Ritz value estimates and applications in Mathematical Physics

DISSERTATION
zur
Erlangung des Grades eines Dr. rer. nat.
des Fachbereiches Mathematik
der Fernuniversität in Hagen

vorgelegt von

Luka Grubišić
aus Zagreb (Kroatien)
Hagen 2005

Eingereicht im Januar 2005
Erstgutachter und Mentor: Prof. Dr. Krešimir Veselić
Zweitgutachter: Prof. Dr. Werner Kirsch, Ruhr-Universität Bochum
Prof. Dr. Klaus Neymeyr, Universität Rostock

Tag der mündlichen Prüfung 03.05.2005
Vorsitzende der Prüfungskommission: Prof. Dr. Andrei Duma
Prüfer: Prof. Dr. Krešimir Veselić
Prof. Dr. Werner Kirsch, Ruhr-Universität Bochum

Contents

Notation

Here we give a list of notations which are used in this thesis:

$\mathcal{Q}(h)$	the domain of the symmetric form h
$\overline{\mathcal{Q}}^{\,\mathcal{H}}$	the closure of the set $\mathcal{Q}$ in the topology of the space $\mathcal{H}$
$\mathcal{H}^*$	the topological dual of the Hilbert space $\mathcal{H}$
$\overline{\mathcal{Q}}^{\,t}$	the closure of the set $\mathcal{Q}$ in the topology induced by the positive definite form t (naturally $\mathcal{Q} \subset \mathcal{Q}(t)$)
$\overline{\mathbf{A}}$	the closure of the operator $\mathbf{A}$
$\mathcal{D}(\mathbf{H})$	the domain of the operator $\mathbf{H}$
$\mathsf{inv}(\mathbf{H})$	the inverse image of the mapping $\mathbf{H}$
$\mathsf{ran}(\mathbf{H})$	the range of the operator $\mathbf{H}$
$\mathsf{ker}(\mathbf{H})$	the kernel (null space) of the operator $\mathbf{H}$
$\sigma(\mathbf{H})$	the spectrum of the operator $\mathbf{H}$
$\sigma_d(\mathbf{H})$	the discrete spectrum of the self adjoint operator $\mathbf{H}$
$\sigma_{ess}(\mathbf{H})$	the essential spectrum of the self adjoint operator $\mathbf{H}$
$\lambda_e(\mathbf{H})$	the minimum of the essential spectrum of the self adjoint operator $\mathbf{H}$
$\mathbf{H}\,\vert_{\mathcal{M}}$	the restriction of the operator $\mathbf{H}$ onto the subspace $\mathcal{M} \cap \mathcal{D}(H)$
$\mathbf{H} = \int \lambda \, dE_{\mathbf{H}}(\lambda)$	the spectral decomposition of the self adjoint operator $\mathbf{H}$
$E_{\mathbf{H}}(\lambda)$	the right continuous spectral family associated to the operator $\mathbf{H}$
$P_\perp$	the projection $\mathbf{I} - P$
$\mathrm{spr}(A)$	the spectral radius of the bounded operator A
$\partial\Omega$	the boundary of the set Ω
χ, χ_Ω	the characteristic function of the set Ω
$L^2(\Omega)$	the second order Lebesgue space of functions defined on Ω
$H^p(\Omega)$	the p-the order Sobolev space of functions defined on Ω
$C^p(\Omega)$	the space of p-times continuously differentiable functions in Ω
div	the divergence operator
∇	the gradient operator
$\triangle$	the Laplace operator
tr	the trace
span $\mathcal{X}$	the linear span of the subset $\mathcal{X}$ of the space $\mathcal{H}$
dim $\mathcal{X}$	the dimension of the Hilbert space $\mathcal{X}$
$\Theta(\mathcal{X}, \mathcal{Y})$	the maximal canonical angle between the subspaces $\mathcal{X}$ and $\mathcal{Y}$
$\Theta_p(\mathcal{X}, \mathcal{Y})$	the maximal principal angle between the subspaces $\mathcal{X}$ and $\mathcal{Y}$
$\mathbb{R}_+$	the set of nonnegative real numbers
$*$	the adjoint on $\mathbb{C}$, the transpose on $\mathbb{R}$
x_i	the i-th component of the vector $x \in \mathbb{R}^n$ (or $\mathbb{C}^n$)
$\lfloor x \rfloor$	the largest $q \in \mathbb{Z}$ such that $q \leq x \in \mathbb{R}$
$\simeq$	is represented by (see [21])
$:=$	is defined by

Conventions

As a general rule we use bold symmetric capital letters ($\mathbf{H}, \mathbf{A}, \mathbf{V}, \ldots$) to denote self adjoint operators in a Hilbert space. Normal script capital letters ($T, B, K, M, \ldots$) will denote bounded operators and matrices. Calligraphic capital letters will be used to denote the Hilbert spaces ($\mathcal{X}, \mathcal{U}, \mathcal{H}, \mathcal{M}, \ldots$). The elementary functions will be denoted by normal script letters ($\sin, \cos, f, \ldots$) when appearing in displayed equations and by sanserif letters ($\mathsf{sin}, \mathsf{cos}, \mathsf{f}, \ldots$) when they appear as a part of an inline formula.

Acknowledgement

First and foremost I would like to thank my supervisor *Prof. Dr. Krešimir Veselić* for creating a working atmosphere where I was allowed to fly away with my own ideas, always knowing that there is a safety net for me to land into. His patience to sit through many seminars and informal discussions helped me to mature both professionally and personally. I am also grateful for his generous support during my years in Hagen.

Furthermore, I wish to express my gratitude to all those who have advised me during the course of the preparation of this thesis. Specifically, I would like to thank:

Prof. Dr. Klaus Neymeyr, for his prompt acceptance to review my thesis, his encouragement and advice at various stages of the research that went into this work as well as for his kind invitation to Universität Rostock.

Prof. Dr. Werner Kirsch, for his willingness to review my thesis, his encouragement in the final stages of the preparation of this manuscript as well as for his comments and his readiness to discuss my research on more than one occasion.

Prof. Dr. Zlatko Drmač, for introducing me to the beautiful field of Matrix Perturbation Theory. This thesis was partly motivated by a suggestion during one of his seminars at the University of Zagreb.

Dr. Josip Tambača, for his encouragement of our joint work on eigenvalue problems for the Curved Rod model, as well as for introducing me to the world of lower-dimensional approximations in the Theory of Elasticity.

Dr. Ivan Veselić, for his constant readiness to discuss various aspects of my research as well as for his suggestions on several instances during the preparation of this manuscript. His generous invitation to TU-Chemnitz was a source of many valuable contacts.

And last but not the least, *Prof. Dr. Vjeran Hari*, for his encouragement and firm support at all stages of my mathematical education. His steadiness was an inspiration at more than one occasion.

I am specially grateful to colleagues from the Department of Mathematics, University of Zagreb for taking my share of teaching load during my stay at Fernuniversität in Hagen.

Marina showed her deepest appreciation of my research, as well as her affection, by her eagerness to proof-read a densely written mathematical manuscript without being a mathematician herself.

At the end, I would like to thank my parents for their unwavering support and encouragement at all stages of my education.

Chapter 1

Overview

In recent years the perturbation theory of symmetric (hermitian) matrices has seen a great advance. Many interesting results were motivated by the needs of developers of mathematical software. Traditionally, the matrix perturbation estimates were derived by specializing corresponding (appropriate) operator results. In the matrix case, however, one has an additional advantage, the ability to perform simple computational experiments. As a result, the matrix perturbation theory has further developed as a separate field. More importantly, with the help of the insight of experiment the theory has reached a level of maturity and elegance.

Our main motivation is to revisit the perturbation theory of positive self adjoint operators with the techniques and the experience of the modern numerical linear algebra. The payoff should be twofold. First, the matrix results bring important formal motivation to operator problems and give rise to interesting results in new (more general) but similar setting. Second, the operator estimates will in the end be computed by matrix procedures. Erasing the borders between the operator and matrix theories enables us to better reuse known results.

Before we proceed with the introduction, a word to the reader. Sometimes the terms will have to be used in a discussion before they are formally introduced in the text. To ease the navigation through this thesis, as well as to prevent misunderstandings, we have provided Index at the end of the text.

One of the oldest methods to study a complex mathematical system is to consider it as a perturbation of a simpler system, whose properties are explicitly known. Basic reference in the study of the perturbation theory of linear operators is still the Kato's book [41]. Standard methods for the approximation of the eigenvalues of positive self adjoint matrices (operators) are based on the assumption that the matrix (operator) has

the following additive structure

$$H = H_0 + H_1. \tag{1.0.1}$$

Here H_0 is assumed to be the matrix on whose spectral properties we have extensive information and H is the matrix whose spectral properties we would like to investigate, cf. [10]. Assuming there is a sequence of low(er) rank (finite dimensional) projections such that $P_k \to \mathbf{I}$ one defines the sequence of matrices (operators)

$$H_k = H_0 + P_k H_1 P_k.$$

Since the matrices H_k have a special form we can hope to establish estimates of the eigenvalues of the matrix H_k from those of the matrix H_0 ,e.g. Weinstein–Aronszajn, Bazley–Fox methods, and then use the fact that $H_k \to H$ to assess the spectrum of H.

To apply this methods H must have the structure (1.0.1), which is not always the case. A similar approach, applicable to more general operators, has been formulated by Kahan in [40]. For a given P_k, Kahan constructs the operator

$$H'_k = P_k H P_k + (\mathbf{I} - P_k) H (\mathbf{I} - P_k). \tag{1.0.2}$$

As opposed to the construction (1.0.1), where the additive structure of the operator H was assumed, Kahan adaptively constructs the operator H'_k and applies a perturbation argument to assess the spectral properties of H_k from those of H'_k.

A tradeoff of this adaptability is that the spectrum of H'_k is, in general, only partially computable, e.g. the eigenvalues of the matrix $P_k H P_k|_{\mathsf{ran}(P_k)}$ are the Ritz values of the matrix H from the subspace $\mathsf{ran}(P_k)$ and can be computed by a finite dimensional procedure, whereas the part $(\mathbf{I} - P_k)H(\mathbf{I} - P_k)$ is infinite dimensional and in general unknown. To build a feasible perturbation argument, based on the operator H'_k, we have to assume some "mild" a priori information on the location of the spectrum of $(\mathbf{I} - P_k)H(\mathbf{I} - P_k)$. As opposed to (1.0.1) we have assumed only partial information on H'_k is computable, but the obtained perturbation method is adaptive. Furthermore, it will turn out that the assumptions on $(\mathbf{I} - P_k)H(\mathbf{I} - P_k)$ are not unnatural in the context of a study of $H'_k \to H$.

Kahan's rationale has been successfully applied to the problem of estimating the eigenvalues of self adjoint operators in [22] and to the problem of estimating eigenvectors in [21]. In both cases $\mathsf{ran}(P_k)$ was required to be somewhat more regular than necessary, e.g. it was not allowed for $\mathsf{ran}(P_k)$ to be a projection onto the subspace made up of linear elements when H is a second order elliptic operator.

Our method to compute subspace approximation estimates is influenced by two recently published works for finite matrices [26] and [45]. Sharp estimates obtained in these works

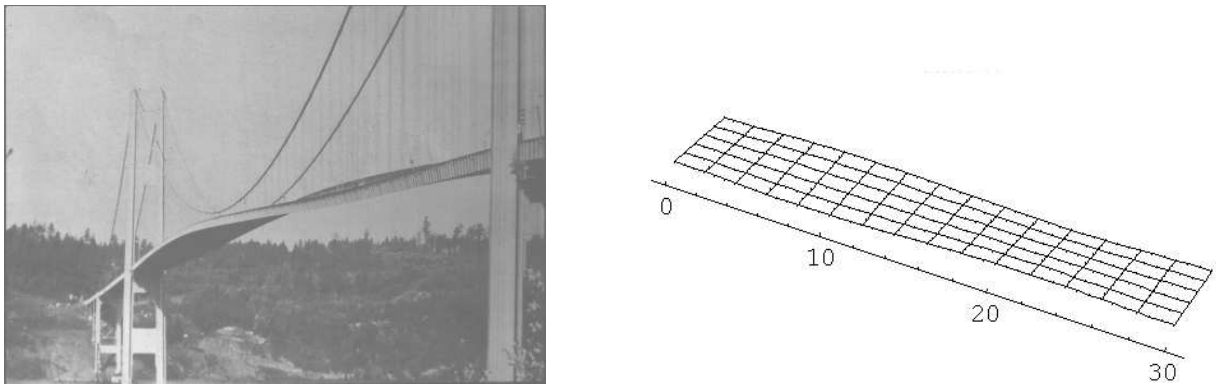

Figure 1.1: *Modelling the Tacoma bridge disaster:* Dangerous vibrations of the bridge, displayed on the picture, can efficiently be modelled by one of the natural modes of the network of curved rods. For details see [56, Tambača].

are based on the maximal angle Θ between the subspaces spanned by LX and $L^{-*}X$ where X are (orthonormal) test vectors and $H = L^*L$ is the given matrix. The Ritz values are the eigenvalues of the matrix $\Xi = X^*HX$. The geometric argumentation enters the eigenvalue estimation through the formulae, cf. [26, 28],

$$\max_x \frac{|x^*(H - H')x|}{x^*H'x} = \sin\Theta(LX, L^{-*}X), \tag{1.0.3}$$

$$\max_x \frac{|x^*(H - H')x|}{x^*Hx} = \frac{\sin\Theta(LX, L^{-*}X)}{1 - \sin\Theta(LX, L^{-*}X)}, \tag{1.0.4}$$

where

$$H' \;=\; PHP + P_\perp HP_\perp = X\ \Xi\ X^* + P_\perp HP_\perp$$

is the "block diagonal part" of H with respect to the projections $P = XX^*$ and $P_\perp = \mathbf{I} - P$. On the other hand, the standard theory from [21, 22] uses

$$\max_x |x^*(H - H')x| = \|R\| < \infty, \tag{1.0.5}$$

where

$$R = HX - X\Xi = HX - H'X$$

is the *residual* of the test subspace $\mathsf{ran}(X)$. We will slightly stretch the terminology and call both approaches "residual". The residual measures from (1.0.3) and (1.0.4) will be colloquially called *energy-scaled residual measures*.

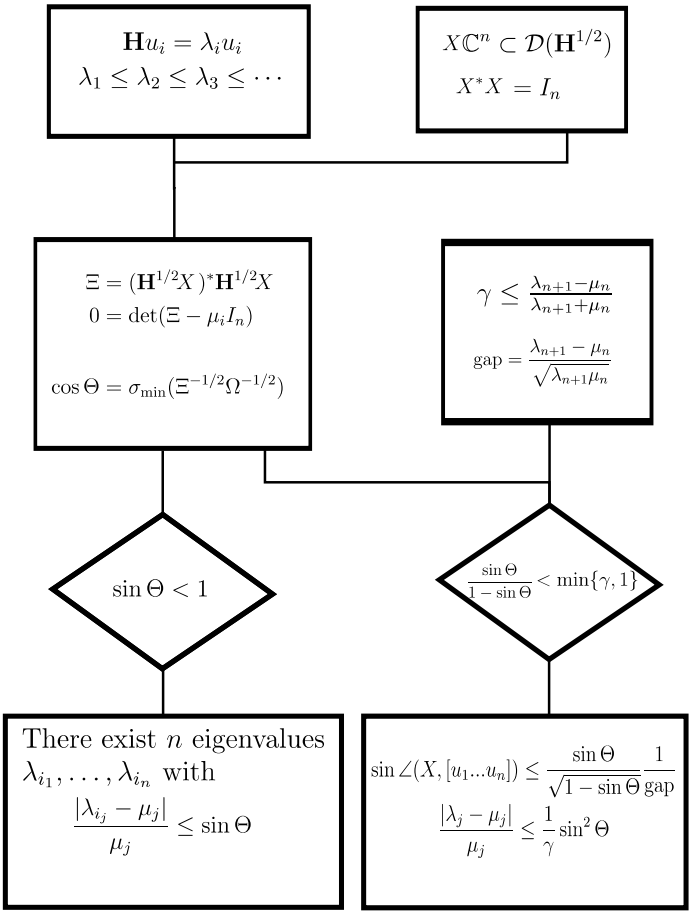

Figure 1.2: A diagrammatic overview of the new perturbation estimates — an interplay between $\Xi = (\mathbf{H}^{1/2}X)^*\mathbf{H}^{1/2}X$ and $\Omega = X^*\mathbf{H}^{-1}X$

Estimates obtained from (1.0.5) are of the "absolute" type, i.e.

$$|\lambda - \mu| \leq \|R\|,$$

whereas the estimates obtained from (1.0.3)–(1.0.4) are of the "relative" type

$$\frac{|\lambda - \mu|}{\mu} \leq \sin\Theta, \qquad \frac{|\lambda - \mu|}{\lambda} \leq \frac{\sin\Theta}{1 - \sin\Theta}.$$

The restriction $\|R\| < \infty$, necessary for (1.0.5) to give useful information in the unbounded operator setting, incurs $\mathsf{ran}(X) \subset \mathcal{D}(H)$. For (1.0.3) and (1.0.3) to be applicable we only need to assume

$$\sin\Theta(LX, L^{-*}X) < 1.$$

This "residual measure" will give nontrivial information even when $\mathsf{ran}(X) \subset \mathcal{D}(L) = \mathcal{D}(H^{1/2})$ is such that $\mathsf{ran}(X) \not\subset \mathcal{D}(H)$.

Notably, both approaches to measure the "residual" share the property:

- $\mathsf{sin}\Theta(LX, L^{-*}X) = 0$ if and only if $\mathsf{ran}(X)$ is the invariant subspace of H
- $\|R\| = 0$ if and only if $\mathsf{ran}(X)$ is the invariant subspace of H.

An important feature of our theory is that it gives an abstract framework for the consideration of both the eigenvalue and eigenvector estimates (see Figure 1.2). The case studies, that will be performed on various model problems from mathematical physics, will demonstrate that the obtained bounds are sharp, see Section 2.7.

We insist on the use of symmetric forms, rather than to work with unbounded operators that are defined by them, for the following reasons:

- Symmetric forms simplify the calculations. More importantly, they allow the test vectors to belong to the form domain of H. This naturally includes linear finite elements for the second order elliptic differential operators.
- The obtained estimates are of the "relative type" whereas the subspace bound is based upon the "relative gap" between the relevant groups of eigenvalues (see Figure 1.2 for the definition of the "relative gap"). The relative gap separates well two close eigenvalues that are themselves small and is therefore particularly suitable for dealing with the lower part of the spectrum of a positive definite operator (see also the diagram on Figure 1.2).
- The energy-scaled measure of the residual is tightly connected with the "dual" norm (or "-1"–norm) of the classical residual $r = Hx - \|H^{1/2}x\|^2 x$, $\|x\| = 1$. We will prove that

$$\sin\Theta(H^{-1/2}x, H^{1/2}x) \leq \frac{\|r\|_{H^{-1}}}{\|H^{1/2}x\|} \leq \frac{\sin\Theta(H^{-1/2}x, H^{1/2}x)}{1 - \sin\Theta(H^{-1/2}x, H^{1/2}x)}, \tag{1.0.6}$$

  where $\|r\|_{H^{-1}} = \sqrt{\langle r, H^{-1}r\rangle}$.

The suggested framework is fairly abstract, so we will present various applications of the new approximation estimates to illustrate our method. As a first application we will consider spectral asymptotics of the family of operators

$$H_\eta = H_b + \eta^2 H_1, \qquad \eta \text{ large}. \tag{1.0.7}$$

Whenever the family (1.0.7) is considered, $\mathsf{ker}(H_1)$ is assumed to be a nontrivial subspace of the environment Hilbert space. As η grows large H_η^{-1} tends to

$$H_\infty^\dagger = (P_{\mathsf{ker}(H_1)} H_b P_{\mathsf{ker}(H_1)})^\dagger,$$

where $\dagger$ denotes the generalized inverse. Assume λ^∞ is an eigenvalue of H_∞, it is then (obviously) a Ritz value of the operator H_η. We will apply the new Ritz value approximation estimates to assess the quality of λ^∞ as an approximate eigenvalue of the operator H_η (for large η). Convergence rate estimates were not studied until recently, see [17] and the references therein. We have stated our results as an abstract approach to spectral asymptotics for the large coupling limit. Our estimates are derived from the local "resolvent" formula

$$\sin^2\Theta(H_\eta^{-1/2}\mathcal{X}, H_\eta^{1/2}\mathcal{X}) = \max_{x\in\mathcal{X}} \frac{(x, H_\eta^{-1}x) - (x, H_\infty^\dagger x)}{(x, H_\eta^{-1}x)}, \qquad H_\infty\mathcal{X} \subset \mathcal{X}.$$

The operators we consider as model problems are used in the modelling of media with a high contrast in the material properties as well as for the analysis of the lower dimensional models of physical phenomena, see [15, 17, 23] for applications in Quantum Mechanics and [51, 58] for applications in Theory of Elasticity. We also identify a class of regular perturbations $\eta^2 H_1$ and formulate a residual based approach to the spectral asymptotics of (1.0.7).

As a second application, we will consider the problem of assessing the quality of finite element approximations to the eigenvalues of nonnegative definite self adjoint operators. Formula (1.0.6) relates $\mathsf{sin}\Theta$-approach to the known (spectral) residual estimates for positive definite operators, cf. [24, 48]. In the finite element literature one usually finds results of the type: *Let x, $\|x\| = 1$ be a test vector, let $\mu = \left(H^{1/2}x, H^{1/2}x\right)$ be the Ritz value. If μ approximates the eigenvalue λ then*

$$\frac{|\lambda - \mu|}{\lambda} \le c\|r\|_\tau.$$

Here, c is a constant of moderate size, τ a finite–element approximation and $\|r\|_\tau$ is a measure of the residual $r = Hx - \mu x$. In fact, (cf. [42])

$$\|Hx - \mu x\|_{H^{-2}} = \sqrt{\langle r, H^{-2}r\rangle} \le c\|r\|_\tau$$

is a bound on the "-2"-norm of the residual r. The Ritz value bound is accompanied by the corresponding subspace error estimate. The subspace error estimate is a function of

$c\|r\|_\tau$ and a subspace stability factor (in the terminology of the paper [42]). The subspace stability factor implicitly contains some measure of a spectral gap and a constant depending on the geometry of the domain. Important feature of our analysis of finite element spectral approximations is a clear separation of the contribution of the perturbation theory of positive operators (localization of the approximated eigenvalues) from the consideration of regularity issues (the geometry of the domain). Furthermore, the applicability of our bounds is not limited to differential operators only.

In these applications we will be measuring residuals to assess the quality of the Rayleigh–Ritz approximations to a part of the spectrum of a positive definite operator. The difference is that in the second case the test space is finite dimensional and in the first case it can be, and usually is, infinite dimensional. However, in both cases we will observe the *decoupling* of the energy space in two subspaces. The subspace (containing the Rayleigh–Ritz test space) which captures all of the important information necessary to measure the residual and its complement whose influence can be bounded away.

Chapter 2

Perturbation approach to the Rayleigh–Ritz method

In this section we develop a perturbation approach to the Rayleigh–Ritz approximations. The idea to represent the eigenvalues (vectors), which we do not know (but want to approximate), as a perturbation of the Ritz values (vectors) which we have computed, goes back to Kahan [40]. The perturbation argument enables us to solve two problems in one go: We determine which part of the spectrum of the operator (infinitely many eigenvalues) is being approximated by the Ritz values (finitely many) and we obtain the approximation estimates. This idea was further developed in [21, 22]. However, in both of these works it was assumed that the test space must belong to the operator domain. We remove this stringent regularity assumption on the test space. In order to do so, we have developed a new perturbation theory particularly suited to the eigenvalue problem in the variational formulation.

2.1 The notation and preliminaries

The environment in this chapter will be a Hilbert space $\mathcal{H}$, with the scalar product $(\cdot,\ \cdot)$. The scalar product is antilinear in the first variable and linear in the second. We start with a closed symmetric form $h(\cdot,\cdot)$ which is additionally assumed to be *positive*

$$h[u] = h(u,u) \geq 0, \qquad u \in \mathcal{Q}(h). \tag{2.1.1}$$

In the sequel when we say nonnegative form h, we shall always mean the closed symmetric form h which satisfies (2.1.1). The form h shall be called *positive definite* when it is closed, symmetric and there exists $m_h > 0$ such that

$$h[u] = h(u,u) \geq m_h \|u\|^2, \qquad u \in \mathcal{Q}(h).$$

There is also an equivalent operator version of these definitions. The self adjoint operator $\mathbf{H}$ is called *positive* if

$$(u, \mathbf{H}u) \geq 0, \qquad u \in \mathcal{D}(\mathbf{H}).$$

Subsequently, $\mathbf{H}$ is called *positive definite* if there exists $m_{\mathbf{H}} > 0$ such that

$$(u, \mathbf{H}u) \geq m_{\mathbf{H}} \|u\|^2, \qquad u \in \mathcal{D}(\mathbf{H}).$$

In this chapter we assume $\overline{\mathcal{Q}}^{\,\mathcal{H}} = \mathcal{H}$, but later we shall also allow $\overline{\mathcal{Q}}^{\,\mathcal{H}}$ to be any nontrivial subspace of $\mathcal{H}$. For nonnegative self adjoint operators one defines, with the help of the spectral theorem, the usual functional calculus. We write the spectral representation of the nonnegative operator $\mathbf{H}$ as

$$\mathbf{H} = \int \lambda \, \mathrm{d}E_{\mathbf{H}}(\lambda),$$

where $E_{\mathbf{H}}(\lambda)$ is the *spectral measure* . When there can be no confusion we write $E(\lambda)$.

The representation theorem for positive forms [41, pp. 331] implies that there exists a self adjoint operator $\mathbf{H}$ such that $\mathcal{D}(\mathbf{H}^{1/2}) = \mathcal{Q}(h)$ and

$$h(u, v) = (\mathbf{H}^{1/2}u, \mathbf{H}^{1/2}v), \qquad u, v \in \mathcal{Q}(h).$$

Following [32] we call $\mathcal{D}(\mathbf{H})$ the *operator domain* of $\mathbf{H}$ and $\mathcal{Q}(\mathbf{H}) = \mathcal{D}(\mathbf{H}^{1/2})$ the *quadratic form domain* of $\mathbf{H}$. We write $\mathcal{D}$ and $\mathcal{Q}$ when there can be no confusion. With the help of the spectral theorem we see that

$$\mathcal{D}(\mathbf{H}) = \{u \in \mathcal{H} : \|\mathbf{H}u\|^2 = \int \lambda^2 \, \mathrm{d}(E(\lambda)u, u) < \infty\},$$
$$\mathcal{Q}(\mathbf{H}) = \{u \in \mathcal{H} : h[u] = \|\mathbf{H}^{1/2}u\|^2 = \int \lambda \, \mathrm{d}(E(\lambda)u, u) < \infty\}.$$

Sometimes we shall write $h = \int \lambda \mathrm{d}(E_{\mathbf{H}}(\lambda)\cdot, \cdot)$ when we want to emphasize the spectral measure generated by the nonnegative operator defined by the form h.

In general, when dealing with the forms in a Hilbert space, we shall follow the terminology of Kato, cf. [41]. In one point we will depart from the conventions in [41]. A positive form

$$h(u, v) = (\mathbf{H}^{1/2}u, \mathbf{H}^{1/2}v)$$

will be called *nonnegative definite* when $\lambda_e(\mathbf{H}) > 0$. Analogously, the positive operator $\mathbf{H}$ such that $\lambda_e(\mathbf{H}) > 0$ will be also called *nonnegative definite*. We will often say nonnegative, meaning the nonnegative definite. Now, we give definitions of some terms that will frequently be used, cf. [32, 41].

Definition 2.1.1. Let h be a positive definite form in $\mathcal{H}$. A sesquilinear form a, which need not be closed, is said to be *h-bounded*, if $\mathcal{Q}(h) \subset \mathcal{Q}(a)$ and there exists $\eta \geq 0$

$$|a[u]| \leq \eta h[u] \qquad u \in \mathcal{Q}(h).$$

Since h is positive definite the space $(\mathcal{Q}(h), h)$ can be considered as a Hilbert space. The form a, which is h-bounded, defines a bounded operator on the space $(\mathcal{Q}(h), h)$.

Definition 2.1.2. A bounded operator $A : \mathcal{H} \to \mathcal{U}$ is called *degenerate* if $\mathsf{ran}(A)$ is finite dimensional.

Definition 2.1.3. If $\mathbf{H}$ is a self adjoint operator and P a projection, to say that P commutes with $\mathbf{H}$ means that $u \in \mathcal{D}(\mathbf{H})$ implies $Pu \in \mathcal{D}(\mathbf{H})$ and

$$\mathbf{H}Pu = P\mathbf{H}u, \qquad u \in \mathcal{D}(\mathbf{H}).$$

Definition 2.1.4. Let $\mathbf{H}$ and $\mathbf{A}$ be nonnegative operators. We define the *order relation* $\leq$ between the nonnegative operators by saying that

$$\mathbf{A} \leq \mathbf{H}$$

if and only if $\mathcal{Q}(\mathbf{H}) \subset \mathcal{Q}(\mathbf{A})$ and

$$\|\mathbf{A}^{1/2}u\| \leq \|\mathbf{H}^{1/2}u\|, \qquad u \in \mathcal{Q}(\mathbf{H}),$$

or equivalently

$$a[u] \leq h[u], \qquad u \in \mathcal{Q}(h),$$

when a and h are nonnegative forms defined by the operators $\mathbf{A}$ and $\mathbf{H}$ and $\mathbf{A} \leq \mathbf{H}$.

Definition 2.1.5. Let h_n, $n \in \mathbb{N}$, be a sequence of positive definite forms. We say that the sequence h_n is uniformly positive definite if there exists a positive definite form s, such that $s \leq h_n$, $n \in \mathbb{N}$.

The main principle we shall use to develop the perturbation theory will be the *monotonicity of the spectrum* with regard to the order relation between nonnegative operators. This principle can be expressed in many ways. The relevant results, which are scattered over the monographs [32, 41], are summed up in the following theorem, see also [44, Corollary A.1].

Theorem 2.1.6. *Let $\mathbf{A} = \int \lambda \, dE_{\mathbf{A}}(\lambda)$ and $\mathbf{H} = \int \lambda \, dE_{\mathbf{H}}(\lambda)$ be nonnegative operators in $\mathcal{H}$ and let $\mathbf{A} \leq \mathbf{H}$. By $0 \leq \mu_1 \leq \mu_2 \leq \cdots < \lambda_e(\mathbf{A})$ and $0 \leq \lambda_1 \leq \lambda_2 \leq \cdots < \lambda_e(\mathbf{H})$ denote the discrete eigenvalues of $\mathbf{A}$ and $\mathbf{H}$, then*

1. $\lambda_e(\mathbf{A}) \leq \lambda_e(\mathbf{H})$

2. $\mathsf{dim}\ E_{\mathbf{H}}(\gamma) \leq \mathsf{dim}\ E_{\mathbf{A}}(\gamma)$, *for every* $\gamma \in \mathbb{R}$

3. $\mu_k \leq \lambda_k, \qquad k = 1, 2, \cdots$.

We close this introductory section with the well known theorem about the perturbation of the *essential spectrum*.

Theorem 2.1.7. *Let* $\mathbf{H}$ *and* $\mathbf{A}$ *be positive definite operators. If the operator*

$$\mathbf{H}^{-1} - \mathbf{A}^{-1}$$

is compact then $\sigma_{ess}(\mathbf{H}) = \sigma_{ess}(\mathbf{A})$.

2.2 The generalized inverse and angle between the subspaces

There are many ways to express that $u \in \mathcal{Q}(h)$ is an eigenvector of the operator $\mathbf{H}$. We will give a geometric characterization of this property. Assume that $\|u\| = 1$ and $\mu = h[u]$. An elementary trigonometric argument yields

$$\|\mathbf{H}^{1/2}u - \mu\mathbf{H}^{-1/2}u\| = 0 \Leftrightarrow \sin\Theta(\mathbf{H}^{1/2}u, \mathbf{H}^{-1/2}u) = 0. \tag{2.2.1}$$

(2.2.1) implies that u is an eigenvector of $\mathbf{H}$ if and only if $\mathsf{sin}\Theta(\mathbf{H}^{1/2}u, \mathbf{H}^{-1/2}u) = 0$. The ability to assess the size of $\mathsf{sin}\Theta(\mathbf{H}^{1/2}u, \mathbf{H}^{-1/2}u)$ will be central to the analysis of the Rayleigh–Ritz method in this thesis.

In this section we give the background information on the angles between two finite dimensional subspaces of a Hilbert space as given in [21, 41, 62]. Basic results on generalized inverses of (unbounded) operators defined between two Hilbert spaces will be presented as well. These results will be applied to the problem of computing $\mathsf{sin}\Theta(\mathbf{H}^{1/2}\mathcal{X}, \mathbf{H}^{-1/2}\mathcal{X})$ for the given positive definite $\mathbf{H}$ and some finite dimensional $\mathcal{X} \subset \mathcal{Q}(\mathbf{H})$.

Closed subspaces of the Hilbert space $\mathcal{H}$ can always be represented as images of the appropriate orthogonal projections. We shall mix the notation for projections and their images when appropriate. For instance, we shall speak about the dimension of the projection P meaning the dimension of the range of the projection P. In the case in which P is finite dimensional, we have another representation for the subspace $\mathsf{ran}(P)$. For a given n-dimensional subspace $\mathsf{ran}(P) \subset \mathcal{Q}$ there exists an isometry $X : \mathbb{C}^n \to \mathcal{H}$ such that

$\mathsf{ran}(P) = \mathsf{ran}(X)$, where $P = XX^*$. Therefore, $\mathsf{ran}(X)$ is an alternative representation of the n-dimensional subspace $\mathsf{ran}(P)$. The isometry X will be called the basis of the subspace $\mathsf{ran}(P)$. We shall freely use both representation of the finite dimensional subspace. $P_X = XX^*$ will generically denote the orthogonal projection on the space $\mathsf{ran}(X)$ (for some isometry $X : \mathbb{C}^n \to \mathcal{H}$).

Let $\mathsf{ran}(P)$ and $\mathsf{ran}(Q)$ be two finite dimensional subspaces of the Hilbert space $\mathcal{H}$. The function $\angle$ that measures the separation of the pair of subspaces $\mathsf{ran}(P)$ and $\mathsf{ran}(Q)$ will be called an *angle function* if it satisfies the following properties

1. $\angle(P,Q) \geq 0$ and
   $\angle(P,Q) = 0$ if and only if $\mathsf{ran}(P) \subset \mathsf{ran}(Q)$ or $\mathsf{ran}(Q) \subset \mathsf{ran}(P)$.
2. $\angle(P,Q) = \angle(Q,P)$
3. $\angle(P,Q) \leq \angle(P,R) + \angle(R,Q)$ if
   $\mathsf{dim}(\mathsf{ran}(P)) \leq \mathsf{dim}(\mathsf{ran}(R)) \leq \mathsf{dim}(\mathsf{ran}(Q))$ or
   $\mathsf{dim}(\mathsf{ran}(P)) \geq \mathsf{dim}(\mathsf{ran}(R)) \geq \mathsf{dim}(\mathsf{ran}(Q))$
4. $\angle(UP,UQ) = \angle(P,Q)$, for any unitary U.

In this thesis we will use the following angle functions, see [62],

$$\Theta(P,Q) = \arcsin\max\{\|P(\mathbf{I}-Q)\|, \|Q(\mathbf{I}-P)\|\} \tag{2.2.2}$$

$$\Theta_p(P,Q) = \arcsin\min\{\|P(\mathbf{I}-Q)\|, \|Q(\mathbf{I}-P)\|\} \tag{2.2.3}$$

The function $\Theta(P,Q)$ from (2.2.2) will be called the *maximal canonical angle* between the subspaces P and Q. The function $\Theta_p(P,Q)$ from (2.2.3) will be called the *maximal principal angle* between the subspaces P and Q.

The following theorem of Kato describes the relation between two finite dimensional subspaces P and Q, cf. [41, Theorem I-6.34].

Theorem 2.2.1. *Let P and Q be two orthogonal projections such that*

$$\|P(\mathbf{I}-Q)\| < 1.$$

Then there are following alternatives. Either

1. $\mathsf{ran}(P)$ *and* $\mathsf{ran}(Q)$ *are isomorphic and*

$$\|P(\mathbf{I}-Q)\| = \|Q(\mathbf{I}-P)\| = \|P-Q\| \qquad \textit{or}$$

2. $\mathsf{ran}(P)$ *is isomorphic to the true subspace of* $\mathsf{ran}(Q)$ *and*

$$\|Q(\mathbf{I}-P)\| = \|P-Q\| = 1.$$

From this theorem we get an insight in the behavior of the canonical and the principal angles that were defined by (2.2.2) and (2.2.3).

