u_S$ corresponding to (2.32) and (2.33) satisfy the error estimate*

$$\|u - u_S\|_X \leq \frac{A}{\beta} \inf_{v \in S} \|u - v\|_X, \tag{2.34}$$

*where $A := \|a\|_{\mathbb{C} \leftarrow X \times \mathbb{X}}$. Finally, let $H$, $\varphi$ and $u_\varphi$ be as in Theorem 2.51. Then, the error estimate*

$$\|u - u_S\|_H \leq A \|u - u_S\|_X \sup_{\varphi \in H \setminus \{0\}} \inf_{v \in S \setminus \{0\}} \frac{\|u_\varphi - v\|_X}{\|\varphi\|_H} \tag{2.35}$$

*holds.*

## 2.11 Weak Solutions

Throughout this section, we assume that $\Omega \subset \mathbb{R}^d$ is an open, bounded set with Lipschitz boundary $\Gamma := \partial\Omega$ and exterior normal field $\mathbf{n}$. We will consider scalar, linear, elliptic differential equations of second order. The general form is

$$-\sum_{1 \leq i,j \leq d} \frac{\partial}{\partial x_i} \left( A_{i,j} \frac{\partial u}{\partial x_j} \right) + \sum_{i=1}^{d} b_i \frac{\partial u}{\partial x_i} + cu = f \qquad \text{in } \Omega, \tag{2.36}$$

where the precise assumptions on $f$, $c$, $\mathbf{b} = (b_1, b_2, \ldots, b_d)^\intercal$, and $\mathbf{A} = (A_{i,j})_{i,j=1}^d$ will be formulated in Assumption 2.55 and Definition 2.56.

The differential equation has to be equipped with boundary conditions and we will consider three different types of boundary conditions

- (homogenous) **Dirichlet boundary conditions:** $u = 0$ on $\Gamma$,
- (inhomogenous) **Neumann boundary conditions:** $\langle \mathbf{An}, \operatorname{grad} u \rangle = g$ on $\Gamma$,
- **mixed Dirichlet-Neumann boundary conditions:** $u = 0$ on $\Gamma_D$ and $\langle \mathbf{An}, \operatorname{grad} u \rangle = g$ on $\Gamma_N$.

Here, we assume that $\Gamma_D \cap \Gamma_N = \emptyset$ and $\Gamma = \Gamma_D \cup \Gamma_N$. For mixed boundary conditions we will assume that $\Gamma_D$ has positive $(d-1)$-dimensional measure. The restriction to homogeneous Dirichlet boundary condition is not essential but avoids technical difficulties.

Let $u \in C^2(\Omega)$ be a solution of (2.36) with homogenous Dirichlet boundary conditions and $v \in C_0^\infty(\Omega)$. Multiplication of (2.36) with $v$, integration over $\Omega$ and application of Gauß' integral theorem leads to

$$\begin{aligned} \int_\Omega fv &= -\int_\Omega v \operatorname{div}(\mathbf{A} \operatorname{grad} u) + \int_\Omega \langle \mathbf{b}, \operatorname{grad} u \rangle v + \int_\Omega cuv \\ &= \int_\Omega (\langle \operatorname{grad} v, \mathbf{A} \operatorname{grad} u \rangle + \langle \mathbf{b}, \operatorname{grad} u \rangle v + cuv). \end{aligned} \tag{2.37}$$

Since $C_0^\infty(\Omega)$ is dense in $H_0^1(\Omega)$ it follows that $u \in H_0^1(\Omega)$ satisfies

$$\int_\Omega (\langle \nabla v, \mathbf{A} \nabla u \rangle + \langle \mathbf{b}, \nabla u \rangle v + cuv) = \int_\Omega fv \qquad \forall v \in H_0^1(\Omega). \tag{2.38}$$

Vice versa, the relation (2.37) implies that a solution of (2.38) satisfies the differential equation (2.36) provided it is *sufficiently smooth*, more precisely, is in $C^2(\Omega)$. In this sense, problem (2.38) is equivalent to the differential equation (2.36) with homogeneous Dirichlet boundary conditions.

If we consider in the previous argument also functions $v \in C^\infty(\overline{\Omega})$, then, there arise additional boundary terms $\int_\Gamma \frac{\partial u}{\partial \tilde{\mathbf{n}}} v$ in (2.37), where $\tilde{\mathbf{n}} := \mathbf{An}$. If $u$ satisfies the Neumann boundary conditions we may substitute them in the integrand:

$$\int_\Gamma \frac{\partial u}{\partial \tilde{\mathbf{n}}} v = \int_\Gamma gv.$$

Hence, in this case, we will modify equation (2.38) by the additional term $\int_\Gamma gv$ on the right-hand side. Before we define the *weak solutions*, we will formulate the basic assumptions on the coefficients.

**Assumption 2.55** *The coefficients* $\mathbf{A}$, $\mathbf{b}$, $c$ *in (2.38) satisfy*

*1.*

$$\begin{aligned} &\mathbf{A} \in \mathbf{L}^\infty\left(\Omega, \mathbb{R}^{d\times d}\right) \wedge \forall \mathbf{x} \in \Omega : \mathbf{A}(\mathbf{x}) = \mathbf{A}^\intercal(\mathbf{x}) \\ &0 < a := \inf_{\mathbf{x}\in\Omega} \lambda_{\min}(\mathbf{x}) \le \sup_{\mathbf{x}\in\Omega} \lambda_{\max}(\mathbf{x}) =: A < \infty, \end{aligned}$$

*where* $\lambda_{\min}(\mathbf{x})$ *denotes the smallest eigenvalue of* $\mathbf{A}(\mathbf{x})$ *and* $\lambda_{\max}(\mathbf{x})$ *the largest one.*

*2.*

$$\mathbf{b} \in \mathbf{L}^\infty\left(\Omega, \mathbb{R}^d\right) \wedge \operatorname{div} \mathbf{b} \in L^\infty(\Omega).$$

*3.*

$$c \in L^\infty(\Omega).$$

These considerations lead to the following definition.

**Definition 2.56**

1. *$u \in H_0^1(\Omega)$ is a **weak solution** of the differential equation (2.36) with homogeneous Dirichlet boundary condition if it satisfies*

$$\int_\Omega \langle \mathbf{A}\nabla u, \nabla v\rangle + \langle \mathbf{b}, \nabla u\rangle v + cuv = \int_\Omega fv \qquad \forall v \in H_0^1(\Omega).$$

2. *$u \in H_D^1(\Omega) := \left\{\varphi \in H^1(\Omega) : \varphi|_{\Gamma_D} = 0\right\}$ is a **weak solution** of the differential equation (2.36) with mixed boundary conditions if*

$$\int_\Omega \langle \mathbf{A}\operatorname{grad} u, \operatorname{grad} v\rangle + \langle \mathbf{b}, \operatorname{grad} u\rangle v + cuv = \int_\Omega fv + \int_{\Gamma_N} gv \qquad \forall v \in H_D^1(\Omega).$$

3. *$u \in H^1(\Omega)$ is a **weak solution** of the differential equation (2.36) with Neumann boundary conditions if*

$$\int_\Omega \langle \mathbf{A}\operatorname{grad} u, \operatorname{grad} v\rangle + \langle \mathbf{b}, \operatorname{grad} u\rangle v + cuv = \int_\Omega fv + \int_\Gamma gv \qquad \forall v \in H^1(\Omega).$$

**Theorem 2.57 (existence and uniqueness)**

1. *If $-\frac{1}{2}\operatorname{div}\mathbf{b} + c \geq 0$ is satisfied, then, the differential equation (2.36) with homogeneous Dirichlet boundary conditions has a unique weak solution.*

2. *If $-\frac{1}{2}\operatorname{div}\mathbf{b} + c \geq 0$ and $\langle \mathbf{b}, \mathbf{n}\rangle \geq 0$ on $\Gamma_N$, then, the differential equation (2.36) with mixed boundary conditions has a unique weak solution.*

3. *If $c \geq c_0 > 0$, $-\frac{1}{2}\operatorname{div}\mathbf{b} + c \geq 0$ and $\langle \mathbf{b}, \mathbf{n}\rangle \geq 0$ on $\Gamma$, then, the differential equation (2.36) with Neumann boundary conditions has a unique weak solution.*

4. *If $c = 0$, $-\frac{1}{2}\operatorname{div}\mathbf{b} \geq 0$ and $\langle \mathbf{b}, \mathbf{n}\rangle = 0$ on $\Gamma$ and $\int_\Omega f + \int_\Gamma g = 0$, then, the differential equation (2.36) with Neumann boundary conditions has a unique weak solution $u$ with $\int_\Omega u = 0$.*

The following example shows that the *regularity assumption* $u \in H^2(\Omega)$ for weak solutions can be expected in general only under additional assumptions on the boundary $\Gamma$.

**Example 2.58** *Let $0 < \alpha < 2\pi$ and $\Omega_\alpha$ denote the segment*

$$\Omega_\alpha := \left\{\mathbf{x} = r\begin{pmatrix}\cos\varphi\\ \sin\varphi\end{pmatrix} \in \mathbb{R}^2 : 0 < r < 1, 0 < \varphi < \alpha\right\}.$$

*Define the function $v \in \Omega_\alpha \to \mathbb{R}$ by*

$$v(\mathbf{x}) = r^{\pi/\alpha}\sin\frac{\pi\varphi}{\alpha} \quad \text{with} \quad \mathbf{x} = r(\cos\varphi, \sin\varphi)^\intercal.$$

*Then, for any* $\mathbf{x} \in \Omega_\alpha$,

$$\Delta v(\mathbf{x}) = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial v}{\partial r}\right) + \frac{1}{r^2}\frac{\partial v^2}{\partial \varphi^2} = 0.$$

*Let* $w \in C_0^\infty(\mathbb{R}^2, \mathbb{R})$ *with* $\operatorname{supp} w \subset B\left(0, \frac{2}{3}\right)$ *and* $w = 1$ *on* $\overline{B\left(0, \frac{1}{3}\right)}$.

*Define*

$$u := wv, \qquad f := \Delta\left(v\left(1 - w\right)\right).$$

*Then, there holds*

$$\begin{aligned} -\Delta u &= f \quad && \text{in } \Omega_\alpha, \\ u &= 0 && \text{auf } \partial\Omega_\alpha. \end{aligned}$$

*Obviously, we have* $(1 - w)\, v \in C^\infty(\mathbb{R}^2, \mathbb{R})$ *and, hence,* $f \in C^\infty\left(\overline{\Omega}_\alpha\right)$. *The function* $u$ *satisfies* $u \in C^\infty(\Omega_\alpha)$. *Because of* $u = v$ *in* $B\left(0, \frac{1}{3}\right)$ *we have*

$$u \notin C^\infty\left(\overline{\Omega}_\alpha\right).$$

*An easy calculation shows*

$$u \in C^k\left(\overline{\Omega}_\alpha\right) \Longleftrightarrow 0 < \alpha \leq \frac{\pi}{k}, \qquad k \geq 1$$

*and*

$$D^k u \in L^2(\Omega_\alpha) \Longleftrightarrow 0 < \alpha < \frac{\pi}{k-1}, \qquad k \geq 2.$$

*Hence, for given* $\alpha$, *we can* ***not*** *expect estimates of the form*

$$\|u\|_{C^{k+2}\left(\overline{\Omega}_\alpha\right)} \leq c_k \|f\|_{C^k\left(\overline{\Omega}_\alpha\right)}$$

*and*

$$\|u\|_{H^{k+2}(\Omega_\alpha)} \leq c_k' \|f\|_{H^k(\Omega_\alpha)},$$

*as they would hold for ordinary differential equations.*

**Theorem 2.59 (regularity theorem)** *Let* $\Gamma$ *be a* $C^1$ *manifold or let* $\Omega$ *be convex and* $f \in L^2(\Omega)$. *Besides the Assumption 2.55 we assume that* $c \in C\left(\overline{\Omega}, \mathbb{R}_{\geq 0}\right)$, $\mathbf{b} = (b_1, b_2, \ldots, b_d)^\intercal \in C^1\left(\overline{\Omega}, \mathbb{R}^d\right)$, *and* $\mathbf{A} = (A_{i,j})_{i,j=1}^d \in C^1\left(\overline{\Omega}, \mathbb{R}^{d\times d}\right)$. *In the case of the mixed Neumann problem we assume that there exists a function* $u_g \in H^2(\Omega)$ *so that* $g = u_g|_{\Gamma_N}$.