Corollary 2.2.2. *Let* P *and* Q *be two orthogonal projections such that* $\mathsf{dim}(\mathsf{ran}(P)) \leq \mathsf{dim}(\mathsf{ran}(Q))$ *and let*

$$\|P(\mathbf{I}-Q)\| < 1.$$

Then there are following alternatives. Either

1. $\mathsf{dim}(\mathsf{ran}(P)) = \mathsf{dim}(\mathsf{ran}(Q))$ *and*

$$\sin\Theta(P,Q) = \sin\Theta_p(P,Q) = \|P-Q\| < 1 \qquad \textit{or}$$

2. $\mathsf{dim}(\mathsf{ran}(P)) < \mathsf{dim}(\mathsf{ran}(Q))$ *and*

$$\sin\Theta_p(P,Q) = \|P(\mathbf{I}-Q)\| < 1.$$

For most of our needs, Theorem 2.2.1 describes the relation between the finite dimensional subspaces $\mathsf{ran}(P)$ and $\mathsf{ran}(Q)$ in sufficient detail. However, sometimes it will be necessary to analyze the structure of the finite dimensional projections $P_V = VV^*$ and $P_U = UU^*$ in further detail. To this end we define the *canonical angles* $\theta_1, \ldots, \theta_n$ between the spaces $\mathsf{ran}(U)$ and $\mathsf{ran}(V)$ as

$$\theta_i = \arccos\sigma_i, \quad i = 1, \ldots, n, \tag{2.2.4}$$

where $\sigma_1, \ldots, \sigma_n$ are the singular values of the matrix

$$V^*U \in \mathbb{C}^{n\times n}.$$

The canonical angles are related to the angle function (2.2.2) through the formula

$$\sin\Theta(P_V, P_U) = \max_i \sin\theta_i.$$

We also define the *acute principal angles* $\theta_1^p \leq \theta_2^p \leq \cdots \leq \theta_k^p$, where $k \leq n$, as those canonical angles θ_i which satisfy the condition $0 < \theta_i < \pi/2$. Subsequently, we obtain a connection to the angle function (2.2.3) through the formula

$$\sin\Theta_p(P_V, P_U) = \max_i \sin\theta_i^p.$$

In dealing with the projections and degenerate operators it is useful to have a notion of the generalized inverse. We will define the generalized inverse of a bounded operator on $\mathcal{H}$ or of the closed densely defined operator in $\mathcal{H}$ following [46], see also [41, Chapter IV.5].

Definition 2.2.3. Let $\mathbf{T} : \mathcal{H} \to \mathcal{U}$ be a closed operator such that $\overline{\mathcal{D}(\mathbf{T})} = \mathcal{H}$. The operator $\mathbf{T}^\dagger : \mathcal{U} \to \mathcal{H}$ is defined by

$$\begin{aligned} \mathcal{D}(\mathbf{T}^\dagger) &= \mathsf{ran}(\mathbf{T}) \oplus \mathsf{ran}(\mathbf{T})^\perp \\ \mathbf{T}^\dagger u &= (\mathbf{T} \mid_{\mathsf{ker}(\mathbf{T})^\perp})^{-1} P_{\mathsf{ran}(\mathbf{T})} u, \qquad u \in \mathcal{D}(\mathbf{T}^\dagger) \end{aligned}$$

and it is called the *Moore–Penrose generalized inverse* of T.

The properties of the generalized inverse[1] are analyzed in the monograph [46]. In particular we use the following characterization.

Theorem 2.2.4 (see [46, Theorem I.5.7]). *Let $\mathbf{T} : \mathcal{H} \to \mathcal{U}$ be the closed operator and let $\overline{\mathcal{D}(\mathbf{T})} = \mathcal{H}$, then $\mathbf{T}^\dagger$ is the unique closed operator such that*

$$\begin{aligned} \mathbf{T}^\dagger \mathbf{T} \mathbf{T}^\dagger &= \mathbf{T}^\dagger, \qquad \text{on } \mathcal{D}(\mathbf{T}^\dagger) \\ \mathbf{T}\mathbf{T}^\dagger &= P_{\mathsf{ran}(\mathbf{T})}\big|_{\mathcal{D}(\mathbf{T}^\dagger)} \\ \mathbf{T}^\dagger \mathbf{T} &= P_{\mathsf{ker}(\mathbf{T})^\perp}\big|_{\mathcal{D}(\mathbf{T})} \end{aligned}$$

where $P_{\mathcal{M}}$ is the orthogonal projection on $\mathcal{M}$.

Assume $\mathbf{H}$ is a nonnegative operator then $\mathbf{H} = \int \lambda \, dE(\lambda)$ and $\mathbf{H}^\dagger$ is also nonnegative. The operator $\mathbf{H}^\dagger$ has the spectral decomposition

$$\mathbf{H}^\dagger = \int \frac{1}{\lambda} \, dE(\lambda), \qquad \mathcal{D}(\mathbf{H}^\dagger) = \{u \in \mathcal{H} : \int \frac{1}{\lambda^2} \, d(E(\lambda)u, u) < \infty\},$$

and the functional calculus implies

$$\mathbf{H}^{\dagger 1/2} = \mathbf{H}^{1/2\dagger}.$$

With the following theorem of Kato we close the preliminary discussion of the generalized inverses, cf. [41, Theorem IV.5.2].

[1] The generalized inverses can also be defined in more general settings. Their properties are also analyzed in [46].

Theorem 2.2.5. *The closed operator $\mathbf{T} : \mathcal{H} \to \mathcal{U}$ has the closed range if and only if there exists $\gamma > 0$ such that*

$$\|\mathbf{T}u\| \geq \gamma \|(\mathbf{I} - P_{\mathsf{ker}(\mathbf{T})})u\| \ , \qquad u \in \mathcal{D}(\mathbf{T}).$$

Proof. Let $\mathbf{T}$ have the closed range, then the operator $\mathbf{T}^\dagger$ is closed and everywhere defined, therefore bounded. For a given $u \in \mathcal{D}(\mathbf{T})$ Theorem 2.2.4 implies

$$\|\mathbf{T}^\dagger \mathbf{T}u\| = \|(\mathbf{I} - P_{\mathsf{ker}(\mathbf{T})})u\|$$

and then

$$\|\mathbf{T}u\| \geq \frac{1}{\|\mathbf{T}^\dagger\|} \|(\mathbf{I} - P_{\mathsf{ker}(\mathbf{T})})u\|.$$

The second part of the statement is obvious. □

Theorems 2.2.4 and 2.2.5 show the relation between the Moore–Penrose generalized inverses and orthogonal projections in a Hilbert space. This is precisely the reason why the generalized inverses will be useful in our study.

In the case in which T is also a degenerate operator more is known. Let $T = BC$, where

$$\begin{aligned} B &: \mathcal{H}'' \to \mathcal{H}', \quad \text{injective}, \\ C &: \mathcal{H} \to \mathcal{H}'', \quad \text{surjective}, \end{aligned}$$

and $\mathcal{H}, \mathcal{H}', \mathcal{H}''$ are any Hilbert spaces. The generalized inverse $T^\dagger$ is given (cf. [34, Gantmacher, Ch I, §5]) by

$$T^\dagger = C^*(CC^*)^{-1}(B^*B)^{-1}B^*. \tag{2.2.5}$$

The following lemma is a generalization of a result by Drmač [26] for finite matrices. The proof, which appeared in [37], is a minor modification of the original finite dimensional proof by Drmač. We give it for the sake of completeness.

Lemma 2.2.6. *Let $\mathcal{H}, \mathcal{H}'$ be Hilbert spaces and $\mathbf{R} : \mathcal{H} \to \mathcal{H}'$ a closed, densely defined linear operator satisfying*

$$\|\mathbf{R}x\| \geq \delta \|x\|, \quad \delta > 0.$$

Let $X : \mathbb{C}^n \to \mathcal{H}$ be a degenerate isometry with $\mathsf{ran}(X) \subset \mathcal{D}(\mathbf{R})$. Let

$$Y = \mathbf{R}X, \quad Z = \mathbf{R}\mathbf{H}^{-1}X, \quad \mathbf{H} = \mathbf{R}^*\mathbf{R}$$

and suppose[2]

$$\mathsf{ran}(Y) \cap (\mathsf{ran}(Z))^\perp = (\mathsf{ran}(Y))^\perp \cap \mathsf{ran}(Z) = \{0\}. \tag{2.2.6}$$

[2]It is sometimes said that the subspaces $\mathsf{ran}(Y)$ and $\mathsf{ran}(Z)$, which satisfy (2.2.6), are in the *acute position*, cf. [21].

Then there is an orthonormal basis in $\mathcal{K} = \mathsf{ran}(Y) \dot{+} \mathsf{ran}(Z)$ *such that in* $\mathcal{K} \oplus \mathcal{K}^{\perp}$ *the operator* YZ^* *is represented by*

$$YZ^* = \left[\begin{array}{ccc|c} \mathbf{I}_k & & & \\ & \bigoplus_{i=1}^{l} \begin{bmatrix} 1 & \tan\theta_i \\ 0 & 0 \end{bmatrix} & & 0 \\ & & 0 & \\ \hline & 0 & & 0 \end{array}\right], \tag{2.2.7}$$

where $\theta_1, \ldots, \theta_l$ *are those angles between* $\mathsf{ran}(Y)$ *and* $\mathsf{ran}(Z)$*, which are different from* 0 *and* $\frac{\pi}{2}$*.*

Proof. Note that $\mathbf{RH}^{-1}$ is everywhere defined and bounded. We have

$$\mathbf{R}^{\dagger} = \overline{\mathbf{H}^{-1}\mathbf{R}^*} = (\mathbf{RH}^{-1})^*,$$

moreover, [41, Ch. V], we conclude $\mathsf{ran}(\mathbf{R}^{\dagger *}) \subseteq \mathcal{D}(\mathbf{R}^*)$. Thus, $Y, Z : \mathbb{C}^n \to \mathcal{H}'$ are bounded and injective. We will prove the identities

$$\begin{aligned} Z^*Y &= \mathbf{I} \qquad (\text{on } \mathbb{C}^n) && (2.2.8)\\ YZ^* &= (P_Z P_Y)^{\dagger}, && (2.2.9) \end{aligned}$$

where P_Z, P_Y are the orthogonal projections onto $\mathsf{ran}(Z)$, $\mathsf{ran}(Y)$, respectively. For any x, y we have $\mathbf{R}^{\dagger *}Xy \in \mathcal{D}(\mathbf{R}^*)$,

$$\begin{aligned} (Yx, Zy) &= \big(\mathbf{R}Xx, \mathbf{R}^{\dagger *}Xy\big) \\ &= \big(Xx, \mathbf{R}^*\mathbf{R}^{\dagger *}Xy\big) = \big(Xx, \mathbf{R}^*\mathbf{R}\mathbf{H}^{-1}Xy\big) \\ &= (Xx, Xy) = (x, y) \end{aligned}$$

and (2.2.8) follows. Furthermore, since $\mathsf{ran}(Y)$ and $(\mathsf{ran}(Z))^{\perp}$ have the zero intersection the operator

$$P_Z Y = ZZ^{\dagger}Y,$$

is injective, whereas $Y^{\dagger}$ is surjective. Thus, we compute

$$\begin{aligned} (P_Z P_Y)^{\dagger} &= (ZZ^{\dagger}YY^{\dagger})^{\dagger} \\ &= (Y^{\dagger})^{\dagger}[(ZZ^{\dagger}Y)^*(ZZ^{\dagger}Y)]^{-1}(ZZ^{\dagger}Y)^* \\ &= Y[(Z^{\dagger}Y)^*Z^*Z(Z^{\dagger}Y)]^{-1}(Z^{\dagger}Y)^*Z^* \\ &= Y(Z^{\dagger}Y)^{-1}(Z^*Z)^{-1}(Z^{\dagger}Y)^{-*}(Z^{\dagger}Y)^*Z^* \\ &= Y(Z^{\dagger}Y)^{-1}(Z^*Z)^{-1}Z^* \\ &= Y(Z^*ZZ^{\dagger}Y)^{-1}Z^* \\ &= Y(Z^*Y)^{-1}Z^*. \end{aligned}$$

Since both P_Z and P_Y are degenerate — the finite dimensional subspace $\mathcal{K} = \mathsf{ran}(X) \dotplus \mathsf{ran}(Y)$ reduces both of them — Wedin's theorem [62] guarantees the existence of an orthonormal basis in $\mathcal{K}$ such that in $\mathcal{K} \oplus \mathcal{K}^\perp$ projections P_Z, P_Y are represented as

$$P_Z = \left[\begin{array}{cc|c} \mathbf{I}_k & & \\ & \bigoplus_{i=1}^{l} \Phi & \\ & & \Delta_Z \\ \hline & & 0 \end{array}\right], \tag{2.2.10}$$

$$P_Y = \left[\begin{array}{cc|c} \mathbf{I}_k & & \\ & \bigoplus_{i=1}^{l} \Psi(\theta_i) & \\ & & \Delta_Y \\ \hline & & 0 \end{array}\right], \tag{2.2.11}$$

with

$$\Phi = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad \Psi(\theta) = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta \end{bmatrix}.$$

By (2.2.5) we have

$$\begin{aligned} &\left(\begin{bmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{bmatrix} \begin{bmatrix} \cos\theta_i & 0 \\ 0 & 0 \end{bmatrix} \right)^\dagger \\ &= \left(\begin{bmatrix} \cos\theta_i \\ \sin\theta_i \end{bmatrix} \begin{bmatrix} \cos\theta_i & 0 \end{bmatrix} \right)^\dagger \\ &= \begin{bmatrix} \cos\theta_i \\ 0 \end{bmatrix} \frac{1}{\cos^2\theta} \cdot 1 \cdot \begin{bmatrix} \cos\theta_i & \sin\theta_i \end{bmatrix} = \begin{bmatrix} 1 & \tan\theta_i \\ 0 & 0 \end{bmatrix}. \end{aligned}$$

Now by (2.2.9) the conclusion (2.2.7) follows. □

Another class of operators for which the generalized inverse can be given by a simple formula are partial isometries. A bounded operator $W : \mathcal{H} \to \mathcal{U}$ is called *partially isometric* if there exists a closed subspace $\mathcal{M} \subset \mathcal{H}$ such that

$$\|Wu\| = \|P_{\mathcal{M}} u\|, \qquad u \in \mathcal{H}.$$

This is equivalent to

$$W^*W = P_{\mathcal{M}}.$$

The set $\mathcal{M} = \mathsf{ran}(W^*) \subset \mathcal{H}$ is called the initial set of the partial isometry W and $\mathsf{ran}(W) \subset \mathcal{U}$ is called the final set. Since $\mathsf{ker}(W^*) \oplus \mathsf{ran}(W)$ we see

$$WW^* = P_{\mathsf{ran}(W)},$$

so W^* is also the partial isometry with the initial set $\mathsf{ran}(W)$. We shall also use the notation

$$W^*W = P_{W^*}, \qquad WW^* = P_W.$$

It is obvious

$$W^* = W^\dagger$$

and we have the following lemma.

Lemma 2.2.7. *A bounded operator $W : \mathcal{H} \to \mathcal{U}$ is partially isometric if and only if*

$$WW^*W = W.$$

Proof. The assumption $W = WW^*W$ implies

$$(W^*W)(W^*W) = W^*(WW^*W) = W^*W,$$

therefore $P = W^*W$ is an orthogonal projection, so W is partially isometric. If the bounded operator W is partially isometric then $W^*W = P$ is an orthogonal projection and $Pu = 0$ implies $Wu = 0$. So,

$$WPu = Wu, \qquad u \in \mathcal{H}$$

and $WW^*W = W$ follows. □

Lemma 2.2.8. *Let V and W be two partial isometries, then*

$$\|P_V P_W\| = \|VP_W\| = \|V^*W\|.$$

Proof. Using Lemma 2.2.7 we compute

$$\begin{aligned}\|P_V P_W\|^2 &= \mathrm{spr}(P_W P_V P_W) = \mathrm{spr}(WW^*VV^*WW^*) \\ &= \mathrm{spr}(W^*VV^*WW^*W) = \mathrm{spr}(W^*VV^*W) = \|V^*W\|.\end{aligned}$$

Since, for bounded operators A, B, C, the identity

$$\mathrm{spr}(ABC) = \mathrm{spr}(CAB)$$

holds. □

In the preparation for the following chapters, we will state another property of the partially isometric operators. The following lemma follows directly from [41, Theorem IV.5.13]

Lemma 2.2.9. *Let $W : \mathcal{H} \to \mathcal{H}'$ be partially isometric and let* $\dim(\mathsf{ker}(W)) < \infty$*, then*

$$\dim(\mathsf{ran}(W)^\perp) = \dim(\mathsf{ker}(W^*)).$$

2.3 Geometrical properties of the Ritz value perturbation

In this section we will present a perturbation approach to the Rayleigh–Ritz approximation of the spectrum of a positive definite operator. The nonnegative definite case is technically more complex and warrants a separate section. Although this section is devoted to the positive definite case, some of the statements and definitions will be given in full generality in which they will be later used in the text.

Let $0 \leq h$ be a nonnegative form and let $\mathsf{ran}(X) \subset \mathcal{Q}(h)$ be the n-dimensional test space. The matrix

$$\Xi_{h,X} = (\mathbf{H}^{1/2}X)^*\mathbf{H}^{1/2}X \in \mathbb{C}^{n\times n}$$

will be called the *Rayleigh quotient* associated to the basis X. When there can be no confusion, we shall denote the Rayleigh quotient by Ξ and drop the indexes. The eigenvalues of the matrix Ξ will be numbered in the ascending order

$$\mu_1 \leq \mu_2 \leq \cdots \leq \mu_n. \tag{2.3.1}$$

We call the numbers μ_i the Ritz values of the operator $\mathbf{H}$ (form h) from the subspace $\mathsf{ran}(X)$. This definition is correct since the eigenvalues of the matrix Ξ do not depend on the choice of the basis X. In the rest of this chapter we will use $P = XX^*$ to denote the projection onto the range of the isometry $X : \mathbb{C}^n \to \mathcal{H}$.

For the given h and $\mathsf{ran}(X) \subset \mathcal{Q}(h)$, $P = XX^*$, we define the symmetric forms δh and h' using the formulae

$$\delta h(u,v) = h(Pu,(\mathbf{I}-P)v) + h((\mathbf{I}-P)u,Pv), \qquad u,v \in \mathcal{Q}(h) \tag{2.3.2}$$

$$h'(u,v) = h(Pu,Pv) + h((\mathbf{I}-P)u,(\mathbf{I}-P)v), \qquad u,v \in \mathcal{Q}(h). \tag{2.3.3}$$

Obviously, (2.3.2) and (2.3.3) imply

$$h'(u,v) = h(u,v) - \delta h(u,v), \qquad u,v \in \mathcal{Q}(h). \tag{2.3.4}$$

In what follows we will describe the properties of the symmetric form h' and the operator $\mathbf{H}'$ it generates.

Lemma 2.3.1. *Let the nonnegative form h and the subspace $\mathsf{ran}(X) \subset \mathcal{Q}$ be given. The form h' is closed and nonnegative. If h is positive definite, then so is h'.*

PROOF. The operator $\mathbf{H}^{1/2}(\mathbf{I}-P)$ is closed and so is the form

$$h((\mathbf{I}-P)\cdot,(\mathbf{I}-P)\cdot)=(\mathbf{H}^{1/2}(\mathbf{I}-P)\cdot,\mathbf{H}^{1/2}(\mathbf{I}-P)\cdot).$$

On the other hand $h(P\cdot,P\cdot)$ is a bounded form. Altogether, h' is a closed nonnegative symmetric form. If there exists $\delta>0$ such that

$$\delta\|u\|^2\le h[u],\qquad u\in\mathcal{Q}$$

then also

$$\delta\|u\|^2\le h'[u],\qquad u\in\mathcal{Q}.$$ □

Lemma 2.3.1 assures us that the construction (2.3.3) defines a nonnegative operator $\mathbf{H}'$ and

$$h'(u,v)=(\mathbf{H}'^{1/2}u,\mathbf{H}'^{1/2}v),\qquad u,v\in\mathcal{Q}(h).$$

It is little less obvious that the Ritz values μ_i are among the eigenvalues of the operator $\mathbf{H}'$. This property, which shall be made precise in the following lemma, is the basic feature that enables us to establish the perturbation approach to the problem of assessing the accuracy of the Rayleigh–Ritz approximations of the spectrum.

Lemma 2.3.2. *Let the nonnegative definite form h and the subspace $\mathsf{ran}(X)\subset\mathcal{Q}$ be given. Let $\mathbf{H}$ be the nonnegative definite operator defined by the form h. Then $\sigma_{ess}(\mathbf{H})=\sigma_{ess}(\mathbf{H}')$ and*

$$\mathbf{H}'X=X\Xi,\tag{2.3.5}$$

for $\Xi=(\mathbf{H}^{1/2}X)^*\mathbf{H}^{1/2}X\in\mathbb{C}^{n\times n}$. (2.3.5) is equivalent to the statement that $P=XX^*$ commutes with $\mathbf{H}'$.

PROOF. We will now show that the subspace $\mathsf{ran}(X)$ reduces $\mathbf{H}'$. Indeed, for $y\in\mathcal{Q}$, $x\in\mathbb{C}^n$ we have

$$\begin{aligned}h'(y,Xx)&=(\mathbf{H}^{1/2}y,\mathbf{H}^{1/2}Xx)-(\mathbf{H}^{1/2}(\mathbf{I}-P)y,\mathbf{H}^{1/2}Xx)\\&=(\mathbf{H}^{1/2}XX^*y,\mathbf{H}^{1/2}Xx)\\&=(\Xi X^*y,x).\end{aligned}$$

This is equivalent to

$$(\mathbf{H}'^{1/2}y,\mathbf{H}'^{1/2}Xx)=(y,X\Xi x),\quad y\in\mathcal{Q},\ x\in\mathbb{C}^n.$$

It implies $\mathsf{ran}(X) \subset \mathcal{D}(\mathbf{H}')$ and

$$(y, \mathbf{H}'Xx - X\Xi x) = 0$$

for all $y \in \mathcal{H}$, $x \in \mathbb{C}^n$. Hence,

$$\mathbf{H}'X = X\Xi \tag{2.3.6}$$

which is equivalent to the statement that P commutes with $\mathbf{H}'$ (see Definition 2.1.3). We now prove that $\sigma_{ess}(\mathbf{H}) = \sigma_{ess}(\mathbf{H}')$. Assume h is a positive definite form, then Lemma 2.3.1 implies that h' is a positive definite form, too. From (2.3.4) we obtain

$$\delta h(\mathbf{H}^{-1}u, \mathbf{H}'^{-1}v) = (\mathbf{H}'^{-1}u - \mathbf{H}^{-1}u, v), \qquad u, v \in \mathcal{H}.$$

On the other hand

$$\delta h(\mathbf{H}^{-1}u, \mathbf{H}'^{-1}v) = (\mathbf{H}^{1/2}P\mathbf{H}^{-1}u, \mathbf{H}^{1/2}P_\perp\mathbf{H}'^{-1}v) + (\mathbf{H}^{1/2}P_\perp\mathbf{H}^{-1}u, \mathbf{H}^{1/2}P\mathbf{H}'^{-1}v)$$

defines a compact operator. Theorem 2.1.7 implies $\sigma_{ess}(\mathbf{H}) = \sigma_{ess}(\mathbf{H}')$ and the statement of the theorem is proved for a positive definite h. In the general case, take $\alpha > 0$. The form $\widetilde{h}(u, v) = h(u, v) + \alpha(u, v)$ is positive definite. Furthermore, we establish

$$\widetilde{h}'(u, v) = \alpha(u, v) + h'(u, v)$$
$$\delta\widetilde{h}(u, v) = \delta h(u, v),$$

so $\sigma_{ess}(\widetilde{\mathbf{H}}) = \sigma_{ess}(\widetilde{\mathbf{H}}')$. The conclusion $\sigma_{ess}(\mathbf{H}) = \sigma_{ess}(\mathbf{H}')$ follows by the spectral mapping theorem. □

Corollary 2.3.3. *Let the nonnegative definite form h and the subspace $\mathsf{ran}(X) \subset \mathcal{Q}$ be given. The projections P and $P_{\mathsf{ran}(\mathbf{H}')}$ commute and $\mathsf{ker}(\mathbf{H}') \subset \mathsf{ker}(\mathbf{H})$.*

Proof. This corollary is a direct consequence of (2.3.3) and the preceding theorem. □

Remark 2.3.4. For positive definite h Lemma 2.3.2 describes the operator $\mathbf{H}'$ in sufficient detail. For a general nonnegative h the operator $\mathbf{H}'$ has somewhat more complex structure. Finer properties of the operator $\mathbf{H}'$, constructed in the case in which h is a general nonnegative form, will be discussed in Section 2.3.1.

We now concentrate on the positive definite case.

Theorem 2.3.5. *Let the subspace $\mathsf{ran}(X) \subset \mathcal{Q}$ be given and let h be positive definite. Assume $\mathsf{sin}\Theta(\mathbf{H}^{1/2}X, \mathbf{H}^{-1/2}X) = \mathsf{sin}\Theta < 1$, then*

$$(1 - \sin\Theta)h'[u] \le h[u] \le (1 + \sin\Theta)h'[u], \qquad u \in \mathcal{Q}(h) \tag{2.3.7}$$

$$(1 - \frac{\sin\Theta}{1 - \sin\Theta})h[u] \le h'[u] \le (1 + \frac{\sin\Theta}{1 - \sin\Theta})h[u], \qquad u \in \mathcal{Q}(h). \tag{2.3.8}$$

PROOF. The product $\mathbf{H}^{1/2}\mathbf{H}'^{-1/2}$ is well defined since $\mathcal{Q} = \mathcal{D}(\mathbf{H}^{1/2}) = \mathcal{D}(\mathbf{H}'^{1/2})$. This implies that the form

$$\delta h_s(x, y) = \delta h(\mathbf{H}'^{-1/2}x, \mathbf{H}'^{-1/2}y)$$

defines the bounded operator δH_s. After the substitutions $u = \mathbf{H}'^{-1/2}x$, $v = \mathbf{H}'^{-1/2}y$ we obtain

$$\max_{u,v\in\mathcal{Q}(h)} \frac{|\delta h(u, v)|}{\sqrt{h'[u]h'[v]}} = \|\delta H_s\|. \tag{2.3.9}$$

We now show $\|\delta H_s\| = \sin\Theta$. Set

$$V = \mathbf{H}^{1/2}P\mathbf{H}'^{-1/2} \tag{2.3.10}$$

$$W = \mathbf{H}^{1/2}P_\perp\mathbf{H}'^{-1/2}, \tag{2.3.11}$$

with $P_\perp = \mathbf{I} - P$. Relation (2.3.4) implies

$$\begin{aligned}\delta h(\mathbf{H}'^{-1/2}u, \mathbf{H}'^{-1/2}v) &= h(P_\perp\mathbf{H}'^{-1/2}u, P\mathbf{H}'^{-1/2}v) + h(P\mathbf{H}'^{-1/2}u, P_\perp\mathbf{H}'^{-1/2}v)\\ &= (Wu, Vv) + (Vu, Wv),\end{aligned} \tag{2.3.12}$$

which can be written as

$$\delta H_s = V^*W + W^*V. \tag{2.3.13}$$

The equations (2.3.10)–(2.3.13) yield

$$VW^* = 0 \tag{2.3.14}$$

$$WV^* = 0 \tag{2.3.15}$$

$$\|\delta H_s\| = \|W^*VV^*W + V^*WW^*V\| = \|V^*W\|. \tag{2.3.16}$$

As the next step we establish that V and W are partial isometries such that

$$\mathsf{ran}(V) = \mathsf{ran}(\mathbf{H}^{1/2}P) \tag{2.3.17}$$

$$\mathsf{ran}(W)^\perp = \mathsf{ran}(\mathbf{H}^{-1/2}P). \tag{2.3.18}$$

The proof will follow from Lemma 2.3.2. It runs along the same lines in both cases, so we will only present the proof for W. Take some $u, v \in \mathcal{H}$, then

$$\begin{aligned}(Wu, Wv) &= (\mathbf{H}^{1/2}P_\perp\mathbf{H}'^{-1/2}u, \mathbf{H}^{1/2}P_\perp\mathbf{H}'^{-1/2}v)\\ &= h(P_\perp\mathbf{H}'^{-1/2}u, P_\perp\mathbf{H}'^{-1/2}v) = h'(P_\perp\mathbf{H}'^{-1/2}u, P_\perp\mathbf{H}'^{-1/2}v) = (P_\perp u, v),\end{aligned}$$

so $W^*W = P_\perp$. This proves that W is a partial isometry.

Relation (2.3.17) is obvious, since

$$\mathsf{ran}(\mathbf{H}^{1/2}P\mathbf{H}'^{-1/2}) = \mathsf{ran}(\mathbf{H}^{1/2}P)$$

is guaranteed by the assumption $\mathsf{ran}(P) \subset \mathcal{Q}(h)$ and the injectivity of $\mathbf{H}'^{-1/2}$.

The proof of (2.3.18) requires a bit more work. One computes

$$W^*\mathbf{H}^{-1/2}P = \mathbf{H}'^{-1/2}P_\perp\mathbf{H}^{1/2}\mathbf{H}^{-1/2}P = 0,$$

which implies

$$\mathsf{ran}(\mathbf{H}^{-1/2}P) \subset \mathsf{ker}(W^*) = \mathsf{ran}(W)^\perp.$$

On the other hand

$$W^* = P_\perp A, \tag{2.3.19}$$

where $A = \overline{\mathbf{H}'^{-1/2}\mathbf{H}^{1/2}} : \mathcal{H} \to \mathcal{H}$ is a homeomorphism (of linear topological vector spaces), so

$$\mathsf{dim\ ker}(W^*) = \mathsf{dim\ ker}(P_\perp) = \mathsf{dim\ ran}(P) = \mathsf{dim\ ran}(\mathbf{H}^{-1/2}P)$$

and (2.3.18) is established. The assumption $\mathsf{sin}\Theta < 1$ and Lemma 2.2.8 guarantee

$$\sin\Theta = \|V^*W\|.$$

Finally, using (2.3.9) we establish

$$(1 - \sin\Theta)h'[u] \leq h[u] \leq (1 + \sin\Theta)h'[v],$$

which is the statement (2.3.7).

It is a well known fact that given some $0 < \lambda, \mu$ and $0 < \eta < 1$ the implication

$$\frac{|\lambda - \mu|}{\mu} \leq \eta \Rightarrow \frac{|\lambda - \mu|}{\lambda} \leq \frac{\eta}{1 - \eta} \tag{2.3.20}$$

holds. Since h and h' are positive definite forms, the relation (2.3.8) is proved. □

Take any positive definite form h, then

$$h(u, v) = (\mathbf{H}^{1/2}u, \mathbf{H}^{1/2}v) \tag{2.3.21}$$

is only one of the possible operator representations of the form h. All of the preceding results are independent of the choice of the representation $h(u, v) = (\mathbf{R}u, \mathbf{R}v)$, since

$$\sin\Theta = \max_{u,v\in\mathcal{Q}} \frac{|\delta h(u, v)|}{\sqrt{h'[u]h'[v]}} \tag{2.3.22}$$

and h' depends only on h and $\mathsf{ran}(P)$. We will elaborate on this in the following remark. Before we proceed let us consider an example.

Example 2.3.6. Let $-\partial_{xx}$ be considered as the self adjoint operator with

$$\mathcal{D}(-\partial_{xx}) = \{u \in H^2[0,1] : \ u(0) = u(1) = 0\}.$$

The partial integration establishes that $-\partial_{xx}$ is defined by the positive definite form

$$h(u,v) = \int_0^1 \partial_x u \ \partial_x v \ dx, \qquad u, v \in \mathcal{Q}(-\partial_{xx}) = H_0^1[0,1]. \tag{2.3.23}$$

The operator $\partial_x u$, $u \in H_0^1[0,1]$ is closed, but not self adjoint, therefore (2.3.23) is an alternative representation (factorization), to the "square root" representation (2.3.21) of the operator $-\partial_{xx}$.

Remark 2.3.7. All of the representations of the form h are in a sense equivalent. Let $\mathbf{R} : \mathcal{H} \to \mathcal{H}'$ be a closed operator such that

$$h(x,y) = (\mathbf{R}x, \mathbf{R}y) = \left(\mathbf{H}^{1/2}x, \mathbf{H}^{1/2}y\right) \tag{2.3.24}$$

and $\mathcal{Q} = \mathcal{D}(\mathbf{R}) = \mathcal{D}(\mathbf{H}^{1/2})$, then by [41, Ch. VI.7]

$$\mathbf{R} = U\mathbf{H}^{1/2}, \quad \mathbf{R}^* = \mathbf{H}^{1/2}U^*, \tag{2.3.25}$$

where U is the isometry from $\mathcal{H}'$ onto $\mathsf{ran}(\mathbf{R})$. Independence of the estimate (2.3.7) from the representation (2.3.24) could have also been proved by the unitary invariance of the canonical angle and (2.3.25).

Formula (2.3.22) is an important corollary of Theorem 2.3.5. In the next theorem we prove

$$\frac{\sin\Theta}{1-\sin\Theta} = \max_{u,v\in\mathcal{Q}} \frac{|\delta h(u,v)|}{\sqrt{h[u]h[v]}}. \tag{2.3.26}$$

Equations (2.3.22) and (2.3.26) demonstrate that the constants $\mathsf{sin}\Theta$ and $\frac{\mathsf{sin}\Theta}{1-\mathsf{sin}\Theta}$ in (2.3.7) and (2.3.8) cannot be improved upon.

The following lemma is a generalization of a corresponding result from [26, Drmač] for finite matrices. The proof will be based on Lemma 2.2.6 and is taken out of the joint paper [37].

Lemma 2.3.8. *Let the form h be positive definite and let the forms h' and δh be as in (2.3.4), then*

$$\max_{u,v\in\mathcal{Q}} \frac{|\delta h(u,v)|}{\sqrt{h[u]h[v]}} = \frac{\sin\Theta}{1-\sin\Theta} \tag{2.3.27}$$

holds. Here $\mathsf{sin}\Theta = \mathsf{sin}\Theta(\mathbf{H}^{1/2}X, \mathbf{H}^{-1/2}X)$, where $\mathsf{ran}(X) \subset \mathcal{Q}$ was the subspace used to define h' and δh.

PROOF. We will prove the theorem by a direct evaluation of the maximum in (2.3.27). For, $x, y \in \mathcal{H}$ we have $\mathbf{H}^{-1/2}x, \mathbf{H}^{-1/2}y \in \mathcal{Q}$ and

$$\begin{aligned}\delta h(\mathbf{H}^{-1/2}x, \mathbf{H}^{-1/2}y) &= \left(\mathbf{H}^{1/2}(\mathbf{I}-XX^*)\mathbf{H}^{-1/2}x, \mathbf{H}^{1/2}XX^*\mathbf{H}^{-1/2}y\right) \\ &\quad + \left(\mathbf{H}^{1/2}XX^*\mathbf{H}^{-1/2}x, \mathbf{H}^{1/2}(\mathbf{I}-XX^*)\mathbf{H}^{-1/2}y\right) \\ &= ((\mathbf{I}-YZ^*)x, YZ^*y) + (YZ^*x, (\mathbf{I}-YZ^*)y)\,.\end{aligned}$$

By Lemma 2.2.6 we obtain

$$\delta h(\mathbf{H}^{-1/2}x, \mathbf{H}^{-1/2}y) = \left(\left[\begin{array}{c|c|c} 0 & & \\ \hline & \bigoplus_{i=1}^{l} \begin{bmatrix} 0 & -\tan\theta_i \\ -\tan\theta_i & -2\tan^2\theta_i \end{bmatrix} & \begin{matrix}0\\0\end{matrix} \\ \hline & \begin{matrix}0 & 0\end{matrix} & 0 \end{array}\right] x, y\right), \tag{2.3.28}$$

so

$$\max_{x,y\in\mathcal{H}} \frac{\left|\delta h(\mathbf{H}^{-1/2}x, \mathbf{H}^{-1/2}y)\right|}{\|x\|\|y\|} = \max_i \| \begin{bmatrix} 0 & -\tan\theta_i \\ -\tan\theta_i & -2\tan^2\theta_i \end{bmatrix} \|_2 = \frac{\sin\Theta}{1-\sin\Theta}.$$

This can equivalently be written as

$$\max_{u,v\in\mathcal{Q}} \frac{|\delta h(u,v)|}{\sqrt{h[u]h[v]}} = \frac{\sin\Theta}{1-\sin\Theta}$$

which implies the conclusion of the lemma. □

2.3.1 The nonnegative definite case

In the nonnegative case we have to provide an alternative definition for a subspace that will play the role of $\mathsf{ran}(\mathbf{H}^{-1/2}X)$. We have shown $W = \mathbf{H}^{1/2}P_\perp\mathbf{H}'^{-1/2}$ to be a partial isometry such that

$$\mathcal{W} = \mathsf{ran}(\mathbf{H}^{1/2}P_\perp)^\perp = \mathsf{ran}(W)^\perp = \mathsf{ran}(\mathbf{H}^{-1/2}X).$$

The left part of the equality is also well defined in the case in which $\mathbf{H}^{1/2}$ is not invertible, so we set

$$\mathcal{W} = \mathsf{ran}(\mathbf{H}^{1/2}P_\perp)^\perp.$$

The construction (2.3.4) was performed with the assumption that h is nonnegative definite and $\mathsf{ran}(X) \subset \mathcal{Q}$. Lemma 2.3.2 says $\sigma_{ess}(\mathbf{H}) = \sigma_{ess}(\mathbf{H}')$ so $\mathbf{H}'^{\dagger 1/2}$ is a bounded operator and

$$V = \mathbf{H}^{1/2}P\mathbf{H}'^{\dagger 1/2}, \tag{2.3.29}$$

$$W = \mathbf{H}^{1/2}P_\perp\mathbf{H}'^{\dagger 1/2} \tag{2.3.30}$$

are everywhere defined. Corollary 2.3.3 enables us to conclude that $\mathsf{ran}(V) = \mathsf{ran}(\mathbf{H}^{1/2}P)$ and $\mathsf{ran}(W) = \mathsf{ran}(\mathbf{H}^{1/2}P_\perp)$, so we set

$$\mathcal{V} = \mathsf{ran}(V), \qquad \mathcal{W} = \mathsf{ran}(W)^\perp. \tag{2.3.31}$$

Lemma 2.3.1 states that given a positive definite $\mathbf{H}$ the constructed operator $\mathbf{H}'$ must always be positive definite. In general nonnegative situation we have only the result of Corollary 2.3.3. It established that $\mathbf{H}'$ is a nonnegative definite operator and that $\mathsf{ker}(\mathbf{H}') \subset \mathsf{ker}(\mathbf{H})$. This does not give sufficient information on the structure of $\mathbf{H}'$. Formulae like (2.3.7)–(2.3.8) are meaningful in the nonnegative definite case, too. They, however, invariably imply $\mathsf{ker}(\mathbf{H}) = \mathsf{ker}(\mathbf{H}')$. We, therefore, proceed is two steps. First, we establish a general (theoretical) condition on the subspace $\mathcal{X} = \mathsf{ran}(P)$ which guarantees that $\mathsf{ker}(\mathbf{H}) = \mathsf{ker}(\mathbf{H}')$. As the second step we give a practical computational formula.