*Then, the weak solution* $u$ *of the elliptic differential equation with homogeneous or mixed or Neumann boundary conditions satisfies* $u \in H^2(\Omega)$ *and the a-priori estimate*

$$\|u\|_{H^2(\Omega)} \leq c\left\{\|f\|_{L^2(\Omega)} + \|u_g\|_{H^2(\Omega)}\right\}$$

*holds. The constant* $c$ *depends only on* $\Omega$ *and on the coefficients* $c$, $\mathbf{b}$, $\mathbf{A}$ *in the differential equation.*

# 3 Elliptic Eigenvalue Problems and their Discretization

## 3.1 Eigenvalue Problems for Elliptic Partial Differential Operators

In Definition 2.56, we have formulated the elliptic boundary value problem in the abstract variational form: Let $H, U$ two Hilbert spaces where the embedding $H \subset U$ is compact and let a sesquilinear form $a : H \times H \to \mathbb{C}$ and a linear form $\ell \in H'$ be given which satisfy the assumptions of Theorem 2.53. The variational problem then is given by seeking $u \in H$ such that

$$a(u, v) = \ell(v) \qquad \forall v \in H.$$

In the setting of Definition 2.56(1) we have $H = H_0^1(\Omega)$, $U = L^2(\Omega)$, $\ell(v) := (f, v)_U$ for some given $f \in L^2(\Omega)$, and

$$a(u, v) := \int_\Omega \langle \mathbf{A}\nabla u, \nabla v\rangle + \langle \mathbf{b}, \nabla u\rangle v + cuv.$$

In order to formulate a variational eigenvalue problem we assume that a further sesquilinear form $d : H \times H \to \mathbb{C}$ is given which satisfies

$$\begin{array}{ll} \text{a)} & \forall u, v \in H : d(u, v) = \overline{d(v, u)} \\ \text{b)} & \|d\|_{\mathbb{C} \leftarrow H \times H} =: C_d < \infty \\ \text{c)} & \forall u \in H \setminus \{0\} : d(u, u) > 0 \\ \text{d)} & \forall (u_j)_{j\in\mathbb{N}} \subset H \text{ with } \|u_j\|_H \le C \\ & \text{there exists } (u_{j_k})_{k\in\mathbb{N}} \text{ which is Cauchy w.r.t. } d(\cdot,\cdot)^{1/2}. \end{array} \tag{3.1}$$

The variationally formulated eigenvalue problem is given by

$$\text{find pairs } (\lambda, u) \in \mathbb{C} \times H \setminus \{0\} \quad \text{such that} \quad a(u, v) = \lambda d(u, v) \quad \forall v \in H. \tag{3.2}$$

Next, we will formulate this problem in terms of a compact operator. Theorem 2.53 implies that for any $f \in H$, the problem:

$$\text{find } u \in H : \quad a(u, v) = d(f, v) \quad \forall v \in H \tag{3.3}$$

has a unique solution. Hence, we may define the *solution operator* $T : H \to H$ by

$$a(Tf, v) = d(f, v) \qquad \forall f, v \in H. \tag{3.4}$$

**Lemma 3.1** *Let $a = a_0 + a_1$ satisfies the assumption as in Theorem 2.53 and assume that $d$ satisfies (3.1). Then, the operator $T$ is compact.*

**Proof.** The Lax-Milgram lemma implies that the problem

$$\text{find } u_f \in H \qquad \text{such that } a_0(u_f, v) = d(f, v)$$

has a unique solution for all $f \in H$. The solution operator is denoted by $K_d$, i.e., $a_0(K_d, v) = d(f, v)$ for all $f, v \in H$. The operator $K_d$ is compact. In a similar fashion one shows that the is a compact operator which satisfies

$$a_0(K_1 f, v) = a_1(f, v) \qquad \forall f, v \in H.$$

Hence, (3.3) can be rewritten in the form

$$(I + K_1)\, u = K_d f.$$

Since the homogenous equation $(I + K_1)\, w = 0$ has only the trivial solution (due to (2.31)) we may apply the Fredholm alternative to see that $T = (I + K_1)^{-1} K_d$ is a continuous operator in $H$. Because $K_d$ is compact, Lemma 2.20 implies that $T$ compact. ■

Note that (3.3) is equivalent to

$$\text{find } (\mu, u) \in \mathbb{C} \times H \setminus \{0\} : Tu = \mu u \tag{3.5}$$

in the sense of

$$(\lambda, u) \text{ is an eigenpair of (3.2)} \iff \left(\frac{1}{\lambda}, u\right) \text{ is an eigenpair of (3.5).}$$

## 3.2 Galerkin Finite Element Method for Eigenvalue Problems

We assume that the reader has basic knowledge in the finite element method and recall here only the main steps for the construction of finite element spaces.

### 3.2.1 Construction of Finite Element Spaces

Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with piecewise smooth boundary $\Gamma := \partial\Omega$. Let $\mathcal{T} = \{\tau_i : 1 \le i \le q\}$ denote a finite element triangulation[4] where all elements are regular images of the $d$-dimensional unit simplex $\widehat{\tau}$.

**Assumption 3.2**

1. *The elements $\tau \in \mathcal{T}$ are closed subsets of $\Omega$ with pairwise disjoint interior and $\overline{\Omega} = \bigcup_{\tau \in \mathcal{T}} \tau$.*

2. *The triangulation $\mathcal{T}$ has no hanging nodes.*

3. *The element maps of elements sharing edges, faces or higher-dimensional simplices at their surface induce the same parametrization on that edge, face, or higher-dimensional analogon.*

4. *Let $h_\tau := \operatorname{diam} \tau$ and let $h_{\mathcal{T}} := \max\{h_\tau, \tau \in \mathcal{T}\}$ denote the mesh width. Any $\tau \in \mathcal{T}$ is the image of the $d$-dimensional unit simplex, i.e., $\tau = F_\tau \widehat{\tau}$. Each element map $F_\tau$ can be written as $F_\tau = R_\tau \circ A_\tau$ where $A_\tau$ is an affine map and the maps $R_\tau$, $A_\tau$ satisfy for constants $C_{\text{affine}}$, $C_{\text{metric}} > 0$ independent of $h_\tau$*

$$\|A'_\tau\|_\infty \le C_{\text{affine}} h_\tau, \qquad \left\|(A'_\tau)^{-1}\right\|_\infty \le C_{\text{affine}} h_\tau^{-1}, \qquad \left\|(R'_\tau)^{-1}\right\|_{L^\infty(A_\tau(\widehat{\tau}))} \le C_{\text{metric}}.$$

[4] We use the notation "triangulation" independent of the spatial dimension $d$ and not only for the case $d = 2$.

**Remark 3.3** *Triangulations satisfying Assumption 3.2 can be obtained by patchwise construction of the mesh: Let $\mathcal{T}^{\text{macro}}$ be a fixed triangulation (with possibly curved elements) with element maps which resolve the geometry. If the finer triangulation $\mathcal{T}$ is obtained by quasi-uniform refinements of the reference element $\widehat{\tau}$ and by mapping the subdivisions of the reference element with the macro element maps, then, the resulting element maps satisfy Assumption 3.2.*

Finite element spaces are composed by local polynomials and are subject to some global smoothness and boundary conditions. Let $\mathbb{P}_p$ denote the space of polynomials in $d$ variables of total degree $p$. Then the finite element space for the mesh $\mathcal{T}$ and polynomial degree $p$ is given by

$$S := S_{\mathcal{T}}^p := \left\{ u \in C^0\left(\overline{\Omega}\right) \mid \forall \tau \in \mathcal{T} : u|_\tau \circ F_\tau \in \mathbb{P}_p \right\} \cap H.$$

### 3.2.2 Galerkin Discretization

The Galerkin discretization of the eigenvalue problem is given by

$$\text{find } \left(\lambda^S, u^S\right) \in \mathbb{C} \times S \setminus \{0\} \text{ such that } a\left(u^S, v\right) = \lambda^S d\left(u^S, v\right) \qquad \forall v \in S. \tag{3.6}$$

By introducing the finite element basis $\varphi_i^S$, $1 \le i \le N$, for $S$, this system can be transformed to a generalized algebraic eigenvalue problem of the form

$$\text{find } \left(\lambda^S, \mathbf{u}^S\right) \in \mathbb{C} \times \mathbb{C}^N \setminus \{0\} \text{ such that } \mathbf{A}\mathbf{u}^S = \lambda^S \mathbf{M}\mathbf{u}^S, \tag{3.7}$$

where

$$\mathbf{A} := \left(a\left(\varphi_j^S, \varphi_i^S\right)\right)_{1 \le i,j \le N}, \quad \mathbf{M} := \left(d\left(\varphi_j^S, \varphi_i^S\right)\right)_{1 \le i,j \le, N}$$

and the equivalence

$$\left(\lambda^S, \mathbf{u}^S\right) \in \mathbb{C} \times \mathbb{C}^N \setminus \{0\} \text{ is an eigenpair of (3.7)} \iff \left(\lambda^S, P\mathbf{u}^S\right) \in \mathbb{C} \times S \setminus \{0\} \text{ is an eigenpair of (3.6)},$$

where the prolongation $P : \mathbb{C}^N \to S$ is given, for $\mathbf{u} = (u_i)_{i=1}^N \in \mathbb{C}^N$, by

$$P\mathbf{u} = \sum_{i=1}^N u_i \varphi_i^S.$$

In these notes we do not discuss the efficient solution of the algebraic eigenvalue problem (3.7) but refer to the lectures of D. Kressner instead.

### 3.2.3 Approximation Properties of Finite Element Spaces

The study of approximation properties of finite elements is a standard topic in any mathematical course on finite elements. Let $H$ denote the space which is employed for the variational formulation of second order elliptic partial differential operators, e.g.., $H = H_0^1(\Omega)$ in the case described in Definition 2.56(1). Let $S = S_{\mathcal{T}}^p$ denote a conforming finite element space, i.e., $S \subset H$ and assume that the solution of the boundary value problem is in a more regular space $W \subset H$, where the subspace $W$, typically, is a higher order Sobolev space or a weighted Sobolev space. In this case one constructs an explicit interpolation operator $\Pi_S : W \to S$ and proves an estimate of the form

$$\|u - \Pi_S u\|_H \le C_u h_{\mathcal{T}}^\alpha,$$

where $h_{\mathcal{T}}$ denotes the mesh width of the triangulation $\mathcal{T}$ and the maximal value of $\alpha \in ]0, p]$ is related to the smoothness of the solution. We always assume that the triangulation $\mathcal{T}$ satisfies Assumption 3.2 and all constants below may depend on the constants $C_{\text{affine}}$, $C_{\text{metric}}$ introduced therein.

**Proposition 3.4** *Let $\Omega$ be a Lipschitz domain with piecewise analytic boundary and let $\mathcal{T}$ be a triangulation of $\Omega$ which satisfies Assumption 3.2.*

*a. Let $t > 1$ be such that the embedding $H^t(\Omega) \hookrightarrow C^0(\overline{\Omega})$ is continuous (cf. Theorem 2.49). Then there exists a continuous interpolation $I_{\mathcal{T}}^p : H^t(\Omega) \to S_{\mathcal{T}}^p$ with*

$$\|\varphi - I_{\mathcal{T}}^p \varphi\|_{H^s(\Omega)} \leq C\, h_{\mathcal{T}}^{\min\{t,p+1\}-s} \|\varphi\|_{H^t(\Omega)}, \qquad s \in \{0,1\}, \tag{3.8}$$

*where the constant $C$ only depends on $p$ and on the constants $C_{\text{affine}}$, $C_{\text{metric}}$ from Assumption 3.2.*

*b. Let $0 \leq s \leq t \leq 1$. Then, there exists a continuous operator $Q_{\mathcal{T}} : H^t(\Gamma) \to S_{\mathcal{T}}^p$ such that, for every $\varphi \in H^t(\Omega)$, we have*

$$\|\varphi - Q_{\mathcal{T}}\varphi\|_{H^s(\Omega)} \leq C h_{\mathcal{T}}^{t-s} \|\varphi\|_{H^t(\Omega)}.$$

*The operator $Q_{\mathcal{T}}$ is stable for $0 \leq s \leq 1$*

$$\|Q_{\mathcal{T}}\|_{H^s(\Omega) \leftarrow H^s(\Omega)} \leq C.$$

From this theorem it becomes clear, that the investigation of the regularity properties of the exact solution is essential for proving a priori error estimates for finite element approximations.

The development of a regularity theory for eigenvalue problems which guarantees the smoothness of the eigenfunctions in terms of Sobolev spaces is beyond the scope of these lecture notes. Here, we will state some relevant results and give links to the literature.

We consider the eigenvalue problem (3.2) and always assume that the setting is as described in Definition 2.56(1) and that Assumption 2.55 and the assumptions of Theorem 2.57 and (3.1) are satisfied.

Let $(\lambda, u) \in \mathbb{C} \times H_0^1(\Omega)$ (with normalization $\|u\|_{H^1(\Omega)} = 1$) denote an eigenpair of (3.2) and consider $\lambda$ in what follows as a fixed parameter. Our assumptions on the bilinear form imply $\lambda \neq 0$, so that the function $u$ is the unique solution of the elliptic equation

$$\begin{aligned} -\operatorname{div}(A \operatorname{grad} u) + \langle b, \operatorname{grad} u\rangle + cu &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

with $f := \lambda u$. In order to prove smoothness of $u$ in higher order Sobolev spaces we have to formulate corresponding smoothness assumptions on the data, i.e., on $\Omega$, $A, b$, and $c$.

**Assumption 3.5**

*a. $\Omega$ is a bounded domain of class $C^{1,1}$.*

*b. The diffusion matrix satisfies $A \in \mathbf{C}^{0,1}\left(\overline{\Omega}, \mathbb{R}_{\text{sym}}^{d\times d}\right)$,*

*c. The coefficients $b, c$ satisfy $b \in \mathbf{L}^\infty\left(\Omega, \mathbb{R}^d\right)$ and $c \in L^\infty(\Omega)$.*

**Theorem 3.6** *Let Assumptions 2.55 and 3.5, the assumptions of Theorem 2.57 and (3.1) be satisfied. Assume that $d(\cdot,\cdot)^{1/2}$ is uniformly equivalent to the $L^2(\Omega)$-norm. Then, there exists a constant $C_{\mathrm{reg}}$ which only depends on the data but not on $\lambda$ such that*

$$\|u\|_{H^2(\Omega)} \leq C_{\mathrm{reg}}\sqrt{\lambda}.$$

This theorem follows from, e.g., from [14, Theorem 2.4.2.7]. We see that the smoothness properties of the eigenfunctions are linked to the smoothness of the coefficients and the domain. By raising the smoothness assumptions on the data in an appropriate way one can prove *shift theorems*, i.e., regularity of the eigenfunctions in higher order Sobolev spaces (see, e.g., [15, Theorem 11.2.22]). We do not go into the details here but end up this section by an example which shows that eigenvalue problems in some cases have better regularity behavior as predicted by the regularity theory for elliptic problems.

Recall that we have constructed an example (cf. Example 2.58 for the segment $\Omega_\alpha$), where the behavior of the solution $u$ for a smooth (!) right-hand side at the origin is given by

$$r^{\pi/\alpha}\sin\frac{\pi\varphi}{\alpha}.$$

This implies, e.g. for $\alpha = 3/2\pi$, that $u \in H^s\left(\Omega_{\frac{3}{2}\pi}\right)$ for any $s < 1 + \frac{2}{3}$ but $u \notin H^2\left(\Omega_{\frac{3}{2}\pi}\right)$.

In the following example we show that the corresponding eigenvalue problem can have a better regularity in some cases.

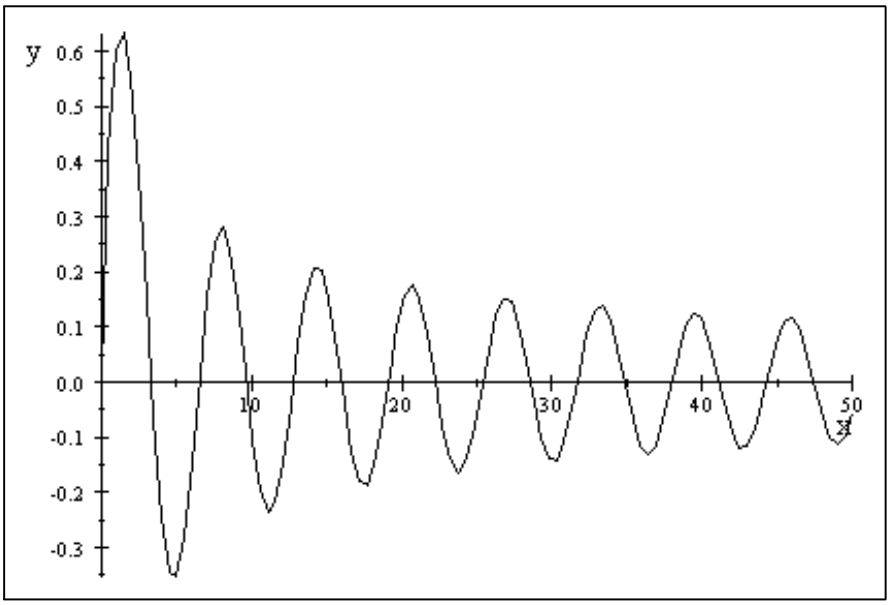
**Example 3.7** *Let $0 < \alpha < 2\pi$ and $\Omega_\alpha$ denote again the segment*

$$\Omega_\alpha := \left\{\mathbf{x} = r\begin{pmatrix}\cos\varphi\\ \sin\varphi\end{pmatrix} \in \mathbb{R}^2 : 0 < r < 1, 0 < \varphi < \alpha\right\}.$$

*We consider the eigenvalue problem*

$$\begin{array}{rl} -\Delta u = \lambda u & \text{in } \Omega_\alpha,\\ u = 0 & \text{on } \partial\Omega_\alpha. \end{array} \tag{3.9}$$

*Let $J_\nu$ denote the Bessel functions of order $\nu$ (cf. [1, Sec. 9]). For every $k \in \mathbb{N}_{\geq 1}$, the function $J_{\frac{\pi k}{\alpha}}(\lambda)$*




*Bessel function $J_{2/3}(\lambda)$.*

*has infinity many zeroes $\lambda_{k,i}$, $i \in \mathbb{N}$ (cf. Fig. 3.7). The eigenvalues of (3.9) are $\lambda_{k,i}$ with corresponding eigenfunction (in polar coordinates)*

$$\hat{e}_{k,i}(r,\varphi) = J_{\frac{\pi k}{\alpha}}(\lambda_{k,i} r) \sin \frac{\pi k}{\alpha} \varphi.$$

*Again, we choose $\alpha = \frac{3}{2}\pi$ and obtain*

$$\hat{e}_{k,i}(r,\varphi) = J_{\frac{2}{3}k}(\lambda_{k,i} r) \sin\left(\frac{2}{3} k \varphi\right).$$

*For small arguments we have*

$$J_{\frac{2}{3}k}(\lambda_{k,i} r) \sim \frac{(\lambda_{k,i} r)^{\frac{2}{3}k}}{\Gamma\left(\frac{2}{3}k + 1\right)}.$$

*Hence, we see that, e.g. for $k = 1$, the corresponding eigenfunctions have the same (low) regularity as the solution of the elliptic problem, while, for $k = 3$, $J_2(\lambda_{k,i} r)$ is smooth.*

# 4 Error Analysis for the Selfadjoint Eigenvalue Problem

## 4.1 Setting

In order to apply the perturbation theory for compact operators we reformulate the discrete eigenvalue problem as a discrete version of (3.5). Let $T_S : S \to S$ denote the solution operator for the problem

$$\text{for given } f \in S \text{ find } u^S \in S \text{ such that } a\left(u^S, v\right) = d(f, v) \qquad \forall v \in S,$$

i.e., $u_S = T_S f$. Let $Q_S : H \to S$ denote the Galerkin projection, i.e.,

$$a(Q_S u, v) = a(u, v) \qquad \forall v \in S.$$

Note that then, $T_S = Q_S T|_S$. It is easy to see that the eigenvalue problem (3.6) is equivalent to

$$\text{find } \left(\mu^S, u^S\right) \in \mathbb{C} \times S \setminus \{0\} \qquad T_S u^S = \mu^S u^S \tag{4.1}$$

in the sense that

$$\left(\lambda^S, u^S\right) \text{ is an eigenpair of (3.6)} \iff \left(\frac{1}{\lambda^S}, u^S\right) \text{ is an eigenpair of (4.1).}$$

The error analysis for symmetric bilinear forms is significantly simpler as for the non-symmetric case. In this section, we will restrict to bilinear forms which satisfy the assumptions of the Lax-Milgram lemma (Theo. 2.50): For a real Hilbert space $H$, let $a : H \times H \to \mathbb{R}$ be a sesquilinear form which satisfies for positive constants $\alpha, C_c > 0$

$$a(u, v) = a(v, u) \qquad \forall u, v \in H, \tag{4.2a}$$

$$\|a\|_{\mathbb{R} \leftarrow H \times H} = C_c < \infty. \tag{4.2b}$$

$$a(u, u) \geq \alpha \|u\|_H^2 \qquad \forall u \in H. \tag{4.2c}$$

We assume that the norm in $H$ is chosen as $\|u\|_H := a(u,u)^{1/2}$ for all $u \in H$. Let a further bilinear form $d : H \times H \to \mathbb{R}$ be given which satisfies condition (3.1). In this section, we consider the eigenvalue problem

$$\text{find pairs } (\lambda, u) \in \mathbb{R} \times H \setminus \{0\} \quad a(u,v) = \lambda d(u,v) \quad \forall v \in H. \tag{4.3}$$

From Remark 2.40, we conclude that all eigenvalues of (4.3) are tacitly real because the symmetry of $a$ and $d$ implies that $T$ – the solution operator (cf. (3.4)) for $a, d$ which satisfy (4.2) and (3.1) – is selfadjoint:

$$a(Tf, v) = d(f, v) = d(v, f) = a(Tv, f) = a(f, Tv) \qquad \forall v, f \in H.$$

Since $T$ is compact, Theorem 2.34 implies that $\sigma(T) \setminus \{0\}$ consists of countably many (finitely or infinitely many) eigenvalues with 0 as the only possible accumulation point. Note that $T$ is positive definite

$$a(Tu, u) = d(u,u) \overset{(3.1)}{>} 0 \qquad \forall u \in H \setminus \{0\}. \tag{4.4}$$

Hence all eigenvalues of $T$ are positive and we order them according to

$$\mu_1 > \mu_2 > \ldots > 0.$$

Let $k_j$ denote the multiplicity of $\mu_j$ and let $N_j$ denote the corresponding eigenspace

$$N_j := \operatorname{span}\{u_{j,\ell} : 1 \le \ell \le k_j\},$$

where the $u_{j,\ell}$ are pairwise orthonormal and satisfy $Tu_{j,\ell} = \mu_j u_{j,\ell}$. Finally, we introduce the space

$$N_{1,j} := \bigoplus_{\ell=1}^{j} N_j.$$

Let $S \subset H$ denote a finite dimensional subspace of dimension $n := \dim S$. We apply the Galerkin discretization (3.6) corresponding to this subspace. The key rôle for the error analysis of the *eigenvalue* approximations plays their characterization via the Rayleigh-Ritz method. We order the spectrum of $T_S$ also in a decreasing way by taking into account their multiplicities and group them according to the corresponding multiplicities of the exact eigenvalues

$$\underbrace{\mu^S_{1,1} \ge \mu^S_{1,2} \ge \ldots . \mu^S_{1,k_1}}_{k_1 \text{ eigenvalues}} \ge \underbrace{\mu^S_{2,1} \ge \mu^S_{2,2} \ge \ldots \ge \mu^S_{2,k_2}}_{k_2 \text{ eigenvalues}} \ge \ldots \ge \underbrace{\mu^S_{m,1} \ge \ldots \ge \mu^S_{m,\tilde{k}_m}}_{\tilde{k}_m \text{ eigenvalues}} > 0,$$

i.e., the last group contains a number $\tilde{k}_m$ of discrete eigenvalues which differ from the multiplicity $k_m$ of the exact eigenvalue $\lambda_m$ only for the case that the dimension $n$ of $S$ does **not** satisfy $\dim S = \sum_{\ell=1}^{m} k_j$. We set