The subspaces $\mathcal{W}$ and $\mathcal{V}$ need not have the same dimension, so we will have to use the principal angle to compare them, cf. Theorem 2.2.1. In what follows we show that

$$\sin\Theta_p(\mathcal{V}, \mathcal{W})$$

takes the role of $\mathsf{sin}\Theta(\mathbf{H}^{1/2}X, \mathbf{H}^{-1/2}X)$ in the nonnegative version of Theorem 2.3.5. In the case when $\mathbf{H}^{1/2}$ is invertible (2.3.18) implies $\mathcal{V} = \mathsf{ran}(\mathbf{H}^{1/2}X)$ and $\mathcal{W} = \mathsf{ran}(\mathbf{H}^{-1/2}X)$. The subspaces $\mathbf{H}^{-1/2}X$ and $\mathbf{H}^{1/2}X$ have the same dimension, so Corollary 2.2.2 yields

$$\sin\Theta_p(\mathcal{V}, \mathcal{W}) = \sin\Theta(\mathbf{H}^{1/2}X, \mathbf{H}^{-1/2}X).$$

We establish the properties of V and W and give a characterization of the subspace $\mathcal{W}$ in the following lemma.

Lemma 2.3.9. *Let $\mathcal{X} = \mathsf{ran}(P)$, $V = \mathbf{H}^{1/2}P\mathbf{H}'^{\dagger 1/2}$ and $W = \mathbf{H}^{1/2}P_\perp\mathbf{H}'^{\dagger 1/2}$. Then*

$$V^*V = P_{\mathsf{ran}(\mathbf{H}'P)} \tag{2.3.32}$$

$$W^*W = P_{\mathsf{ran}(\mathbf{H}'P_\perp)} \tag{2.3.33}$$

$$WV^* = 0 \tag{2.3.34}$$

$$VW^* = 0 \tag{2.3.35}$$

and

$$\mathsf{inv}(\mathbf{H}^{1/2})\mathcal{X} = \mathcal{W}, \tag{2.3.36}$$

where $\mathcal{W}$ is from (2.3.31) and

$$\mathsf{inv}(\mathbf{H}^{1/2})\mathcal{X} = \{x : \mathbf{H}^{1/2}x \in \mathcal{X}\}$$

denotes the inverse image of the subspace $\mathcal{X}$ under the mapping $\mathbf{H}^{1/2}$.

PROOF. The relations (2.3.32)–(2.3.35) follow analogously as in the proof of Theorem 2.3.5. It only remains to prove (2.3.36).

We first show that $\mathsf{inv}(\mathbf{H}^{1/2})\mathcal{X} \subset \mathcal{W} = \mathsf{ran}(W)^{\perp}$. Take any $u \in \mathsf{inv}(\mathbf{H}^{1/2})\mathcal{X}$, then

$$\mathbf{H}^{1/2}u = z \in \mathcal{X}.$$

This implies

$$0 = (z, P_{\perp}\mathbf{H}'^{\dagger 1/2}v) = (u, \mathbf{H}^{1/2}P_{\perp}\mathbf{H}'^{\dagger 1/2}v), \qquad v \in \mathcal{H}$$

which proves $u \in \mathsf{ran}(W)^{\perp} = \mathcal{W}$.

The other inclusion follows in two steps. Take $u \in \mathcal{W}$, then

$$(u, \mathbf{H}^{1/2}P_{\perp}\mathbf{H}'^{\dagger 1/2}v) = 0, \qquad v \in \mathcal{H}.$$

On the other hand, the subspace

$$\mathsf{ran}(P_{\perp}\mathbf{H}'^{\dagger 1/2})^{\perp} = \mathsf{ran}(P_{\perp}P_{\mathsf{ran}(\mathbf{H}')})^{\perp} \subset \mathcal{D}(\mathbf{H}^{1/2})$$

is finite dimensional, so we conclude $u \in \mathcal{D}(\mathbf{H}^{1/2})$. Corollary 2.3.3 implies

$$0 = (\mathbf{H}^{1/2}u, P_{\perp}P_{\mathsf{ran}(\mathbf{H}')}v) = (\mathbf{H}^{1/2}u, P_{\mathsf{ran}(\mathbf{H}')}P_{\perp}v) = (\mathbf{H}^{1/2}u, P_{\perp}v), \qquad v \in \mathcal{H},$$

which proves $\mathbf{H}^{1/2}u \in \mathcal{X}$. With this conclusion we have established (2.3.36). □

As a direct consequence of Corollary 2.2.2 and (2.3.36) we obtain the following result.

Corollary 2.3.10. *Let* $\mathcal{X} = \mathsf{ran}(P)$, $V = \mathbf{H}^{1/2}P\mathbf{H}'^{\dagger 1/2}$ *and* $W = \mathbf{H}^{1/2}P_{\perp}\mathbf{H}'^{\dagger 1/2}$. *Then*

$$\|P_{\mathcal{V}}P_{\mathcal{W}^{\perp}}\| \leq \|P_{\mathcal{V}^{\perp}}P_{\mathcal{W}}\|,$$

so

$$\sin\Theta_p(\mathbf{H}^{1/2}\mathcal{X}, \mathsf{inv}(\mathbf{H}^{1/2})\mathcal{X}) = \|V^*W\|. \tag{2.3.37}$$

It would be pleasing to use $\mathbf{H}^{1/2\dagger}$ in the place of $\mathsf{inv}(\mathbf{H}^{1/2})$. This is only possible under additional restrictions on the subspace $\mathsf{ran}(P)$. To get better feeling for the meaning of $\mathsf{sin}\Theta_p(\mathbf{H}^{1/2}\mathcal{X}, \mathsf{inv}(\mathbf{H}^{1/2})\mathcal{X})$ consider the following example.

Example 2.3.11. Take

$$H = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \qquad \mathcal{X} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

then

$$H' = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

is, unlike H, a positive definite matrix. Now,

$$H^{1/2} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}, \qquad H^{1/2\dagger} = \begin{bmatrix} \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \end{bmatrix}$$

and we compute

$$\mathsf{ran}(V) = \mathrm{span}\{[1 \quad 1]^*\}, \qquad \mathsf{ran}(W)^\perp = \mathrm{span}\{[-1 \quad 1]^*\}$$

which proves that in this case $\mathsf{sin}\Theta_p(\mathsf{ran}(V), \mathsf{ran}(W)^\perp) = 1$ and

$$\mathsf{ran}(W)^\perp = \mathsf{ker}(H) \neq \mathsf{ran}(\mathbf{H}^{1/2\dagger}P).$$

Instead of advocating the use of the general formula (2.3.36) we will establish a form of compatibility condition under which we may use the generalized inverse of $\mathbf{H}^{1/2}$ to check the statement of the theorems.

The next result is a nonnegative analogue of Theorem 2.3.5. It will enable us to, in effect, "deflate away" the kernel of the nonnegative form h and reduce the problem to the positive definite case. The only additional restriction we have to impose on the nonnegative form h is that it satisfies the assumptions of Lemma 2.3.2.

Theorem 2.3.12. *Let the subspace $\mathcal{X} = \mathsf{ran}(P) \subset \mathcal{Q}$ be given and let h be a nonnegative form. Assume $\mathsf{sin}\Theta_p(\mathbf{H}^{1/2}\mathcal{X}, \mathsf{inv}(\mathbf{H}^{1/2})\mathcal{X}) = \mathsf{sin}\Theta_p < 1$, then*

$$(1 - \sin\Theta_p)h'[u] \leq h[u] \leq (1 + \sin\Theta_p)h'[u], \qquad u \in \mathcal{Q}(h), \tag{2.3.38}$$

$$(1 - \frac{\sin\Theta_p}{1 - \sin\Theta_p})h[u] \leq h'[u] \leq (1 + \frac{\sin\Theta_p}{1 - \sin\Theta_p})h[u], \qquad u \in \mathcal{Q}(h). \tag{2.3.39}$$

Proof. The proof is similar to the proof of Theorem 2.3.5. Let h' and δh be as in (2.3.4). Set δH_s to be the operator defined by the form

$$\delta h_s(x, y) = \delta h(\mathbf{H}'^{\dagger 1/2}x, \mathbf{H}'^{\dagger 1/2}y), \qquad x, y \in \mathcal{H}.$$

The form δh_s is closed and everywhere defined, so δH_s is a bounded operator. We obviously have $\mathsf{ker}(\mathbf{H}'^{\dagger 1/2}) = \mathsf{ker}(\mathbf{H}') \subset \mathsf{ker}(\delta H_s)$, so $P_{\mathsf{ran}(\mathbf{H}')}$ commutes with the operator δH_s. With the use of Corollary 2.3.3 one computes, analogously as in Theorem 2.3.5,

$$\begin{aligned} \delta h(\mathbf{H}'^{\dagger 1/2}x, \mathbf{H}'^{\dagger 1/2}y) &= h(P_\perp \mathbf{H}'^{\dagger 1/2}x, P\mathbf{H}'^{\dagger 1/2}y) + h(P\mathbf{H}'^{\dagger 1/2}x, P_\perp \mathbf{H}'^{\dagger 1/2}y) \\ &= (Wx, Vy) + (Vx, Wy), \end{aligned}$$

so

$$\delta H_s = V^*W + W^*V.$$

Since $\mathbf{H}'^{1/2}\mathbf{H}'^{\dagger 1/2} = P_{\mathsf{ran}(\mathbf{H}')}$ we obtain

$$\max_{u,v\in\mathsf{ran}(\mathbf{H}')\cap\mathcal{Q}} \frac{|\delta h(u,v)|}{\sqrt{h'[u]h'[v]}} = \|\delta H_s\| = \|V^*W\|. \tag{2.3.40}$$

Corollary 2.3.10 implies that the assumption $\mathsf{sin}\Theta_p < 1$, in fact, reads

$$\sin\Theta_p = \|V^*W\| < 1.$$

With this in hand, we have established

$$(1-\sin\Theta_p)h'[u] \le h[u] \le (1+\sin\Theta_p)h'[u], \qquad u\in\mathcal{Q}(h),$$

which implies $\mathsf{ker}(\mathbf{H}') = \mathsf{ker}(\mathbf{H})$. The relation (2.3.39) follows by the same argument as the one used in Theorem 2.3.5. □

The main insight into the structure of the operator $\mathbf{H}'$, gained from Theorem 2.3.12, is summed up in the following corollary.

Corollary 2.3.13. *Take a nonnegative form h and a subspace $\mathcal{X} = \mathsf{ran}(P) \subset \mathcal{Q}$. If $\mathsf{sin}\Theta_p(\mathbf{H}^{1/2}\mathcal{X}, \mathsf{inv}(\mathbf{H}^{1/2})\mathcal{X}) < 1$ then $\mathsf{ran}(\mathbf{H}') = \mathsf{ran}(\mathbf{H})$.*

Corollary 2.3.13 gives precise meaning to the statement "deflate away". Set $\mathcal{R} = \mathsf{ran}(\mathbf{H}) = \mathsf{ran}(\mathbf{H}')$ and $\mathcal{N} = \mathsf{ker}(\mathbf{H}) = \mathsf{ker}(\mathbf{H}')$. The projections $P_{\mathcal{N}}$ and P commute, so

$$P_{\mathcal{N}\cap\mathsf{ran}(P)} = P_{\mathcal{N}}P, \qquad \widetilde{P} = P - P_{\mathcal{N}\cap\mathsf{ran}(P)}$$

are both orthogonal projections. A direct calculation shows

$$\widetilde{\mathcal{X}} := \mathsf{ran}(\widetilde{P}) = \mathsf{ran}(P) \ominus (\mathcal{N}\cap\mathsf{ran}(P)) = \mathsf{ran}(\mathbf{H}')\cap\mathsf{ran}(P) = \mathsf{ran}(\mathbf{H}'P).$$

The form

$$\widetilde{h}(u,v) = h(P_{\mathcal{R}}u, P_{\mathcal{R}}v)$$

is positive definite in $\mathcal{R}$ and $\mathsf{ran}(\widetilde{P}) \subset \mathcal{Q}(\widetilde{h})\cap\mathcal{R}$. Now, apply the construction (2.3.2)—(2.3.4) to the form $\widetilde{h}$ and the projection $\widetilde{P}$. By $\widetilde{\mathbf{H}} : \mathcal{R}\to\mathcal{R}$ denote the operator defined by the form $\widetilde{h}$ in $\mathcal{R}$, then $\mathsf{ran}(\widetilde{P}) \subset \mathcal{R}$ and

$$\widetilde{h}'(u,v) = h'(P_{\mathcal{R}}u, P_{\mathcal{R}}v).$$

We conclude that

$$\sin\Theta(\widetilde{\mathbf{H}}^{1/2}\widetilde{\mathcal{X}}, \widetilde{\mathbf{H}}^{-1/2}\widetilde{\mathcal{X}}) = \sin\Theta_p(\mathbf{H}^{1/2}\mathcal{X}, \mathsf{inv}(\mathbf{H}^{1/2})\mathcal{X}) < 1$$

and $\widetilde{h}$ and $\widetilde{P}$ satisfy the assumptions of Theorem 2.3.5. If we were to "a priori" assume $\mathsf{ran}(\mathbf{H}') = \mathsf{ran}(\mathbf{H})$, then this argument would give an alternative proof of Theorem 2.3.12. "Deflate away" means that we assume we were given $\widetilde{h}$ and $\widetilde{P}$ as input.

Remark 2.3.14. Another consequence of Corollary 2.3.13 is that we can invoke Lemma 2.3.8 to conclude that the constant $\frac{\sin\Theta_p}{1-\sin\Theta_p}$ (in (2.3.39)) cannot be sharpened. Furthermore, Example 2.3.11 shows that the assumption

$$\sin\Theta_p(\mathbf{H}^{1/2}\mathcal{X}, \mathsf{inv}(\mathbf{H}^{1/2})\mathcal{X}) < 1$$

is a necessary requirement to establish the inequalities (2.3.38) and (2.3.39) as well as to guarantee that $\mathsf{ran}(\mathbf{H}) = \mathsf{ran}(\mathbf{H}')$ (equivalently $\mathsf{ker}(\mathbf{H}) = \mathsf{ker}(\mathbf{H}')$).

Important special case

The assumption that P and $P_{\mathsf{ker}(\mathbf{H})}$ commute and Corollary 2.3.3 yield $\mathsf{ker}(\mathbf{H}) = \mathsf{ker}(\mathbf{H}')$ and $\mathsf{ran}(\mathbf{H}) = \mathsf{ran}(\mathbf{H}')$. This implies

$$\mathsf{inv}(\mathbf{H}^{1/2})\mathcal{X} = \mathbf{H}^{1/2\dagger}\mathcal{X}. \tag{2.3.41}$$

The projections P and $P_{\mathsf{ker}(\mathbf{H})}$ certainly commute when $\mathsf{ker}(\mathbf{H}) \perp \mathsf{ran}(P)$ or when[3] $\mathsf{ker}(\mathbf{H}) \subset \mathsf{ran}(P)$. This discussion is summed up in the following corollary.

Corollary 2.3.15. *Assume $P = XX^*$ and $P_{\mathsf{ker}(\mathbf{H})}$ commute and let $\mathsf{sin}\Theta_p(\mathbf{H}^{1/2}X, \mathbf{H}^{1/2\dagger}X) < 1$, then*

$$(1 - \sin\Theta_p)h'[u] \leq h[u] \leq (1 + \sin\Theta_p)h'[u], \qquad u \in \mathcal{Q}(h) \tag{2.3.42}$$

$$(1 - \frac{\sin\Theta_p}{1-\sin\Theta_p})h[u] \leq h'[u] \leq (1 + \frac{\sin\Theta_p}{1-\sin\Theta_p})h[u], \qquad u \in \mathcal{Q}(h). \tag{2.3.43}$$

Remark 2.3.16. To assess the restriction that P and $P_{\mathsf{ker}(\mathbf{H})}$ should commute, consider the definition of the relatively accurate approximation of the number $\lambda \in \mathbb{R}_+$. $\mu \in \mathbb{R}_+$ is relatively accurate approximation of $\lambda \in \mathbb{R}_+$, if

1. $\lambda = \mu$, when $\lambda = 0$

[3]The other situation when P and $P_{\mathsf{ker}(\mathbf{H})}$ commute is when $\mathsf{ran}(P) \subset \mathsf{ker}(\mathbf{H})$, this situation is however trivial and we have tacitly left it out.

2. $\frac{|\lambda-\mu|}{\mu} < 1$, when $\lambda \neq 0$.

This implies that we can expect to compute "relatively accurate" Ritz value approximation of the spectrum of the nonnegative definite operator $\mathbf{H}$ only in the case when we have computed a basis for $\mathsf{ker}(\mathbf{H})$, cf. [1].

Remark 2.3.14 implies that we may assume that the condition of Corollary 2.3.15 were $\mathsf{ker}(\mathbf{H}) \perp \mathsf{ran}(P)$. To compute the basis of the set $\mathsf{inv}(\mathbf{H}^{1/2})\mathcal{X}$ we need to repeatedly solve the equation

$$\mathbf{H}^{1/2}u = x_i, \qquad i = 1, ..., \mathsf{dim}(\mathcal{X}).$$

The vectors x_i are assumed to be a basis for $\mathcal{X}$. The restriction that $\mathsf{ker}(\mathbf{H}) \perp \mathsf{ran}(P)$ amounts to nothing more than to impose a compatibility condition on x_i (e.g. think of the Laplacian with Neumann boundary conditions).

2.3.2 A first approximation estimate

Theorem 2.1.6 and Lemma 2.3.2 yield the first eigenvalue estimates. The next theorem will give an eigenvalue estimate with the minimum of the restrictions on the subspace $\mathsf{ran}(X) \subset \mathcal{Q}$. Sharper bounds are possible when we impose additional assumptions on $\mathsf{ran}(X)$. Even this (first order) estimate will compare favorably with other higher order bounds that can be found in the literature, cf. Section 2.7.

Theorem 2.3.17. *Let $0 \leq h$ and let the n-dimensional subspace $\mathsf{ran}(P) \subset \mathcal{Q}$, $P = XX^*$, be given. Define*

$$\Xi = (\mathbf{H}^{1/2}X)^*\mathbf{H}^{1/2}X, \qquad \Xi \in \mathbb{C}^{n\times n}$$

and assume $\mu_n < \lambda_e(\mathbf{H})$. Here, the Ritz values are numbered as in (2.3.1). If $\mathsf{ran}(P)$ is such that $\mathsf{sin}\Theta_p < 1$ then there are n eigenvalues of the operator $\mathbf{H}$, counting the eigenvalues according to their multiplicities, such that

$$|\lambda_{i_j} - \mu_j| \leq \mu_j \sin\Theta_p, \qquad j = 1, \ldots, n, \tag{2.3.44}$$

$$|\lambda_{i_j} - \mu_j| \leq \lambda_{i_j} \frac{\sin\Theta_p}{1 - \sin\Theta_p}, \qquad j = 1, \ldots, n, \tag{2.3.45}$$

where $i_{(\cdot)} : \mathbb{N} \to \mathbb{N}$ is a permutation.

Proof. Corollary 2.3.13 readily implies the conclusion (2.3.44) for the Ritz values $\mu_j = 0$, $j = 1, \ldots, \mathsf{dim}(\mathsf{ker}(\Xi))$. Therefore, we may safely assume that h is a positive definite form.

Lemma 2.3.2 implies $\sigma_{ess}(\mathbf{H}) = \sigma_{ess}(\mathbf{H}')$, so the assumption $\mu_n < \lambda_e(\mathbf{H})$ guarantees that μ_n is a discrete eigenvalue of $\mathbf{H}'$. Theorem 2.3.12 established

$$(1 - \sin\Theta_p)h'[u] \leq h[u] \leq (1 + \sin\Theta_p)h'[u], \qquad u \in \mathcal{Q}(h)$$
$$(1 - \frac{\sin\Theta_p}{1 - \sin\Theta_p})h[u] \leq h'[u] \leq (1 + \frac{\sin\Theta_p}{1 - \sin\Theta_p})h[u], \qquad u \in \mathcal{Q}(h).$$

The conclusion follows directly from Theorem 2.1.6. □

2.4 Localizing the approximated eigenvalues

There is a multitude of ways to match the computed Ritz values to a part of the spectrum of the operator $\mathbf{H}$ of the same multiplicity. These approaches usually differ with regard to the allowed amount of additional information about the spectrum of the operator $\mathbf{H}$. Here, we present two possible answers to that problem.

Theorem 2.3.17 can be interpreted as a first localization result. It gives an estimate of the infimum of

$$\max_{j=1,\ldots,n} \frac{|\lambda_{i_j} - \mu_j|}{\mu_j}$$

over all of the permutations $i_{(\cdot)} : \mathbb{N} \to \mathbb{N}$. So, we would be correct in stating that the Ritz values are approximating the eigenvalues of $\mathbf{H}$ that are closest to $\sigma(\Xi)$.

Having only limited additional infirmation we got a limited answer. We know that there is a collection of eigenvalues of operator $\mathbf{H}$, having the joint multiplicity n, that is being approximated by the Ritz values from the subspace $\mathsf{ran}(X)$. The information we have on the location of those eigenvalues in the spectrum of $\mathbf{H}$ is only that they are the eigenvalues closest to computed Ritz values.

Only when we have additional information about the location of the part of the spectrum *we do not want to approximate*, we can guarantee that we are approximating the part of the spectrum we are interested in. A best known example of such estimates is a well known Temple–Kato inequality. Let $\lambda_1 < \lambda_2$ and let $u \in \mathcal{D}(\mathbf{H})$ be a unit vector such that $(u, \mathbf{H}u) < \gamma \leq \lambda_2$, then

$$(u, \mathbf{H}u) \geq \lambda_1 \geq (u, \mathbf{H}u) - \frac{(\mathbf{H}u, \mathbf{H}u) - (u, \mathbf{H}u)^2}{\gamma - (u, \mathbf{H}u)} \ . \tag{2.4.1}$$

For the proof see [50]. The estimate (2.4.1) is valid for a general self adjoint operator $\mathbf{H}$. As a result, under the appropriate assumptions on the location of the "unwanted" part of the spectrum, we remove the regularity constraint that test vector u be in $\mathcal{D}(H)$ and

obtain sharp bounds for the matching cluster of eigenvalues. In the last section of this chapter we will demonstrate that on some examples our bound considerably outperforms the estimate (2.4.1).

We now give a theorem that determines those eigenvalues of the operator $\mathbf{H}$, given by a symmetric form h, which are approximated by the Ritz values associated with the test subspace $\mathsf{ran}(X) \subset \mathcal{Q}$. Before we proceed with the formulation of the theorem we state a well known fact that given $0 < \lambda, \mu$ and $\sin\Theta_p < 1$ the relation

$$\frac{|\lambda - \mu|}{\mu} \leq \sin\Theta_p < 1$$

implies the relation

$$\frac{|\lambda - \mu|}{\lambda} \leq \frac{\sin\Theta_p}{1 - \sin\Theta_p} =: \eta_{\Theta_p}. \tag{2.4.2}$$

Note that

$$\sin\Theta_p \leq \eta_{\Theta_p} = \frac{\sin\Theta_p}{1 - \sin\Theta_p} \leq 2\sin\Theta_p.$$

Theorem 2.4.1. *Take a nonnegative form h and the subspace $\mathsf{ran}(X) \subset \mathcal{Q}$. Assume $r = \dim(\mathsf{ker}(\mathbf{H})) \leq n$, set $P = XX^*$ and let h' be as in (2.3.3). By*

$$\mu_1 \leq \cdots \leq \mu_n$$

denote the eigenvalues of the matrix

$$\Xi = (\mathbf{H}^{1/2}X)^*\mathbf{H}^{1/2}X, \qquad \Xi \in \mathbb{C}^{n\times n}.$$

If

$$\eta_{\Theta_p} := \frac{\sin\Theta_p}{1 - \sin\Theta_p} < \min\{\gamma_r, 1\}, \tag{2.4.3}$$

where $\gamma_r = \min\limits_{\substack{k=1,\ldots,n \\ p=n+1,\ldots,\infty}} \dfrac{\lambda_p - \mu_k}{\lambda_p + \mu_k}$ is supposed to be positive, then

$$|\lambda_i - \mu_i| \leq \mu_i \sin\Theta_p, \quad i = 1, ..., n. \tag{2.4.4}$$

Proof. The assumption (2.4.3) and Theorem 2.3.12 imply $\mathsf{ker}(\mathbf{H}) \subset \mathsf{ran}(X)$. Also, by Theorem 2.3.12 we have $\mathsf{ker}(\mathbf{H}) = \mathsf{ker}(\mathbf{H}')$, so we are allowed to "deflate away" the kernel of $\mathbf{H}$. Therefore, set $P_1 = P_{\mathsf{ran}(\mathbf{H}'P)}$ and proceed as if h were positive definite and $P = P_1$. Let

$$0 < \widetilde{\mu}_1 \leq \widetilde{\mu}_2 \leq \cdots \leq \widetilde{\mu}_n \leq \cdots$$

be the eigenvalues of $\mathbf{H}'$. The estimates

$$1 - \sin\Theta_p \leq \frac{\lambda_i}{\widetilde{\mu}_i} \leq 1 + \sin\Theta_p, \quad i \in \mathbb{N},$$
$$1 - \eta_{\Theta_p} \leq \frac{\widetilde{\mu}_i}{\lambda_i} \leq 1 + \eta_{\Theta_p}, \quad i \in \mathbb{N}.$$

are a consequence of Theorem 2.1.6 . Alternatively we write this assertion as

$$\frac{|\lambda_i - \widetilde{\mu}_i|}{\widetilde{\mu}_i} \leq \sin\Theta_p, \quad i \in \mathbb{N}, \tag{2.4.5}$$
$$\frac{|\lambda_i - \widetilde{\mu}_i|}{\lambda_i} \leq \frac{\sin\Theta_p}{1 - \sin\Theta_p}, \quad i \in \mathbb{N}. \tag{2.4.6}$$

The assumption (2.4.3) implies $\mu_n < \lambda_e(\mathbf{H}')$. Lemma 2.3.2 and Theorem 2.3.17 guarantee that there exists a permutation $i_{(\cdot)} : \mathbb{N} \to \mathbb{N}$ such that

$$\mu_{r+k-1} = \widetilde{\mu}_{i_k}, \quad k = 1, ..., m.$$

Note that $k < d$ implies $i_k < i_d$ for $k, d \leq m$. Now (2.4.5) and (2.4.6) imply

$$\frac{|\lambda_{i_k} - \widetilde{\mu}_i|}{\widetilde{\mu}_i} \leq \sin\Theta_p, \qquad k = 1, ..., m, \tag{2.4.7}$$
$$\frac{|\lambda_{i_k} - \widetilde{\mu}_i|}{\lambda_{i_k}} \leq \frac{\sin\Theta_p}{1 - \sin\Theta_p}, \qquad k = 1, ..., m. \tag{2.4.8}$$

To prove the theorem we show a slightly stronger assertion, namely,

$$\mu_k < \widetilde{\mu}_p, \quad p = m+1, ..., \infty, \quad k = 1, ..., m \tag{2.4.9}$$
$$\lambda_{i_k} < \lambda_p, \quad p = m+1, ..., \infty, \quad k = 1, ..., m. \tag{2.4.10}$$

In other words, we show that $i_k \leq m, k = 1, ..., m$ which together with (2.4.5) implies

$$\frac{|\lambda_i - \mu_i|}{\mu_i} \leq \sin\Theta_p, \quad i = 1, ..., m.$$

Let us prove the first statement $\mu_k \neq \widetilde{\mu}_p,\ p = m+1, ..., \infty$. Choosing $k \in \{1, ..., m\}$, we have

$$\begin{aligned}
\frac{\widetilde{\mu}_p - \mu_k}{\mu_k} &\geq \frac{\lambda_p - \mu_k}{\mu_k} - \frac{|\widetilde{\mu}_p - \lambda_p|}{\mu_k} \\
&\geq \frac{\lambda_p - \mu_k}{\mu_k + \lambda_p}\frac{\mu_k + \lambda_p}{\mu_k} - \frac{|\widetilde{\mu}_p - \lambda_p|}{\lambda_p}\frac{\lambda_p}{\mu_k} \\
&> \gamma(1 + \frac{\lambda_p}{\mu_k}) - \gamma\frac{\lambda_p}{\mu_k} = \gamma > 0
\end{aligned}$$

which proves (2.4.9), whereas (2.4.10) follows from

$$\begin{aligned}\frac{\lambda_p - \lambda_{i_k}}{\mu_k} &\geq \frac{\lambda_p - \mu_k}{\mu_k} - \frac{|\mu_k - \lambda_{i_k}|}{\mu_k} \\ &\geq \frac{\lambda_p - \mu_k}{\mu_k + \lambda_p}\frac{\mu_k + \lambda_p}{\mu_k} - \gamma \\ &> \gamma(1 + \frac{\lambda_p}{\mu_k}) - \gamma = \gamma\frac{\lambda_p}{\mu_k} > 0.\end{aligned}$$

□

If we are provided with the information that

$$\frac{\sin\Theta_p}{1 - \sin\Theta_p} < \min\{\gamma_l, \gamma_r, 1\},$$

where

$$\gamma_r = \min_{\substack{k=1,\dots,n \\ p=q+n,\dots,\infty}} \frac{\lambda_p - \mu_k}{\lambda_p + \mu_k} \quad \text{and} \quad \gamma_l = \min_{\substack{k=1,\dots,n \\ p=1,\dots,q-1}} \frac{\mu_k - \lambda_p}{\lambda_p + \mu_k},$$

then

$$\mu_1 \leq \cdots \leq \mu_n$$

approximate the "inner" eigenvalues

$$\lambda_q \leq \lambda_{q+2} \cdots \leq \lambda_{q+n-1}.$$

This statement is made precise in the following theorem.

Theorem 2.4.2. *Take a nonnegative form h and a subspace $\mathsf{ran}(X) \subset \mathcal{Q}$. By*

$$\mu_1 \leq \cdots \leq \mu_n$$

denote the eigenvalues of the matrix $\Xi = (\mathbf{H}^{1/2}X)^*\mathbf{H}^{1/2}X$, $\Xi \in \mathbb{C}^{n\times n}$. If

$$\frac{\sin\Theta_p}{1 - \sin\Theta_p} < \min\{\gamma_l, \gamma_r, 1\}, \tag{2.4.11}$$

where $\gamma_r := \min\limits_{\substack{k=1,\dots,n \\ p=q+n,\dots,\infty}} \dfrac{\lambda_p - \mu_k}{\lambda_p + \mu_k}$ and $\gamma_l := \min\limits_{\substack{k=1,\dots,n \\ p=1,\dots,q-1}} \frac{\mu_k - \lambda_p}{\lambda_p + \mu_k}$,

then $\mathsf{ran}(P) \subset \mathsf{ran}(\mathbf{H}')$ and

$$\frac{|\lambda_{i+q-1} - \mu_i|}{\mu_i} \leq \sin\Theta_p, \quad i = 1, \dots, n\ .$$

PROOF. The assumption (2.4.11) and Theorem 2.3.12 and Corollary 2.3.13 imply $\mathsf{ran}(\mathbf{H}') = \mathsf{ran}(\mathbf{H})$ and

$$\mathsf{ran}(P) \subset \mathsf{ran}(\mathbf{H}).$$

The rest of the proof follows analogously as in the proof of Theorem 2.4.1. □

Remark 2.4.3. Theorems 2.4.1 and 2.4.2 imply that we can divide the spectrum of the operator $\mathbf{H}$ in two disjoint parts: the part that is being approximated by the $\sigma(\Xi)$ and the rest of the spectrum. To understand this statement assume that the conditions of Theorem 2.4.1 hold. In this case both of the "block diagonal" forms

$$h(u,v) = h(E(\lambda_n)u, E(\lambda_n)v) + h(E(\lambda_n)_\perp u, E(\lambda_n)_\perp v) \simeq \begin{bmatrix} \Lambda & \\ & \mathbf{\Lambda}_c \end{bmatrix},$$
$$h'(u,v) = h(Pu,Pv) + h(P_\perp u, P_\perp v) \simeq \begin{bmatrix} \Xi & \\ & \mathbf{\Xi}_c \end{bmatrix}$$

have "diagonal blocks" with disjoint spectra. We have assumed $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and $\Xi = \mathrm{diag}(\mu_1, \ldots, \mu_n)$ and $\mathbf{\Xi}_c$ and $\mathbf{\Lambda}_c$ were unbounded operators defined by the forms h' and h in the spaces $\mathsf{ran}(P_\perp)$ and $\mathsf{ran}(E(\lambda_n)_\perp)$. In fact, we will colloquially call h' the *block diagonal part of the operator* $\mathbf{H}$ *with respect to the subspace* $\mathsf{ran}(P)$. We will use the notation h_P to denote h' in situations when it is not clear with respect to which test space $\mathsf{ran}(P)$ was this construction performed.

2.5 Eigenvector and invariant subspace estimates

For the computed Ritz values

$$0, 0, \ldots, 0, \mu_{r+1}, \mu_{r+2}, \ldots, \mu_n$$

Theorem 2.3.17 guarantees the existence of the eigenvalues

$$\lambda_{i_1} \leq \lambda_{i_2} \leq \cdots \leq \lambda_{i_n},$$

that are being approximated by the Ritz values (provided $\mathsf{sin}\Theta_p < 1$) in the sense of

$$|\lambda_{i_j} - \mu_j| \leq \mu_j \sin\Theta_p, \qquad j = 1, \ldots, n.$$

Assume $v_1, \ldots, v_n$ are mutually orthogonal eigenvectors that belong to the eigenvalues $\lambda_{i_1} \leq \lambda_{i_2} \leq \cdots \leq \lambda_{i_n}$. If the conditions of Theorems 2.4.1 and 2.4.2 are satisfied Remark 2.4.3 assures us that

$$\mathrm{span}\{v_1, ..., v_n\} = \mathsf{ran}(E(\{\lambda_{i_1}, \lambda_{i_2}, \cdots, \lambda_{i_n}\})).$$

Here we have assumed that $\mathbf{H} = \int \lambda \, \mathrm{d}E(\lambda)$. To ease the presentation we generically use

$$\widehat{E} = E(\{\lambda_{i_1}, \lambda_{i_2}, \cdots, \lambda_{i_n}\})$$

to denote the projection on the subspace that is selected by a result like Theorem 2.4.1.

We will give two answers to the problem of eigenvector estimates. One is in the case in which we are approximating the contiguous group of eigenvalues (as guaranteed by Theorems 2.4.1 and 2.4.2) and the other is in the general case of the group of eigenvalues as delivered by Theorem 2.3.17. In the first case we present the estimates for $\|\widehat{E} - P\|$, whereas in the second case we give estimates for the individual eigenvectors $\|v_i - u_i\|$, $i = 1, ..., n$. Here $\mathbf{H}'u_i = \mu_i u_i$ and u_i are assumed to be of norm one and mutually orthogonal.

The central role in the analysis of the eigenvector approximations will be played by the following lemma.

Lemma 2.5.1. *Let h be a nonnegative form and let $0 \notin \sigma_{ess}(\mathbf{H})$. Take $\mathsf{ran}(P) \subset \mathcal{Q}$ such that $\sin\Theta_p < 1$ and define*

$$s(x, y) = \delta h(\mathbf{H}^{\dagger 1/2}x, \mathbf{H}'^{\dagger 1/2}y), \qquad x, y \in \mathcal{H}.$$

The form s defines a bounded operator S and

$$S = \mathbf{H}^{1/2}\mathbf{H}'^{\dagger 1/2} - \overline{\mathbf{H}^{\dagger 1/2}\mathbf{H}'^{1/2}} \tag{2.5.1}$$

$$|(x, Sy)| = |s(x, y)| \leq \frac{\sin\Theta_p}{\sqrt{1 - \sin\Theta_p}}\|x\|\|y\|, \quad x, y \in \mathcal{H}. \tag{2.5.2}$$

Proof. The closed graph theorem implies that the operator

$$S = \mathbf{H}^{1/2}\mathbf{H}'^{\dagger 1/2} - \overline{\mathbf{H}^{\dagger 1/2}\mathbf{H}'^{1/2}}$$

is bounded. Also, $\mathsf{ker}(\mathbf{H}) = \mathsf{ker}(\mathbf{H}') = \mathsf{ker}(S)$ and $P_{\mathsf{ker}(S)}$ commutes with S. It is sufficient to prove the estimate for $x, y \in \mathsf{ran}(\mathbf{H})$. The inequality (2.3.40) gives

$$|\delta h(\mathbf{H}^{\dagger 1/2}x, \mathbf{H}'^{\dagger 1/2}y)| \leq \sin\Theta_p \|y\| \; h'[\mathbf{H}^{\dagger 1/2}x]^{1/2}.$$

Analogously, (2.3.38) implies

$$\|\mathbf{H}'^{1/2}\mathbf{H}^{\dagger 1/2}\| \leq \frac{1}{\sqrt{1 - \sin\Theta_p}} \, . \tag{2.5.3}$$

Altogether, the estimate (2.5.2) follows. □

The operator S has the special structure. Assume $\mathbf{H}'u = \mu u$ and $\mathbf{H}v = \lambda v$, then

$$\begin{aligned}(v, Su) &= \lambda^{1/2}(v,u)\mu^{1/2} - \lambda^{-1/2}(v,u)\mu^{1/2} \\ &= \frac{\lambda-\mu}{\sqrt{\lambda\mu}}(v,u)\ . \end{aligned} \tag{2.5.4}$$

The equation (2.5.4) introduces the distance function

$$\frac{\lambda-\mu}{\sqrt{\lambda\mu}}$$

that measures the distance between the Ritz values and the spectrum of the operator $\mathbf{H}$. This distance function will feature in the important role in the estimates that follow. More involved analysis will be necessary to utilize the structure[4] of the operator S to obtain the invariant subspace approximation estimates. The next theorem extends the scope, as well as strengthens the eigenvector estimate from [37, 45] and is even new in the matrix case. It can be seen as the eigenvector companion result of Theorem 2.3.17.

Theorem 2.5.2. *Let h be a nonnegative form, and let $\mathsf{ran}(P) \subset \mathcal{Q}$ be such that it satisfies the assumptions of Theorem 2.3.17. Let $u_1, \ldots, u_n$ the mutually orthogonal eigenvectors belonging the eigenvalues $\mu_1, \ldots, \mu_n$ of $\mathbf{H}'P$, then there exist mutually orthogonal eigenvectors $v_1, \ldots, v_n$ of $\mathbf{H}$, belonging to the eigenvalues $\lambda_{i_1}, \ldots \lambda_{i_n}$ and*

$$\|v_j - u_j\| \leq \frac{\sqrt{2}\ \sin\Theta_p}{\sqrt{1-\sin\Theta_p}} \max_{k\neq j} \frac{\sqrt{\mu_j\lambda_{i_k}}}{|\lambda_{i_k}-\mu_j|}. \tag{2.5.5}$$

The eigenvalues λ_{i_j}, $j = 1, \ldots, r$ are numbered in the ascending order as given by Theorem 2.3.17.

Proof. Assume $\mu_1 = \cdots = \mu_r = 0$. Corollary 2.3.13 implies that $u_i \in \mathsf{ker}(\mathbf{H})$ for $i = 1, \ldots, r$ so we take

$$v_i = u_i, \quad i = 1, \cdots, r.$$

For v_j, $j = r+1, \ldots n$ take any orthonormal set of eigenvectors belonging to the eigenvalues λ_{i_j}, $j = r+1, \ldots, n$. Since both u_i and v_i, for $i = r+1, \ldots, n$ are perpendicular to $\mathsf{ker}(\mathbf{H})$ we may assume that $\mathbf{H}$ is positive definite and we are only given u_i, $i = r+1, \ldots n$ as test vectors. Take s from Lemma 2.5.1 and use (2.5.1) to compute

$$\begin{aligned} s(v_k, u_j) &= \delta h(\mathbf{H}^{-1/2}v_k, \mathbf{H}'^{-1/2}u_j) \\ &= \left(v_k, \mathbf{H}^{1/2}\mathbf{H}'^{-1/2}u_j\right) - \left(\mathbf{H}'^{1/2}\mathbf{H}^{-1/2}v_k, u_j\right) \\ &= (\lambda_{i_k}^{1/2}\mu_j^{-1/2} - \lambda_{i_k}^{-1/2}\mu_j^{1/2})\,(v_k, u_j) \end{aligned}$$

[4] More about the subspace estimates in the next subsection.

and

$$\begin{aligned}\sum_{k\neq j} |\,(v_k,u_j)\,|^2 &\leq \max_{k\neq j}\frac{\lambda_{i_k}\mu_j}{(\lambda_{i_k}-\mu_j)^2}\sum_{k\neq j}|s(v_k,u_j)|^2\\ &\leq \max_{k\neq j}\frac{\lambda_{i_k}\mu_j}{(\lambda_{i_k}-\mu_j)^2}\ \|S^*u_j\|^2\\ &\leq \max_{k\neq j}\frac{\lambda_{i_k}\mu_j}{(\lambda_{i_k}-\mu_j)^2}\frac{\sin^2\Theta_p}{1-\sin\Theta_p}.\end{aligned}$$

Scaling v_j, u_j so that $(v_j,u_j)\geq 0$, we obtain

$$\begin{aligned}\|v_j-u_j\| &= \sqrt{2}\sqrt{1-(v_j,u_j)} = \sqrt{2}\sqrt{1-\sqrt{1-\sum_{k\neq j}|\,(v_k,u_j)\,|^2}}\\ &\leq \sqrt{2}\sqrt{1-\sqrt{1-\max_{k\neq j}\frac{\lambda_{i_k}\mu_j}{(\lambda_{i_k}-\mu_j)^2}\frac{\sin^2\Theta_p}{1-\sin\Theta_p}}}\\ &\leq \frac{\sqrt{2}\ \sin\Theta_p}{\sqrt{1-\sin\Theta_p}}\max_{k\neq j}\frac{\sqrt{\mu_j\lambda_{i_k}}}{|\lambda_{i_k}-\mu_j|}.\end{aligned}$$

This proves the lemma in the case in which $\sigma_e(\mathbf{H})=\emptyset$. In the general case we use the formula

$$\frac{\sqrt{\lambda_e\mu_j}}{\lambda_e-\mu_j}\left|\left(E_{\mathbf{H}^{1/2}}([\sqrt{\lambda_e},\infty\rangle)u_j,Su_j\right)\right| \geq \left|\left(E_{\mathbf{H}^{1/2}}([\sqrt{\lambda_e},\infty\rangle)u_j,u_j\right)\right|$$

and analogous argument. □

2.5.1 The weak Sylvester equation

Let $V:\mathbb{C}^n\to\mathcal{H}$ be an isometry such that $\widehat{E}=VV^*$ and $\widehat{E}_\perp := I-VV^*$. For $\Lambda=V^*\mathbf{H}V$ we compute

$$\mathbf{H}V = V\Lambda.$$

On the other hand, Lemma 2.3.2 states

$$\mathbf{H}'X = X\Xi,$$

for $\Xi = X^*\mathbf{H}'X$. The expressions $\mathbf{H}V=V\Lambda$, $\mathbf{H}'X=X\Xi$ and $\mathbf{H}v=\lambda v$, $\mathbf{H}'u=\mu u$ are suggestively similar.

The subspaces $\mathsf{ran}(V)$ and $\mathsf{ran}(X)$ have the same dimension so

$$\sin\Theta(V,P) = \|\widehat{E}-P\| = \|\widehat{E}_\perp P\|.$$

To estimate $\|\widehat{E}_\perp P\|$ one typically starts from, cf. [11, 21],

$$G = \widehat{E}_\perp(\mathbf{H} - \mathbf{H}')P = \mathbf{H}\widehat{E}_\perp P - \widehat{E}_\perp P\mathbf{H}'. \tag{2.5.6}$$

In the case in which G and $\mathbf{H}'P$ are bounded operators and $\mathbf{H}\widehat{E}_\perp\Big|_{\mathsf{ran}(\mathbf{H}\widehat{E}_\perp)\cap\mathcal{D}(\mathbf{H})}$ and $\mathbf{H}'P|_P$ have disjoint spectra the standard theory of [11, 21] establishes

$$\delta\|\widehat{E}_\perp P\| \leq \|G\|.$$

The number δ is a measure of the separation of the spectra of $\mathbf{H}\widehat{E}_\perp\Big|_{\mathsf{ran}(\mathbf{H}\widehat{E}_\perp)\cap\mathcal{D}(\mathbf{H})}$ and $\mathbf{H}'P|_P$. However, G is a bounded operator if and only if $\mathsf{ran}(P) \subset \mathcal{D}(\mathbf{H})$, which is an assumption we have not made. If $\mathsf{ran}(P) \subset \mathcal{Q}(h)$, then only

$$Q = \mathbf{H}^{1/2}\widehat{E}_\perp P\mathbf{H}'^{-1/2} - \mathbf{H}^{-1/2}\widehat{E}_\perp P\mathbf{H}'^{1/2} \tag{2.5.7}$$

$$= \widehat{E}_\perp(\mathbf{H}^{1/2}\mathbf{H}'^{-1/2} - \overline{\mathbf{H}^{-1/2}\mathbf{H}'^{1/2}})P = \widehat{E}_\perp SP \tag{2.5.8}$$

is a bounded operator. From (2.5.7) and (2.5.8) we see that $T = \widehat{E}_\perp P$ satisfies the equation

$$((\mathbf{H}\widehat{E}_\perp)^{1/2}v, T(\mathbf{H}'P)^{\dagger 1/2}u) - ((\mathbf{H}\widehat{E}_\perp)^{\dagger 1/2}v, T(\mathbf{H}P)^{1/2}u) = (v, Qu) = (v, Su), \tag{2.5.9}$$
$$v \in \mathsf{ran}(\mathbf{H}\widehat{E}_\perp) \cap \mathcal{D}(\mathbf{H}^{1/2}), \quad u \in \mathsf{ran}(P).$$

The solution T can be seen as the bounded operator from $\mathsf{ran}(P)$ to $\mathsf{ran}((\mathbf{H}\widehat{E}_\perp)^{1/2}) \cap \mathcal{D}(\mathbf{H}'^{1/2})$. The equation (2.5.9) is a bit confusing, since the coefficients operators are self adjoint operators that are only nontrivial in some true subspace of the environment Hilbert space. This has necessitated the use of generalized inverses.

Let us simplify the situation and outline the general picture. We have an unbounded positive definite operator $\mathbf{A}$ and a bounded positive definite operator M. They are defined in, possibly, different subspaces of the environment Hilbert space $\mathcal{H}$. Thus, $\mathcal{H}_M = \mathsf{ran}(M)$ is (of necessity) a closed subspace of $\mathcal{H}$ and likewise

$$\overline{\mathcal{D}(\mathbf{A}^{1/2})}^{\,\mathcal{H}} = \mathsf{ran}(\mathbf{A}^{1/2}) = \mathcal{H}_\mathbf{A}.$$

Let the bounded operator $Q : \mathcal{H}_M \to \mathcal{H}_\mathbf{A}$ be given, then we are looking for the bounded operator $T : \mathcal{H}_M \to \mathcal{H}_\mathbf{A}$ such that

$$(\mathbf{A}^{1/2}v, TM^{-1/2}u) - (\mathbf{A}^{-1/2}v, TM^{1/2}u) = (v, Qu)\ , \qquad v \in \mathcal{D}(\mathbf{A}^{1/2}),\ u \in \mathcal{H}_M. \tag{2.5.10}$$

Formally, we say that T solves the equation

$$\mathbf{A}T - TM = \mathbf{A}^{1/2}QM^{1/2}. \tag{2.5.11}$$

Here $\mathbf{A}^{1/2}QM^{1/2}$ is naturally only a formal expression and does not represent a genuine operator. In the case in which $G = \mathbf{A}^{1/2}QM^{1/2}$ be a genuine operator equation (2.5.11) becomes the rigorous equation

$$\mathbf{A}T - TM = G,$$

called the *Sylvester equation*. Relation (2.5.10) represents a weakly formulated Sylvester equation. The following theorem slightly generalizes the corresponding result from the joint paper [37] and corrects a technical glitch in one of the proofs.

Theorem 2.5.3. *Let* $\mathbf{A}$ *and* M *be positive definite operators in* $\mathcal{H}_{\mathbf{A}}$ *and* $\mathcal{H}_M$*, respectively and let* Q *be a bounded operator from* $\mathcal{H}_M$ *into* $\mathsf{ran}(\mathbf{A}^{1/2}) = \mathcal{H}_{\mathbf{A}}$*. If* M *is bounded and*

$$\|M\| < \frac{1}{\|\mathbf{A}^{-1}\|} \tag{2.5.12}$$

then the weakly formulated Sylvester equation

$$\left(\mathbf{A}^{1/2}v, TM^{-1/2}u\right) - \left(v, \mathbf{A}^{-1/2}TM^{1/2}u\right) = (v, Qu) \tag{2.5.13}$$

has a unique solution T*, given by* $\tau(v,u) = (v,Tu)$ *and*

$$\tau(v,u) = -\frac{1}{2\pi}\int_{-\infty}^{\infty}(\mathbf{A}^{1/2}v, (\mathbf{A} - \mathrm{i}\zeta - d)^{-1}Q(M - \mathrm{i}\zeta - d)^{-1}M^{1/2}u)d\zeta, \tag{2.5.14}$$

where d *is any number satisfying*

$$\|M\| < d < \frac{1}{\|\mathbf{A}^{-1}\|}\ . \tag{2.5.15}$$

Proof. The uniqueness means that

$$\left(\mathbf{A}^{1/2}v, WM^{-1/2}u\right) - \left(v, \mathbf{A}^{-1/2}WM^{1/2}u\right) = 0, \tag{2.5.16}$$

for $u \in \mathcal{H}_M$, $v \in \mathcal{D}(\mathbf{A}^{1/2})$, has the only bounded solution $W = 0$. Let

$$E_n = \int_0^n d\ E_{\mathbf{A}^{1/2}}(\lambda),$$

then in particular

$$\left(\mathbf{A}^{1/2}v, E_nWM^{-1/2}u\right) - \left(v, \mathbf{A}^{-1/2}E_nWM^{1/2}u\right) = 0,$$

for $u \in \mathcal{H}_M$, $v \in \mathcal{D}(\mathbf{A}^{1/2}) \cap E_n\mathcal{H}$. Define the cut–off function

$$f_n(x) = \begin{cases} x, & D \le x \le n \\ n, & n \le x \end{cases}$$

with $D = 1/\|\mathbf{A}^{-1}\|$. The operator $f_n(\mathbf{A}^{1/2})$ is bicontinuous and

$$f_n(\mathbf{A}^{1/2})E_nWM^{-1/2} - f_n(\mathbf{A}^{1/2})^{-1}E_nWM^{1/2} = 0. \tag{2.5.17}$$

Since $f_n(\mathbf{A}^{1/2})$ and $M^{1/2}$ are bounded and positive definite operators, the standard Sylvester equation (2.5.17) has the unique solution

$$E_nW = 0, \qquad n \in \mathbb{N}\ . \tag{2.5.18}$$

This is a consequence of the standard theory of the Sylvester equation with bounded coefficients, see [11, 21]. The statement (2.5.18) implies $W = 0$.

Now for the existence. We use the spectral integral $\mathbf{A} = \int \lambda\, dE(\lambda)$ to compute

$$\begin{aligned}
\int_{-\infty}^{\infty} \|(\mathbf{A} + i\zeta - d)^{-1}\mathbf{A}^{1/2}v\|^2\, d\zeta &= \int_{-\infty}^{\infty} (\mathbf{A}^{1/2}v, \left|\mathbf{A} - i\zeta - d\right|^{-2}\mathbf{A}v)\, d\zeta \\
&= \int_{-\infty}^{\infty} d\zeta \int_D \frac{\lambda\, d(E(\lambda)\mathbf{A}^{1/2}v, \mathbf{A}^{1/2}v)}{(\lambda - d)^2 + \zeta^2} \\
&= \int_D^{\infty} \lambda\, d(E(\lambda)\mathbf{A}^{1/2}v, \mathbf{A}^{1/2}v) \int_{-\infty}^{\infty} \frac{d\zeta}{(\lambda - d)^2 + \zeta^2} \\
&= \int_D^{\infty} \frac{\pi\lambda\, d(E(\lambda)\mathbf{A}^{1/2}v, \mathbf{A}^{1/2}v)}{\lambda - d} \\
&= \pi(\mathbf{A}(\mathbf{A} - d)^{-1}v, v).
\end{aligned} \tag{2.5.19}$$

Analogously, one establishes

$$\int_{-\infty}^{\infty} \|(M - i\zeta - d)^{-1}M^{1/2}u\|^2\, d\zeta = \pi(M(d - M)^{-1}u, u). \tag{2.5.20}$$

The convergence of these integrals justifies the following computation. Set

$$\tau(v, u) = -\frac{1}{2\pi}\int_{-\infty}^{\infty} (\mathbf{A}^{1/2}v, (\mathbf{A} - \mathrm{i}\zeta - d)^{-1}Q(M - \mathrm{i}\zeta - d)^{-1}M^{1/2}u)d\zeta$$

and then compute using (2.5.19) and (2.5.20)

$$\begin{aligned}
|\tau(v, u)|^2 &= \frac{1}{(2\pi)^2}\Big[\int_{-\infty}^{\infty} ((\mathbf{A} + \mathrm{i}\zeta - d)^{-1}\mathbf{A}^{1/2}v, Q(M - \mathrm{i}\zeta - d)^{-1}M^{1/2}u)d\zeta\Big]^2 \\
&\le \frac{\|Q\|^2}{(2\pi)^2}\Big[\int_{-\infty}^{\infty} \|(\mathbf{A} + i\zeta - d)^{-1}\mathbf{A}^{1/2}v\|\ \|(M - i\zeta - d)^{-1}M^{1/2}u\|d\zeta\Big]^2 \\
&\le \frac{\|Q\|^2}{4}(\mathbf{A}(\mathbf{A} - d)^{-1}v, v)(M(d - M)^{-1}u, u).
\end{aligned}$$

This in turn implies that the operator

$$\tau(v,u) = (v, Tu)$$

is a bounded operator and also gives the meaning to the formula (2.5.14).

Now we will prove that this T satisfies the equation (2.5.13). Note that

$$\mathbf{A}(\mathbf{A}-\rho-d)^{-1} = \mathbf{I} + (\rho+d)(\mathbf{A}-\rho-d)^{-1}, \qquad \rho \notin \sigma(\mathbf{A})$$

and then take $v \in \mathcal{D}(\mathbf{A})$ to compute

$$\begin{aligned}
(\mathbf{A}^{1/2}v, TM^{-1/2}u) - (\mathbf{A}^{-1/2}v, TM^{1/2}u) &= \\
&= -\frac{1}{2\pi}\Big[\int_{-\infty}^{\infty} (\mathbf{A}v, (\mathbf{A}-\mathrm{i}\zeta-d)^{-1}Q(M-\mathrm{i}\zeta-d)^{-1}u)\, d\zeta \\
&\qquad - \int_{-\infty}^{\infty} (v, (\mathbf{A}-\mathrm{i}\zeta-d)^{-1}Q(M-\mathrm{i}\zeta-d)^{-1}Mu)\, d\zeta\Big] \\
&= -\frac{1}{2\pi}\Big[\text{v.p.}\int_{-\infty}^{\infty} (v, Q(M-i\zeta-d)^{-1}u)\, d\zeta \\
&\qquad + \int_{-\infty}^{\infty} (i\zeta+d)((\mathbf{A}-i\zeta-d)^{-1}v, Q(M-i\zeta-d)^{-1}u)\, d\zeta \\
&\qquad - \int_{-\infty}^{\infty} (i\zeta+d)((\mathbf{A}-i\zeta-d)^{-1}v, Q(M-i\zeta-d)^{-1}u)\, d\zeta \\
&\qquad - \text{v.p.}\int_{-\infty}^{\infty} ((\mathbf{A}-i\zeta-d)^{-1}v, Qu)\, d\zeta\Big] \\
&= (v, Qu).
\end{aligned}$$

By a usual density argument we conclude that the operator T satisfies (2.5.13). □

Allowing for a more general relation between $\sigma(M)$ and $\sigma(\mathbf{A})$.

An analogue of Theorem 2.5.3 holds, if the assumption (2.5.15) is replaced by a more general one, namely that the interval

$$\big[\|M^{-1}\|^{-1}, \|M\|\big]$$

be contained in the resolvent set of the operator $\mathbf{A}$. We omit the proof of the following result.

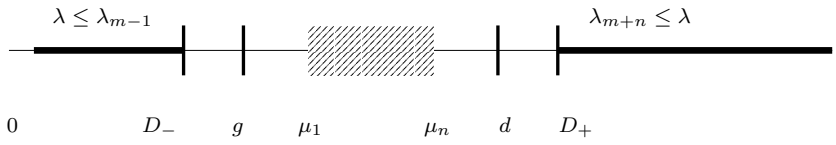

Figure 2.1: The spectral gaps

Theorem 2.5.4. *Let the operators* $\mathbf{A}$*,* M *and* Q *be as in Theorem 2.5.3, and let their spectra be arranged as on Figure 2.1, then (in the sense of (2.5.14))*

$$\begin{aligned} T &= -\frac{1}{2\pi}\int_{-\infty}^{\infty} \mathbf{A}^{1/2}(\mathbf{A} - \mathrm{i}\zeta - d)^{-1}Q(M - \mathrm{i}\zeta - d)^{-1}M^{1/2}d\zeta \\ &\quad + \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathbf{A}^{1/2}(\mathbf{A} - \mathrm{i}\zeta - g)^{-1}Q(M - \mathrm{i}\zeta - g)^{-1}M^{1/2}d\zeta, \end{aligned}$$

where d, g *are chosen from the right and left spectral gap, see Figure 2.1, is the solution of the weak Sylvester equation (2.5.13).*

2.5.2 Invariant subspace estimates

To obtain invariant subspace estimates only a portion of the theory from the preceding section will be necessary. Let, for now, h be a positive definite form. Take an n–dimensional subspace $\mathsf{ran}(P) \subset \mathcal{Q}(h)$, where $P = XX^*$, and let h' be as given by (2.3.3). In the equation (2.5.9) we have already seen the connection between the Sylvester equation and the subspace estimates. The general assumptions of the subspace theorems will correspond to the matching theorems (Theorems 2.4.1 and 2.4.2). The reason is that we can talk about the subspace estimates *only after we have localized the approximated eigenvalues.*

We will use the equation (2.5.9) in somewhat simpler form. We want to compute an estimate of

$$\sin\Theta(P, \widehat{E}) = \|P - \widehat{E}\| = \|\widehat{E}_\perp X\|,$$

where $\mathsf{ran}(\widehat{E})$ is the subspace selected by any of the Theorems 2.4.1 or 2.4.2. From (2.5.9) it follows

$$\begin{aligned} ((\mathbf{H}\widehat{E}_\perp)^{1/2}v, \widehat{E}_\perp PX\Xi^{-1/2}x) - ((\mathbf{H}\widehat{E}_\perp)^{\dagger 1/2}v, \widehat{E}_\perp PX\Xi^{1/2}x) = (v, SXx), \qquad (2.5.21) \\ v \in \mathsf{ran}(\mathbf{H}\widehat{E}_\perp) \cap \mathcal{D}(\mathbf{H}^{1/2}), x \in \mathbb{C}^n. \end{aligned}$$

To simplify the presentation we set

$$\mathbf{A} = \mathbf{H}\widehat{E}_\perp\Big|_{\mathsf{ran}(\mathbf{H}\widehat{E}_\perp)\cap\mathcal{D}(\mathbf{H}^{1/2})}$$

and use Theorem 2.5.3 or Corollary 2.5.4 to obtain the estimates.

Theorem 2.5.5. *Let $X, \Xi, \mathbf{A}, \widehat{E}_\perp, D, \sin\Theta$ be as defined above, then*

$$\|\widehat{E}_\perp X\| \le \frac{\sin\Theta}{\sqrt{1-\sin\Theta}} \frac{\sqrt{D\|\Xi\|}}{D-\|\Xi\|}. \tag{2.5.22}$$

Proof. $\widehat{E}_\perp X = T$ satisfies the equation (2.5.13) where $Q = SX$. By Lemma 2.5.1 we can estimate the norm of T from (2.5.14):

$$\begin{aligned} |\,(Tx,y)\,|^2 &\le \frac{1}{4\pi^2}\beta(y)\alpha(x)\|SX\|^2, \\ \beta(y) &= \int_{-\infty}^{\infty} \big(\mathbf{A}((\mathbf{A}-d)^2+\zeta^2)^{-1}y, y\big)\, d\zeta, \\ \alpha(x) &= \int_{-\infty}^{\infty} \big((\Xi(\Xi-d)^2+\zeta^2)^{-1}x, x\big)\, d\zeta. \end{aligned}$$

Using the spectral integral $\mathbf{A} = \int \lambda\, dE(\lambda)$ we obtain

$$\begin{aligned} \beta(y) &= \int_{-\infty}^{\infty} d\zeta \int_D^{\infty} \frac{\lambda\, d\,(E(\lambda)y,y)}{(\lambda-d)^2+\zeta^2} \\ &= \int_D^{\infty} \lambda\, d\,(E(\lambda)y,y) \int_{-\infty}^{\infty} \frac{d\zeta}{(\lambda-d)^2+\zeta^2} \\ &= \int_D^{\infty} \frac{\pi\lambda\, d\,(E(\lambda)y,y)}{\lambda-d} = \big(\mathbf{A}(\mathbf{A}-d)^{-1}y, y\big)\,\pi \end{aligned}$$

and similarly

$$\alpha(x) = \big(\Xi(d-\Xi)^{-1}x, x\big)\,\pi.$$

Together with Lemma 2.5.1 this gives

$$\begin{aligned} |\,(Tx,y)\,| &\le \frac{\sin\Theta}{\sqrt{1-\sin\Theta}} \frac{1}{2} \sqrt{(\mathbf{A}(\mathbf{A}-d)^{-1}y,y)}\sqrt{(\Xi(d-\Xi)^{-1}x,x)} \\ &\le \frac{\sin\Theta}{2\sqrt{1-\sin\Theta}} \sqrt{\frac{D\|\Xi\|}{(D-d)(d-\|\Xi\|)}}\|x\|\|y\| \end{aligned} \tag{2.5.23}$$

for any $\|\Xi\| < d < D$. The optimal d equals $\frac{(D+\|\Xi\|)}{2}$ and

$$\|\widehat{E}_\perp X\| \le \frac{\sin\Theta}{\sqrt{1-\sin\Theta}} \frac{\sqrt{D\|\Xi\|}}{D-\|\Xi\|}. \qquad \square$$

In the case in which h is only nonnegative and $\mathsf{sin}\Theta_p < 1$ we have shown that $\mathsf{ran}(\mathbf{H}) = \mathsf{ran}(\mathbf{H}')$. Formula (2.3.44) allows us to conclude that $\mathcal{N} = \mathsf{ran}(X) \cap \mathsf{ker}(\mathbf{H})$ is contained in $\mathsf{ran}(\widehat{E})$. This implies that both

$$\widetilde{E} = E - P_{\mathcal{N}}, \qquad \widetilde{P} = P - P_{\mathcal{N}}$$

are orthogonal projections and

$$\|\widehat{E} - P\| = \|\widetilde{E} - \widetilde{P}\|.$$

Since $\mathsf{ran}(\widetilde{P}) \subset \mathsf{ran}(\mathbf{H})$ and $\mathsf{ran}(\widetilde{E}) \subset \mathsf{ran}(\mathbf{H})$ we can reduce the problem to the positive definite case.

Theorem 2.5.6. *Let h be a nonnegative form, and let $\mathsf{ran}(X) \subset \mathcal{Q}$, $P = XX^*$, be such that $\sin\Theta_p < 1$ and $0 < \min\{\delta_l, \delta_r\}$, then*

$$\sin\Theta(\mathsf{ran}(X), \mathsf{ran}(\widehat{E})) = \|P_X - \widehat{E}\| \leq \frac{\sin\Theta_p}{\sqrt{1 - \sin\Theta_p}} \left(\frac{1}{\delta_r} + \frac{1}{\delta_l}\right). \tag{2.5.24}$$

Here we have taken

$$\delta_r = \min_{\substack{k=1,\dots,n \\ p=q+n,\dots,\infty}} \frac{\lambda_p - \mu_k}{\sqrt{\lambda_p \mu_k}} \qquad \textit{and} \qquad \delta_l = \min_{\substack{k=1,\dots,n \\ p=1,\dots,q-1}} \frac{\mu_k - \lambda_p}{\sqrt{\lambda_p \mu_k}}.$$

If $q = 0$ then $\frac{1}{\delta_l} = 0$ and $\delta_l = 1$ by the definition.

Proof. The assumption $0 < \min\{\delta_l, \delta_r\}$ implies $0 < \min\{\gamma_l, \gamma_r\}$, so Corollary 2.3.13 gives

$$\mathsf{ker}(\mathbf{H}) \perp \mathsf{ran}(X)$$

for $q \neq 0$ and

$$\mathsf{ker}(\mathbf{H}) \subset \mathsf{ran}(X)$$

for $q = 0$. Theorem 2.3.12 says that we have reduced the problem to the positive definite case without losing any generality. The same proof as in the Theorem 2.5.3 yields the estimate

$$\|\widehat{E}_\perp P\| \leq \frac{\sin\Theta_p}{\sqrt{1 - \sin\Theta_p}} \left(\frac{\sqrt{D_+ \|\Xi\|}}{D_+ - \|\Xi\|} + \frac{\sqrt{\|\Xi^{-1}\|^{-1} D_-}}{\|\Xi^{-1}\|^{-1} - D_-}\right). \tag{2.5.25}$$

Now, Lemma 2.5.1 and (2.5.25) prove the statement of the theorem. □

Remark 2.5.7. Note that for $0 < \lambda, \mu$

$$\frac{\lambda - \mu}{2\sqrt{\lambda\mu}} \geq \frac{\lambda - \mu}{\lambda + \mu}$$

and

$$\frac{\sin\Theta_p}{\sqrt{1 - \sin\Theta_p}} \leq \frac{\sin\Theta_p}{1 - \sin\Theta_p}.$$

Therefore the assumption

$$\frac{\sin\Theta_p}{1 - \sin\Theta_p} \leq \min\{\delta_l, \delta_r, 1\}$$

implies that both Theorem 2.4.2 and Theorem 2.5.6 hold.

2.6 Higher order estimates

The assumptions of Theorem 2.4.2 are graphically displayed on Figure 2.1. It is an established rule of thumb, that when the eigenvectors are being approximated linearly, then the eigenvalues are being approximated quadratically. In this section we demonstrate that under the assumptions of Theorem 2.4.2 this intuition is correct. Some of our reasoning has been motivated by the techniques from [24, 48], which deal with finite matrices.

Theorem 2.6.1. *Let* $\mathbf{H}$ *be a positive definite operator and let us assume that the eigenvalues are so ordered that*

$$\lambda_1 \leq \cdots \leq \lambda_{m-1} < \lambda_m = \cdots = \lambda_{m+n-1} < \lambda_{m+n} \leq \lambda_{m+n+1} \leq \cdots \ .$$

Let $\mathsf{ran}(P) \subset \mathcal{Q}$, $P = XX^*$ *be such a subspace that the inequality*[5]

$$\eta_\Theta = \frac{\sin\Theta}{1 - \sin\Theta} < \min\{\gamma_s, 1\}, \quad \gamma_s = \min\{\frac{\mu_1 - \lambda_{m-1}}{\lambda_{m-1} + \mu_1}, \frac{\lambda_{m+n} - \mu_n}{\lambda_{m+n} + \mu_n}\} \tag{2.6.1}$$

holds for the Ritz values

$$\mu_1 \leq \mu_2 \leq \cdots \leq \mu_n.$$

Then we have

$$\frac{|\lambda_m - \mu_i|}{\lambda_m} \leq \frac{1}{\gamma_s}\eta_\Theta^2, \qquad i = 1, \ldots, n \ . \tag{2.6.2}$$

[5]Obviously, $\gamma_s = \min\{\gamma_r, \gamma_l\}$, where γ_l and γ_r are defined in Theorem 2.4.2. Index s comes from symmetric.

PROOF. The assumption (2.6.1) implies that the subspace $\mathsf{ran}(X)$ satisfies the requirements of Theorem 2.4.2, hence

$$\frac{|\lambda_{m+i-1} - \mu_i|}{\lambda_{m+i-1}} \leq \eta_\Theta, \qquad i = 1, \ldots, n \ .$$

Take $\widetilde{u} \in \mathsf{ran}(X)$, $\|\widetilde{u}\| = 1$, and set $\widetilde{\mu} = h[\widetilde{u}]$. The operator $\mathbf{H}$ is positive definite, so

$$\widetilde{\mu} = h[\widetilde{u}] = \int \lambda \ d(E(\lambda)\widetilde{u}, \widetilde{u}).$$

Furthermore, the positive definiteness implies [6]

$$(\lambda_m - \lambda)(\lambda_{m+n} - \lambda) \geq 0, \qquad \lambda \in \sigma(\mathbf{H}),$$

so

$$\frac{1}{\lambda} \geq \frac{\lambda_m + \lambda_{m+n} - \lambda}{\lambda_m \lambda_{m+n}}, \qquad \lambda \in \sigma(\mathbf{H}).$$

Integrating both sides of the equation we obtain

$$\int \frac{1}{\lambda} \ \ d(E(\lambda)\widetilde{u}, \widetilde{u}) \geq \frac{\lambda_m + \lambda_{m+n} - \widetilde{\mu}}{\lambda_m \lambda_{m+n}}. \tag{2.6.3}$$

Assume $\widetilde{\mu} \in [\lambda_m, \lambda_{m+n}\rangle$ and multiply the equation (2.6.3) by $\widetilde{\mu}^2$. Then add $-\widetilde{\mu}$ to both sides to obtain

$$\begin{aligned}
\widetilde{\mu} \int \left(\frac{\widetilde{\mu}}{\lambda} - 1\right) \ \ d(E(\lambda)\widetilde{u}, \widetilde{u}) &\geq \widetilde{\mu}^2 \ \frac{\lambda_m + \lambda_{m+n} - \widetilde{\mu}}{\lambda_m \lambda_{m+n}} - \widetilde{\mu} \\
&= \widetilde{\mu} \ \frac{\widetilde{\mu}(\lambda_m + \lambda_{m+n}) - \widetilde{\mu}^2 - \lambda_m \lambda_{m+n}}{\lambda_m \lambda_{m+n}} \\
&= \widetilde{\mu} \ \frac{\widetilde{\mu}(\widetilde{\mu} - \lambda_m)(\lambda_{m+n} - \widetilde{\mu})}{\lambda_m \lambda_{m+n}} \ .
\end{aligned}$$

[6] The idea for this line of proof has been taken from [48].