$$\forall 1 \le i < m : \quad \mu^S_i := \mu^S_{i,k_i} \quad \text{and} \quad \mu^S_m := \mu^S_{m,\tilde{k}_m}$$

and define the index sets

$$\forall 1 \le j < m : \quad J_j := \{(j,\ell) : 1 \le \ell \le k_j\} \quad \text{and} \quad J_m := \left\{(m,\ell) : 1 \le \ell \le \tilde{k}_m\right\}$$

$$J_S := \bigcup_{j=1}^{m} J_j.$$

According to Theorem 2.39 we may choose for every eigenvalue $\mu_{j,\ell}^S$ a function $u_{j,\ell}^S \in S$ such that $\left\{u_{j,\ell}^S : (j,\ell) \in J_S\right\}$, forms an orthonormal (w.r.t. the $a(\cdot,\cdot)$ scalar product) basis of $S$ and $Tu_{j,\ell}^S = \mu_{j,\ell}^S u_{j,\ell}^S$. We set

$$N_j^S := \operatorname{span}\left\{u_{j,\ell}^S : (j,\ell) \in J_j\right\}.$$

Note that the definition of the spaces $N_j^S$, in general, depends on the ordering of the eigenvectors if the multiplicity of some eigenvalue is larger than 1. Finally we set

$$N_{1,j}^S := \bigoplus_{\ell=1}^{j} N_\ell^S.$$

From the viewpoint of numerical discretization, the choice and the (precise) dimension of the finite element space $S$, typically, is independent of the multiplicities of the exact eigenvalues because these are not known beforehand. Once $S$ is chosen, the natural question is how well the $n$ eigenvalues and eigenvectors of (3.6) resp. (3.7) approximate the continuous ones. For this *error analysis*, we define $m = m(n)$, $n := \dim S$, as the smallest integer such that the sum of the multiplicities satisfy

$$\sum_{j=1}^{m} k_j \geq n \tag{4.5}$$

and set

$$\forall 1 \leq j < m \qquad \tilde{N}_{1,j} := N_{1,j} \quad \text{and} \quad \tilde{N}_{1,m} := \tilde{N}_m \oplus \left(\bigoplus_{\ell=1}^{m-1} N_\ell\right), \tag{4.6}$$

where $\tilde{N}_m := \operatorname{span}\left\{u_{m,\ell} : 1 \leq \ell \leq \tilde{k}_m\right\}$. Also here the choice of $\tilde{N}_m$ depends on the ordering of the eigenvectors corresponding to eigenvalue $\lambda_m$ if $\tilde{k}_m \neq k_m$.

For comparing the closeness of a subspace $V \subset H$ to a subspace $W \subset H$ we will use the quantity

$$\Theta(V,W) := \max_{v \in V \setminus \{0\}} \frac{\operatorname{dist}(v,W)}{\|v\|_H}. \tag{4.7}$$

**Remark 4.1** *Note that for spaces $U, V \subset H$ of same dimension $\dim U = \dim V = n < \infty$ (as for $\tilde{N}_{1,m}$ and $S$) it holds*

$$\Theta(U,V) = \Theta(V,U).$$

**Proof.** In any case we have $\Theta(U,V) \leq 1$. Let $P_U : H \to U$ and $P_V : H \to V$ denote the $H$-orthogonal projections onto $U$ resp. $V$. If $\Theta(U,V) = 1$ we have

$$\max_{u \in U \setminus \{0\}} \frac{\|u - P_V u\|_H}{\|u\|_H} = 1$$

and hence there exists some $u \perp V$. For $n = 1$, it is obvious that $\Theta(U,V) = \Theta(V,U)$ and we assume $n > 1$ in the following. Let $u^\perp := \{w \in U : a(w,u) = 0\}$. It is clear that $\dim u^\perp = n-1$ and, hence, we may choose some $\tilde{v} \in V \setminus \{0\}$ with $\tilde{v} \perp u^\perp$. This function $v$ satisfies

$$\max_{v \in V \setminus \{0\}} \frac{\|v - P_U v\|_H}{\|v\|_H} \geq \frac{\inf_{u \in U} \|\tilde{v} - u\|_H}{\|\tilde{v}\|_H} = \frac{\inf_{u \in u^\perp} \|\tilde{v} - u\|_H}{\|\tilde{v}\|_H} = 1$$

and, thus, $\Theta(V,U) = 1$.

It remains to consider the statement for the case $\Theta(U,V) < 1$ and $\Theta(V,U) < 1$ which follows from [17, Lemma 2.21]. ■

## 4.2 Estimates for Eigenvalues of Selfadjoint Operators

In the remaining part of the chapter we consider the following setting.

**Assumption 4.2** *The bilinear forms $a, d$ satisfy (4.2), (3.1) and $T$ is defined by (3.4) (so that $T$ is compact, self-adjoint and positive definite (cf. (4.4))).*

**Definition 4.3 (Ritz value)** *Let*

$$\mu_1 := \sup_{u \in H \setminus \{0\}} \frac{d(u,u)}{a(u,u)}$$

*and, recursively, for $n = 1, 2, ...$*

1. *let*

$$\mu_n := \sup_{\substack{v \in H: \quad a(v,v)=1 \\ a(v,u_1)=0 \\ a(v,u_2)=0 \\ \vdots \\ a(v,u_{n-1})=0}} d(v,v) \tag{4.8}$$

2. *if there is an element $v \in H$ which satisfies*

$$\begin{aligned} a(v,u_1) = a(v,u_2) = \ldots = a(v,u_{n-1}) = 0 \\ a(v,v) = 1 \quad \textit{and} \quad d(v,v) = \mu_n \end{aligned}$$

*we set $u_n := v$.*

**Theorem 4.4** *The recursive Definition 4.3 leads to a nonincreasing sequence of eigenvalues $\mu_1 \geq \mu_2 \geq \ldots$ and eigenvectors $u_1, u_2, \ldots$ which terminates if and only if for some $n$ the supremum in (4.8) is not attained or $H$ has dimension $n-1$.*

**Proof. a)** We first prove: If there is an element $u_1 \in H$ for which $\mu_1 a(u_1,u_1) = d(u_1,u_1)$ then

$$Tu_1 = \mu_1 u_1.$$

Note that the quadratic functional $\gamma(u,v) := \mu_1 a(u,v) - d(u,v)$ is symmetric, bilinear, and positive semidefinite

$$\gamma(v,v) \geq 0$$

and, hence, $\gamma(v,v)^{1/2}$ defines a seminorm on $H$. Schwarz's inequality implies

$$|\mu_1 a(v,u_1) - d(v,u_1)|^2 = \gamma(v,u_1)^2 \leq \gamma(v,v)\,\gamma(u_1,u_1) = 0 \qquad \forall v \in H$$

because $\gamma(u_1,u_1) = \mu_1 a(u_1,u_1) - d(u_1,u_1) = 0$. We set $v = \mu_1 u_1 - Tu_1$ and obtain

$$a(\mu_1 u_1 - Tu_1, \mu_1 u_1 - Tu_1) = a(v, \mu_1 u_1 - Tu_1) = \mu_1 a(v,u_1) - d(v,u_1) = 0$$

so that $\mu_1 u_1 = Tu_1$. By multiplying $u_1$ by $\|u_1\|_H^{-1}$ we obtain $a(u_1,u_1) = 1$.

**b)** Since the eigenspaces are orthogonal (cf. (2.25)) we may define the Hilbert space $V_1 := (\text{span}\{u_1\})^\perp := \{v \in H : a(v, u_1) = 0\}$. We now have

$$\mu_2 := \sup_{v \in V_1} \frac{d(v,v)}{a(v,v)}.$$

This time we define $\gamma_2 : V_1 \times V_1 \to \mathbb{R}$ by $\gamma_2(u,v) := \mu_2 a(u,v) - d(u,v)$ and the restriction to $V_1$ ensures that $\gamma_2$ is positive semindefinite on $V_1$. If there is some $u_2$ such that $\gamma_2(u_2, u_2) = 0$ then $u_2$ satisfies the equation $\gamma(v, u_2) = 0$ for all $v \in V_1$. Hence,

$$a(v, Tu_2 - \mu_2 u_2) = d(v, u_2) - \mu_2 a(v, u_2) = -\gamma_2(v, u_2) = 0 \qquad \forall v \in V_1. \tag{4.9}$$

Now

$$\begin{aligned} a(Tu_2 - \mu_2 u_2, u_1) &= a(Tu_2, u_1) - \mu_2 a(u_2, u_1) = a(u_2, Tu_1) - \mu_2 a(u_2, u_1) \\ &= (\mu_1 - \mu_2)\, a(u_2, u_1) = 0 \end{aligned}$$

since $u_2 \in V_1$ (cf. (2.25)). Hence, (4.9) is valid for all $v \in H$ and we may choose $v = Tu_2 - \mu_2 u_2$ to see that $Tu_2 = \mu_2 u_2$.

**c)** For $n > 2$, the statement of the theorem follows by induction. ■

If the multiplicity of an eigenvalue, say, $\mu_\ell$ is $\kappa_\ell > 1$, the Ritz value $\mu_\ell$ occurs several times, i.e., $\mu_\ell = \mu_{\ell+1} + ... + \mu_{\ell+\kappa_\ell-1}$, with different corresponding eigenvectors $u_\ell$, $u_{\ell+1}, \ldots u_{\ell+\kappa_\ell-1}$. Note that any linear combination of these eigenvectors is again an eigenvector for $\mu_\ell$.

To show that the definition of the Ritz values (Def. 4.3) gives us all the eigenvalues down to the point where they terminate we prove the following result.

**Theorem 4.5** *There are no eigenvalue of $T$ above $\mu_1$ and if $\mu$ is an eigenvalue that satisfies $\mu > \mu_n$ for some $n$, then $\mu = \mu_k$ for some $k < n$, and the corresponding eigenvector is a linear combination for those eigenvectors $u_i$ which correspond to the eigenvalue $\mu_k$.*

**Proof.** We first observe that if $Tu = \mu u$ then $d(u,u) = \mu a(u,u)$. Hence, by definition of $\mu_1$, $\mu \le \mu_1$. Now if $\mu > \mu_n$ for some $n$, then there is a $k \le n-1$ such that

$$\mu_{k+1} < \mu \le \mu_k.$$

If $\mu < \mu_k$ we see from (2.25) that $a(u, u_1) = ... = a(u, u_k) = 0$ and, hence, by the definition of the Ritz values that $\mu \le \mu_{k+1}$, which gives a contradiction. Therefore $\mu = \mu_k$.

If $u$ is not a linear combination for those $u_i$ which correspond to $\mu_k$, we can produce a linear combination of $u$ and these $u_i$ which is orthogonal to $u_1, \ldots, u_k$ and which is again an eigenvector corresponding to $\mu$. But this would again make $\mu \le \mu_{k+1}$ and thus give a contradiction. ■

An important monotonicity principle for the eigenvalue problem (3.6) is stated in the next theorem.

**Theorem 4.6 (Rayleigh-Ritz method)** *Let $\mu_1 \ge \mu_2 \ge \ldots \ge$ be the Ritz values for the Rayleigh quotient $d(u,u)/a(u,u)$ in $H$. Let $\varphi_i^S$, $1 \le i \le k$, denote $k$ linear independent vectors in $H$. Let the eigenvalues of the matrix eigenvalue problem (3.7) be*

$$\mu_1' \ge \mu_2' \ge \ldots \ge \mu_k'.$$

*Then*

$$\mu_i' \le \mu_i \qquad \forall 1 \le i \le k.$$

This theorem is a consequence of the *first monotonicity principle* and the *Poincaré principle* which we will state and prove first.