On the other hand,

$$\begin{aligned}
\widetilde{\mu}\int(\frac{\widetilde{\mu}}{\lambda}-1)\;d(E(\lambda)\widetilde{u},\widetilde{u}) &= -\widetilde{\mu}\left((\widetilde{u},\widetilde{u})-\widetilde{\mu}\;(\widetilde{u},\mathbf{H}^{-1}\widetilde{u})\right)\\
&= \|\mathbf{H}^{1/2}\widetilde{u}-\widetilde{\mu}\mathbf{H}^{-1/2}\widetilde{u}\|^2\\
&= \left(\max_{\|y\|=1}\left|(\mathbf{H}^{1/2}\widetilde{u}-\widetilde{\mu}\mathbf{H}^{-1/2}\widetilde{u},y)\right|\right)^2\\
&= \left(\max_{\|y\|=1}\left|(\mathbf{H}^{1/2}\widetilde{u},y)-(\widetilde{\mu}\;\widetilde{u},\mathbf{H}^{-1/2}y)\right|\right)^2\\
&= \left(\max_{\|y\|=1}\left|h(\widetilde{u},\mathbf{H}^{-1/2}y)-h'(\widetilde{u},\mathbf{H}^{-1/2}y)\right|\right)^2\\
&= \left(\max_{\|y\|=1}\left|\delta h(\widetilde{u},\mathbf{H}^{-1/2}y)\right|\right)^2\\
&= \max_{v\in\mathcal{Q}}\frac{(|\delta h(\widetilde{u},v)|)^2}{h[v]}
\end{aligned}$$

and

$$\int(\frac{\widetilde{\mu}}{\lambda}-1)\;d(E(\lambda)\widetilde{u},\widetilde{u}) = \max_{v\in\mathcal{Q}}\frac{(|\delta h(\widetilde{u},v)|)^2}{h[v]h[\widetilde{u}]} \le \left(\max_{v,u\in\mathcal{Q}}\frac{|\delta h(u,v)|}{\sqrt{h[v]h[u]}}\right)^2 = \eta_\Theta^2\;.$$

Finally, one obtains

$$\frac{\widetilde{\mu}-\lambda_m}{\lambda_m} \le \frac{\lambda_{m+n}}{\lambda_{m+n}-\widetilde{\mu}}\;\eta_\Theta^2 \le \frac{\lambda_{m+n}+\widetilde{\mu}}{\lambda_{m+n}-\widetilde{\mu}}\;\eta_\Theta^2\;. \tag{2.6.4}$$

Analogously for some $\widetilde{\mu}\in\langle\lambda_{m-1},\lambda_m]$, we obtain

$$\frac{\lambda_m-\widetilde{\mu}}{\lambda_m} \le \frac{\lambda_{m-1}}{\widetilde{\mu}-\lambda_{m-1}}\;\eta_\Theta^2 \le \frac{\lambda_{m-1}+\widetilde{\mu}}{\widetilde{\mu}-\lambda_{m-1}}\;\eta_\Theta^2\;. \tag{2.6.5}$$

Now, (2.6.4) and (2.6.5) yield

$$\frac{|\lambda_m-\widetilde{\mu}|}{\lambda_m} \le \max\left\{\frac{\lambda_{m+n}+\widetilde{\mu}}{\lambda_{m+n}-\widetilde{\mu}},\frac{\lambda_{m-1}+\widetilde{\mu}}{\widetilde{\mu}-\lambda_{m-1}}\right\}\;\eta_\Theta^2 \le \frac{1}{\gamma_s}\;\eta_\Theta^2$$

for any $\widetilde{u}\in\mathsf{ran}(X)$ of norm one. Specially, the conclusion of the theorem follows. □

Remark 2.6.2. All of the Ritz values from the subspace $\mathsf{ran}(X)$ are the convex combinations of

$$\mu_1 \le \mu_2 \le \cdots \le \mu_n.$$

It is obvious that there are Ritz values $\widetilde{\mu}$ from the subspace $\mathsf{ran}(X)$ that approximate the eigenvalue λ_m of multiplicity n better than some of the μ_i do. The Ritz value

$$\widetilde{\mu} = \sum_i \frac{1}{n}\mu_i$$

is a good candidate for such an approximation.

Corollary 2.6.3. *Let* $\mathbf{H}$ *be a nonnegative operator such that* $0 \notin \sigma_{ess}(\mathbf{H})$ *and we assume that the eigenvalues of the operator* $\mathbf{H}$ *are so ordered that*

$$\lambda_1 \leq \cdots \leq \lambda_{m-1} < \lambda_m = \cdots = \lambda_{m+n-1} < \lambda_{m+n} \leq \lambda_{m+n+1} \leq \cdots \ .$$

Let $\mathsf{ran}(X) \subset \mathcal{Q}$ *be such a subspace that the inequality*

$$\eta_{\Theta_p} < \min\{\gamma_s, 1\}, \quad \gamma_s = \min\{\frac{\mu_1 - \lambda_{m-1}}{\lambda_{m-1} + \mu_1}, \frac{\lambda_{m+n} - \mu_n}{\lambda_{m+n} + \mu_n}\} \tag{2.6.6}$$

is satisfied for $\lambda_m > 0$ *and for the Ritz values*

$$\mu_1 \leq \mu_2 \leq \cdots \leq \mu_n \ ,$$

then we have

$$\frac{|\lambda_m - \mu_i|}{\lambda_m} \leq \frac{1}{\gamma_s}\eta^2_{\Theta_p}, \qquad i = 1, \ldots, n \ . \tag{2.6.7}$$

The proof of the corollary follows from the observation that (2.6.6) implies $\mathsf{ran}(X) \perp \mathsf{ker}(\mathbf{H}) = \mathsf{ker}(\mathbf{H}')$. The proof is the same as the proof of Theorem 2.6.1, since $\mathsf{ran}(X) \subset \mathsf{ran}(\mathbf{H})$ and $\mathbf{H}$ is positive definite in $\mathsf{ran}(\mathbf{H})$, by assumption.

In Theorem 2.6.1 we have derived an estimate of the error in the eigenvalue approximation relative to the eigenvalue being approximated. It is preferable to have an estimate of the error relative to the Ritz value, since the Ritz value is the quantity we have computed.

Theorem 2.6.4. *Take a nonnegative definite form* h *and the* n*-dimensional subspace* $\mathsf{ran}(P) \subset \mathcal{Q}$, $P = XX^*$. *Let the eigenvalues of the operator* $\mathbf{H}$ *be so ordered that*

$$\lambda_1 \leq \cdots \leq \lambda_{q-1} < D_- \leq \lambda_q = \cdots = \lambda_{q+n-1} \leq D_+ < \lambda_{q+n} \leq \lambda_{q+n+1} \leq \cdots \tag{2.6.8}$$

and let also

$$D_- \leq \mu_1 \leq \cdots \leq \mu_n \leq D_+$$

hold for the eigenvalues of the matrix $\Xi = (\mathbf{H}^{1/2}X)^*\mathbf{H}^{1/2}X \ \in \mathbb{C}^{n\times n}$. *If*

$$\frac{\sin\Theta_p}{1 - \sin\Theta_p} < \min\{\frac{D_+ - \lambda_{q+n}}{D_+ + \lambda_{q+n}}, \frac{\lambda_{q-1} - D_-}{\lambda_{q-1} + D_-}\} = \gamma_s \tag{2.6.9}$$

then

$$\frac{|\lambda_q - \mu_i|}{\mu_i} \leq \frac{1}{\gamma_s} \sin^2 \Theta_p, \quad i = 1, ..., n \ .$$

PROOF. Assume that the form h' is defined as in Theorem 2.3.12. Set $\mathcal{X} = \mathsf{ran}(P)$ and

$$\mathcal{E} = \mathsf{ran}(E(\lambda_{q+n+1}) - E(0))$$

and then take $\mathcal{K} = \mathcal{X} + \mathcal{E}$. Define the matrix—i.e. the finite dimensional operator— H by compressing the form $h[P_{\mathcal{K}}\cdot, P_{\mathcal{K}}\cdot]$ on $\mathcal{K}$. The matrix H is a positive definite matrix, since according to Theorem 2.3.12 the assumption $\mathsf{sin}\Theta_p < 1$ implies

$$\mathsf{ker}(\mathbf{H}) \perp \mathcal{X} \quad \text{and} \quad \mathsf{ker}(\mathbf{H}) \perp \mathcal{E}.$$

Now, the matrix Ξ can be obtained as the further compression of the form $h[P_{\mathcal{K}}\cdot, P_{\mathcal{K}}\cdot]$ to the subspace $\mathcal{X}$. In other words, the Ritz values of the operator $\mathbf{H}$ from the space $\mathcal{X}$ coincide with the Ritz values of the matrix H from the space $\mathcal{X}$. Equivalently, by Ξ_c we denote the operator defined by the compression of the form $h[P_{\mathcal{K}}\cdot, P_{\mathcal{K}}\cdot]$ to the subspace $\mathcal{X}^\perp$, where $\mathcal{X} \oplus \mathcal{X}^\perp = \mathcal{K}$. Therefore, we can represent the matrix H, in the appropriately chosen basis of $\mathcal{X} \oplus \mathcal{X}^\perp = \mathcal{K}$, as

$$H = \begin{bmatrix} \Xi & K^* \\ K & \Xi_c \end{bmatrix}.$$

Since $\mathcal{E}$ reduces the operator $\mathbf{H}$ and $\mathcal{E} \subset \mathcal{K}$ we conclude that the spectrum of the matrix H is the same as the spectrum of the matrix

$$\begin{bmatrix} \Lambda_{q+n+1} & \\ & \Delta_c \end{bmatrix},$$

where $\Lambda_{q+n+1} = \mathbf{H}P_{\mathcal{E}} \mid_{\mathcal{E}}$ and $D_+ \ \mathbf{I} < \Delta_c$. This also follows from Min–max Theorem, if we note that $u_1, \ldots, u_{q+n+1} \in \mathcal{K}$ and $\lambda_i(\mathbf{H}) = h[P_{\mathcal{K}} u_i]$, $i = 1, ..., q+n+1$, so

$$\begin{aligned} \lambda_1(\mathbf{H}) &= \min_{x \in \mathcal{H}} \frac{h[x]}{\|x\|} = \min_{x \in \mathcal{K}} \frac{h[P_{\mathcal{K}}\ x]}{\|x\|} = \lambda_1(H) \\ \lambda_2(\mathbf{H}) &= \min_{x \in \mathcal{H}, x \perp u_1} \frac{h[x]}{\|x\|} = \min_{x \in \mathcal{K}, x \perp u_1} \frac{h[P_{\mathcal{K}}\ x]}{\|x\|} = \lambda_2(H) \\ \vdots \qquad & \qquad \vdots \qquad\qquad\qquad\qquad \vdots \end{aligned}$$

Therefore, the assumption (2.6.9) is the assumption about the matrix H.

By H' denote the matrix defined by compressing the form $h'[P_{\mathcal{K}}\cdot, P_{\mathcal{K}}\cdot]$ on the subspace $\mathcal{K}$. It is straight forward to establish (P and $P_{\mathcal{K}}$ commute) that in the same basis of $\mathcal{K}$

$$H' = \begin{bmatrix} \Xi & 0 \\ 0 & \Xi_c \end{bmatrix}.$$

Set $K_S = \Xi_c^{-1/2} K \Xi^{-1/2}$ then

$$|x^*(H - H')y| \le \|K_S\| \sqrt{x^* H' x \; y^* H' y}, \quad x, y \in \mathcal{K}. \tag{2.6.10}$$

On the other hand, we compute

$$\begin{aligned} |x^*(H - H')y| &= |h[P_{\mathcal{K}} x, P_{\mathcal{K}} y] - h'[P_{\mathcal{K}} x, P_{\mathcal{K}} y]| \\ &\le \sin\Theta_p \sqrt{h'[P_{\mathcal{K}} x, P_{\mathcal{K}} x] h'[P_{\mathcal{K}} y, P_{\mathcal{K}} y]} \\ &= \sin\Theta_p \sqrt{x^* H' x \; y^* H' y}, \end{aligned}$$

so $\|K_S\| \le \sin\Theta_p$. Now (2.6.8) and (2.6.10) and Theorem 2.4.2 imply

$$\begin{aligned} \frac{|\lambda_q - \mu_i|}{\mu_i} &\le \sin\Theta_p < \gamma_s, \qquad i = 1, \ldots, n, \\ \frac{|\lambda_j - \mu_j(\Xi_c)|}{\mu_j(\Xi_c)} &\le \sin\Theta_p < \gamma_s, \qquad j = 1, \ldots, q-1 \\ \frac{|\lambda_{j+n}(H) - \mu_j(\Xi_c)|}{\mu_j(\Xi_c)} &\le \sin\Theta_p < \gamma_s, \qquad j = q, \ldots, \dim(\mathcal{K}) - n. \end{aligned} \tag{2.6.11}$$

Subsequently, see Figure 2.2, it follows

$$\min_{\mu \in \sigma(\Xi_c)} \frac{|\mu - \lambda_q|}{\mu} \ge \min\{\frac{D_+ - \lambda_{q+n}}{D_+ + \lambda_{q+n}}, \frac{\lambda_{q-1} - D_-}{\lambda_{q-1} + D_-}\} = \gamma_s, \tag{2.6.12}$$

so $\mathbf{I} - \lambda_q \Xi_c^{-1}$ is an invertible matrix.

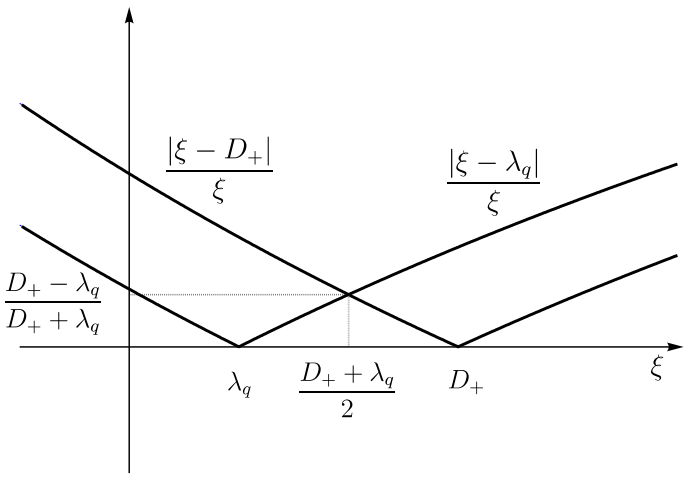

Figure 2.2: The relative gap function

Since $\mathbf{I}-\lambda_q\Xi_c^{-1}$ is a regular matrix, we can use the so called Wilkinson's trick[7] to derive quadratic estimates. The matrix

$$H_S(\lambda_q) = H'^{-1/2}(H-\lambda_q\mathbf{I})H'^{-1/2} = \begin{bmatrix} \mathbf{I}-\lambda_q\Xi^{-1} & K_S^* \\ K_S & \mathbf{I}-\lambda_q\Xi_c^{-1} \end{bmatrix}$$

has the same rank as the matrix $\mathbf{I}-\lambda_q\Xi_c^{-1}$ and

$$H_S(\lambda_q) = \begin{bmatrix} \mathbf{I} & K_S^*(\mathbf{I}-\lambda_q\Xi_c^{-1})^{-1} \\ 0 & \mathbf{I} \end{bmatrix} \begin{bmatrix} (\mathbf{I}-\lambda\Xi^{-1}) - K_S^*(\mathbf{I}-\lambda_q\Xi_c^{-1})^{-1}K_S & 0 \\ 0 & (\mathbf{I}-\lambda_q\Xi_c^{-1}) \end{bmatrix} \begin{bmatrix} \mathbf{I} & 0 \\ (\mathbf{I}-\lambda_q\Xi_c^{-1})^{-1}K_S & \mathbf{I} \end{bmatrix}.$$

The matrices $H_S(\lambda_q)$ and

$$\begin{bmatrix} (\mathbf{I}-\lambda_q\Xi^{-1}) - K_S^*(\mathbf{I}-\lambda_q\Xi_c^{-1})^{-1}K_S & 0 \\ 0 & (\mathbf{I}-\lambda_q\Xi_c^{-1}) \end{bmatrix}$$

are congruent so they have the same rank. This can only take place if

$$\mathbf{I}-\lambda_q\Xi^{-1} = K_S^*(\mathbf{I}-\lambda_q\Xi_c^{-1})^{-1}K_S.$$

Finally from (2.6.12) we obtain

$$\frac{|\lambda_q-\mu_i|}{\mu_i} \leq \frac{1}{\gamma_s}\sin^2\Theta_p, \quad i=1,...,n$$

and the theorem is proved. □

2.7 A computational example

Our perturbation reasoning has yielded a host of estimates for spectral elements. To ease the interpretation of the results we develop a procedure to compute $\mathsf{sin}\Theta_p$ and compare our bounds with other competing estimates on a model problem.

2.7.1 Computing the $\mathsf{sin}\Theta$ for given h and $\mathsf{ran}(X)$

First, let us consider the problem of computing $\mathsf{cos}\Theta(\mathsf{ran}(B),\mathsf{ran}(F)) = \mathsf{cos}\Theta(B,F)$ where $B:\mathbb{C}^n\to\mathcal{H}$ and $F:\mathbb{C}^n\to\mathcal{H}$ are any bounded operators with the maximal rank.

[7] See [49, p. 183]. The finite dimensional part of this proof is essentially contained in [28].

The operators $Q_B = B(B^*B)^{-1/2}$ and $Q_F = F(F^*F)^{-1/2}$ are obviously isometric. According to Section 2.2 (in particular definition (2.2.4)), the cosines of the canonical angles between $\mathsf{ran}(B)$ and $\mathsf{ran}(F)$ are equal to the singular values of the matrix

$$C(B,F) = Q_B^* Q_F = (B^*B)^{-1/2} B^* F (F^*F)^{-1/2} \in \mathbb{C}^{n\times n}. \tag{2.7.1}$$

For the application to the problem of the Ritz value estimates we are dealing with the special B and F. Assume h is positive definite, $X : \mathbb{C}^n \to \mathcal{H}$ is isometric and $P = XX^*$. The equation (2.7.1) now reads

$$C(\mathbf{H}^{1/2}X, \mathbf{H}^{-1/2}X) = ((\mathbf{H}^{1/2}X)^*\mathbf{H}^{1/2}X)^{-1/2}(X^*\mathbf{H}^{-1}X)^{-1/2} = \Xi^{-1/2}\Omega^{-1/2}. \tag{2.7.2}$$

The matrix $\Xi = (\mathbf{H}^{1/2}X)^*\mathbf{H}^{1/2}X$ is called the *Rayleigh quotient* of the subspace $\mathsf{ran}(X)$. Analogously the matrix $\Omega = X^*\mathbf{H}^{-1}X$ will be called the *harmonic Rayleigh quotient*.

By c_i, $i = 1, ..., n$ denote the singular values of

$$C(\mathbf{H}^{1/2}X, \mathbf{H}^{-1/2}X) = \Xi^{-1/2}\Omega^{-1/2}.$$

One way to obtain the sines of the canonical angles is to use the formula

$$s_i = \sqrt{1 - c_i^2}, \qquad i = 1, ..., n\ . \tag{2.7.3}$$

However, (2.7.3) is notoriously unstable as the computational procedure to evaluate s_i, cf. [27]. The right way to obtain the computationally robust formulas (in particular with regard to the computations in the floating point arithmetic) for the canonical sines would be to develop the procedure that does not require the matrices Ξ and Ω, but operators B and F. This is highly nontrivial even when we are dealing with large matrices, not to mention unbounded operators which are our principal concern. The problem of the robust computation of the sines of the canonical angles between the finite dimensional subspaces of the infinite dimensional Hilbert space will be not further considered in this thesis. Formula (2.7.3) will be sufficient for the theoretical case studies we plan to carry out.

The singular value problem for the matrix $C(\mathbf{H}^{1/2}X, \mathbf{H}^{-1/2}X)$ can be reformulated as the generalized eigenvalue problem

$$\Xi^{-1}x_i = c_i^2\Omega x_i, \quad i = 1, ..., n\ . \tag{2.7.4}$$

The squares of the sines s_i^2, $i = 1, ..., n$ are all the eigenvalues of the symmetric *matrix*

pencil[8] $(\Omega - \Xi^{-1}, \Omega)$ and in particular

$$\sin^2\Theta = \max_{x\neq 0,\ x\in\mathbb{C}^n} \frac{|x^*(\Omega-\Xi^{-1})x|}{x^*\Omega x} = \max_{x\in\mathcal{X}\setminus\{0\}} \frac{|(x,\mathbf{H}^{-1}x)-(x,\mathbf{H}'^{-1}x)|}{(x,\mathbf{H}^{-1}x)}\ . \tag{2.7.5}$$

The relation (2.7.5) is particularly interesting. It reveals the nature of the estimate $\mathsf{sin}\Theta$. In (2.7.5) we see that the difference between (the inverses of) the operators is measured on the test space $\mathcal{X}$ **only**. Let us consider, for the moment, the invariant subspace estimate (estimate (2.5.22))

$$\|P-\widehat{E}\| \leq \frac{\sin\Theta}{\sqrt{1-\sin\Theta}}\frac{\sqrt{D\|\Xi\|}}{D-\|\Xi\|}.$$

The projection $\widehat{E}$ in (2.5.22) has been defined with the help of Theorem 2.4.1. This estimate of $\|P-\widehat{E}\|$ has been computed with the help of the compression of the inverse of the operator $\mathbf{H}$ on the test space $\mathsf{ran}(P)$. An alternative approach would have been to use the integral representation (for some appropriate Γ)

$$P=\int_\Gamma (\mathbf{H}'-\xi\mathbf{I})^{-1}\,d\xi, \qquad \widehat{E}=\int_\Gamma (\mathbf{H}-\xi\mathbf{I})^{-1}\,d\xi.$$

To estimate the norm of

$$P-\widehat{E}=\int_\Gamma ((\mathbf{H}'-\xi\mathbf{I})^{-1}-(\mathbf{H}-\xi\mathbf{I})^{-1})\,d\xi \tag{2.7.6}$$

one would need an extensive information on the family of operators (resolvents)

$$(\mathbf{H}'-\xi\mathbf{I})^{-1}, \qquad (\mathbf{H}-\xi\mathbf{I})^{-1}, \qquad \xi\in\Gamma.$$

On the other hand, we require information on the **parts** of the inverses

$$\Xi^{-1}=X^*\mathbf{H}'^{-1}X, \qquad \Omega=X^*\mathbf{H}^{-1}X$$

in the test space, **only**. Furthermore, any estimate of $\|P-\widehat{E}\|$, which can be obtained from (2.7.6), will contain some form of a measure of the distance between the Ritz values and the unwanted component of the spectrum. Such information depends on the choice of the curve Γ. As a comparison we offer an optimal choice of the distance function (cf. Theorem 2.5.5) and provide an estimate

$$\|P-\widehat{E}\| \leq \frac{\sin\Theta}{\sqrt{1-\sin\Theta}}\frac{\sqrt{D\|\Xi\|}}{D-\|\Xi\|}$$

which is featuring computable quantities, only.

[8] (A,M) will be used to denote the matrix pencil (A and M are assumed to be symmetric matrices), cf. [53].

2.7.2 A Sturm-Liouville problem with coupled boundaries

We compare our bounds with other explicit estimates (i.e. bounds that are free from unknown quantities) found in literature. The comparison will be carried out as a case study on a model problem.

Our subspace theorem is compared with the Davis–Kahan $\sin\Theta$ theorem and the Temple–Kato eigenvector bound, cf. [18, 21]. On the other hand, the Ritz value bound is compared with the Temple–Lehmann and the Temple–Kato inequalities, cf. [18, 20, 50] and (2.4.1). The model problem is (cf. [50])

$$\begin{aligned} -z'' - \alpha\, z &= \omega\ z, \\ e^{\mathrm{i}\theta} z(0) &= z(2\pi), \\ e^{\mathrm{i}\theta} z'(0) &= z'(2\pi), \end{aligned} \tag{2.7.7}$$

where $\theta \in [0, \pi]$ and $\alpha \in \mathbb{R}$ is a constant fixed so that all the eigenvalues are positive. The solution to problem (2.7.7) is given by the pairs

$$\begin{aligned} \omega_{\pm k} &= \left(\pm k + \frac{\theta}{2\pi}\right)^2 - \alpha, \quad z_{\pm k}(t) = e^{-\mathrm{i}\left(\pm k + \frac{\theta}{2\pi}\right)t}, \quad k \in \mathbb{N} \\ \omega_0 &= \left(\frac{\theta}{2\pi}\right)^2 - \alpha, \qquad\quad z_0(t) = e^{-\mathrm{i}\frac{\theta}{2\pi}t}. \end{aligned}$$

On Figure 2.3 we see increasingly ordered eigenvalues of the family of problems (2.7.7) displayed as functions of θ. For $\theta = \pi$ we have an eigenvalue problem that has all the eigenvalues of multiplicity two. By varying the parameter θ in a "neighborhood" of π we construct eigenvalue problems that have as tightly clustered eigenvalues as we desire. "Relative" distance functions are not shift invariant. For every $\theta \in [0, \pi]$ we can choose a shift α so as to make the two lowermost eigenvalues well separated in a "relative" sense.

For the parameters θ and α we choose

$$\theta = \frac{9999\,\pi}{10000}, \qquad \alpha = 0.2499 \tag{2.7.8}$$

and compute

$$\min_{\substack{i=1,2 \\ p\neq 1,2}} \frac{|\lambda_i - \lambda_p|}{\sqrt{\lambda_i \lambda_p}} = 115.459, \quad \frac{|\lambda_1 - \lambda_2|}{\sqrt{\lambda_1 \lambda_2}} = 1.15466$$

and

$$\min_{\substack{i=1,2 \\ p\neq 1,2}} |\lambda_i - \lambda_p| = 1.9998, \quad |\lambda_1 - \lambda_2| = 10^{-4}.$$

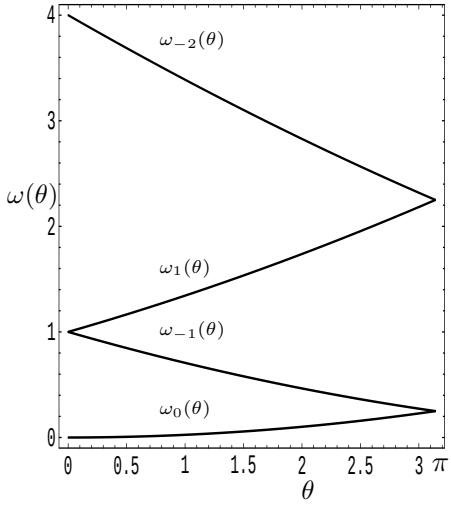

Figure 2.3: Increasingly ordered eigenvalues of the family of problems (2.7.7) as functions of θ.

The eigenvalue problem (2.7.7) with this choice of parameters α and θ will demonstrate why in some situations it is preferable to have bounds that are functions of the "relative" gaps rather than the "absolute" ones.

We will rewrite the problem (2.7.10) in abstract form. With α and θ as in (2.7.8) the form

$$(\mathbf{H}^{1/2}z, \mathbf{H}^{1/2}v) = h(z,v) = \int_0^{2\pi} \overline{z'}v' - \alpha \int_0^{2\pi} \overline{z}v \tag{2.7.9}$$

is positive definite with the domain

$$\mathcal{Q}(h) = \{f : f, f' \in L^2[0, 2\pi], e^{\mathrm{i}\theta} f(0) = f(2\pi)\}.$$

In a weak formulation (2.7.7) reads

$$h(z,v) = \lambda\,(z,v)\,, \qquad z, v \in \mathcal{Q}(h). \tag{2.7.10}$$

Obviously, the eigenvalues and eigenvectors of Problem 2.7.7 are the same as the eigenvalues and eigenvectors of the operator $\mathbf{H}$. Assuming the usual ordering of eigenvalues of $\mathbf{H}$, we get $\lambda_1 = \omega_0(\theta)$, $\lambda_2 = \omega_{-1}(\theta)$, $\lambda_3 = \omega_1(\theta)$ (and so on). Similarly, in the notation we have employed so far, we have $v_1 = z_0$, $v_2 = z_{-1}$, $v_3 = z_1$. For the equidistant subdivision

$$0 = x_0 < x_1 < \cdots < x_n < x_{n+1} = 2\pi, \quad x_{i+1} - x_i = \frac{1}{n+1},$$

we introduce the finite dimensional function space

$$\begin{aligned}\mathcal{V}_n^3 \;=\; &\big\{f : f \in C^1[0,2\pi], e^{\mathrm{i}\theta} f(0) = f(2\pi),\\ &\qquad f \text{ is cubic on } \langle x_{j-1}, x_j\rangle, j = 1,\dots,n+1\big\} \subset \mathcal{D}(\mathbf{H}).\end{aligned}$$

By Π_n^3 denote the interpolation operator

$$\Pi_n^3 \;:\; C^1[0,2\pi] \to \mathcal{V}_n^3.$$

Let the matrices

$$X_n = [\Pi_n^3 v_1], \qquad Y_n = [\Pi_n^3 v_1\, \Pi_n^3 v_2]$$

be understood as operators. For the test subspaces we take $X_n\mathbb{C}$, $Y_n\mathbb{C}^2$, $n \in \mathbb{N}$. Obviously,

$$\lim_n (\Pi_n^3 v_j) = v_j, \quad j = 1, 2\,,$$

so we analyze the error bounds as functions of n.

Green function of the operator $\mathbf{H}u = -u'' - \alpha u$, defined by the form (2.7.10), is

$$G(t_1,t_2) = \frac{\mathrm{i}}{2\sqrt{\alpha}} \left(e^{\mathrm{i}\sqrt{\alpha}|t_1-t_2|} + \frac{e^{\mathrm{i}(t_1-t_2)\sqrt{\alpha}}}{-1+e^{-2\mathrm{i}\pi\sqrt{\alpha}-\mathrm{i}\theta}} + \frac{e^{\mathrm{i}(t_2-t_1)\sqrt{\alpha}}}{-1+e^{-2\mathrm{i}\pi\sqrt{\alpha}+\mathrm{i}\theta}} \right). \tag{2.7.11}$$

In this case we can use the formula

$$\big(\mathbf{H}^{-1}u, v\big) = \int_0^{2\pi} dt_1 \int_0^{2\pi} G(t_1,t_2)\overline{u(t_2)}v(t_1)\,dt_2 \tag{2.7.12}$$

to compute the elements of Ω.

The results of the numerical experiments are presented on Figure 2.4 where:

Graph (a): $(\times)$ denotes our lower bound for the λ_1 obtained from the subspace $X_n\mathbb{C}$ using Theorem 2.3.8

$$\lambda_1 \geq (1 - \sin\Theta)\big(\mathbf{H}^{1/2}X_n, \mathbf{H}^{1/2}X_n\big)$$

and $(\blacklozenge)$ denotes a lower bound for the λ_1 obtained from the *Temple–Kato inequality*

$$\lambda_1 \geq \big(\mathbf{H}^{1/2}X_n, \mathbf{H}^{1/2}X_n\big) - \frac{(\mathbf{H}X_n, \mathbf{H}X_n) - \big(\mathbf{H}^{1/2}X_n, \mathbf{H}^{1/2}X_n\big)^2}{\lambda_2 - (\mathbf{H}^{1/2}X_n, \mathbf{H}^{1/2}X_n)}.$$

The dashed line represents the value of λ_1. Our desire not to get negative lower estimates for the eigenvalues of positive definite operators can be clearly seen on the picture.

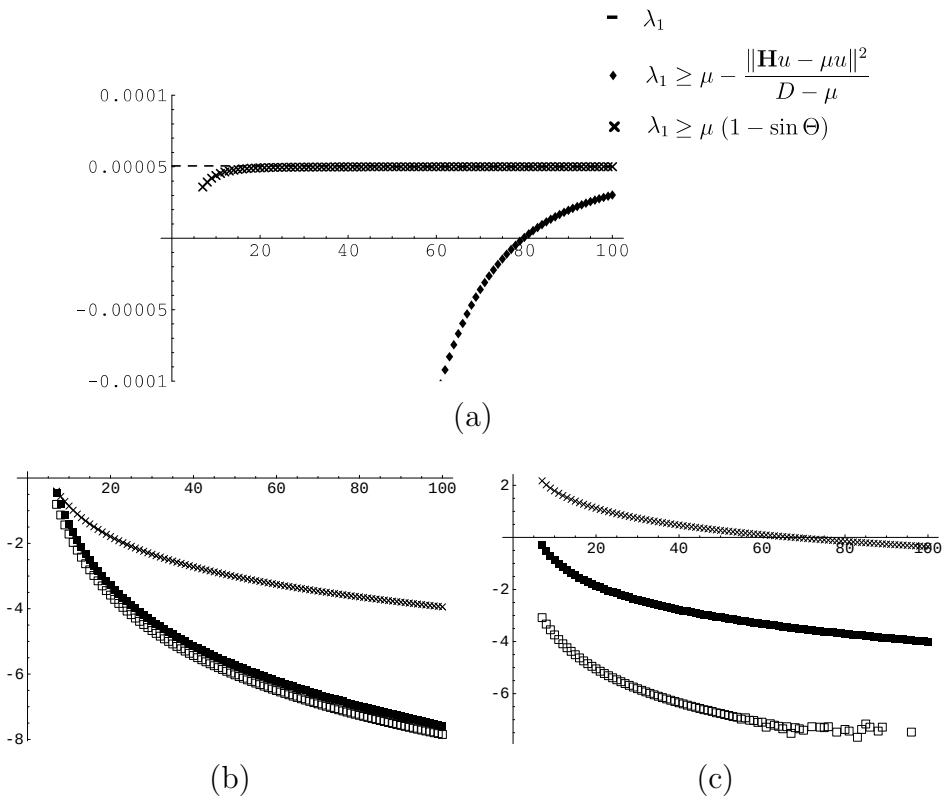

Figure 2.4: The eigenvalue and the eigenvector estimates from the space $X_n\mathbb{C} \subset \mathcal{V}_n^3$ as functions of n.

Graph (b): (□) denotes the logarithm of the true error, (■) denotes the logarithm of the quadratic estimate from Theorem 2.6.4 and (×) denotes the logarithm of the bound from Theorem 2.3.8. An uncanny accuracy of the quadratic estimate can easily be spotted on the graph.

Graph (c): (■) denotes the logarithm of the subspace bound from Theorem 2.5.5

$$\sin\Theta(E(\lambda_1), X_n) \leq \frac{\sin\Theta}{\sqrt{1-\sin\Theta}} \frac{\sqrt{(\mathbf{H}^{1/2}X_n, \mathbf{H}^{1/2}X_n)\,\lambda_2}}{\lambda_2 - (\mathbf{H}^{1/2}X_n, \mathbf{H}^{1/2}X_n)},$$

(×) denotes the logarithm of the *Davis–Kahan bound*

$$\sin\Theta(E(\lambda_1), X_n) \leq \frac{\|r_{X_n}\|}{\lambda_2 - (\mathbf{H}^{1/2}X_n, \mathbf{H}^{1/2}X_n)}$$

whereas the *Temple–Kato eigenvector bound*

$$\sin\Theta(E(\lambda_1), X_n) \leq \frac{2}{\lambda_2 - 0} \sqrt{\left((\mathbf{H}^{1/2}X_n, \mathbf{H}^{1/2}X_n) - \frac{\lambda_2 - 0}{2}\right)^2 + \|r_{X_n}\|^2}$$

turns out to be larger than 1 and it is not graphically represented. The residual vector $r_{X_n} \in \mathcal{H}$ was defined as

$$r_{X_n} = \mathbf{H}X_n - \left(\mathbf{H}^{1/2}X_n, \mathbf{H}^{1/2}X_n\right) X_n.$$

The logarithm of the true error

$$\sin\Theta(E(\lambda_1), X_n)$$

is denoted by ($\square$). The inherent lack of stability of the formula (2.7.3) to floating point perturbations can be observed in the lower right corner of the graph. The evaluation of (2.7.3) was done in double precision and (2.7.3) begins to wobble as $\sqrt{1-s^2}$ approaches 10^{-8}. This is an expected behavior, since we cannot evaluate $1-s^2$, in double precision, more accurately than the machine precision 10^{-16}.

Remark 2.7.1. In both the Temple–Kato and the Davis–Kahan bounds

$$\lambda_2 - \left(H^{1/2}X_n, H^{1/2}X_n\right)$$

should be understood as the best possible estimate of the spectral gap

$$\min_{\lambda\in\sigma(H)\setminus\lambda_1} \left|\left(H^{1/2}X_n, H^{1/2}X_n\right) - \lambda\right|.$$

All measures of the spectral gap, that appear in the subspace theorems, are estimated in the same fashion.

The right matching

The subspace approximation theorem will be tested on the subspaces $Y_n\mathbb{C}^2$. Specifically, we want to investigate the moment in which we can establish that there are exactly two eigenvalues of $\mathbf{H}$ that are being approximated by the two Ritz values from the space $Y_n\mathbb{C}^2$.

For the operator $\mathbf{H}$, defined by (2.7.9) and α and θ as in (2.7.8), there exists a $D \in \mathbb{R}$ such that

$$\lambda_2 < D < \lambda_3.$$

By $\mu_1^n \leq \mu_2^n$ denote the Ritz values of $\mathbf{H}$ from the subspace $Y_n\mathbb{C}^2$. Assuming $0 < \sin\Theta(Y_n) < 1$, we obtain

$$\frac{\sin\Theta(Y_n)}{1-\sin\Theta(Y_n)} > \frac{\sin\Theta(Y_n)}{\sqrt{1-\sin\Theta(Y_n)}}.$$

Assume further that $n \in \mathbb{N}$ is such that $\frac{\sin\Theta(Y_n)}{1-\sin\Theta(Y_n)}\frac{D+\mu_2^n}{D-\mu_2^n} < 1$, then we can conclude that

$$\frac{\sin\Theta(Y_n)}{\sqrt{1-\sin\Theta(Y_n)}}\frac{\sqrt{D\mu_2^n}}{D-\mu_2^n} \leq \frac{\sin\Theta(Y_n)}{1-\sin\Theta(Y_n)}\frac{D+\mu_2^n}{D-\mu_2^n}.$$

Provided $n \in \mathbb{N}$ is such that

$$\frac{\sin\Theta(Y_n)}{1-\sin\Theta(Y_n)}\frac{D+\mu_2^n}{D-\mu_2^n} < 1, \tag{2.7.13}$$

Theorems 2.4.1 and 2.5.5 guarantee that we have both a good approximation of the desired eigenspace and a good approximation of the accompanying eigenvalues.