**Theorem 4.7 (Poincaré principle)** *Let the eigenvalues $\mu_1 \geq \mu_2 \geq \dots$ be the Ritz values as in Definition 4.3 with the convention that if $\mu_n$ is not attained, we put $\mu_n = \mu_{n+1} = \mu_{n+2} = \cdots$. Then*

$$\mu_n = \sup_{\substack{S=\operatorname{span}\left\{\varphi_i^S : 1\leq i\leq n\right\}\subset H \\ \dim S=n}} \min_{\mathbf{c}=(c_i)_{i=1}^n \subset \mathbb{R}^n\setminus\{0\}} \frac{d\left(\sum_{i=1}^n c_i\varphi_i^S, \sum_{i=1}^n c_i\varphi_i^S\right)}{a\left(\sum_{i=1}^n c_i\varphi_i^S, \sum_{i=1}^n c_i\varphi_i^S\right)}. \tag{4.10}$$

**Proof.** Let $S = \operatorname{span}\left\{\varphi_i^S : 1 \leq i \leq n\right\} \subset H$ with $\dim S = n$. Then, there exists at least one nontrivial linear combination $\varphi = \sum_{i=1}^n b_i\varphi_i^S$ which satisfies the $n-1$ orthogonality condition $a(\varphi, u_1) = \dots a(\varphi, u_{n-1}) = 0$. Hence, by Definition 4.3

$$\frac{d\left(\sum_{i=1}^n b_i\varphi_i^S, \sum_{i=1}^n b_i\varphi_i^S\right)}{a\left(\sum_{i=1}^n b_i\varphi_i^S, \sum_{i=1}^n b_i\varphi_i^S\right)} \leq \mu_n. \tag{4.11}$$

(If the sequence of eigenvalues $\mu_1 \geq \mu_2 \geq \cdots$ terminates at $\mu_k$ for some $k < n$ so that the values $\mu_{k+1} = \cdots = \mu_n$ is not attained, we reach the same conclusion by making $\varphi$ orthogonal to $u_1, \cdots, u_k$.)

Since the $\varphi_i^S$ are linearly independent, the set of $c_i$, for which $a\left(\sum_{i=1}^n c_i\varphi_i^S, \sum_{i=1}^n c_i\varphi_i^S\right) = 1$ holds, is closed and bounded. Hence, $a\left(\sum_{i=1}^n c_i\varphi_i^S, \sum_{i=1}^n c_i\varphi_i^S\right)$ takes on its minimum on this set. We see from (4.11) that

$$\min_{\mathbf{c}=(c_i)_{i=1}^n \subset \mathbb{R}^n\setminus\{0\}} \frac{d\left(\sum_{i=1}^n c_i\varphi_i^S, \sum_{i=1}^n c_i\varphi_i^S\right)}{a\left(\sum_{i=1}^n c_i\varphi_i^S, \sum_{i=1}^n c_i\varphi_i^S\right)} \leq \mu_n$$

for any set of linear independent elements $\varphi_i^S$, $1 \leq i \leq n$. If the eigenvalue $\mu_n$ is attained, the choice $\varphi_i^S = u_i$, $1 \leq i \leq n$, makes the left-hand side equal to $\mu_n$ so that (4.10) is verified. If $\mu_n$ is not attained then it belongs to the essential spectrum. Hence, there exists a sequence $(v_i)_{i\in\mathbb{N}}$ with the properties

$$\begin{aligned} a(v_\ell, v_m) &= \delta_{\ell,m}, \\ d(v_\ell, v_m) &= 0 \qquad \forall \ell \neq m, \\ \lim_{\ell\to\infty} d(v_\ell, v_\ell) &= \mu_n. \end{aligned}$$

We choose the $\varphi_i^S$ from the sequence with the above properties to obtain (4.10). ■

The following mapping principle serves as the basis for several approximation methods.

Let $H_1$ and $H_2$ denote Hilbert spaces with norm $a_1(\cdot,\cdot)^{1/2}$ and $a_2(\cdot,\cdot)^{1/2}$, respectively and let $d_1(\cdot,\cdot)$ and $d_2(\cdot,\cdot)$ denote bounded quadratic functionals on $H_1$ and $H_2$, respectively. We definite $\mu_1^{(1)} \geq \mu_2^{(1)} \geq \dots$ to be the eigenvalues of the Rayleigh quotient $d_1(u,u)/a_1(u,u)$ and $\mu_1^{(2)} \geq \mu_2^{(2)} \geq \dots$ to be the eigenvalues of $d_2(u,u)/a_2(u,u)$ on $H_2$. We shall prove an inequality between these eigenvalues under some conditions.

**Theorem 4.8 (Mapping Principle)** *Let $M$ be a linear transformation from a subspace $S_1$ of $H_1$ into $H_2$. For $i = 1, 2$, let $\rho_i(u) := d_i(u,u)/a_i(u,u)$. Suppose that for some nondecreasing functions $f$ and $g$ the inequalities*

$$f(\rho_1(u)) \leq \rho_2(Mu) \tag{4.12}$$

*and*

$$g\left(\rho_1\left(u\right)\right) \leq \frac{a_2\left(Mu, Mu\right)}{a_1\left(u, u\right)} \tag{4.13}$$

*hold for all nonzero* $u \in S_1$.

*If* $S_1$ *contains the eigenfunctions* $u_1^{(1)}, u_2^{(1)}, \ldots, u_n^{(1)}$ *corresponding to* $\mu_1^{(1)}, \mu_2^{(1)}, \ldots, \mu_n^{(1)}$ *and if*

$$g\left(\mu_n^{(1)}\right) > 0,$$

*then*

$$f\left(\mu_n^{(1)}\right) \leq \mu_n^{(2)}.$$

**Proof.** Let $T_n := \operatorname{span}\left\{u_1^{(1)}, u_2^{(1)}, \ldots, u_n^{(1)}\right\}$, which is, by hypotheses, a subspace of $S_1$. Then

$$d_1\left(u, u\right) \geq \mu_n^{(1)} a_1\left(u, u\right) \qquad \forall u \in T_n.$$

Since $g\left(\mu_n^{(1)}\right) > 0$, we obtain by (4.13) that for any nonzero $u \in T_n$, we have $a_2\left(Mu, Mu\right) > 0$ and, hence, $Mu \neq 0$. This implies that the elements $Mu_1^{(1)}, Mu_2^{(1)}, \ldots, Mu_n^{(1)}$ are linearly independent.

Therefore, by the Poincaré principle

$$\mu_n^{(2)} \geq \min_{u \in T_n \setminus \{0\}} \frac{d_2\left(Mu, Mu\right)}{a_2\left(Mu, Mu\right)} \geq \min_{u \in T_n \setminus \{0\}} f\left(\frac{d_1\left(u, u\right)}{a_1\left(u, u\right)}\right) \geq f\left(\mu_n^{(1)}\right).$$

■

Let $V_1$ be a subspace of $V_2$, set $a_1\left(u, u\right) = a_2\left(u, u\right) = a\left(u, u\right)$ and $d_1\left(u, u\right) = d_2\left(u, u\right) = d\left(u, u\right)$ and let $M$ be the trivial injection $M : V_1 \hookrightarrow V_2$. Then, the hypothesis (4.12) and (4.13) are satisfied when $f\left(\xi\right) := \xi$ and $g\left(\xi\right) \equiv 1$, and the mapping theorem becomes the following theorem.

**Theorem 4.9 (First monotonicity principle)** *Let* $H_2$ *be a Hilbert space with the norm* $a\left(u, u\right)^{1/2}$ *and let* $d\left(u, u\right)$ *be a bounded quadratic functional on* $H_2$. *Let* $H_1 \subset H_2$ *be a subspace. If* $\mu_1^{(2)} \geq \mu_2^{(2)} \geq \ldots$ *are the Ritz values of the Rayleigh quotient* $d\left(u, u\right)/a\left(u, u\right)$ *on* $H_2$ *and* $\mu_1^{(1)} \geq \mu_2^{(1)} \geq \ldots$ *are the Ritz values of the same Rayleigh quotient with* $u$ *restricted to* $V_1$, *then,*

$$\mu_n^{(1)} \leq \mu_n^{(2)}, \qquad n = 1, 2, \ldots.$$

**Proof of Theorem 4.6.** We choose $V_2 = H$ and let $V_1$ be the space spanned by $\varphi_i^S$, $1 \leq i \leq k$, and apply Theorem 4.9. ■

**Theorem 4.10 (Knyazev)** *Let Assumption 4.2 be satisfied. Let* $m$ *be chosen as in (4.5) and let* $\tilde{N}_{1,j}$ *be defined by (4.6). Then, for all* $1 \leq j \leq m$

$$0 \leq \frac{\mu_j - \mu_j^S}{\mu_j} \leq \Theta^2\left(\tilde{N}_{1,m}, S\right), \tag{4.14}$$

*where* $\Theta\left(\tilde{N}_{1,m}, S\right)$ *is as in (4.7).*

Before we will prove this theorem, we will show its application to the Dirichlet problem. Let $H := H_0^1(\Omega)$, $U := L^2(\Omega)$, and

$$a(u,v) := \int_\Omega \langle \mathbf{A}\nabla u, \nabla v\rangle + cuv, \qquad d(u,v) := (u,v)_{L^2(\Omega)},$$

where $A, c$ satisfy the relevant conditions in Assumption 2.55 and Definition 2.57 (for $\mathbf{b} = 0$). Let $S := S_{\mathcal{T}}^p$ denote the finite element space as introduced in Subsection 3.2 with mesh width $h_{\mathcal{T}}$. If we assume that $\tilde{N}_{1,m} \subset H^s(\Omega)$ for some $1 < s \leq p+1$, then, the approximation property can be split into

$$\Theta\left(\tilde{N}_{1,m}, S\right) \leq \left(\sup_{\substack{w\in H^s(\Omega)\\ \|w\|_{H^s(\Omega)}=1}} \inf_{u^S\in S} \left\|w - u^S\right\|_H\right) \times \left(\sup_{\substack{u\in N_{1,m}\\ \|u\|_H=1}} \|u\|_{H^s(\Omega)}\right).$$

The first factor is independent from the fact that we are dealing with eigenvalue problems but related to the approximation property of finite elements for higher order Sobolev spaces. The estimate of the second factor requires a regularity result for eigenvalue problems. Under the conditions of Theorem 3.6 we obtain $s = 2$

$$\sup_{\substack{u\in\tilde{N}_{1,m}\\ \|u\|_{H^1(\Omega)}=1}} \|u\|_{H^2(\Omega)} \leq C\sqrt{\lambda_m}.$$

For the approximation property of the space $S_{\mathcal{T}}^1$ of continuous, piecewise affine finite elements we employ Proposition 3.4 to obtain

$$\sup_{\substack{w\in H^2(\Omega)\\ \|w\|_{H^2(\Omega)}=1}} \inf_{u^S\in S} \left\|w - u^S\right\|_{H^1(\Omega)} \leq Ch_{\mathcal{T}}.$$

Hence,

$$\Theta\left(\tilde{N}_{1,m}, S_{\mathcal{T}}^1\right) \leq C\sqrt{\lambda_m}h_{\mathcal{T}}$$

and the eigenvalue error estimate becomes

$$0 \leq \frac{\mu_j - \mu_j^S}{\mu_j} \leq C\lambda_m h_{\mathcal{T}}^2.$$

**Proof of Theorem 4.10.**

The estimate $\mu_j \geq \mu_j^S$ directly follows from the Theorem 4.6 on the Rayleigh-Ritz method and $\mu_j > 0$ is a consequence of (4.4). In the case that $\Theta\left(S, \tilde{N}_{1,m}\right) = 1$, the right-hand side estimate in (4.14) directly follows from the positivity of the eigenvalues $\mu_j$ and $\mu_j^S$.

It remains to prove the right-hand inequality in (4.14) for $\Theta\left(S, \tilde{N}_{1,m}\right) < 1$ (Recall that $\Theta := \Theta\left(S, \tilde{N}_{1,m}\right) = \Theta\left(\tilde{N}_{1,m}, S\right)$ (cf. Remark 4.1)).

Let $P : H \to \tilde{N}_{1,m}$ be the orthogonal projection w.r.t. the $a(\cdot,\cdot)$ scalar product. The condition

$$\Theta\left(S, \tilde{N}_{1,m}\right) = \max_{u\in S\setminus\{0\}} \frac{\|u - Pu\|_H}{\|u\|_H} < 1$$

implies that for $u \neq 0$ the projection satisfies $Pu \neq 0$. Since $\dim \tilde{N}_{1,m} = \dim S$ we have that $P : S \to \tilde{N}_{1,m}$ is one-to-one. Note that for any $\mu < 0$

$$\gamma(u,v) := d(u,v) - \mu a(u,v)$$

is a scalar product. For any $u \in H$, there exist coefficients $\alpha_{(j,\ell)}$, $(j,\ell) \in J_S$, such that $Pu = \sum_{(j,\ell)\in J_S} \alpha_{j,\ell} u_{j,\ell}$ and, hence,

$$\begin{aligned}\gamma(Pu, u_{j',\ell'}) &= d(Pu, u_{j',\ell'}) - \mu a(Pu, u_{j',\ell'}) = \sum_{(j,\ell)\in J_S} \alpha_{j,\ell} d(u_{j,\ell}, u_{j',\ell'}) - \mu a(u, u_{j',\ell'}) \\ &= \sum_{(j,\ell)\in J_S} \alpha_{j,\ell}\mu_j a(u_{j,\ell}, u_{j',\ell'}) - \mu a(u, u_{j',\ell'}) = \alpha_{j',\ell'}(\mu_j - \mu).\end{aligned}$$

On the other hand, there is some $u^\perp \in \tilde{N}_{1,m}^\perp$ such that $u = Pu + u^\perp$. Thus,

$$\gamma(u, u_{j',\ell'}) = \alpha_{j',\ell'}(\mu_j - \mu)$$

and $P$ is orthogonal also with respect to the $\gamma$-scalar product.