On Figure 2.5 we have displayed the comparison of the true error in the Rayleigh–Ritz approximation from the subspace $Y_n\mathbb{C}^2 \subset \mathcal{V}_n^3$ with our bound. The error in approximation of λ_2 is denoted as ($\star$), while the error in the approximation of λ_1 is denoted as ($\blacklozenge$). The bound, denoted as ($\blacksquare$), follows the error in λ_1, since Theorem 2.3.8 guarantees the existence of the matching between the Ritz values and the part of the spectrum of the same multiplicity.

Figure 2.6 is even more instructive, it illustrates the real strength of our bounds. For the same example it displays

- $\blacklozenge$ the error between $\mu_{\min} = \min\sigma(\Xi)$ and λ_1 (the expected matching),
- $\bigstar$ the error between $\mu_{\min}$ and λ_2 (the wrong matching),
- $\blacksquare$ our bound.

We can observe that

$$\sin\Theta(Y_n) < \frac{|\mu_1-\lambda_2|}{\mu_1}$$

immediately implies the correct matching of μ_1 to λ_1. The connection between $\sin\Theta(Y_n)$ and $\frac{|\mu_1-\lambda_2|}{\mu_1}$ is not surprising when one has (2.6.12) in mind. Note also that regardless of the fact that $Y_n = [\Pi_n^3 v_1 \Pi_n^3 v_2]$, the Ritz value μ_1 is closer to λ_2 than to λ_1 for $n \leq 5$. As $\mathsf{sin}\Theta(Y_n)$ "enters" the spectral gap μ_1 veers away from λ_2 and never comes close to it again. This considerations show that—at least on the model example—our bound is quite sharp and that "matching condition" detects the multiplicity of the eigenvalue well.

Remark 2.7.2. Theorem 2.4.1 is the reason why we have opted for Temple–Kato inequality rather than the Residual theorem of Davis, Kahan and Weinberger [22]. The residual theorem of Davis, Kahan and Weinberger can also be employed for the subspaces $X_n\mathbb{C}$

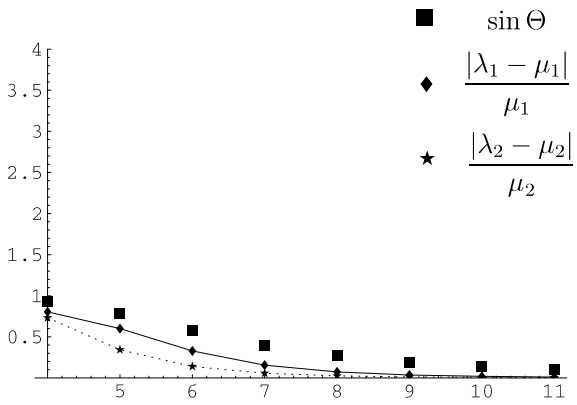

Figure 2.5: The true error and the Ritz value estimate for the approximation from the subspace $Y_n\mathbb{C}^2 \subset \mathcal{V}_n^3$ as a function of n.

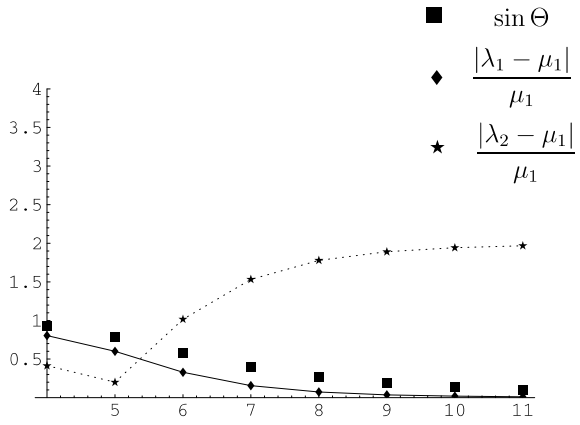

Figure 2.6: Right and wrong matching

and $Y_n\mathbb{C}^2$. However, we would need to compare Davis, Kahan and Weinberger's residual to the "absolute" gap if we were to guarantee that the Ritz values match the lowermost eigenvalues. This conclusion is necessary in order to produce the plots analogous to Figure 2.4. On the other hand, Temple–Kato inequality is a quadratic estimate which utilizes the same residual as the result of Davis, Kahan and Weinberger (cf. [22]). Furthermore, using the additional information contained in the "absolute" gap it provides the bound of the lowermost eigenvalue.

2.7.3 A case study: The linear finite elements for the Sturm–Liouville problem with coupled boundaries

We will now construct a finite element procedure to compute the two lowermost eigenvalues of the Sturm–Liouville problem (2.7.7) where $\theta \in [0, \pi]$ and $\alpha \in \mathbb{R}$ is a constant fixed so that all the eigenvalues are positive. We will chose the same parameters α and θ as in (2.7.8). The matrix Ω from

$$\sin^2 \Theta = \max_{x \neq 0,\ x \in \mathbb{C}^n} \frac{|x^*(\Omega - \Xi^{-1})x|}{x^*\Omega x}$$

will be computed with the help of Green function (2.7.11) and Formula (2.7.12).

We take

$$\begin{aligned} \mathcal{V}_N^1 &= \big\{ f : f \in C[0, 2\pi], \mathrm{e}^{\mathrm{i}\theta} f(0) = f(2\pi), \\ &\qquad f \text{ is linear in } \langle x_{j-1}, x_j \rangle, j = 1, \ldots, N \big\} \subset \mathcal{Q} \end{aligned}$$

as a finite element space. The space $\mathcal{V}_N^1$ is an N–dimensional subspace of $\mathcal{Q}$ with the basis

$$\psi_k(x) = \frac{N}{2\pi} \begin{cases} (x - \frac{(k-1)2\pi}{N}), & \frac{(k-1)2\pi}{N} \leq x \leq \frac{k2\pi}{N} \\ (\frac{(k+1)2\pi}{N} - x), & \frac{k2\pi}{N} \leq x \leq \frac{(k+1)2\pi}{N} \\ 0, & \text{otherwise} \end{cases}, \quad 1 \leq k \leq N-1,$$

$$\psi_N(x) = \frac{N}{2\pi} \begin{cases} \mathrm{e}^{-\mathrm{i}\theta}(x - \frac{(N-1)2\pi}{N}), & \frac{(N-1)2\pi}{N} \leq x \leq 2\pi \\ (\frac{2\pi}{N} - x), & 0 \leq x \leq \frac{2\pi}{N} \\ 0, & \text{otherwise} \end{cases}.$$

The subspace $\mathcal{V}_N^1$ will be represented as the space $\mathbb{C}^N$ with the scalar product given by the matrix

$$(S_N)_{kp} = (\psi_k, \psi_p) = \psi_k^* \psi_p. \tag{2.7.14}$$

By writing $(\psi_k, \psi_p) = \psi_k^* \psi_p$ in (2.7.14) we emphasize the finite dimensional character of $\mathcal{V}_N^1$. This notation will be further used in this context. It can be justified by noting that if $\psi_k^* \in L^{2*} = L^2$, then $\psi_k^* \psi_p$ denotes the value of the functional ψ_k^* on the vector ψ_p. The matrix

$$(H_N)_{kp} = h(\psi_k, \psi_p)$$

represents the form h in the basis $(\psi_k)_{k=1}^N$ of the subspace $\mathcal{V}_N^1$. Let the matrix

$$\Psi_N = \begin{bmatrix} \psi_1 & \cdots & \psi_N \end{bmatrix}$$

be understood as an operator from $\mathbb{C}^N$ into $\mathcal{H}$, then $\Psi_N^*\Psi_N = S_N$. Given the vectors $x_1, ..., x_r \in \mathbb{C}^N$, define the matrix $X = \begin{bmatrix} x_1 & \cdots & x_r \end{bmatrix} \in \mathbb{C}^{N\times r}$. The eigenvalues of

$$\Xi = (X^* S_N X)^{-1/2} X^* H_N X (X^* S_N X)^{-1/2}$$

are the Ritz values of the operator $\mathbf{H}$ associated with the subspace $\mathsf{ran}(\Psi_N X) \subset \mathcal{H}$. It is important to note that if we were to define the isometry[9]

$$\mathbf{X} = \Psi_N X (X^* S_N X)^{-1/2}, \tag{2.7.15}$$

then we would have

$$\Xi = (\mathbf{H}^{1/2}\mathbf{X})^* \mathbf{H}^{1/2}\mathbf{X}. \tag{2.7.16}$$

Let

$$\mu_1 \le \mu_2 \le \cdots \le \mu_r$$

be the eigenvalues, counting the multiplicities, of the matrix Ξ and $u_1, ..., u_r$ the corresponding eigenvectors. The vectors $Xu_1, ..., Xu_r$ are the Ritz vectors of the operator $\mathbf{H}$ belonging to the subspace $\mathsf{ran}(\mathbf{X})$. We also note that the Ritz values are the solution of the generalized eigenvalue problem

$$X^* H_N X \; u = \mu \; X^* S_N X \; u. \tag{2.7.17}$$

Computing $\sin\Theta$ for Problem (2.7.7)

Let us assume we are given the finite dimensional subspace $\mathsf{ran}(\mathbf{X}) \subset \mathcal{Q}$, represented by the isometry $\mathbf{X} : \mathbb{C}^r \to \mathcal{Q}$. We already know that $\Xi = (\mathbf{H}^{1/2}\mathbf{X})^*(\mathbf{H}^{1/2}\mathbf{X})$ and now we introduce the matrix $\Omega = \mathbf{X}^*\mathbf{H}^{-1}\mathbf{X}$ to compute

$$\sin^2\Theta = \max_{x\neq 0,\; x\in\mathbb{C}^n} \frac{|x^*(\Omega - \Xi^{-1})x|}{x^*\Omega x}\,.$$

For $f, g \in \mathbb{C}^N$ we have

$$\begin{aligned}
(\mathbf{X}g, \mathbf{H}^{-1}\mathbf{X}f) &= \int\int G(x-y)(\mathbf{X}f)(y)\overline{(\mathbf{X}g)(x)}\, dy\, dx \\
&= \sum_{p,k} \int\int G(x-y)\widetilde{f}_p\overline{\widetilde{g}}_k \psi_p(y)\overline{\psi_k(x)}\, dy\, dx \\
&= \sum_{p,k} (T_N)_{kp}\widetilde{f}_p\overline{\widetilde{g}}_k = \widetilde{g}\; {}^*T_N\widetilde{f}.
\end{aligned} \tag{2.7.18}$$

[9]Bold script has been used to denote the isometry $\mathbf{X} : \mathbb{C}^r \to \mathcal{H}$, so that it can be distinguished from the matrix $X \in \mathbb{C}^{N\times r}$ (its representation in the space $\mathcal{V}_N^1 \simeq \mathbb{C}^N$). This is a small departure from the convention to reserve the bold script symbols for unbounded operators.

Here $(X(X^*S_NX)^{-1/2}f)_p = \widetilde{f}_p$ and $(X(X^*S_NX)^{-1/2}g)_k = \widetilde{g}_k$, so

$$G = (X^*S_NX)^{-1/2}X^*T_NX(X^*S_NX)^{-1/2}.$$

The matrix T_N, introduced in the relation (2.7.18), is defined by the relation

$$T_N = \Psi_N^*\mathbf{H}^{-1}\Psi_N.$$

Its elements have the property

$$\begin{aligned}(T_N)_{kp} &= \int\int G(x-y)\psi_p(y)\overline{\psi_k(x)}\,dy\,dx \\ &= \int\int G(x-y+\frac{(p-k)2\pi}{N})\psi_1(y)\psi_1(x)\,dy\,dx. \qquad (2.7.19)\end{aligned}$$

So, T_N is a symmetric Toeplitz matrix by a definition and, furthermore, using (2.7.19) we derive a closed formula for the elements of the matrix T_N. Namely, we were able to explicitly express all of the $(T_N)_{kp}$ as functions of α and θ. The formulas were computed symbolically with the use of MATHEMATICA® and are to cumbersome to be displayed here.

Let $\mu_1^N \leq \cdots \leq \mu_N^N$ be the eigenvalues of the matrix $S_N^{-1/2}H_NS_N^{-1/2}$ and $u_1^N, ..., u_N^N$ the corresponding eigenvectors. Define the isometry $\mathbf{X}$ as

$$\mathbf{X} = \Psi_N \begin{pmatrix} u_1^N & u_2^N \end{pmatrix} \left(\begin{pmatrix} u_1^N & u_2^N \end{pmatrix}^* S_N \begin{pmatrix} u_1^N & u_2^N \end{pmatrix} \right)^{-1/2}.$$

According to the Rayleigh–Ritz procedure μ_1^N and μ_2^N are taken as the Rayleigh–Ritz approximations to the λ_1 and λ_2, the two smallest eigenvalues of the operator $\mathbf{H}$, from the finite element space $\mathcal{V}_N^1$.

Remark 2.7.3. It is important to note that

$$\Xi = (\mathbf{H}^{1/2}\mathbf{X})^*\mathbf{H}^{1/2}\mathbf{X}$$

is well defined for any $\mathbf{X}\mathbb{C}^2 \subset \mathcal{Q}$, not just for the $\mathbf{X}\mathbb{C}^2$ defined by the Ritz vectors from $\mathcal{V}_N^1$. In order to apply Theorem 2.4.1 we only need the estimate of the "relative" gap

$$\min_{\substack{i=1,2 \\ p \neq 0,-1}} \frac{\lambda_p - \mu_i}{\lambda_p + \mu_i}.$$

Here μ_1 and μ_2 are the eigenvalues of Ξ.

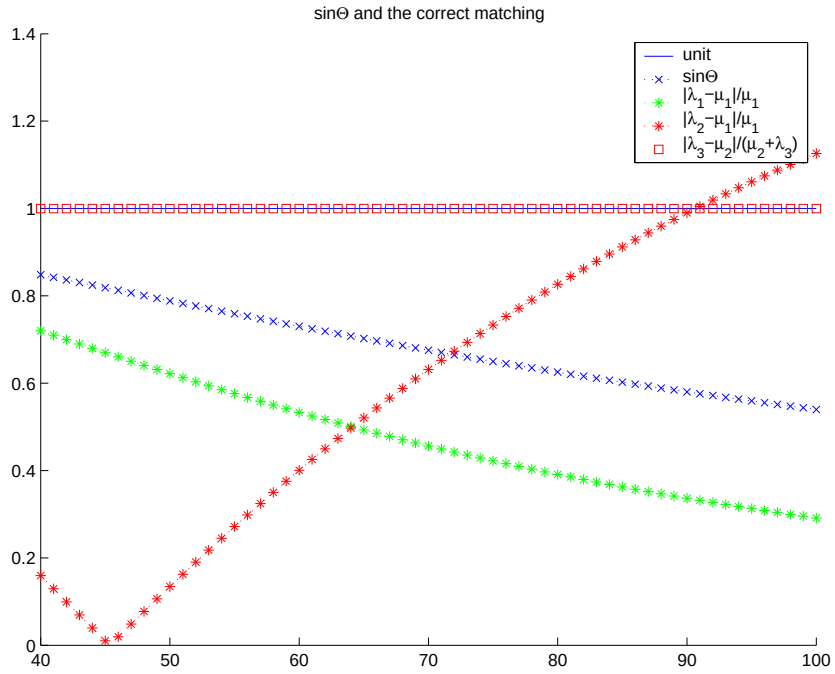

Figure 2.7: The matching of the Ritz values for the finite element approximations.

In this case we can solve the matrix problem (2.7.17) directly. The Ritz values and the Ritz vectors (assuming $\alpha = 0$) are given by the formula

$$\mu_{(N,k)} = 6\frac{N^2}{4\pi^2}\frac{1-\cos\left[\frac{2\pi}{N}\left((-1)^k\lfloor\frac{k}{2}\rfloor - \frac{\theta}{2\pi}\right)\right]}{2+\cos\left[\frac{2\pi}{N}\left((-1)^k\lfloor\frac{k}{2}\rfloor - \frac{\theta}{2\pi}\right)\right]}\,, \tag{2.7.20}$$

$$u_{(N,k)} = [1 \quad e^{\mathrm{i}\left[\frac{\theta}{2\pi}+(-1)^{k+1}\lfloor\frac{k}{2}\rfloor\right]\frac{2\pi}{N}} \quad \dots \quad e^{\mathrm{i}\left[\frac{\theta}{2\pi}+(-1)^{k+1}\lfloor\frac{k}{2}\rfloor\right]\frac{2\pi(N-1)}{N}}]^*, \tag{2.7.21}$$

for $k = 1, \dots, N$. In the usual notation we have

$$\mu_1^N = \mu_{(N,1)} - \alpha, \qquad u_1^N = u_{(N,1)}.$$

The appearance of matching can be observed again. The accuracy of the quadratic estimates on this example, which cannot be handled by other estimates that were displayed on Figure 2.4, is even more striking. We investigate the inequality

$$\frac{|\lambda_1 - \mu_1^N|}{\mu_1^N} \le \frac{1}{\frac{|\lambda_3-\mu_2^N|}{\lambda_3+\mu_2^N}}\sin^2\Theta,$$

where we have taken $\theta = \pi$ and $\alpha = 0.2499$ in order that we formally satisfy the assumptions of Theorem 2.6.4. It is also possible to derive a similar result that would hold in

N	$\frac{\lvert\lambda_1 - \mu_1^N\rvert}{\mu_1^N}$	$\frac{1}{\frac{\lvert\lambda_3-\mu_2^N\rvert}{\lambda_3+\mu_2^N}} \sin^2\Theta$
40	5.624338644959740e-001	5.625623859061157e-001
50	4.513258011710761e-001	4.514080600129017e-001
60	3.635477753543245e-001	3.636049274679543e-001
70	2.956031150163320e-001	2.956451329092697e-001
80	2.431654227984744e-001	2.431975758884184e-001
90	2.024616236466049e-001	2.024869983332471e-001
100	1.705536747027966e-001	1.705742750519106e-001

Figure 2.8: The quadratic estimates for finite element approximations

the case $\theta = 9999\pi/10000$, cf. Theorem 3.3.8. We will not further pursue the problem of quadratic estimates for finite element spectral approximation in the presence of eigenvalue clusters.

2.8 Conclusion

A method to compute an estimate of the accuracy of the subspace approximation method is presented. It can also be used to obtain accurate lower estimates of the desired group of eigenvalues. The bounds have to be viewed as a combination of the Ritz value bound, which gives an existence of the matching of the Ritz values and eigenvalues, and the subspace bound, which describes the nature of that matching. The main features of our theory are:

- We allow any subspace $\mathsf{ran}(X) \subset \mathcal{D}(h)$ to be taken as a test space.
- Our bounds contain computable quantities only.
- Our estimate of a subspace error is a function of the computable quantity $\mathsf{sin}\Theta$ and a relative gap between the Ritz values and the "unwanted" component of the spectrum.
- The key quantity in our estimates is a "local resolvent" formula (2.7.5), which is valid for a general positive definite form h.

- Our theory also applies to operators which are not represented in a differential form.

The contributions in this chapter

This chapter forms the theoretical backbone of this thesis. Some of the proofs have been influenced by the results from the joint paper [37]. We will now highlight the main contributions independently:

- The perturbation approach to eigenvalue estimates from [21, 22, 26, 28] has been generalized to enable the analysis of the (practical) Rayleigh–Ritz method for unbounded operators in a weak form, see Section 2.3.

- The problem of the localization of the approximated eigenvalues is solved from the viewpoint of the perturbation theory, see Section 2.4.

- Both individual eigenvector and invariant subspace estimates have been derived. The new results generalize and extend the matrix results from [45] as well as operator results from [21], see Section 2.5.

- The weakly formulated Sylvester equation was introduced in a joint paper [37]. The results from [37] are slightly improved so that now we can consider weakly formulated Sylvester equations with more general operator coefficients, see Section 2.5.1.

- The higher order eigenvalue estimates for finite matrices from [28] and [48] have been extended to apply to unbounded nonnegative definite operators in a Hilbert space, see Section 2.6.

- On an example of a Sturm-Liouville eigenvalue problem with coupled boundaries we compare the new eigenvalue estimates with the Temple–Kato inequality (see [50]) and the eigenvector estimates with the Davis–Kahan $\sin\Theta$ theorem (see [21]). This demonstrates the sharpness of our estimates on a nontrivial example, see Section 2.7.

Chapter 3

Spectral asymptotics for large coupling limits

In this chapter we will present applications of the perturbation estimates to problems in Mathematical Physics. In general, these problems will be reduced to a study of the family of positive definite operators, formally written as,

$$\mathbf{H}_\eta = \mathbf{H}_b + \eta^2 \mathbf{H}_e.$$

We will also identify a class of regular perturbations $\eta^2 \mathbf{H}_e$, which allow sharp residual based analysis. Before we proceed with the presentation of our results, we will make precise the applications that motivated this study.

The applications of the abstract theory from Chapter 2 to the eigenvalue problems in the Theory of Elasticity were a joint work with Josip Tambača, Zagreb — see [36, 59] and Section 3.4.

3.1 Introduction

We will establish convergence estimates for the spectral problems for a class of positive definite forms

$$h_\eta(u, v) = h_b(u, v) + \eta^2 h_e(u, v), \quad \eta \text{ large} . \tag{3.1.1}$$

Here we take $\mathcal{H}$ to be the environment Hilbert space and $h_b + h_e$ to be a positive definite form in $\mathcal{H}$. Family (3.1.1) can always be considered as a perturbation of $h_b + h_e$ (after an obvious change of variable η) rather than as a perturbation of h_b. Therefore, we may assume h_b is positive definite and $\mathcal{Q}(h_b) \subset \mathcal{Q}(h_e)$ without affecting the level of generality. There in another additional assumption. In all that follows $\overline{\mathsf{ker}(h_e)}$ is a nontrivial subspace of $\mathcal{H}$.

From [52, 64] we conclude that the forms h_η converge (in the *strong resolvent sense*[1]) to the closed form

$$h_\infty(u,v) = h_b(u,v) \quad u,v \in \mathcal{Q}_\infty := \mathsf{ker}(h_e) \cap \mathcal{Q}(h_b)$$

and we also have, cf. [51, 52, 64],

$$\mathcal{Q}_\infty = \{f \in \bigcap_\eta \mathcal{Q}(h_\eta) : \sup_\eta h_\eta[f] < \infty\}.$$

Since we will be considering the whole family of operators $\mathbf{H}_\eta$, additional notation will be introduced to ease the understanding. By

$$\lambda_1^\eta \le \cdots \le \lambda_n^\eta \le \cdots < \lambda_e(\mathbf{H}_\eta)$$

we denote the increasingly ordered eigenvalues of the operator $\mathbf{H}_\eta$ and by

$$\lambda_1^\infty \le \cdots \le \lambda_n^\infty \le \cdots < \lambda_e(\mathbf{H}_\infty)$$

the eigenvalues of the operator $\mathbf{H}_\infty$. The corresponding spectral families will be $E_\eta(\cdot)$ and $E_\infty(\cdot)$.

In the subsequent theory we will consider the form h_∞ as a well known object. As a consequence, the estimates will be formulated in terms of objects defined by the form h_∞. When we say "convergence rate estimates" for the spectral problems (3.1.1), we mean estimates for the rate of convergence of

$$\frac{|\lambda_i^\eta - \lambda_i^\infty|}{\lambda_i^\infty} \to 0, \tag{3.1.2}$$

$$\|E_\eta(D) - E_\infty(D)\| \to 0. \tag{3.1.3}$$

The fact the convergence in (3.1.2) and (3.1.3) was established in [51, 64]. In particular, results of [64] give a complete theory of the behavior of eigenvalues and spectral families under the strong resolvent convergence. However, convergence rate estimates were not provided in any of the mentioned works.

The driving motivation in [36] was to provide the convergence rate estimates for eigenvalues (and spectral families) of 1D approximations in the Theory of Elasticity, and thus complement the convergence results from [51, 58]. The energy norm estimates for various 1D models from [58, 59] were the main tools needed to complete this task.

[1]The notion of the strong resolvent convergence for forms was introduced in [52, Simon]. This notion of the convergence of forms will be considered in the next section.

One of the contributions in this chapter will be an "algebraization" of the main theorem from [59] — i.e. we will introduce a notion of a residuum vector to the analysis from [59]. This will allow us to better assess the sharpness of the obtained estimates.

Assuming $\mathbf{H}_b$ and $\mathbf{H}_e$ are the operators defined by the forms h_b and h_e, we show (see Section 3.4) that if

$$\|(\mathbf{H}_e^{1/2}\mathbf{H}_b^{-1/2})^\dagger\| < \infty, \tag{3.1.4}$$

then

$$\|\mathbf{H}_\eta^{-1} - \mathbf{H}_\infty^\dagger\| = O(\frac{1}{\eta^2}). \tag{3.1.5}$$

Since $\mathbf{H}_\eta$ are uniformly positive definite this implies that $\mathbf{H}_\eta \to \mathbf{H}_\infty$ converges in the *norm resolvent sense*[2] .

Condition (3.1.4) is an additional *regularity requirement* on the perturbation $\eta^2 h_e$. It is equivalent to the *Babuška–Brezzi inf–sup condition*

$$\sup_{\mathbf{v}\in\mathcal{Q}(h_b)} \frac{|(q, \mathbf{H}_e^{1/2}\mathbf{v})|}{\|\mathbf{H}_b^{1/2}\mathbf{v}\|} \geq \frac{1}{k}\, \|P_{\mathcal{Q}_\infty} q\|, \qquad q \in \mathcal{H},$$

but more appropriate to our form approach. Furthermore, it simplifies the computation of the constants appearing in our convergence estimates. Also, note the following equivalence

$$\|(\mathbf{H}_e^{1/2}\mathbf{H}_b^{-1/2})^\dagger\| < \infty \Leftrightarrow \mathsf{ran}(\mathbf{H}_e^{1/2}\mathbf{H}_b^{-1/2}) \quad \text{is closed in} \quad \mathcal{H}.$$

The analysis of the family (3.1.1), when (3.1.4) is not satisfied, is inherently more difficult. Even in the case in which $\mathbf{H}_e = P$, where P is a projection on a closed subspace of $\mathcal{H}$, estimating the rate of convergence of $\mathbf{H}_\eta^{-1} \to \mathbf{H}_\infty^\dagger$ is a complex problem, cf. [54].

Assume now $\mathcal{O} \subset \mathbb{R}^n$ is bounded and connected set with sufficiently smooth boundary. If we consider $\mathcal{H} = L^2(\mathbb{R}^n)$, $\mathbf{H}_b = -\triangle$ and $\mathbf{H}_e = P_{L^2(\mathcal{O})}$, then advanced probabilistic techniques, heavily dependent on the properties of these particular $\mathbf{H}_b$ and $\mathbf{H}_e$, yield convergence rate estimates for $\mathbf{H}_\eta^{-1} \to \mathbf{H}_\infty^\dagger$, see [23].

Let now $\mathcal{A} \subset \mathcal{O}$ be a connected set with sufficiently smooth boundary which is compactly contained in $\mathcal{O}$. We take $\mathcal{H} = L^2(\mathcal{O})$, $\mathbf{H}_b = -\triangle$ with Dirichlet boundary conditions and $\mathbf{H}_e = P_{L^2(\mathcal{A})}$. In this special case boundary layer techniques yield convergence rate estimates for $\mathbf{H}_\eta^{-1} \to \mathbf{H}_\infty^\dagger$, see [17].

Further analysis of the behavior of the family (3.1.1), when the condition (3.1.4) is not satisfied, falls outside the scope of this thesis. Instead, we will concentrate on an abstract framework for obtaining spectral estimates in a situation when we are given a result like (3.1.5).

[2]For the definition of the norm resolvent convergence see [41, 52, 64]. In the case of uniformly positive definite families the norm resolvent convergence is equivalent to $\|\mathbf{H}_\eta^{-1} - \mathbf{H}_\infty^\dagger\| \to 0$.

Simple model problems

Problem 3.1. Consider the family of positive definite forms

$$h_\eta(u,v) = h_b(u,v) + \eta^2 h_e(u,v) = \int_0^\infty u'v'\ dx + \eta^2 \int_1^\infty uv\ dx, \quad u,v \in H_0^1(\mathbb{R}_+). \quad (3.1.6)$$

By $\mathbf{H}_\eta$ denote the positive definite operator defined by the form h_η in (3.1.6). We are interested in the eigenvalues of the operator $\mathbf{H}_\eta$ for large η. Here, $H_0^1(\mathbb{R}_+)$ denotes the subspace of the first order Sobolev space $H^1(\mathbb{R}_+)$ consisting of functions with zero trace on the boundary.

Problem 3.1 is also called the problem for the Schrödinger operator with a *square-well potential*, cf. [41, Example VII.3.3]. This is only an academic example of a typical problem from that class, see [23]. When considered on the finite domain, as it was done in [17], it also has an important application in engineering. We will further discuss the results from [17] at the end of the chapter.

Problem 3.2. Consider the family of positive definite forms

$$h_\eta(u,v) = h_b(u,v) + \eta^2 h_e(u,v) = \int_0^2 u'v'\ dx + \eta^2 \int_1^2 u'v'\ dx, \quad u,v \in H_0^1[0,2]. \quad (3.1.7)$$

By $\mathbf{H}_\eta$ denote the positive definite operator defined by the form h_η from (3.1.7). We are interested in the eigenvalues of the operator $\mathbf{H}_\eta$ for large η. Here, $H_0^1[0,2]$ denotes the first order Sobolev space with zero trace on the boundary.

Problem 3.2 is the eigenvalue problem for the vibration of a highly inhomogeneous string. Again, we are only considering an academic example where we can efficiently compute all information we need.

If we identify the functions from $H_0^1[0,\alpha]$, $\alpha > 0$, with their extension by zero to the whole of $\mathbb{R}_+$, then we can write

$$H_0^1[0,\alpha] \subset H_0^1[0,\beta] \subset H_0^1(\mathbb{R}_+), \qquad 0 < \alpha < \beta. \quad (3.1.8)$$

Let $\chi_{[0,1]}$ be the characteristic function of the interval $[0,1]$ and let $\chi_{[0,1]^c} = 1 - \chi_{[0,1]}$. Keeping (3.1.8) in mind, we conclude that

$$\mathbf{H}_\eta = -\partial_{xx} + \eta^2 \chi_{[0,1]^c}, \qquad \mathcal{D}(\mathbf{H}_\eta) = H^2(\mathbb{R}_+) \cap H_0^1(\mathbb{R}_+)$$

is the operator in Problem 3.1 and in Problem 3.2 we are considering

$$\mathbf{H}_\eta = -\partial_x (1 + \eta^2 \chi_{[0,1]^c}) \partial_x, \qquad \mathcal{D}(\mathbf{H}_\eta) = H^2[0,2] \cap H_0^1[0,2].$$

It is known ([38, 47, 64]) that the forms h_η converge—in both cases—to the form

$$h_\infty(u,v) = \int_0^1 u'v'\ dx, \quad u,v \in H_0^1\,[0,1]$$

in the norm resolvent sense. Apparently the first result of this type was obtained in [47], where the norm-convergence of the resolvent was established as a consequence of the pointwise positivity of the Green functions. The results for the convergence of more general families h_η were obtained in [52, 64].

The first eigenpair of the operator $\mathbf{H}_\infty$ is $(\pi^2, \sqrt{2}\mathsf{sin}(\pi x))$. The function

$$u_1(x) = \begin{cases} \sqrt{2}\sin(\pi x), & 0 \le x \le 1 \\ 0, & 1 \le x \end{cases} \tag{3.1.9}$$

is in $H_0^1(\mathbb{R}_+)$ and also in $H_0^1[0,2]$. Therefore, it can be used as a test function for an approximation of the lowest eigenvalue of both operators $\mathbf{H}_\eta$ (for large η). In both cases we compute the Ritz value

$$h_\eta(u_1,u_1) = \pi^2.$$

According to (2.7.5) we obtain

$$\sin^2\Theta_\eta := \sin\Theta(\mathbf{H}_\eta^{-1/2}u_1, \mathbf{H}_\eta^{1/2}u_1) = \frac{(u_1,\mathbf{H}_\eta^{-1}u_1) - (u_1,\mathbf{H}_\infty^\dagger u_1)}{(u_1,\mathbf{H}_\eta^{-1}u_1)}.$$

When $\sin\Theta_\eta < 1$, Theorem 2.3.17 guarantees existence of an eigenvalue $\lambda_{i_1}^\eta$ such that

$$\frac{|\lambda_{i_1}^\eta - \lambda_1^\infty|}{\lambda_1^\infty} \le \sin\Theta_\eta\ .$$

A direct computation shows that

$$\begin{aligned}
&(u_1,\mathbf{H}_\eta^{-1}u_1 - \mathbf{H}_\infty^\dagger u_1) \\
&\quad = \int_0^1 \left[\int_0^x 2\left(\frac{y\ (1+\eta\ (1-x))}{1+\eta} - y\ (1-x)\right)\sin(\pi\,y)\sin(\pi\,x)\ dy\right. \\
&\quad \left. + \int_x^1 2\left(\frac{x\ \ (1+\eta\ (1-y))}{1+\eta} - x\ \ (1-y)\right)\ \sin(\pi\,y)\sin(\pi\,x)\ dy\right]\ dx \\
&\quad = \frac{2}{(1+\eta)\pi^2} = O(\eta^{-1})
\end{aligned} \tag{3.1.10}$$

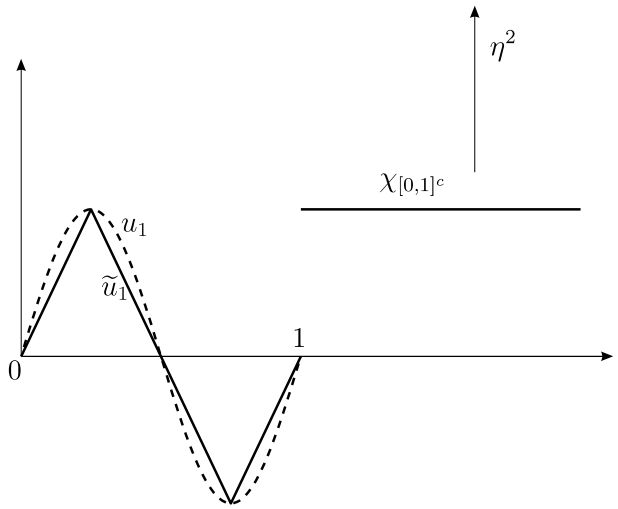

Figure 3.1: Various test functions for $\mathbf{H}_\infty$ and $\mathbf{H}_\eta$, η large.

in the case of (3.1.6) and in the case of (3.1.7) we compute

$$\begin{aligned}
&(u_1,\mathbf{H}_\eta^{-1}u_1 - \mathbf{H}_\infty^\dagger u_1)\\
&\quad= \int_0^1 \left[\int_0^x 2\left(\frac{y\ (1+(1+\eta^2)\ (1-x))}{2+\eta^2} - y\ (1-x)\right)\sin(\pi\, y)\sin(\pi\, x)\ dy\right.\\
&\quad+ \left.\int_x^1 2\left(\frac{x\ \ (1+(1+\eta^2)\ (1-y))}{2+\eta^2} - x\ \ (1-y)\right)\ \sin(\pi\, y)\sin(\pi\, x)\ dy\right]\ dx\\
&\quad= \frac{2}{(2+\eta^2)\pi^2} = O(\eta^{-2}).
\end{aligned} \tag{3.1.11}$$

This establishes that in both cases $\mathsf{sin}\Theta_\eta \to 0$, so Theorem 2.3.17 will be applicable for $\eta \geq 1$ such that

$$\frac{(u_1,\mathbf{H}_\eta^{-1}u_1) - (u_1,\mathbf{H}_\infty^\dagger u_1)}{(u_1,\mathbf{H}_\eta^{-1}u_1)} = \frac{2}{1+\eta} < 1$$

in Problem 3.1 and for $\eta \geq 1$ such that

$$\frac{(u_1,\mathbf{H}_\eta^{-1}u_1) - (u_1,\mathbf{H}_\infty^\dagger u_1)}{(u_1,\mathbf{H}_\eta^{-1}u_1)} = \frac{2}{2+\eta^2} < 1$$

in Problem 3.2 .

The difference between Problems 3.1 and 3.2 is in the nature of the behavior at infinity of the function

$$\eta \mapsto \frac{(u_1,\mathbf{H}_\eta^{-1}u_1 - \mathbf{H}_\infty^\dagger u_1)}{(u_1,\mathbf{H}_\eta^{-1}u_1)}. \tag{3.1.12}$$

In the second case the perturbation $\mathbf{H}_e$ satisfies (3.1.4) and consequently we get a $O(\eta^{-2})$ estimate (3.1.12) at infinity. On the other hand, the family $\mathbf{H}_\eta$, from Problem 3.1, does not satisfy (3.1.4) and we have a lower order convergence of $\mathbf{H}_\eta^{-1} \to \mathbf{H}_\infty^\dagger$, as $\eta \to \infty$.