Let $u \in S \setminus \{0\}$ so that also $Pu \in \tilde{N}_{1,m} \setminus \{0\}$. The Rayleigh quotient for $d(\cdot,\cdot)$ and $a(\cdot,\cdot)$ is denoted by $\rho(u) := d(u,u)/a(u,u)$. Note that from Pythagoras' theorem we get

$$\begin{aligned}a(Pu, Pu) &= a(u,u) - a(Pu - u, Pu - u) \geq a(u,u) - \|Pu - u\|_H^2 \\ &\geq a(u,u)\left(1 - \Theta^2\right)\end{aligned}$$

and, because $P$ is an $\gamma$-orthogonal projection, we have

$$\gamma(Pu, Pu) \leq \gamma(u,u).$$

This leads to

$$\rho(Pu) - \mu = \frac{\gamma(Pu,Pu)}{a(Pu,Pu)} \leq \frac{\gamma(u,u)}{(1-\Theta^2)\, a(u,u)} = \frac{\rho(u) - \mu}{1 - \Theta^2}.$$

Hence, we may apply the mapping principle (Theorem 4.8) with $S_1 := H_1 := \tilde{N}_{1,m}$, $H_2 := S$, $\rho_1(u) := \rho_2(u) := \rho(u)$, $f(\xi) := (1-\Theta^2)\,\xi$, $g(\xi) := 1$, $M = P^{-1}$, $a_1 = a_2 = a$, $d_1 = d_2 = \gamma$, $\mu_\ell^{(1)} := \mu_\ell$, $\mu_\ell^{(2)} := \mu_\ell^S$ to obtain

$$\left(1 - \Theta^2\right)(\mu_\ell - \mu) \leq \left(\mu_\ell^S - \mu\right).$$

Since both sides depend continuously on $\mu < 0$ (for sufficiently small $\mu$) the estimate holds also for $\mu = 0$ and we obtain

$$\frac{\mu_\ell - \mu_\ell^S}{\mu_\ell} \leq \Theta^2.$$

■

We finish off this section with some remarks concerning various generalizations of Theorem 4.10.

In [12] the estimate

$$0 \leq \frac{\mu_j - \mu_j^S}{\mu_j} \leq \Theta^2\left(\tilde{N}_{1,j}, S\right)$$

has been proved for $1 \leq j \leq n := \dim S$, where $\mu_j$, $\mu_j^S$, $1 \leq j \leq n$, denote the largest $n$ eigenvalues of $T$ resp. $T_S$ by taking into account their multiplicity. This can be expressed in terms of the eigenvalues $\lambda_j := \mu_j^{-1}$, $\lambda_j^S := \left(\mu_j^S\right)^{-1}$ as

$$0 \leq \frac{\lambda_j^S - \lambda_j}{\lambda_j^S} \leq \Theta_j^2 \quad \text{with} \quad \Theta_j := \Theta^2\left(\tilde{N}_{1,j}, S\right).$$

For $\Theta_j < 1$ we derive from this the estimate for the relative error

$$\frac{\lambda_j^S - \lambda_j}{\lambda_j} \leq \frac{\Theta_j^2}{1 - \Theta_j^2}.$$

We made the assumptions that the operator $T$ is positive definite just to simplify the arguments. Typically, a scalar shift $\tilde{T} := T + \alpha I$ and $\tilde{T}_S := T_S + \alpha I$ can be chosen so that $\tilde{T}$, $\tilde{T}_S$ are positive definite. The eigenfunctions do not change and the eigenvalues are simply shifted by $\alpha$ and the adaption of the error analysis is straightforward.

## 4.3 Estimates of Eigenvector Approximations for Selfadjoint Eigenproblems

We come to the estimates of the eigenvector approximation. We consider the continuous problem in the form (3.5) $Tu = \mu u$, where $T$ is a compact operator and the discrete problem in the form (4.1) $T_S u^S = \mu^S u^S$. The eigenspace corresponding to a continuous eigenvalue $\mu$ is denoted by $N(\mu) \subset H$ and $N_S\left(\mu^S\right) \subset S$ is the eigenspace corresponding to a discrete eigenvalue $\mu^S$.

We will prove the following convergence theorem only for the case that all eigenvalues of $T$ have multiplicity 1, i.e.,

$$\mu_1 > \mu_2 > \ldots > 0. \tag{4.15}$$

**Theorem 4.11 (Saad)** *Let (4.15) be satisfied. Let $(\mu_i, u_i)$, $1 \leq i \leq \dim S$ be the i-th eigenpair of (3.2) with normalization $\|u_i\|_H = 1$. Let $d_{i,S} := \min\left\{\left|\mu_i - \mu^S\right| : \mu^S \in \sigma(T_S) \setminus \left\{\mu_i^S\right\}\right\}$. Then, there exists some $u_i^S \in N_S\left(\mu_i^S\right)$ such that*

$$\left\|u_i - u_i^S\right\|_H \leq \left(1 + \frac{\|(I - P_S) T P_S\|_{H \leftarrow H}^2}{d_{i,S}^2}\right)^{1/2} \inf_{v \in S} \|u_i - v\|_H, \tag{4.16}$$

*where $P_S$ denotes the $a(\cdot,\cdot)$-orthogonal projection onto $S$.*

**Proof. 1.** We first prove the statement: There exists $u_i^S \in N_S\left(\mu_i^S\right)$ such that

$$\left\|P_S u_i - u_i^S\right\|_H \leq \frac{r_S}{d_{i,S}} \|(I - P_S) u_i\|_H, \tag{4.17}$$

where $r_S := \|(I - P_S) T P_S\|_{H \leftarrow H}$.

Let $\mu_1^S, \ldots, \mu_m^S$ denote the distinct eigenvalues of $T_S$ and let $P_i^S$ denote the associated eigenprojection $P_i^S : H \to N_S\left(\mu_i^S\right)$ characterized by

$$a\left(P_i^S u, v\right) = a(u, v) \quad \forall v \in N_S\left(\mu_i^S\right).$$

From Theorem 2.34 and (2.26) for finite dimensional spaces, it follows

$$P_i^S P_j^S = \delta_{i,j} P_j^S \quad \text{and} \quad \sum_{j=1}^{m} P_j^S = P_S. \tag{4.18}$$

Hence $(P_S T - \mu_i I)\, P_S u_i = (P_S T - \mu_i I) \sum_{j=1}^{m} P_j^S u_i$ and

$$(P_S T - \mu_i I)\, P_S u_i = \sum_{j=1}^{m} \left(\mu_j^S - \mu_i\right) P_j^S u_i.$$

Multiplying the two sides by $I - P_i^S$ results in

$$\left(I - P_i^S\right)(P_S T - \mu_i I)\, P_S u_i = \sum_{j=1}^{m} \left(\mu_j^S - \mu_i\right)\left(I - P_i^S\right) P_j^S u_i. \tag{4.19}$$

In view of (4.18) this last term is equal to

$$\sum_{j\neq i}^{m} \left(\mu_j^S - \mu_i\right) P_j^S u_i.$$

Taking the norms of the two sides of equation (4.19) gives

$$\left\|\left(I - P_i^S\right)(P_S T - \mu_i I)\, P_S u_i\right\|_H^2 = \sum_{j\neq i} \left(\mu_j^S - \mu_i\right)^2 \left\|P_j^S u_i\right\|_H^2. \tag{4.20}$$

For the right-hand side we get the inequality

$$\sum_{j\neq i} \left(\mu_j^S - \mu_i\right)^2 \left\|P_j^S u_i\right\|_H^2 \geq d_{i,S}^2 \sum_{j\neq i} \left\|P_j^S u_i\right\|_H^2. \tag{4.21}$$

But (4.18) shows that

$$\sum_{j\neq i} \left\|P_j^S u_i\right\|_H^2 = \left\|\left(P_S - P_i^S\right) u_i\right\|_H^2.$$

For the left-hand side of (4.20) we get

$$\begin{aligned}
\left\|\left(I - P_i^S\right)(P_S T - \mu_i I)\, P_S u_i\right\|_H^2 &\leq \left\|I - P_i^S\right\|_H^2 \left\|P_S\left(T - \mu_i I\right) P_S u_i\right\|_H^2 \\
&\leq \left\|P_S\left(T - \mu_i I\right)\left(u_i - \left(I - P_S\right) u_i\right)\right\|_H^2 \\
&= \left\|P_S\left(T - \mu_i I\right)\left(I - P_S\right)\left(I - P_S\right) u_i\right\|_H^2 \\
&\leq \left\|P_S\left(T - \mu_i I\right)\left(I - P_S\right)\right\|_{H\leftarrow H}^2 \left\|\left(I - P_S\right) u_i\right\|_H^2.
\end{aligned}$$

Note that $\left\|P_S\left(T - \mu_i I\right)\left(I - P_S\right)\right\|_{H\leftarrow H} = \left\|P_S T\left(I - P_S\right)\right\|_{H\leftarrow H} = \left\|\left(I - P_S\right) T P_S\right\|_{H\leftarrow H}$ because all operators are self-adjoint. Thus,

$$\left\|\left(I - P_i^S\right)(P_S T - \mu_i I)\, P_S u_i\right\|_H^2 \leq r_n^2 \left\|\left(I - P_S\right) u_i\right\|_H^2. \tag{4.22}$$

Now, using (4.20), (4.21), and (4.22) yields the stated inequality (4.17).

**2.** Inequality (4.16) is obtained from the decomposition

$$\left(I - P_i^S\right) u_i = \left(I - P_S\right) u_i + \left(P_S - P_i^S\right) u_i,$$

where the two vectors in the right-hand side are orthogonal. Thus

$$\left\| \left(I - P_i^S\right) u_i \right\|_H^2 = \left\| \left(I - P_S\right) u_i \right\|_H^2 + \left\| \left(P_S - P_i^S\right) u_i \right\|_H^2$$

which, by (4.17), gives (4.16) and completes the proof. ■

Estimate (4.16) only makes sense if $d_{i,S} > 0$. This condition can be replaced by a stronger condition which employs the error estimate for the eigenvalue approximation. For $j \neq i$, Theorem 4.10 (with $\Theta := \Theta\left(\tilde{N}_{1,m}, S\right)$) implies the estimate

$$\begin{aligned} \left|\mu_i - \mu_j^S\right| &\geq \left|\mu_i - \mu_j\right| - \left|\mu_j - \mu_j^S\right| \geq \left|\mu_i - \mu_j\right| - \mu_j \Theta^2 \\ &\geq \left|\mu_i - \mu_j\right| \left(1 - \Theta^2\right) - \mu_i \Theta^2. \end{aligned}$$

Hence,

$$d_{i,S} \geq \left(1 - \Theta^2\right) d\left(T, \mu_i\right) - \left|\mu_i\right| \Theta^2,$$

where

$$d\left(T, \mu_i\right) := \min\left\{\left|\mu_i - \mu\right| : \mu \in \sigma\left(T\right) \setminus \left\{\mu_i\right\}\right\}$$

denotes the spectral gap of $\mu_i$ for the operator $T$. By

$$d_{\mathrm{rel}}\left(T, \mu_i\right) := \frac{d\left(T, \mu_i\right)}{\mu_i}$$

we denote its relative version. In the following, we will rewrite the gap in terms of the eigenvalues $\lambda_i = \mu_i^{-1}$ of (3.2). Note that the set of eigenvalues of (3.2) equals $\sigma\left(T^{-1}\right)$.