The case study that was just performed can be described as leading to a "pseudo spectral" method. We have used the operator

$$(\mathbf{H}_\infty^{1/2} u, \mathbf{H}_\infty^{1/2} v) = h_\infty(u,v) = \int_0^1 u'v' \, dx, \quad u, v \in H_0^1[0,1],$$

defined on the finite interval, to analyze the operator $\mathbf{H}_\eta$ that is defined on the unbounded interval $\mathbb{R}_+ = [0, \infty\rangle$. The eigenvalue problem for the operator $\mathbf{H}_\infty$ was completely solvable, so we have used the eigenfunctions of the operator $\mathbf{H}_\infty$ to define a test space for the operator $\mathbf{H}_\eta$. Analogously, we could have used other test functions from $H_0^1[0,1]$ to analyze the operator $\mathbf{H}_\eta$. For instance, assume we have used the linear finite elements to compute an approximation $\widetilde{u}_1$ of the function u_1, see Figure 3.1. Theorem 2.3.17 can be invoked if we find a way to estimate $\sin\Theta(\mathbf{H}_\eta^{-1/2}\widetilde{u}_1, \mathbf{H}_\eta^{1/2}\widetilde{u}_1)$, cf. Section 4.2.4.

To establish eigenvector estimates (and higher order eigenvalue estimates) we will need to do a bit more work. Establishing such estimates will be the main contribution of this chapter. Among other things, this will give us estimates of the error between the used Ritz vectors, that have bounded supports, and the approximated eigenvectors, that have unbounded supports. This is to say that within this framework we can analyze a use of finite elements from a bounded domain to compute spectral approximations of an operator $\mathbf{H}_\eta$, that lives on an unbounded domain. Investigating the efficiency of a numerical method that is build on these considerations remains a task for the future[3]. In general, applicability of a numerical method for the operator $\mathbf{H}_\eta$ (for large η), that is based on good properties of the operator $\mathbf{H}_\infty$, will essentially depend on the rate of the convergence of $\mathbf{H}_\eta^{-1} \to \mathbf{H}_\infty^\dagger$. To this end we identify a regular class of positive definite forms (3.1.1) where we can guarantee a higher order convergence of $\mathbf{H}_\eta^{-1} \to \mathbf{H}_\infty^\dagger$. This is another contribution in this chapter.

3.2 The convergence of nondensely defined positive definite forms

In order to be able to handle the problems of the type (3.1.1), we shall need to work with operators that are not densely defined, cf. Problems 3.1 and 3.2. We use the notion

[3]For a discussions of some finite element approximation procedures see Chapter 4.

of the pseudo inverse of the operator $\mathbf{H}$ that is assumed to be self adjoint in the space $\overline{\mathcal{D}(\mathbf{H})} \subset \mathcal{H}$. A definition from [64] will be used. The *pseudo inverse* of the operator $\mathbf{H}$ is the self adjoint operator $\mathbf{H}^\dagger$ defined by

$$\begin{aligned} \mathcal{D}(\mathbf{H}^\dagger) &= \mathsf{ran}(\mathbf{H}) \oplus \mathcal{D}(\mathbf{H})^\perp, \\ \mathbf{H}^\dagger(u+v) &= \mathbf{H}^{-1}u, \qquad u \in \mathsf{ran}(\mathbf{H}),\ v \in \mathcal{D}(\mathbf{H})^\perp. \end{aligned}$$

It follows that $\mathbf{H}^\dagger = \mathbf{H}^{-1}$ in $\overline{\mathsf{ran}(\mathbf{H})}$. Note that we did not assume $\mathbf{H}^\dagger$ to be bounded. The operator $\mathbf{H}^\dagger$ will be bounded if an only if $\mathsf{ran}(\mathbf{H})$ is closed in $\mathcal{H}$. The operator $\mathbf{H}^\dagger$ could have also been defined by the spectral calculus, since

$$\mathbf{H}^\dagger = f(\mathbf{H}), \qquad f(\lambda) = \begin{cases} 0, & \lambda = 0, \\ \frac{1}{\lambda}, & \lambda \neq 0. \end{cases}$$

In [64] Weidmann has given a short survey of the properties of the pseudo inverse of the nondensely defined operator $\mathbf{H}$. In particular, let $\mathbf{H}_1$ and $\mathbf{H}_2$ be two nonnegative operators in $\overline{\mathcal{D}(\mathbf{H}_1)}$ and $\overline{\mathcal{D}(\mathbf{H}_2)}$ respectively, then

$$\|\mathbf{H}_1^{1/2}u\| \leq \|\mathbf{H}_2^{1/2}u\| \Leftrightarrow \|\mathbf{H}_2^{1/2\dagger}u\| \leq \|\mathbf{H}_1^{1/2\dagger}u\|. \tag{3.2.1}$$

Analogously, let h_1 and h_2 be two closed, not necessarily densely defined, positive definite forms and let $\mathbf{H}_1$ and $\mathbf{H}_2$ be the self adjoint operators defined by h_1 and h_2 in $\overline{\mathcal{Q}(h_1)}$ and $\overline{\mathcal{Q}(h_2)}$. We say $h_1 \leq h_2$ when $\mathcal{Q}(h_2) \subset \mathcal{Q}(h_1)$ and

$$h_1[u] = \|\mathbf{H}_1^{1/2}u\|^2 \leq h_2[u] = \|\mathbf{H}_2^{1/2}u\|^2, \qquad u \in \mathcal{Q}(h_2). \tag{3.2.2}$$

Equivalently, we write $\mathbf{H}_1 \leq \mathbf{H}_2$ when $h_1 \leq h_2$. Now, we can write the fact (3.2.1) as

$$\mathbf{H}_1 \leq \mathbf{H}_2 \Longleftrightarrow \mathbf{H}_2^\dagger \leq \mathbf{H}_1^\dagger. \tag{3.2.3}$$

An important feature of the family (3.1.1) is that the limiting form h_∞ is closed and

$$\begin{aligned} h_\infty(u,v) &= \lim_\eta h_\eta(u,v), \qquad u,v \in \mathcal{Q}_\infty, \\ \mathbf{H}_\infty^\dagger &= \text{s-lim}_\eta\ \mathbf{H}_\eta^\dagger\ . \end{aligned}$$

In fact, according to [52] the form h_∞, obtained as the limit of the monotone increasing family of positive definite forms, is always closed and defines a self adjoint operator in $\overline{\mathcal{Q}(h_\infty)}$, cf. [64].

The general framework for the description of families of converging positive definite forms will be the following theorem from [64, Weidmann].

Theorem 3.2.1. *Let s_n, h_n, u_n and h_∞ be closed symmetric forms in $\mathcal{H}$ such that they are all uniformly positive definite.*

1. *If $s_n \geq s_{n+1} \geq h_\infty$ and*

$$\mathcal{Q}(h_\infty) = \overline{\bigcup_{n\in\mathbb{N}} \mathcal{Q}(s_n)}^{\,h_\infty},$$
$$h_\infty(u,v) = \lim_{n\to\infty} s_n(u,v), \qquad u,v \in \bigcup_{n\in\mathbb{N}} \mathcal{Q}(s_n)$$

   then $\mathbf{H}_\infty^\dagger = s\text{–}\lim_n \mathbf{S}_n^\dagger$.

2. *If $u_n \leq u_{n+1} \leq h_\infty$ and*

$$\mathcal{Q}(h_\infty) = \left\{ f \in \bigcap_{n\in\mathbb{N}} \mathcal{Q}(h_n) \;:\; \sup u_n[f] < \infty \right\},$$
$$h_\infty(u,v) = \lim_{n\to\infty} u_n(u,v), \qquad u,v \in \mathcal{Q}(t)$$

   then $\mathbf{H}_\infty^\dagger = s\text{–}\lim_n \mathbf{U}_n^\dagger$.

3. *If u_n and s_n are as before and $u_n \leq h_n \leq s_n$ also holds, then*

$$h_\infty(u,v) = \lim_{n\to\infty} h_n(u,v), \qquad u,v \in \mathcal{Q}(t),$$
$$\mathbf{H}_\infty^\dagger = \mathrm{s} - \lim_{\eta\to\infty} \mathbf{H}_n^\dagger.$$

In [63] it has been proved that for any family of sesquilinear forms h_η which satisfies the conditions of Theorem 3.2.1, we have

$$\|E_\eta(D) - E_\infty(D)\| \to 0, \qquad D \notin \sigma(\mathbf{H}_\infty) \text{ and } D < \lambda_e(\mathbf{S}) \tag{3.2.4}$$

and the eigenvalues of the operators $\mathbf{H}_\eta$ (assuming $\mathbf{S} \leq \mathbf{H}_\infty$) converge to the eigenvalues of the operator $\mathbf{H}_\infty$, together with their multiplicities. To be more precise, we provide the following theorem.

Theorem 3.2.2. *Let h_η be a sequence of positive definite forms that satisfies any of the assumptions of Theorem 3.2.1. Let there also be the positive definite form s such that $h_n \geq s$ and $\lambda_e(\mathbf{S}) > 0$. Then*

$$\|E_\eta(D) - E_\infty(D)\| \to 0, \qquad D < \lambda_e(\mathbf{S}), D \notin \sigma(\mathbf{H}_\infty). \tag{3.2.5}$$

The main part of the proof is contained in the following lemma that is implicit in [63].

Lemma 3.2.3. *Let h_η, h_∞ and s be closed, positive definite forms. Assume that the operator $\mathbf{S}$, defined by the form s, has all of its essential spectrum in $[\lambda_e(\mathbf{S}), \infty\rangle$. Assume $h_\eta \geq s$ for all η and assume that $\mathbf{H}_\infty^\dagger = \text{s-lim}_\eta\ \mathbf{H}_\eta^\dagger$, then*

$$\mathsf{dim}\ E_\eta(D) \leq \mathsf{dim}\ E_\infty(D), \qquad D < \lambda_e(\mathbf{S}),\ D \notin \sigma(\mathbf{H}_\infty). \tag{3.2.6}$$

In the case in which h_η is a monotone increasing family of forms we do not need to suppose the existence of the form s. The statement (3.2.6) is now

$$\mathsf{dim}\ E_\eta(D) \leq \mathsf{dim}\ E_\infty(D), \qquad D < \lambda_e(\mathbf{H}_\infty), D \notin \sigma(\mathbf{H}_\infty).$$

To identify the problems that have to be tackled we will outline the proof of Theorem 3.2.2 from [63]:

An outline of the Weidmann's proof of Theorem 3.2.2

The assumption $h_\eta \geq s$ and (3.2.3) imply that $\mathbf{H}_\eta^\dagger$ are bounded and $\lambda_e(\mathbf{H}_\infty^\dagger) = 1/\lambda_e(\mathbf{H}_\infty) \leq 1/\lambda_e(\mathbf{S})$. Since $\mathbf{H}_\infty^\dagger = \text{s-lim}_\eta\ \mathbf{H}_\eta^\dagger$ we have, see [29, 52],

$$E_\infty(D) = \text{s} - \lim_{\eta\to\infty} E_\eta(D), \tag{3.2.7}$$

for $D < \lambda_e(\mathbf{S})$ and D not an eigenvalue of $\mathbf{H}_\infty$. There is a theorem of Kato (see [41]) which states that (3.2.7), together with

$$\mathsf{dim}\ E_\eta(D) \leq \mathsf{dim}\ E_\infty(D), \tag{3.2.8}$$

implies

$$\|E_\eta(D) - E_\infty(D)\| \to 0.$$

Since Lemma 3.2.3 implies (3.2.8) the statement (3.2.5) is proved. □

We are primarily interested in the perturbation families (3.1.1). However, everything proved will, with minor modifications, hold for any family of positive definite forms that satisfies the assumptions of Theorem 3.2.1. We will state the results in the most general form when that does not induce additional notational overhead. It is our aim to prove Theorem 3.2.2 by a use of theorems from Section 2.5.2 and Lemma 3.2.3 alone. The main gain will be the rigorous estimate of the rate of convergence

$$\|E_\eta(D) - E_\infty(D)\| \to 0.$$

Remark 3.2.4. Simon has developed a notion of the pseudo resolvent, parallel to the notion of the pseudo inverse, for a given closed, nondensely defined form h, see [52]. Based on this resolvent there is a functional calculus for the real functions which vanish at infinity. Due to the convergence result of [63], we could use the integral representation of the spectral projection to compute the estimates of the rate of convergence of $\|E_\eta(D) - E_\infty(D)\| \to 0$. However, that would require extensive information on the resolvents of h_η and h_∞. We show that it is sufficient to study the pseudo inverse only on the test space from $\mathcal{Q}_\infty$, which we are given as input (see Section 2.7).

Since $\mathbf{H}_\eta$, from Theorem 3.2.1, is a family of uniformly positive definite operators the conclusion

$$\mathbf{H}_\infty^\dagger = \mathrm{s} - \lim_{\eta\to\infty} \mathbf{H}_\eta^\dagger \tag{3.2.9}$$

is equivalent to the *strong resolvent convergence* from Remark 3.2.4. All of the families that will be considered in the rest of this thesis have this property. Therefore, when we want to prove that uniformly positive definite family converges in the strong resolvent sense, we will be proving (3.2.9).

3.3 Convergence rate estimates for the perturbation family $h_b + \eta^2 h_e$

The perturbation argument that stood behind the reasoning in Chapter 2 will be particularly suitable to analyze the family (3.1.1). We will review the main line of argument to ease the transition from the Ritz value estimates to the consideration of spectral asymptotics for large coupling limits.

Let h_η be a sequence of positive definite forms that satisfies the assumptions of Theorem 3.2.1. For the given n-dimensional space $\mathsf{ran}(P) \subset \mathcal{Q}_\infty := \mathcal{Q}(h_\infty)$, $P = XX^*$, we construct the forms

$$h'_\eta(u,v) = h_\eta(Pu, Pv) + h_\eta(P_\perp u, P_\perp v), \quad u, v \in \mathcal{Q} \tag{3.3.1}$$

and

$$\delta h_\eta = h_\eta - h'_\eta. \tag{3.3.2}$$

Let $\mathbf{H}'_\eta$ be the operator defined by the positive definite form h'_η, then

$$\sin\Theta_\eta(X) := \max_{u,v\in\mathcal{Q},u,v\neq 0} \frac{|\delta h_\eta(u,v)|}{\sqrt{h'_\eta[u]h'_\eta[v]}} = \left[\max_{u\in\mathsf{ran}(X)} \frac{(u, \mathbf{H}_\eta^{-1}u - \mathbf{H}_\infty^\dagger u)}{(u, \mathbf{H}_\eta^{-1}u)}\right]^{1/2} \tag{3.3.3}$$

In theorems that follow we will use a matrix formulation of (3.3.3)

$$\sin^2\Theta_\eta(X) = \left[\max_{x\neq 0, x\in\mathbb{C}^n} \frac{|x^*(\Omega_\eta - \Xi_\eta^{-1})x|}{x^*\Omega_\eta x}\right]^{1/2},$$

where

$$\Xi_\eta^{-1} = X^*\mathbf{H}_\eta'^{-1}X, \qquad \Omega_\eta = X^*\mathbf{H}_\eta^{-1}X.$$

Relations (3.3.1), (3.3.2) and (3.3.3) imply

$$(1-\sin\Theta_\eta(X))h_\eta' \leq h_\eta \leq (1+\sin\Theta_\eta(X))h_\eta'. \tag{3.3.4}$$

The eigenvalues of the matrices Ξ_η are among the eigenvalues of the operators $\mathbf{H}_\eta'$, counting them according to their multiplicities.

To ease the presentation of the results we introduce additional notation and convergence rate measures. The Ritz values, the eigenvalues of the matrix Ξ_η, will be

$$\mu_1^\eta \leq \cdots \leq \mu_n^\eta.$$

Theorem 3.2.1 implies

$$\Omega_\eta \to \Omega_\infty, \qquad \Xi_\eta \to \Xi_\infty, \tag{3.3.5}$$

so we can use

$$\alpha_\eta = \max_{x\in\mathbb{C}^n} \frac{|x^*(\Xi_\eta^{-1} - \Xi_\infty^{-1})x|}{x^*\Xi_\infty^{-1}x}, \qquad \beta_\eta = \max_{x\in\mathbb{C}^n} \frac{|x^*(\Omega_\eta - \Omega_\infty)x|}{x^*\Omega_\infty x}$$

to measure the speed of the convergence in (3.3.5).

It was assumed that h_∞ be a well known object, so α_η and β_η, unlike $\mathsf{sin}\Theta_\eta$, measure the speed of convergence relative to the known objects Ω_∞ and Ξ_∞^{-1}.

Assume now that $h_\eta = h_b + \eta^2 h_e$ and h_b, h_e are as in (3.1.1), then for every η

$$\Xi_\eta = \Xi_\infty.$$

The following lemma and Theorem 2.5.5 will enable us to bypass an invocation of the theorem of Kato (statements (3.2.7) and (3.2.8)) in the proof of Theorem 3.2.2. Thus, we will obtain a (new) proof of the convergence theorem directly from the monotonicity principle. As a byproduct we will get the convergence rate estimates, too.

Lemma 3.3.1. *Let $h_\eta(u,v) = \int \lambda \; d\,(E_\eta(\lambda)u, v)$ be the monotone increasing sequence [4] of positive definite forms such that*

$$\lim_{\eta\to\infty} h_\eta(u,v) = h_\infty(u,v) = \int \lambda \; d\,(E_\infty(\lambda)u, v), \qquad u, v \in \mathcal{Q}_\infty,$$

then

$$\mathsf{dim} E_\eta(D) \geq \mathsf{dim} E_\infty(D)$$

for $D < \lambda(\mathbf{H}_\infty)$ and D not an eigenvalue of $\mathbf{H}_\infty$.

Proof. Without reducing the level of generality, we can assume that there exist $n \in \mathbb{N}$ and $D \in \mathbb{R}$ such that

$$\lambda_n^\infty < D < \lambda_{n+1}^\infty. \tag{3.3.6}$$

Now, take an isometry[5] X such that $\mathsf{ran}(X) = \mathsf{ran}(E_\infty(D))$ and define the matrices

$$\Xi_\eta = (\mathbf{H}_\eta^{1/2} X)^* \mathbf{H}_\eta^{1/2} X, \qquad \Omega_\eta = X^* \mathbf{H}_\eta^{-1} X.$$

We now set

$$\alpha_\eta = \max_{x\in\mathbb{C}^n} \frac{|x^*(\Xi_\eta^{-1} - \Xi_\infty^{-1})x|}{x^*\Xi_\infty^{-1}x}, \qquad \beta_\eta = \max_{x\in\mathbb{C}^n} \frac{|x^*(\Omega_\eta - \Omega_\infty)x|}{x^*\Omega_\infty x} \; .$$

We can chose the isometry X, $\mathsf{ran}(X) = \mathsf{ran}(E_\infty(D))$, such that the matrix $\Omega_\infty = \Xi_\infty^{-1}$ has a diagonal form with $1/\lambda_1^\infty \geq \cdots \geq 1/\lambda_n^\infty$ on the diagonal. Also, $\sin\Theta_\infty(X) = 0$ must hold since $\mathsf{ran}(X)$ is an invariant subspace of $\mathbf{H}_\infty$. Theorem 3.2.1 implies that $\alpha_\eta \to 0$ and $\beta_\eta \to 0$, so we may assume $\alpha_\eta < 1$, $\beta_\eta < 1$. Based on [61, Theorem 2.1] we obtain

$$\sin\Theta_\eta(X) \leq \sqrt{\frac{\alpha_\eta + \beta_\eta}{1 + \beta_\eta}} \to 0 \; .$$

The relations (3.3.4) and (3.3.5), together with Theorem 2.3.17 complete the proof. □

Several spinoffs are a consequence of the method of the proof of this lemma.

[4] In fact any sequence from Theorem 3.2.1, will be just as good. In such a case, in the formulation of the theorems, one needs an additional assumptions of Lemma 3.2.3. We leave out the details.

[5] To a certain extent one can say that we shall also investigate a use of the constructions (3.3.1) and (3.3.2) when measuring the "spectral error" for approximations from infinite dimensional subspaces $\mathsf{ran}(X)$, cf. Section 3.4. However, when the limit operator $\mathbf{H}_\infty$, $\mathcal{D}(H_\infty) \subset \mathcal{Q}_\infty$, possesses only discrete eigenvalues we may and shall, without reducing the level of generality, only consider finite dimensional $\mathsf{ran}(X)$. It is just an idea to have in mind when reading the rest of this chapter. We shall not go into a messy business of rigorously defining an infinite dimensional Rayleigh quotient.

Corollary 3.3.2. *Let h_η, h_∞ be the positive definite forms that satisfy the assumptions of Lemma 3.2.3. Take the n-dimensional subspace* $\mathsf{ran}(X) \subset \bigcap_{n\in\mathbb{N}} \mathcal{Q}(h_n)$*, then*

$$\sin\Theta_\eta(X) \leq \sqrt{\frac{\alpha_\eta + \beta_\eta}{1+\beta_\eta}},$$

where

$$\alpha_\eta = \max_{x\in\mathbb{C}^n} \frac{|x^*(\Xi_\eta^{-1} - \Xi_\infty^{-1})x|}{x^*\Xi_\infty^{-1}x}, \qquad \beta_\eta = \max_{x\in\mathbb{C}^n} \frac{|x^*(\Omega_\eta - \Omega_\infty)x|}{x^*\Omega_\infty x} \ .$$

The proof is obvious if we note that Theorem 3.2.1 implies that the space $\overline{\bigcap_{n\in\mathbb{N}} \mathcal{Q}(h_n)}^{\ \mathcal{H}}$ will not be a trivial subspace of $\mathcal{H}$. A more special result is possible in the case of the perturbation family (3.1.1). Most importantly, in this case

$$\Xi_\infty = \Xi_\eta$$

and the eigenvalues of Ξ_∞ are the Ritz values of $\mathbf{H}_\eta$.

Corollary 3.3.3. *Let h_b be a positive definite form and let h_e be such that $\mathcal{Q}(h_b) \subset \mathcal{Q}(h_e)$. By*

$$h_\eta(u,v) = h_b(u,v) + \eta^2 h_e(u,v), \quad u,v \in \mathcal{Q}(h_b)$$

we define the monotone increasing family of positive definite forms. Take the subspace

$$\mathsf{ran}(X) \subset \mathcal{Q}(h_\infty) = \mathcal{Q}(h_b) \cap \ker(h_e),$$

then $\Xi_\eta = \Xi_\infty$ and

$$\sin\Theta_\eta(X) \leq \sqrt{\frac{\beta_\eta}{1+\beta_\eta}}$$

for

$$\beta_\eta = \max_{x\in\mathbb{C}^n} \frac{|x^*(\Omega_\eta - \Omega_\infty)x|}{x^*\Omega_\infty x} \ .$$

Since $\Omega_\infty = \Xi_\infty^{-1}$, β_η is suspiciously similar to $\mathsf{sin}\Theta_\eta$. However, we reiterate that in β_η, unlike in $\mathsf{sin}\Theta_\eta$, we are measuring the convergence relative to the known quantity Ω_∞ (rather than relative to Ω_η).

We now state and prove an extended version of Theorem 3.2.2.

Theorem 3.3.4. *Let $h_\eta(u,v) = \int \lambda \ d\ (E_\eta(\lambda)u, v)$ be the monotone increasing sequence of positive definite forms such that*

$$\lim_{\eta\to\infty} h_\eta(u,v) = h_\infty(u,v) = \int \lambda \ d\ (E_\infty(\lambda)u, v), \qquad u,v \in \mathcal{Q}_\infty.$$

Take $D \in \mathbb{R}$ such that $\lambda_n^\infty < D < \lambda_{n+1}^\infty$, then

$$\frac{|\lambda_j^\eta - \mu_j^\eta|}{\mu_j^\eta} \leq \sin\Theta_\eta(X), \quad j = 1, \ldots, n, \tag{3.3.7}$$

$$\|E_\eta(D) - E_\infty(D)\| \leq \frac{\sqrt{D\mu_n^\eta}}{|D - \mu_n^\eta|} \frac{\sin\Theta_\eta(X)}{\sqrt{1 - \sin\Theta_\eta(X)}} \tag{3.3.8}$$

for η large enough.

PROOF. (*The proof of an extended version of Theorem 3.2.2.*) Take $D \in \mathbb{R}$ such that

$$\lambda_n^\infty < D < \lambda_{n+1}^\infty$$

and set $\mathsf{ran}(X) = \mathsf{ran}(E_\infty(D))$. Lemmata 3.2.3 and 3.3.1 imply that there exists η_0 such that

$$\mathsf{dim}\ E_\eta(D) = \mathsf{dim}\ E_\infty(D), \qquad \eta > \eta_0.$$

An alternative way to say this is to state that

$$\lambda_n^\eta < D < \lambda_{n+1}^\eta, \qquad \eta > \eta_0. \tag{3.3.9}$$

Since $\Xi_\eta \to \Xi_\infty$, we have

$$\frac{D - \mu_n^\eta}{\sqrt{\mu_n^\eta D}} \to \frac{D - \lambda_n^\infty}{\sqrt{D\lambda_n^\infty}}\ .$$

Corollary 3.3.2 implies $\mathsf{sin}\Theta_\eta(X) \to 0$, so we can find $\widehat{\eta}_0 \geq \eta_0$ such that

$$\frac{\sin\Theta_\eta(X)}{\sqrt{1 - \sin\Theta_\eta(X)}} < \frac{D - \mu_n^\eta}{\sqrt{\mu_n^\eta D}}\ , \qquad \eta > \widehat{\eta}_0.$$

Theorem 2.5.5 yields

$$\|E_\eta(D) - E_\infty(D)\| \leq \frac{\sqrt{D\mu_n^\eta}}{D - \mu_n^\eta} \frac{\sin\Theta_\eta(X)}{\sqrt{1 - \sin\Theta_\eta(X)}}\ .$$

The proof of (3.3.7) follows analogously. □

Theorem 3.3.4 is an extension of Theorem 3.2.2 since, as a direct consequence of (3.3.7) and (3.3.8), we have established

$$\|E_\eta(D) - E_\infty(D)\| \to 0.$$

The estimate for the eigenvalue rate of convergence from Theorem 3.3.4 will be sharpened at the end of this section. Theorem 3.3.4 takes on a special form when applied to the family of forms (3.1.1).

Corollary 3.3.5. *Let $h_\eta = h_b + \eta^2 h_e$ be the family of forms as in Corollary 3.3.3. Take $\lambda_n^\infty < D < \lambda_{n+1}^\infty$ the subspace $\mathsf{ran}(X) = \mathsf{ran}(E_\infty(D))$, then*

$$\frac{|\lambda_j^\eta - \lambda_j^\infty|}{\lambda_j^\infty} \leq \sin\Theta_\eta(X), \quad j = 1, \dots, n, \tag{3.3.10}$$

$$\|E_\eta(D) - E_\infty(D)\| \leq \frac{\sqrt{D\lambda_n^\infty}}{D - \lambda_n^\infty} \frac{\sin\Theta_\eta(X)}{\sqrt{1 - \sin\Theta_\eta(X)}} \tag{3.3.11}$$

for η large enough.

In practice, we suggest a use of β_η in the estimates (3.3.10) and (3.3.11). Such inequalities follow directly from Corollary 3.3.3.

Remark 3.3.6. In the proof of Theorem 3.3.4, and in particular in the proof of Lemma 3.3.1, it is apparent that the right convergence requirement for our technique would have been the weak convergence of resolvents at zero. For the sequence of uniformly positive definite forms this would imply the weak convergence of resolvents. However, it is well known, cf. [50], that the weak convergence of resolvents is equivalent to the strong convergence of resolvents.

3.3.1 The quadratic convergence of eigenvalues

Theorem 2.6.4 has established that the assumption like (3.3.9) implies the higher order approximation estimates. Establishing this result for the families (3.1.1) will involve some technical overhead. In what follows it is important that h_η has the structure as in (3.1.1).

We assume that the operator $\mathbf{H}_\infty$ is a well known object. Let λ_m^∞ be an eigenvalue of multiplicity $n \in \mathbb{N}$. Take D_- and D_+ such that

$$\lambda_{m-1}^\infty < D_- < \lambda_m^\infty < D_+ < \lambda_{m+n}^\infty. \tag{3.3.12}$$

To be able to apply Theorem 2.6.4 one should establish that there exists $\eta_0 > 0$ such that

$$\lambda_1^\eta \leq \cdots \leq \lambda_{m-1}^\eta < D_- < \lambda_m^\eta = \cdots = \lambda_{m+n-1}^\eta < D_+ < \lambda_{m+n}^\eta \leq \lambda_{n+n+1}^\eta \leq \cdots, \tag{3.3.13}$$

for $\eta > \eta_0$. However, if $n > 1$ it is not plausible to expect that (3.3.13) will hold in general. Instead, we will get a tight cluster of n eigenvalues (counting the eigenvalues according to their multiplicity) that converge to λ_m^∞. In the case of multiple eigenvalue λ_m^∞ the quadratic convergence of the cluster of eigenvalues will be proved in a generalized sense. To be more precise, we will prove that the mean value of the cluster of n eigenvalues converges quadratically to λ_m^∞. In this section we assume (3.3.12) and set $\mathsf{sin}\Theta_\eta := \mathsf{sin}\Theta_\eta(E_\infty[D_-, D_+])$. As a first step, we will establish the result in the case $n = 1$.

Theorem 3.3.7. *Let the eigenvalues of the operator $\mathbf{H}_\infty$ be so ordered that*

$$\lambda^\infty_{m-1} < D_- < \lambda^\infty_m < D_+ < \lambda^\infty_{m+1}.$$

Define $\gamma_s(D_-, \lambda^\infty_m, D_+) = \min\left\{\frac{\lambda^\infty_m - D_-}{\lambda^\infty_m + D_-}, \frac{D_+ - \lambda^\infty_m}{D_+ + \lambda^\infty_m}\right\}$, *then*

$$\frac{|\lambda^\infty_m - \lambda^\eta_m|}{\lambda^\infty_m} \leq \frac{1}{\gamma_s(D_-, \lambda^\infty_m, D_+)} \sin^2 \Theta_\eta \tag{3.3.14}$$

for η large enough.

Proof. Applying Lemmata 3.2.3 and 3.3.1 twice (once for D_- and once for D_+), we establish

$$D_- < \lambda^\eta_m < D_+$$

for η large enough. The conclusion (3.3.14) is a direct consequence of Theorem 2.6.4. □

When $n > 1$ we will no longer measure the convergence of individual eigenvalues λ^η_{m+i-1}, $i = 1, ..., n$ to λ^∞_m. Instead, assuming that there exists η_0 such that

$$\lambda^\eta_1 \leq \cdots \leq \lambda^\eta_{m-1} < D_- < \lambda^\eta_m \leq \cdots \leq \lambda^\eta_{m+n-1} < D_+ < \lambda^\eta_{m+n} \leq \lambda^\eta_{n+n+1} \leq \cdots \tag{3.3.15}$$

for all $\eta \geq \eta_0$. Assuming $\eta \geq \eta_0$, we define

$$\widehat{\lambda}^\eta_m := \frac{1}{n}\mathrm{tr}(\mathbf{H}_\eta E_\eta[D_-, D_+])$$

and estimate

$$\frac{|\widehat{\lambda}^\eta_m - \lambda^\infty_m|}{\lambda^\infty_m}.$$

The proof will follow from the analytic perturbation theory of Kato, cf. [41]. Yet another interesting feature of the perturbation δh_η will be revealed in the course of the proof. It will shed new light on the quadratic estimates from Section 2.6.

Theorem 3.3.8. *Let the eigenvalues of the operator $\mathbf{H}_\infty$ be so ordered that*

$$\lambda^\infty_{m-1} < \lambda^\infty_m = \cdots = \lambda^\infty_{m+n-1} < \lambda^\infty_{m+n}.$$

Define the measure of the relative separation of λ^∞_m from the rest of the spectrum of $\mathbf{H}_\infty$ as the number

$$\gamma_s(\lambda^\infty_m) = \min\left\{\frac{\lambda^\infty_{m+n} - \lambda^\infty_m}{\lambda^\infty_{m+n} + \lambda^\infty_m}, \frac{\lambda^\infty_m - \lambda^\infty_{m-1}}{\lambda^\infty_m + \lambda^\infty_{m-1}}\right\}.$$

There exists $\eta_0 > 0$ such that for $\eta \geq \eta_0$

$$\frac{|\widehat{\lambda}^\eta_m - \lambda^\infty_m|}{\lambda^\infty_m} < \sin\Theta_\eta \frac{\frac{3\sin\Theta_\eta}{\gamma_c(\lambda^\infty_m)}}{1 - \frac{3\sin\Theta_\eta}{\gamma_s(\lambda^\infty_m)}}. \tag{3.3.16}$$

PROOF. Since $\sin\Theta_\eta \to 0$, an argument analogous to the argument that led to Lemma 3.3.1 together with (3.3.4) implies that we can pick $\eta_0 > 0$ such that for $\eta > \eta_0$

$$\sin\Theta_\eta \le \frac{1}{3}\gamma_s(\lambda_m^\infty) \tag{3.3.17}$$

$$|\lambda_k^\eta - \lambda_m^\infty| \le \frac{1}{3}\gamma_s(\lambda_m^\infty)\lambda_m^\infty, \qquad k = m, m+1, ..., m+n-1, \tag{3.3.18}$$

$$|\zeta - \lambda_k(\mathbf{H}'_\eta)| > \frac{1}{3}\gamma_s(\lambda_m^\infty)\lambda_k(\mathbf{H}'_\eta), \qquad k \notin \{m, m+1, ..., m+n-1\} \tag{3.3.19}$$

where ζ is such that $|\zeta - \lambda_m^\infty| = \frac{1}{3}\gamma_s(\lambda_m^\infty)\lambda_m^\infty$. Assume $\eta > \eta_0$ is fixed, then define the family

$$a(\kappa) = h'_\eta + \kappa\delta h_\eta, \qquad \kappa \in \mathbb{C}. \tag{3.3.20}$$

This is a *holomorphic family of type (B)* (for the definition see [41, Chapter VII]). We know that

$$|\delta h_\eta[u]| < \sin\Theta_\eta h'_\eta[u], \qquad u \in \mathcal{Q}, \tag{3.3.21}$$

so [41, Theorem VII-4.9 and (VII-4.45)] imply that the resolvent

$$R(\kappa, \zeta) = (\mathbf{A}(\kappa) - \zeta\mathbf{I})^{-1}$$

is a convergent power series in κ for $\zeta \in \Gamma$. Here Γ is a circle in the complex plane with the radius $\frac{1}{3}\gamma_s(\lambda_m^\infty)\lambda_m^\infty$ and the center λ_m^∞. The power series for $R(\kappa, \zeta)$ converges for every

$$|\kappa| < r_0 = \frac{1}{\sin\Theta_\eta} \inf_{\zeta\in\Gamma, \lambda\in\sigma(\mathbf{H}'_\eta)} \frac{|\lambda - \zeta|}{\lambda} = \frac{1}{\sin\Theta_\eta}\frac{1}{3}\gamma_s(\lambda_m^\infty). \tag{3.3.22}$$

In particular, assumption (3.3.17) implies that the series converges for $\kappa = 1$.