For $\mu_j > \mu_i$ we obtain

$$\mu_j - \mu_i = \lambda_j^{-1} - \lambda_i^{-1} = \lambda_i^{-1} \frac{\lambda_i - \lambda_j}{\lambda_j} \geq \lambda_i^{-1} \frac{\lambda_i - \lambda_j}{\lambda_i}$$

and, for $\mu_j < \mu_i$, it holds

$$\mu_i - \mu_j = \lambda_i^{-1} \frac{\lambda_j - \lambda_i}{\lambda_j} = \lambda_i^{-1} \frac{\lambda_j - \lambda_i}{\lambda_i + \left(\lambda_j - \lambda_i\right)}.$$

Hence,

$$\left|\mu_i - \mu_j\right| \geq \lambda_i^{-1} \frac{\frac{\left|\lambda_i - \lambda_j\right|}{\lambda_i}}{1 + \frac{\left|\lambda_i - \lambda_j\right|}{\lambda_i}}.$$

Because the function $\frac{x}{1+x}$ is monotonously increasing we get

$$d\left(T, \mu_i\right) \geq \lambda_i^{-1} \frac{d_{\mathrm{rel}}\left(T^{-1}, \lambda_i\right)}{1 + d_{\mathrm{rel}}\left(T^{-1}, \lambda_i\right)}. \tag{4.23}$$

To the best of our knowledge, sharp lower estimates for the spectral gap are not available in the literature. The study of the *asymptotic* distribution of eigenvalues of elliptic operators

goes back to H. Weyl [29] and was refined e.g., by [9, Sec. VI, § 4, Satz 17 and 19], [6], [5], [22, Theorem 13.1]. The main result reads

$$\lim_{t\to\infty} \frac{N(t)}{t^{d/2}} = C_d^{\text{asymptotic}}, \tag{4.24}$$

where $N : \mathbb{R} \to \mathbb{R}$ is a function which satisfies for all eigenvalues $\lambda$ of (3.2)

$$N(\lambda) = \operatorname{card}\left\{\tilde{\lambda} \text{ is an eigenvalue of (3.2) and } \tilde{\lambda} \le \lambda\right\}.$$

This implies that for $t$ large enough, we have

$$N(t) \le C_d t^{d/2}. \tag{4.25a}$$

$C_d$ in (4.25a) is a positive constant which only depends on the space dimension $d$. Since the values of $N$ are fixed only for the (discrete) eigenvalues $\lambda$ we may assume that $N \in C^1(\mathbb{R})$ and that $N$ is strictly monotonously increasing. The gap between an eigenvalue $\lambda$ and the next larger one $\lambda^+$ is

$$\lambda^+ - \lambda = N^{-1}(N(\lambda) + 1) - \lambda.$$

A Taylor argument yields

$$\lambda^+ - \lambda = \frac{1}{N'\left(N^{-1}(N(\lambda) + \xi)\right)} \quad \text{for some } \xi \in [0,1].$$

To the best of knowledge bounds of $N'$ are not available in the literature. If we *assume* the hypotheses

$$\begin{array}{ll} N^{-1}(y) \le \tilde{C}_d y^{2/d} & \forall y \in [N(\lambda), N(\lambda) + 1] \\ N'(t) \le c_d t^{d/2-1} & \forall t \in [\lambda, N^{-1}(N(\lambda) + 1)] \end{array} \tag{4.25b}$$

we obtain

$$\lambda^+ - \lambda \ge \frac{1}{c_d \left(N^{-1}(N(\lambda) + \xi)\right)^{d/2-1}}.$$

For $d = 1, 2$, we conclude from the monotonicity of $N^{-1}$ and $N$ that

$$\lambda^+ - \lambda \ge \frac{\lambda^{1-d/2}}{c_d}$$

holds, while for $d \ge 3$ and $\lambda$ large enough so that $C_d \lambda^{d/2} \ge 1$, we get

$$\begin{aligned} \lambda^+ - \lambda &\ge \frac{1}{c_d \left(N^{-1}(N(\lambda) + 1)\right)^{d/2-1}} \ge \frac{1}{c_d \left(\tilde{C}_d C_d^{2/d} \lambda \left(1 + \frac{1}{C_d \lambda^{d/2}}\right)^{2/d}\right)^{d/2-1}} \\ &\ge \hat{C}_d \lambda^{1-d/2}. \end{aligned}$$

By repeating these arguments for the closest smaller eigenvalue $\lambda^-$, we have derived that the assumptions (4.25) on $\lambda$ imply that the relative spectral gap satisfies

$$d_{\text{rel}}\left(T^{-1}, \lambda\right) \ge \check{c}_d \lambda^{-d/2}. \tag{4.26}$$

From (4.23) we conclude for this situation that

$$d(T,\mu_i) \geq \lambda_i^{-1} \frac{\check{c}_d \lambda_i^{-d/2}}{1+\check{c}_d \lambda_i^{-d/2}} \geq C_{d,T} \lambda_i^{-1-d/2},$$

where $C_{d,T}$ only depends on the spatial dimension and the minimal eigenvalue $\lambda_0$ of (3.2).

This leads to the following corollary of Theorem 4.11.

**Corollary 4.12** *We assume that all eigenvalues of $T$ are simple, i.e., (4.15) holds. Let $(\mu_i, u_i)$, $1 \leq i \leq \dim S$ be the i-th eigenpair of (3.2) with normalization $\|u_i\|_H = 1$. Let the finite element $S$ be chosen such that*

$$\Theta^2\left(\tilde{N}_{1,i}, S\right) \leq \min\left\{c_{d,T}\lambda_i^{-d/2}, 1/2\right\} \quad \text{with} \quad c_{d,T} := \frac{1}{\lambda_0^{-d/2} + 2C_{d,T}^{-1}}.$$

*Then, there exists some $u_i^S \in N_S\left(\mu_i^S\right)$ such that*

$$\left\|u_i - u_i^S\right\|_H \leq \left(1 + 4\frac{\lambda_i^{1+d/2}\left\|(I-P_S)TP_S\right\|_{H\leftarrow H}}{C_{d,T}}\right) \inf_{v\in S}\|u_i - v\|_H .$$

*Let the assumptions of Proposition 3.4 and Theorem 3.6 be satisfied and let (4.26) be valid for $\lambda_i$. In the case of piecewise linear finite elements $S = S_{\mathcal{T}}^1$ we obtain under the condition $\lambda_i^{1+d/2} h_{\mathcal{T}}^2 \ll 1$ the eigenfunction error estimate*

$$\left\|u_i - u_i^S\right\|_H \leq \left(1 + C\lambda_i^{1+d/2} h\right)\sqrt{\lambda_i} h_{\mathcal{T}}.$$

The restriction to simple eigenvalues for the eigenvector error estimates is quite strong. The error estimates have been generalized in [20] and [23] to the case of clustered eigenvalues.

**Theorem 4.13** *Let Assumption 4.2 be satisfied. Let $\mathcal{I}$ denote an invariant subspace of $T$ and $\sigma_{\mathcal{I}}$ the spectrum of $T$ restricted to $\mathcal{I}$. Let $\mathcal{I}_S$ denote an invariant subspace of $T_S$ and $\sigma_{\mathcal{I}_S}$ the spectrum of $T_S|_{\mathcal{I}_S}$. Assume that*[5]

$$\delta(\mathcal{I},\mathcal{I}_S) := \operatorname{dist}\left(\sigma(T_S)\setminus\sigma_{\mathcal{I}_S}, \operatorname{conv}\sigma_{\mathcal{I}}\right) > 0.$$

*Then, for any $u \in \mathcal{I}$ there exists some $u_S \in \mathcal{I}_S$ such that*

$$\|u - u_S\|_H \leq \left(1 + \frac{\left\|(I-P_S)TP_S\right\|_{H\leftarrow H}^2}{\delta^2(\mathcal{I},\mathcal{I}_S)}\right)\Theta(\mathcal{I},S).$$

# 5 Error Analysis for the Non-Selfadjoint Eigenvalue Problem

The error analysis for non-selfadjoint eigenvalue problems is based on the representation of the spectral projections as contour integrals and we follow the approach which was developed in

[5] $\operatorname{conv}(\cdot)$ denotes the convex hull of a set.

[3]. We only sketch the main steps and will not prove all statements. A complete development of this theory can be found in [10], [11], [18], [3].

We restrict here to the case where $T : H \to H$ is a compact operator on a complex Hilbert space $H$ with norm $\|\cdot\|_H$ which satisfies the assumption of Theorem 2.54. The relation between $T$ and $a(\cdot,\cdot)$ is given by (3.4). We assume that the bilinear form $d(\cdot,\cdot)$ satisfies (3.1) and consider the eigenvalue problem (3.2) in the form

$$Tu = \mu u$$

and its discretization by (3.6) in the form

$$T_S u^S = \mu^S u^S.$$

Recall the definition of the index $n_\mu$ of an eigenvalue (cf. Theo. 2.34). We denote by

$$N_*(\mu) := N\left((\mu - T)^{n_\mu}\right)$$

the space of *generalized* eigenfunctions. Next, we will characterize the spectral projection associated with $T$ and $\mu$ by a contour integral. Since all $\mu \in \sigma(T) \setminus \{0\} \subset \mathbb{C}$ are discrete, we may choose a circle $\Gamma \subset \mathbb{C}$ about $\mu$ which lies in the resolvent set $\rho(T)$ and which encloses no other point of $\sigma(T)$. The spectral projection $E = E(\mu) : H \to N_*(\mu)$ is surjective and given by

$$E = \frac{1}{2\pi\,\mathrm{i}} \int_\Gamma R_z(T)\, dz.$$

We choose $S$ "rich" enough (cf.(4.7)), (e.g., the mesh width small enough) such that

$$\Theta(H, S) \text{ is small enough such that } \Gamma \subset \rho(T_S) \tag{5.1}$$

and the discrete spectral projection

$$E_S = E_S(\mu) = \frac{1}{2\pi\,\mathrm{i}} \int_\Gamma R_z(T_S)\, dz$$

converges to $E$ in norm and $\dim R(E_S(\mu)) = \dim R(E(\mu)) = m$. $E_S$ is the spectral projection associated with $T_S$ and the eigenvalues of $T_S$ which lie in $\Gamma$ and is a projection onto the direct sum of the spaces of generalized eigenvectors corresponding to these eigenvalues, i.e.,

$$R(E_S) = \bigoplus_{\substack{\mu^S \in \sigma(T_S) \\ \mu^S \text{ inside } \Gamma}} N\left(\left(\mu^S I - T_S\right)^{n_{\mu^S}}\right).$$

Thus, counting according to the algebraic multiplicities there are $m$ eigenvalues of $T_S$ in $\Gamma$; we denote these by $\mu_1^S, \mu_2^S, \ldots, \mu_m^S$.

**Theorem 5.1** *Let $S$ be such that (5.1) is satisfied. Then, there is a constant $C$ independent of $S$ such that*

$$\Theta(R(E), R(E_S)) \le C \left\| (T - T_S)|_{R(E)} \right\|_{H \leftarrow H}.$$

**Proof.** For $f \in R(E)$ we have

$$\|f - E_S f\|_H = \|(E - E_S) f\|_H = \left\| \frac{1}{2\pi \mathrm{i}} \int_\Gamma (R_z(T) - R_z(T_S)) f dz \right\|_H$$
$$= \left\| \frac{1}{2\pi \mathrm{i}} \int_\Gamma R_z(T_S)(T - T_S) R_z(T) f dz \right\|_H$$

and hence,

$$\|f - E_S f\|_H \leq \frac{\mathrm{Length}(\Gamma)}{2\pi} \sup_{z\in\Gamma} \|R_z(T_S)\|_{H\leftarrow H} \left\| (T - T_S)|_{R(E)} \right\|_{H\leftarrow H} \sup_{z\in\Gamma} \|R_z(T)\|_{H\leftarrow H}.$$

The stability condition on $S$ ensures that $\sup_{z\in\Gamma} \|R_z(T_S)\|_{H\leftarrow H} \leq C$ uniformly as $S \to H$. From this, the assertion follows. ■

**Remark 5.2** *The proof of Theorem 5.1 also shows that*

$$\left\| (E - E_S)|_{R(E)} \right\|_{H\leftarrow H} \leq C \left\| (T - T_S)|_{R(E)} \right\|_{H\leftarrow H}.$$

Although each of the eigenvalues $\mu_1^S, \mu_2^S, \ldots, \mu_m^S$ is close to $\mu$ their *arithmetic* mean is generally a closer approximation to $\mu$. We define

$$\widehat{\mu^S} := \frac{1}{m} \sum_{j=1}^m \mu_j^S.$$

Let $\varphi_1, \ldots, \varphi_m$ be any basis for $R(E)$ and let $\varphi_1', \ldots, \varphi_m'$ be the corresponding dual basis in $R(E)'$ which is the dual space of $R(E)$. We can extend each $\varphi_j'$ to $X$ as follows. Since $X = R(E) \oplus N(E)$, any $f \in X$ can be written as $f = g + h$ with $g \in R(E)$ and $h \in N(E)$. Define $\langle f, \varphi_j' \rangle = \langle g, \varphi_j' \rangle$. Clearly $\varphi_j'$ is bounded, i.e., $\varphi_j' \in X'$. Now $\langle f, (\mu I - T')^{n_\mu} \varphi_j' \rangle = \langle (\mu I - T)^\alpha f, \varphi_j' \rangle$ for $f \in R(E) = N_*(\mu)$ and it vanishes for $f \in N(E)$ since $N(E)$ is invariant for $\mu I - T$. Thus, we have shown that $\varphi_1', \ldots, \varphi_m' \in R(E')$.

**Theorem 5.3** *Let $\varphi_1, \ldots, \varphi_m$ be any basis for $R(E)$ and let $\varphi_1', \ldots, \varphi_m'$ be the dual basis in $R(E')$ as defined above. Then, there is a constant $C$, uniformly as $S \to H$, such that*

$$\left| \mu - \widehat{\mu^S} \right| \leq \frac{1}{m} \sum_{j=1}^m |\langle (T - T_S)\varphi_j, \varphi_j' \rangle| + C \left\| (T - T_S)|_{R(E)} \right\|_{H\leftarrow H} \left\| (T' - T_S')|_{R(E')} \right\|_{H\leftarrow H}.$$

**Proof.** If condition (5.1) is satisfied, the operator $E_S|_{R(E)} : R(E) \to R(E_S)$ is one-to-one since $\|E - E_S\|_{H\leftarrow H} \to 0$ and $E_S f = 0$, $f \in R(E)$ implies

$$\|f\|_H = \|Ef - E_S f\|_H \leq \|E - E_S\|_{H\leftarrow H} \|f\|_H,$$

and $E_S|_{R(E)}$ is surjective since

$$\dim R(E_S) = \dim R(E) = m.$$

Thus, $E_S|_{R(E)}^{-1} : R(E_S) \to R(E)$ is well-defined. We write $E_S^{-1}$ short for $\left( E_S|_{R(E)} \right)^{-1}$. For $\Theta(H, S)$ sufficiently small and $f \in R(E)$ with $\|f\|_H = 1$ we have

$$1 - \|E_S f\|_H = \|Ef\|_H - \|E_S f\|_H \leq \|E - E_S\|_{H\leftarrow H} \leq 1/2$$

and, hence, $\|E_S f\|_H \geq 1/2\,\|f\|_H$. This implies that $\left\|E_S^{-1}\right\|_{H\leftarrow H}$ is bounded uniformly as $S \to H$. We note that $E_S E_S^{-1}$ is the identity on $R(E_S)$ and $E_S^{-1}E_S$ is the identity on $R(E)$. Now, we define

$$\widehat{T_S} := E_S^{-1} T_S\, E_S|_{R(E)} : R(E) \to R(E)\,.$$

Using the fact that $R(E_S)$ is invariant for $T_S$ we see that $\sigma\left(\widehat{T_S}\right) = \left\{\mu_1^S, \ldots, \mu_m^S\right\}$ and that the algebraic (geometric, resp.) multiplicity of any $\mu_j^S$ as an eigenvalue of $\widehat{T_S}$ is equal to its algebraic (geometric, resp.) multiplicity as an eigenvalue of $T_S$. Letting $\hat{T} = T|_{R(E)}$ we see that $\sigma\left(\widehat{T}\right) = \{\mu\}$. Thus trace $\left(\widehat{T}\right) = m\mu$ and trace $\left(\widehat{T_S}\right) = m\widehat{\mu^S}$ and, since $\widehat{T}$ and $\widehat{T_S}$ act on the same space we can write

$$\mu - \widehat{\mu^S} = \frac{1}{m}\,\text{trace}\left(\hat{T} - \widehat{T_S}\right). \tag{5.2}$$

Let $\varphi_1, \ldots, \varphi_m$ be a basis for $R(E)$ and let $\varphi_1', \ldots, \varphi_m'$ be the dual basis to $\varphi_1, \ldots, \varphi_m$. Then, from (5.2) we get

$$\mu - \widehat{\mu^S} = \frac{1}{m}\,\text{trace}\left(\widehat{T} - \widehat{T_S}\right) = \frac{1}{m}\sum_{j=1}^{m}\left\langle\left(\widehat{T} - \widehat{T_S}\right)\varphi_j, \varphi_j'\right\rangle. \tag{5.3}$$

Using the facts that $T_S E_S = E_S T_S$ and $E_S^{-1}E_S$ is the identity on $R(E)$, we have

$$\begin{aligned}\left\langle\left(\widehat{T} - \widehat{T_S}\right)\varphi_j, \varphi_j'\right\rangle &= \left\langle T\varphi_j - E_S^{-1}T_S E_S\varphi_j, \varphi_j'\right\rangle \\ &= \left\langle E_S^{-1}E_S\,(T - T_S)\,\varphi_j, \varphi_j'\right\rangle \\ &= \left\langle (T - T_S)\,\varphi_j, \varphi_j'\right\rangle + \left\langle\left(E_S^{-1}E_S - I\right)(T - T_S)\,\varphi_j, \varphi_j'\right\rangle. \end{aligned}\tag{5.4}$$

Note that $L_S := E_S^{-1}E_S$ is the projection on $R(E)$ along $N(E_S)$. Hence, $L_S'$ is the projection on $N(E_S)^\perp = R(E_S')$ along $R(E)^\perp = N(E')$. Thus

$$\left\langle\left(E_S^{-1}E_S - I\right)(T - T_S)\,\varphi_j, \varphi_j'\right\rangle = \left\langle (L_S - I)\,(T - T_S)\,\varphi_j, (E' - E_S')\,\varphi_j'\right\rangle. \tag{5.5}$$

From (5.5), the boundedness of $L_S$ and Remark 5.2 (applied to $T'$ and $T_S'$) we get

$$\begin{aligned}&\left|\left\langle\left(E_S^{-1}E_S - I\right)(T - T_S)\,\varphi_j, \varphi_j'\right\rangle\right| \\ &\qquad\leq \|L_S - I\|_{H\leftarrow H}\left\|(T - T_S)|_{R(E)}\right\|_{H\leftarrow H}\left\|(E' - E_S')|_{R(E)}\right\|_{H\leftarrow H}\|\varphi_j\|_H\left\|\varphi_j'\right\|_H \\ &\qquad\leq C\left\|(T - T_S)|_{R(E)}\right\|_{H\leftarrow H}\left\|(T' - T_S')|_{R(E')}\right\|_{H\leftarrow H}. \end{aligned}\tag{5.6}$$

Finally (5.3), (5.4), and (5.6) yield the desired result. ■

The following theorem which is proved in [3, Theorem 7.3] shows that error estimate for the eigenvalues itself (instead of the average $\widehat{\mu^S}$) converge at a reduced rate if the index is larger than one.

**Theorem 5.4** *Let $n_\mu$ be the index of $\mu - T$. Let $\varphi_1, \ldots, \varphi_m$ be any basis for $R(E)$ and let $\varphi_1', \ldots \varphi_m'$ be the dual basis. Then there is a constant $C$ such that*

$$\left|\mu - \mu_j^S\right| \leq C\left\{\sum_{i,k=1}^{m}\left\langle (T - T_S)\,\varphi_i, \varphi_k'\right\rangle + \left\|(T - T_S)|_{R(E)}\right\|_{H\leftarrow H}\left\|(T' - T_S')|_{R(E')}\right\|_{H\leftarrow H}\right\}^{\frac{1}{n_\mu}},$$

*for all $j = 1, 2, \ldots, m$.*

# Literatur


[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions.* Applied Mathematics Series 55. National Bureau of Standards, U.S. Department of Commerce, 1972.

[2] H. W. Alt. *Lineare Funktionalanalysis.* Springer-Verlag, 1985.

[3] I. Babuška and J. Osborn. Eigenvalue Problems. In P. Ciarlet and J. Lions, editors, *Handbook of Numerical Analysis, Vol. II: Finite Element Methods (Part 1)*, pages 641–788. Elsevier Science Publishers, Amsterdam, 1991.

[4] S. Brenner and L. Scott. *The Mathematical Theory of Finite Element Methods.* Springer-Verlag, New York, 1994.

[5] F. Brownell. Extended asymptotic eigenvalue distributions for bounded domains in $n$-space. *Pacific Journ. of Math*, 5:483–499, 1955.

[6] F. Brownell. An extension of Weyl's asymptotic law for eigenvalues. *Pacific Journ. of Math*, 5:483–499, 1955.

[7] F. Chatelin. *Spectral Approximation of Linear Operators.* Academic Press, New York, 1983.

[8] P. Ciarlet. *The finite element method for elliptic problems.* North-Holland, 1987.

[9] R. Courant and D. Hilbert. *Methoden der Mathematischen Physik. Bd. 1.* Springer, Berlin, 1924.

[10] N. Dunford and J. Schwartz. *Linear Operators Part I: General Theory.* Interscience Publishers, Inc., New York, 1957.

[11] N. Dunford and J. Schwartz. *Linear Operators Part II: Spectral Theory.* Wiley-Interscience, New York, New York, 1963.

[12] E. Dyakonov. *Optimization in Solving Elliptic Problems.* CRC Press, Boca Raton, FL, 1996.

[13] D. Gilbarg and N. Trudinger. *Elliptic Partial Differential Equations of Second Order.* Springer-Verlag, 1983.

[14] P. Grisvard. *Elliptic Problems in Nonsmooth Domains.* Pitman, Boston, 1985.

[15] W. Hackbusch. *Elliptic Differential Equations.* Springer Verlag, 1992.

[16] H. Heuser. *Funktionalanalysis.* Teubner-Verlag, Stuttgart, 1986.

[17] T. Kato. Perturbation Theory for Nullity, Deficiency and other Quantities of Linear Operators. *J. d'Analyse Mathématique*, 6:261–322, 1958.

[18] T. Kato. *Perturbation theory for linear operators.* Springer-Verlag, Berlin, 1966.

[19] A. Knyazev. Sharp a priori error estimates of the Raleigh-Ritz method without assumptions of fixed sign or compactness. *Mathematical Notes*, 38(5-6):998–1002, 1986.

[20] A. Knyazev. New Estimates for Ritz Vectors. *Math. Comp.*, 66:985–995, 1997.

[21] A. Knyazev and J. Osborn. New a priori FEM error estimates for eigenvalues. *SIAM J. Numer. Anal.*, 48(6):2647–2667, 2006.

[22] S. Levendorskii. *Asymptotic Distribution of Eigenvalues of Differential Operators*. Springer, Berlin, 1990.

[23] E. Ovtchinnikov. Cluster robust error estimates for the Raleigh-Ritz approximation I: Estimates for invariant subspaces. *LAA*, 415(1):167–187, 2006.

[24] B. Parlett. *The Symmetric Eigenvalue Problem*. Prentice Hall, 1980.

[25] Y. Saad. On the Rates of Convergence of the Lanczos and the Block-Lanczos Methods. *SIAM J. Numer. Anal.*, 17:687–706, 1980.

[26] Y. Saad. Projection Methods for Solving Large Sparse Eigenvalue Problems. In B. Kagsrom and A. Ruhe, editors, *Matrix Pencils: Proc. of a Conference held in Pite Havsbad, Swedon, 1982*, Lecture Notes in Mathematics, 973, pages 121–144, Berlin, 1983. Springer.

[27] G. Strang and G. Fix. *An Analysis of the Finite Element Method*. Prentice-Hall, 1973.

[28] H. Weinberger. *Variational Methods for Eigenvalue Approximation*. SIAM, Philadelphia, 1974.

[29] H. Weyl. Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). *Math. Ann.*, 71:441–479, 1912.

[30] K. Yosida. *Functional Analysis*. Springer-Verlag, 1964.