Define

$$\widehat{\lambda}_m^\eta(\kappa) = -\frac{1}{2\pi i\ n}\mathrm{tr}\left(\mathbf{A}(\kappa)\int_\Gamma R(\kappa, \zeta)\ d\zeta\right),$$

then $\widehat{\lambda}_m^\eta(\kappa)$ is a holomorphic function and due to the assumptions we have made

$$|\widehat{\lambda}_m^\eta(\kappa) - \lambda_m^\infty| < \frac{1}{3}\gamma_s(\lambda_m^\infty)\lambda_m^\infty, \qquad |\kappa| < r_0.$$

Cauchy's integral inequality[6] for the coefficients of the Taylor expansion implies

$$|\widehat{\lambda}_{m,\eta}^{(n)}| < \frac{\frac{1}{3}\gamma_s(\lambda_m^\infty)\lambda_m^\infty}{r_0^n}, \qquad n = 1, 2, \cdots$$

[6]For further details see [8, Section 8.1.4] and [41, Section II-3].

where

$$\widehat{\lambda}_m^\eta(\kappa) = \lambda_m^\infty + \kappa\widehat{\lambda}_{m,\eta}^{(1)} + \kappa^2\widehat{\lambda}_{m,\eta}^{(2)} + \kappa^3\widehat{\lambda}_{m,\eta}^{(3)} + \cdots .$$

This yields

$$|\widehat{\lambda}_m^\eta(\kappa) - \lambda_m^\infty - \kappa\widehat{\lambda}_{m,\eta}^{(1)}| < \frac{\frac{1}{3}\gamma_s(\lambda_m^\infty)\lambda_m^\infty}{r_0}\frac{|\kappa|^2}{r_0 - |\kappa|} \leq \frac{\frac{1}{3}\gamma_s(\lambda_m^\infty)\lambda_m^\infty}{r_0^2}\frac{|\kappa|^2}{1 - \frac{|\kappa|}{r_0}}$$

for $|\kappa| < r_0$ and in particular for $\kappa = 1$

$$|\frac{1}{n}\sum_{i=1}^{n}\lambda_{i+m-1}^\eta - \lambda_m^\infty - \widehat{\lambda}_{m,\eta}^{(1)}| < \sin\Theta_\eta\lambda_m^\infty\ \frac{3\sin\Theta_\eta}{\gamma_c(\lambda_m^\infty)}\ \frac{1}{1 - \frac{3\sin\Theta_\eta}{\gamma_s(\lambda_m^\infty)}}.$$

If it were not for $\widehat{\lambda}_{m,\eta}^{(1)}$, the theorem would have been proved. And now comes the trick! It was established, in [41, (VII-4.44)], that

$$\widehat{\lambda}_{m,\eta}^{(1)} = \frac{1}{n}\sum_{j=1}^{n}\delta h_\eta[u_i]$$

where u_j, $j = 1, .., n$ form a basis for $\mathsf{ran}(P) = \mathsf{ran}(E_\infty[D_-D_+])$. Since

$$\delta h_\eta[u] = h_\eta(P_\perp u, Pu) + h_\eta(Pu, P_\perp u) = 0, \qquad u \in \mathsf{ran}(P),$$

we obtain $\widehat{\lambda}_{m,\eta}^{(1)} = 0$ and the desired result follows. □

The fact that $\widehat{\lambda}_{m,\eta}^{(1)} = 0$, for this particular perturbation, was first noticed by Drmač and used (by him) to compute eigenvalue and eigenvector estimates in an unpublished note on Jacobi-Davidson method. Subsequently (on his incentive), it was also used in [35] as a starting point of a further analysis of Jacobi–Davidson iterative scheme for the solution of a partial eigenvalue problem. The proof of Theorem 3.3.8, as well as the obtained estimate, is a generalization and improvement of these matrix results.

Remark 3.3.9. Theorem 3.3.8 sheds a new light on the study of the quadratic estimates from Section 2.6. Let us emphasize that in Theorem 3.3.8 the perturbation error is measured both by the measure of the size of the first nontrivial term in the perturbation expansion and the size of the region in which such expansion is valid. This illustrates the role which is played by the relative gap in our estimates.

3.3.2 A model problem: Schrödinger operator with a square-well potential

The eigenvalue problem for the Schrödinger operator with s square-well potential will illustrate the theory we have built so far.

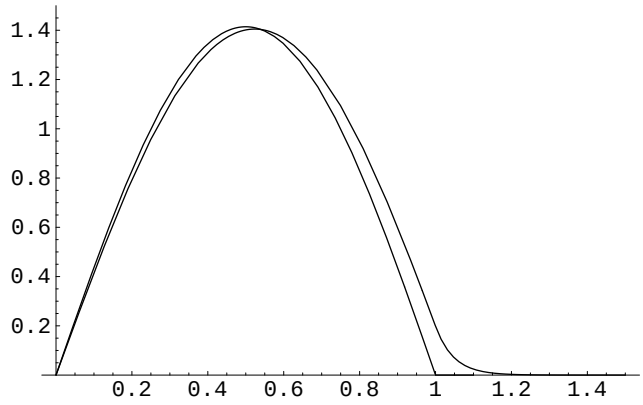

Figure 3.2: First eigenvector of the Schrödinger operator with the square-well potential and with $\eta^2 = 9 + 9\,\mathsf{cot}^2(3)$.

Let us reconsider Problem 3.1 in detail. For the family of operators

$$h_\eta(u,v) = (\mathbf{H}_\eta^{1/2}u, \mathbf{H}_\eta^{1/2}u) = \int_0^\infty u'v'\ dx + \eta^2 \int_1^\infty uv\ dx, \quad u,v \in H^1(\mathbb{R}_+)$$

we will be able to give a formula for both $\mathsf{sin}\Theta_\eta$ and β_η. Furthermore, the function

$$\frac{\lambda_1^\infty - \lambda_1^\eta}{\lambda_1^\infty}$$

can be—in this case—expanded in the Taylor series (as $\eta \to \infty$), see [47]. Therefore, we will be able to assess the sharpness of the estimate from Theorem 3.3.8 on this model problem.

The operator $\mathbf{H}^\infty$, defined by the form

$$h_\infty(u,v) = \int_0^1 u'v'\ dx, \quad u,v \in H_0^1\,[0,1]\,,$$

is the standard negative Laplace operator with zero boundary conditions on $[0,1]$. All of its eigenvalues are of multiplicity one, so Theorem 3.3.7 applies. The eigenvalues of $\mathbf{H}_\infty$ are given by the formula

$$\lambda_i = i^2\pi^2.$$

The accompanying eigenfunctions are $u_i(x) = \sqrt{2}\ \mathsf{sin}(\sqrt{\lambda_i}\ x)$, $x \in [0,1]$. We extend the functions u_i from $[0,1]$, as it was done for u_1 in (3.1.9), by zero to the whole of $\mathbb{R}_+$.

The situation is a bit more complex for $\mathbf{H}_\eta$. The eigenvalues of the operator $\mathbf{H}_\eta$ have to be described implicitly. Let

$$\mathbf{H}_\eta v^\eta = \lambda^\eta v^\eta,$$

then $v^\eta \in C^1(\mathbb{R}_+)$ is

$$v^\eta(x) = \begin{cases} \sin(\sqrt{\lambda^\eta}x), & 0 \le x \le 1 \\ \frac{\sin\sqrt{\lambda^\eta}}{e^{-\sqrt{\eta^2-\lambda^\eta}}}\, e^{-\sqrt{\eta^2-\lambda^\eta}\ x}, & 1 \le x \end{cases}$$

and λ^η is a solution of the equation

$$\sqrt{\eta^2 - \lambda^\eta} = -\sqrt{\lambda^\eta}\cot(\sqrt{\lambda^\eta}). \tag{3.3.23}$$

In [47] it was shown that $\frac{\lambda_1^\infty - \lambda_1^\eta}{\lambda_1^\infty}$ can be represented (for $\eta \to \infty$) by a convergent Taylor series

$$\frac{\lambda_1^\infty - \lambda_1^\eta}{\lambda_1^\infty} = 2\frac{1}{\eta} - 3\frac{1}{\eta^2} + 8\left(\frac{1}{2!} + \frac{1}{4!}\pi^2\right)\frac{1}{\eta^3} - 10\left(\frac{1}{2!} + \frac{4}{4!}\pi^2\right)\frac{1}{\eta^4} + \cdots . \tag{3.3.24}$$

However, the method used to compute this expansion does not say anything about the radius of convergence. In what follows we will illustrate the role played by the inequalities (3.3.17)–(3.3.19) and (3.3.22) when assessing the sharpness of the estimates that can be obtained by "$\sin\Theta_\eta$-method".

To apply Theorem 3.3.7 we need to compute β_η. In the case in which we are approximating only the lowest eigenvalue we have

$$\beta_\eta = \frac{|(u_1, \mathbf{H}_\infty^\dagger u_1) - (u_1, \mathbf{H}_\eta^{-1} u_1)|}{(u_1, \mathbf{H}_\infty^\dagger u_1)}. \tag{3.3.25}$$

We compute $(u_1, \mathbf{H}_\eta^{-1}u_1) - (u_1, \mathbf{H}_\infty^\dagger u_1)$ with the help of the Green functions. From (3.1.10) it follows

$$\beta_\eta = \frac{2}{1+\eta} \le \frac{2}{\eta}, \tag{3.3.26}$$

which is a reasonable estimate for $\eta \to \infty$. Theorem 3.3.7 implies

$$\frac{\lambda_1^\infty - \lambda_1^\eta}{\lambda_1^\infty} \lesssim \frac{1}{\gamma_s(\lambda_1^\infty)}\frac{\beta_\eta}{1+\beta_\eta} \lesssim \frac{10}{3\eta}. \tag{3.3.27}$$

Here $\gamma_s(\lambda_1^\infty)$ is as defined in Theorem 3.3.8.

Since all of the eigenvalues λ_i are simple, the estimates for λ_i, $i > 1$, follow in the same fashion. Let

$$\beta_\eta^i = \frac{|(u_i, \mathbf{H}_\infty^\dagger u_i) - (u_i, \mathbf{H}_\eta^{-1} u_i)|}{(u_i, \mathbf{H}_\infty^\dagger u_i)},$$

then

$$\beta_\eta^i = \frac{2}{(1+\eta)}, \quad i \in \mathbb{N}.$$

Analogously as in (3.3.27) we establish that

$$\frac{|\lambda_i^\eta - \lambda_i^\infty|}{\lambda_i^\infty} \lesssim \frac{1+2\,i+2\,i^2}{1+2\,i}\frac{3\cdot 2}{(1+\eta)},$$

since

$$\gamma_s(\lambda_i^\infty) = \min\left\{\frac{(i+1)^2\pi^2 - i^2\pi^2}{(i+1)^2\pi^2 + i^2\pi^2}, \frac{i^2\pi^2 - (i-1)^2\pi^2}{(i-1)^2\pi^2 + i^2\pi^2}\right\} = \frac{1+2\,i}{1+2\,i+2\,i^2}. \tag{3.3.28}$$

We can now concentrate on the role of the relative gap $\gamma_s(\lambda_i^\infty)$. Roughly, it tells us that for sufficiently large η, the eigenvalues of $\mathbf{H}_\eta$ for which

$$\frac{1+2\,i+2\,i^2}{1+2\,i}\frac{2}{(1+\eta)} \lesssim 1$$

are the ones that are being approximated by the eigenvalues of $\mathbf{H}_\infty$, cf. Theorem 2.4.2, Theorem 2.6.4 and Theorem 3.3.8.

The reason that we were using the "wiggly" sign $\lesssim$ is that only for sufficiently large η we can pick $D_\pm$ such that

$$\lambda_{i-1}^\infty \le D_- < \lambda_i^\infty < D_+ \le \lambda_{i+1}^\infty$$

and $\gamma_s(D_-, \lambda_i^\infty, D_+) \approx \gamma_s(\lambda_i^\infty)$.

If we are to be rigorous, we have to resort to a somewhat conservative reasoning from Theorem 3.3.8. The argument will be presented in several steps. Assume we have chosen to approximate λ_i^η. Thanks to (3.3.23) we can establish that every λ_i^∞ is a simple eigenvalue. Let $\eta_0 > 0$ be such that

$$\sin\Theta_{\eta_0} < \frac{1}{3}\gamma_s(\lambda_i^\infty) \rightsquigarrow \frac{1+2\,i+2\,i^2}{1+2\,i}\frac{2}{(1+\eta_0)} < \frac{1}{3}, \tag{3.3.29}$$

then Lemma 3.3.1 implies

$$\mathsf{dim}\ E_\eta(\lambda_i^\infty + \frac{1}{3}\gamma_s(\lambda_i^\infty)\lambda_i^\infty) \ge \mathsf{dim}\ E_\infty(\lambda_i^\infty + \frac{1}{3}\gamma_s(\lambda_i^\infty)\lambda_i^\infty), \qquad \eta_0 < \eta.$$

From (3.3.23) and (3.3.28) we conclude that, indeed,

$$\mathsf{dim}\ E_\eta(i^2\pi(1+\frac{1}{3}\gamma_s(\lambda_i^\infty))) = \mathsf{dim}\ E_\infty(i^2\pi(1+\frac{1}{3}\gamma_s(\lambda_i^\infty))), \qquad \eta_0 < \eta. \tag{3.3.30}$$

This illustrates the way in which the external information, obtained from (3.3.23), steps in to play the role of Weidmann's lemma (Lemma 3.2.3). Now we may conclude that

$$\frac{|\lambda_i^\eta - \lambda_i^\infty|}{\lambda_i^\infty} \le \frac{(1+2\,i+2\,i^2)}{(1+2\,i)}\frac{3}{\left(1+\sqrt{\frac{1+\eta}{2}}\right)\left(\sqrt{\frac{1+\eta}{2}} - \frac{2\,(1+2\,i+3\,i^2)}{1+2\,i}\right)}, \qquad \eta_0 < \eta. \tag{3.3.31}$$

Without such external information, we have to resort to Lemma 3.2.3. Subsequently, in such a case we can only conclude there exists $\eta_0 > 0$ such that the estimate is rigorous for all $\eta > \eta_0$.

To get a better feeling for the formula (3.3.31) we consider the case $i = 1$. Condition (3.3.29) implies that (3.3.31) holds for $\eta > 9$. Estimate (3.3.31) can be sharpened and simplified if we put further restrictions on η. Let us further assume that

$$\frac{2}{(1+\eta_0)} < 8.1 \cdot 10^{-5} \rightsquigarrow \eta_0 = 246.$$

With this in hand we obtain

$$\frac{|\lambda_1^\eta - \lambda_1^\infty|}{\lambda_1^\infty} \leq \frac{15}{1+\eta}, \quad \eta > 246. \tag{3.3.32}$$

This is obviously a conservative estimate. Furthermore, it was forged by imposing a pessimistic choice of η_0. This should be kept in mind when comparing (3.3.32) with the expansion (3.3.24), which comes without convergence radius estimate.

If we assume $\eta \geq 246$ and squeeze the maximum amount of information out of (3.3.23) we can obtain a good estimate of the spectral gap. A direct application of Theorems 2.6.4 and 2.5.2 yields the estimates (for $i = 1$)

$$\frac{|\lambda_1^\eta - \lambda_1^\infty|}{\lambda_1^\infty} \leq \frac{3.348}{\eta - 1}, \qquad \eta \geq 246, \tag{3.3.33}$$

$$\|u_1 - v_1^\eta\| \leq \frac{1.343\sqrt{\frac{1}{\eta-1}}}{\sqrt{1-\sqrt{\frac{2}{\eta-1}}}}, \qquad \eta \geq 246. \tag{3.3.34}$$

We reiterate that Theorem 3.3.8, as well as assumptions that led to (3.3.31), represent a plausible requirements on the outside information. We have assumed:

1. A result on the distribution of eigenvalues which assures us that for $\eta > \eta_0$ conclusion (3.3.30) holds, cf. [44, 65].

2. Operator $\mathbf{H}_\infty$ is a well known object, so we can use the gaps in the spectrum of $\mathbf{H}_\infty$ to estimate the gaps in the spectrum of $\mathbf{H}_\eta$.

Estimate (3.3.33)–(3.3.34) show that a direct application of the theorems from Section 2 yields better estimates, when possible.

The $H_0^1(\mathbb{R}^r)$ case—perturbations by a deep square-well potential

When considering higher dimensional Schrödinger operators perturbed by a square-well potential, we can apply the theory of [9, 23].

Definition 3.3.10. Let $\mathbf{H}$ be a nonnegative definite operator in $\mathcal{H}$ and let P an orthogonal projection in $\mathcal{H}$. The operator $\mathbf{H}$ is called *local with respect to a projection* P if $\mathsf{ran}(P) \cap \mathcal{D}(\mathbf{H})$ is dense in $\mathsf{ran}(P)$ and if

$$\mathbf{H}v = P\mathbf{H}v, \qquad v \in \mathsf{ran}(P) \cap \mathcal{D}(\mathbf{H})$$

holds.

Assume that $\mathbf{H}_b$ is local with respect to the projection $\mathbf{H}_e = P$. The restriction of the operator $\mathbf{H}$ to the space $\mathsf{ran}(P_\perp)$ is the symmetric operator $\mathbf{A}$ defined by

$$\mathbf{A}f = \mathbf{H}_b f, \qquad f \in \mathcal{D}(\mathbf{H}_b) \cap \mathsf{ran}(P_\perp).$$

Define $\mathbf{H}_\infty$ as the *Friedrichs extension*[7] of the symmetric nonnegative operator $\mathbf{A}$ in $P_\perp\mathcal{H}$. Baumgärtel and Demuth (see [9]) have shown that $\mathbf{H}_\eta \to \mathbf{H}_\infty$ (in the sense of strong resolvent convergence). We will show the way to utilize the theory of [23] to obtain the convergence rates for the eigenvalues.

Take $\mathcal{H} = L^2(\mathbb{R}^r)$ and let $\mathbf{H}_b = -\triangle$ and $\mathcal{Q}(h_b) = H_0^1(\mathbb{R}^r)$. Define

$$(P_e\ f)(x) = \chi_{\mathbb{R}^r\setminus\mathcal{A}}(x)v(x), \quad x \in \mathbb{R}^r$$

where $\mathcal{A} \subset \mathbb{R}^r$ is a bounded connected region and $\chi_{\mathbb{R}^r\setminus\mathcal{A}}$ is the characteristic function of its complement. Other, more general operators $\mathbf{H}_b$ and regions $\mathcal{A} \subset \mathbb{R}^r$ are possible, cf. [23]. Important is that the Friedrichs extension of the operator $\mathbf{H}_\infty\ f = \mathbf{H}_b\ f$, $f \in \mathsf{ran}(P_\mathcal{A}) \cap \mathcal{D}(\mathbf{H})$ should be positive definite.

Assume $\mathcal{A} \subset \mathbb{R}^r$ is (additionally) convex with the *Lipschitz boundary* [8] and define

$$\begin{aligned} h_\eta(u,v) &= (\mathbf{H}_b^{1/2}u, \mathbf{H}_b^{1/2}v) + \eta^2\ (P_e\ u, P_e\ v), & u,v &\in \mathcal{Q}(h_b) = H_0^1(\mathbb{R}^r),\\ h_\infty(u,v) &= (\mathbf{H}_b^{1/2}u, \mathbf{H}_b^{1/2}v), & u,v &\in \mathcal{Q}(h_\infty) = H_0^1(\mathcal{A}). \end{aligned}$$

[7] Friedrichs extension of a nonnegative symmetric operator $\mathbf{A}$ in $\mathcal{H}_\mathbf{A}$ is a minimal self adjoint extension of $\mathbf{A}$ in $\mathcal{H}_\mathbf{A}$.

[8] For a definition of the Lipschitz boundary see [23] and the references therein. The theory of [23] allows for more general $\mathbf{H}_b$ than $-\triangle$. Everything said is also valid for those $\mathbf{H}_b$, provided the forms h_η satisfy the assumptions of our convergence theorems.

The theory of [23] (see also [9]) guarantees the existence of a constant $C_{\mathcal{A}}$ such that

$$\|\mathbf{H}_\infty^\dagger - \mathbf{H}_\eta^{-1}\| \leq \frac{C_{\mathcal{A}}}{\eta^{1/2}} \ . \tag{3.3.35}$$

From (3.3.35) we see that the convergence rate deteriorates for dimension $r > 1$. The convergence properties deteriorate even further if we drop the requirement for $\mathcal{A}$ to be convex, see [23].

Remark 3.3.11. After we have finished the research that led to this thesis, we became aware of Reference [15]. Unfortunately, it was to late to incorporate those results in this work. The approach of [15] should enable us to directly obtain estimates of (3.3.25). This should yield sharper eigenvalue bounds than those which can be obtained through a use of (3.3.35).

Let us now see what kind of information on the behavior of the spectrum follows from an estimate on

$$\|\mathbf{H}_\eta^{-1} - \mathbf{H}_\infty^\dagger\|. \tag{3.3.36}$$

Assume that we have found D_- and D_+ such that

$$\lambda_{m-1}^\infty < D_- < \lambda_m^\infty \qquad\qquad < D_+ < \lambda_{m+n}^\infty \tag{3.3.37}$$

$$\lambda_{m-1}^\eta < D_- < \lambda_m^\eta = \cdots = \lambda_{m+n-1}^\eta < D_+ < \lambda_{m+n}^\eta. \tag{3.3.38}$$

For such D_- and D_+ set $\mathcal{S} = [D_-, D_+]$. The min-max characterization of the discrete spectrum of a bounded self adjoint operator and (3.3.36) imply

$$\left| \frac{1}{\lambda_m^\infty} - \frac{1}{\lambda_{m+n-i}^\eta} \right| = \frac{|\lambda_m^\infty - \lambda_{m+n-i}^\eta|}{\lambda_m^\infty \lambda_{m+n-i}^\eta} \leq \|\mathbf{H}_\eta^{-1} - \mathbf{H}_\infty^\dagger\|, \qquad i = 1, ..., n,$$

which yields the "relative" estimate

$$\frac{|\lambda_m^\infty - \lambda_{m+n-i}^\eta|}{\lambda_m^\infty} \leq \lambda_{m+n-i}^\eta \|\mathbf{H}_\eta^{-1} - \mathbf{H}_\infty^\dagger\| \leq D_+ \|\mathbf{H}_\eta^{-1} - \mathbf{H}_\infty^\dagger\|, \qquad i = 1, ..., n. \tag{3.3.39}$$

For a comparison, Theorem 3.3.8 gives

$$\frac{|\widehat{\lambda}_m^\eta - \lambda_m^\infty|}{\lambda_m^\infty} < \sin^2\Theta_\eta \ \frac{3}{\gamma_c(\lambda_m^\infty)} \ \frac{1}{1 - \frac{3\sin\Theta_\eta}{\gamma_s(\lambda_m^\infty)}},$$

where, using Corollary 3.3.3,

$$\sin^2\Theta_\eta := \sin^2\Theta_\eta(E_\infty(\mathcal{S})) = \max_{x \in \mathsf{ran}(E_\infty(\mathcal{S}))} \frac{(x, \mathbf{H}_\eta^{-1}x) - (x, \mathbf{H}_\infty^\dagger x)}{(x, \mathbf{H}_\eta^{-1}x)}$$

$$\leq \frac{\beta_\eta}{1+\beta_\eta} \leq \frac{\|\mathbf{H}_\eta^{-1} - \mathbf{H}_\infty^\dagger\| D_+}{1 + \|\mathbf{H}_\eta^{-1} - \mathbf{H}_\infty^\dagger\| D_+}.$$

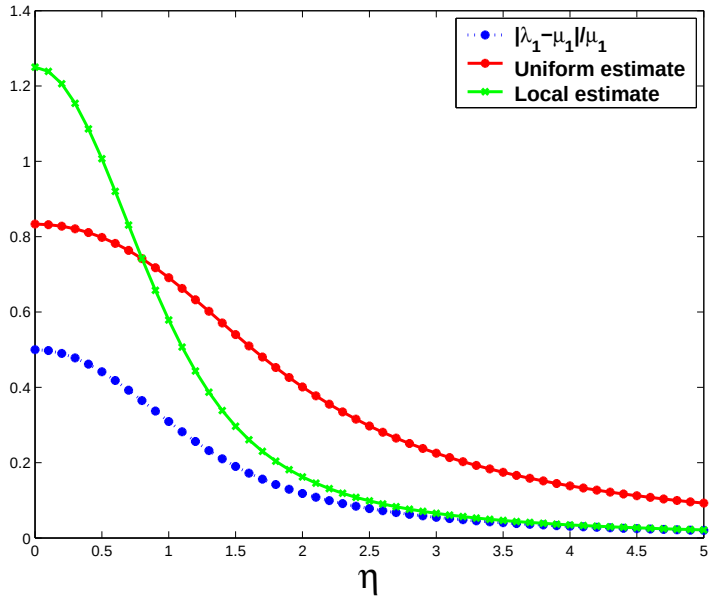

Figure 3.3: Comparing the uniform and the local estimates. *Uniform estimate* depicts the bound (3.3.44), whereas *Local estimate* depicts the bound (3.3.43). We can observe the influence of the gap for $\eta > 1$.

This shows that rather than $\|\mathbf{H}_\eta^{-1} - \mathbf{H}_\infty^\dagger\|$, a better target for the analysis would have been the "relative" quantity

$$\frac{(x, \mathbf{H}_\eta^{-1} x) - (x, \mathbf{H}_\infty^\dagger x)}{(x, \mathbf{H}_\eta^{-1} x)} \tag{3.3.40}$$

for $x \in \mathsf{ran}(E_\infty)$. The measure $\mathsf{sin}^2\Theta_\eta$ is a local quantity, since it measures the discrepancy between $\mathbf{H}_\eta^{-1}$ and $\mathbf{H}_\infty^\dagger$ on the subspace $\mathsf{ran}(E_\infty(\mathcal{S}))$, only. It can, and usually will be, considerably smaller than the global measure $\|\mathbf{H}_\eta^{-1} - \mathbf{H}_\infty^\dagger\|$, cf. Example 3.3.12. On top of that, we have — through a combination of Theorem 3.3.4 and Theorem 2.5.6— the eigenvector estimate

$$\|E_\eta(\mathcal{S}) - E_\infty(\mathcal{S})\| \leq \max\left\{ \frac{\sqrt{\lambda_m^\infty D_-}}{\lambda_m^\infty - D_-}, \frac{\sqrt{D_+ \lambda_m^\infty}}{D_+ - \lambda_m^\infty} \right\} \frac{\sin\Theta_\eta}{\sqrt{1 - \sin\Theta_\eta}}. \tag{3.3.41}$$

Example 3.3.12. To get a feeling for the preceding discussion consider the following 2×2 example. We study the family

$$\mathbf{H}_\eta = \begin{bmatrix} 2 & -1 \\ -1 & 2 + \eta^2 \end{bmatrix}$$

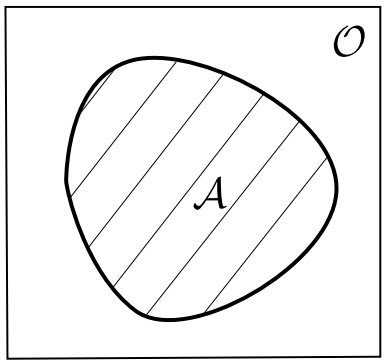

Figure 3.4: *High contrast media*: The bounded domain Ω is decomposed as $\Omega = \mathcal{A} \cup \mathcal{O}$. The limit lives in $\mathcal{O}$.

for η large. One computes, as $\eta \to \infty$,

$$\begin{bmatrix} 2 & -1 \\ -1 & 2+\eta^2 \end{bmatrix}^{-1} \to \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}^{\dagger} = \mathbf{H}_{\infty}^{\dagger}$$

and

$$\sin^2 \Theta(\mathbf{H}_{\eta}^{-1/2} e_1, \mathbf{H}_{\eta}^{1/2} e_1) = \frac{1}{4+2\eta^2}, \qquad e_1 = [1 \quad 0]^*. \tag{3.3.42}$$

Now, $\lambda_1^{\eta} = \frac{1}{2}(4 + \eta^2 - \sqrt{4-\eta^4})$, $\lambda_1^{\infty} = 2$ and Theorem 3.3.7 states

$$\frac{|\lambda_1^{\eta} - \lambda_1^{\infty}|}{\lambda_1^{\infty}} \leq \frac{\frac{1}{2}(4+\eta^2+\sqrt{4-\eta^4})+2}{|\frac{1}{2}(4+\eta^2+\sqrt{4-\eta^4})-2|} \frac{1}{4+2\eta^2}. \tag{3.3.43}$$

We also establish

$$\|\mathbf{H}_{\eta}^{-1} - \mathbf{H}_{\infty}^{\dagger}\| = \frac{5}{2(3+2\eta^2)},$$

which gives, based on (3.3.39),

$$\frac{|\lambda_1^{\eta} - \lambda_1^{\infty}|}{\lambda_1^{\infty}} \leq \frac{1}{2}(4+\eta^2-\sqrt{4-\eta^4}) \frac{5}{2(3+2\eta^2)}. \tag{3.3.44}$$

The comparison between (3.3.43) and (3.3.44) is displayed on Figure 3.3. We see the superior performance of the local estimate based on $\mathsf{sin}^2\Theta_{\eta}$ (and coupled with the relative gap).

Remark 3.3.13. Spectral estimates that are based on a study of a localized quantity like (3.3.40) are not new. Bruneau and Carbou have recently studied (in [17])) the spectral asymptotics for the Helmholtz operator

$$\mathbf{A}_{\eta} = -\triangle + \eta^2 \chi_{\mathcal{A}} \tag{3.3.45}$$

where $\mathcal{D}(\mathbf{A}_\eta) = H^2(\Omega) \cap H_0^1(\Omega)$. The structure of (3.3.45) is such that it is additionally assumed that Ω and $\mathcal{A}$ are two bounded, connected and smooth domains in $\mathbb{R}^2$ and that $\mathcal{A}$ is *compactly contained*[9] in Ω. This model is used in the study of electromagnetic wave guide $\Omega \times \mathbb{R}$, where $\mathcal{A} \times \mathbb{R}$ is a supra-conductor material with very large conductivity, see [17, 33] and the references therein.

Assume that (3.3.37) and (3.3.38) hold and define the operator $\mathbf{A}_\infty$ by requiring that $\mathbf{A}_\infty^\dagger = \mathrm{s} - \lim_\eta \mathbf{A}_\eta^{-1}$. Using the boundary layer techniques, Bruneau and Carbou have provided an asymptotic expansion (in $\eta \to \infty$) of the difference(s)

$$E_\infty(\mathcal{S})(\mathbf{A}_\eta - \xi)^{-1}E_\infty(\mathcal{S}) - E_\infty(\mathcal{S})(\mathbf{A}_\infty - \xi)^{-1}E_\infty(\mathcal{S}), \qquad |\xi - \lambda_m^\infty| = \min\{D_-, D_+\}. \tag{3.3.46}$$

The asymptotic expansion of (3.3.46) was then used, in connection with the finite dimensional perturbation theory from [41], to establish eigenvalue asymptotics. The first order asymptotic expansion for the chosen eigenvalue (e.g. λ_m^∞), as well as the higher order expansions were computed in [17]. The obtained results were fine-tuned to the structure of the operator (3.3.45). Thus, it is not obvious how to separate an abstract framework from the considerations of the special case.

Our estimates are, admittedly, only of the first order. However, we explicitly state an abstract theory which is applicable to a broad class of singularly perturbed problems, cf. Theorem 3.3.4 and Theorem 3.3.8. The technique used to compute the asymptotic expansion of (3.3.46) can be used to compute an estimate of (3.3.40). In addition to that, we offer an explicit eigenvector estimate (3.3.41).

Based on the references that were available to us, as well as based on the references from [17], it appears that ours is a first abstract theory for computing spectral asymptotics (3.1.2)–(3.1.3) in the large coupling limit.

Further discussion of the results like (3.3.35) and (3.3.46) is well beyond the scope of this thesis. Our aim was only to inaugurate a (new) abstract framework for establishing asymptotic eigenvalue and eigenvector estimates when we are provided with a (local) resolvent estimate.

3.4 Spectral asymptotics in the regular case

We will now investigate the family

$$h_\eta(u, v) = h_b(u, v) + \eta^2 h_e(u, v), \quad u, v \in \mathcal{Q} \subset \mathcal{H} \tag{3.4.1}$$

[9] For a definition see [17]. Also, see Figure 3.4.

with an additional *regularity assumption* (3.1.4) on the form h_e. The gist of the analysis are the energy norm convergence estimates that were presented in [36, 59]. The main contribution in this section is an algebraic version of the proof of the main result from [59] and the subsequent (new) corollary that will enable us to assess the sharpness of the technique from [59]. As an important byproduct we introduce a notion of the residual in the analysis. Furthermore, a new criterion for the norm resolvent convergence of families (3.4.1) will be established. The condition is equivalent to the Babuška–Brezzi inf–sup condition used in [36, 59] to obtain energy norm convergence estimates. This condition is also easier to check in the examples we plan on presenting.

The results were originally proved in [36, 59] in the variational setting with the help of the theory of *Lagrange multipliers* from [16]. A reformulation of the proofs in the quadratic form context required some extra work (to find correct analogies). Since the generalized convergence in the resolvent sense was not considered in [36, 59], the new theorems will also enhance the applicability of the obtained convergence rate estimates.

Under the additional assumptions (to the one already made) about the forms h_b and h_e, we prove that $\|\mathbf{H}_\eta^\dagger - \mathbf{H}_\infty^\dagger\| \to 0$ and show that the convergence is of the order η^{-2}. More importantly, we will establish

$$\frac{(x, \mathbf{H}_\eta^{-1} x) - (x, \mathbf{H}_\infty^\dagger x)}{(x, \mathbf{H}_\eta^{-1} x)} \leq \frac{C_x}{\eta^2},$$

where constant C_x depends only on the vector x.

Let us write down the assumptions on the forms h_b and h_e we have made so far. We have assumed that h_b and h_e are closed, nonnegative and densely defined forms in $\mathcal{H}$ and that

$$\mathcal{Q}(h_b) \subset \mathcal{Q}(h_e).$$

The minimal, further, requirement would have been that $h_b + h_e$ be positive definite. However, it was already noted that we may assume h_b is positive definite, without effecting the generality of the result. We also require that $\overline{\mathsf{ker}(h_e)}$ is a nontrivial subspace of $\mathcal{H}$. Note that $\mathcal{Q}(h_b) \subset \mathcal{Q}(h_e)$ implies that the operator $\mathbf{H}_e^{1/2}\mathbf{H}_b^{-1/2}$ is a bounded operator on $\mathcal{H}$. Now we shall additionally assume that

$$\mathsf{ran}(\mathbf{H}_e^{1/2}\mathbf{H}_b^{-1/2}) = \overline{\mathsf{ran}(\mathbf{H}_e^{1/2}\mathbf{H}_b^{-1/2})}^{\,\mathcal{H}}. \tag{3.4.2}$$

This is equivalent to $\|(\mathbf{H}_e^{1/2}\mathbf{H}_b^{-1/2})^\dagger\| < \infty$. The pseudo inverse of bounded operators with the closed range can easily be constructed, cf. [46] and Theorem 2.2.4.

Definition 3.4.1. Let $\mathcal{W}$ be a closed subspace of $\mathcal{H}$ and $f \in \mathcal{H}$. We say that $w \in \mathcal{W}$ is a *Galerkin approximation* (from the subspace $\mathcal{W}$) to the vector $\mathbf{H}^{-1}f$ if

$$h(w, v) = (P_{\mathcal{W}} f, v), \qquad v \in \mathcal{W}. \tag{3.4.3}$$

We call (3.4.3) the *Galerkin condition* and the problem of the existence of such $w \in \mathcal{W}$ the *Galerkin problem*.

The assumption $\mathcal{Q}(h_b) \subset \mathcal{Q}(h_e)$ also implies that $\mathcal{Q}_\infty = \mathsf{ker}(h_e) \cap \mathcal{Q}(h_e)$ is a subspace of the Hilbert space $(\mathcal{Q}(h_b), h_b)$. Since $\mathbf{H}_b^{1/2}$ is the Hilbert space isomorphism of the spaces $(\mathcal{Q}(h_b), h_b)$ and $\mathcal{H}$, the statement: "$\mathcal{Q}_\infty$ is the subspace of $(\mathcal{Q}(h_b), h_b)$," is equivalent to the statement: "$\mathbf{H}_b^{1/2}\mathcal{Q}_\infty$ is a subspace of $\mathcal{H}$". In fact we see that $\mathbf{H}_\infty^\dagger f$, for some $f \in \mathcal{H}$, satisfies the Galerkin condition

$$h_b(\mathbf{H}_\infty^\dagger f, v) = (P_{\mathcal{Q}_\infty} f, v), \qquad v \in \mathcal{Q}_\infty, \tag{3.4.4}$$

since

$$h_b(u, v) = h_\infty(u, v), \qquad u, v \in \mathcal{Q}_\infty.$$

Parallel to the notion of the Galerkin condition for the form h_b, the subspace $\mathcal{Q}_\infty$, and the vector $f \in \mathcal{H}$, there is the notion of the residual of the Galerkin approximation $\mathbf{H}_\infty^\dagger f$. Instead of working with the *Gelfand triple* $\mathcal{Q}(h_b) \hookrightarrow \mathcal{H} \hookrightarrow \mathcal{Q}(h_b)^*$, and thinking about the residual as a functional (an element of $\mathcal{Q}(h_b)^*$), we define the residual as an element of $\mathcal{H}$ and note that $(\mathbf{H}_b^{1/2})^*$ is a Hilbert space isomorphism between the spaces $\mathcal{H}$ and $\mathcal{Q}(h_b)^*$. For $f \in \mathcal{H}$, we define the *middle space residual*

$$r_f \equiv \mathbf{H}_b^{-1/2} f - \mathbf{H}_b^{1/2} \mathbf{H}_\infty^\dagger f. \tag{3.4.5}$$

The middle space residual is the vector such that $r_f \in \mathcal{H}$ and

$$\mathbf{H}_b^{-1/2} f - \mathbf{H}_b^{1/2} \mathbf{H}_\infty^\dagger f = r_f \perp \mathbf{H}_b^{1/2} \mathcal{Q}_\infty. \tag{3.4.6}$$

To check (3.4.5) observe that $\mathbf{H}_\infty^\dagger f \in \mathcal{Q}_\infty$, for every $f \in \mathcal{H}$, and that

$$(\mathbf{H}_b^{1/2} u, \mathbf{H}_b^{1/2} v) = (\mathbf{H}_\infty^{1/2} u, \mathbf{H}_\infty^{1/2} v), \qquad u, v \in \mathcal{Q}_\infty,$$

so finally

$$(\mathbf{H}_b^{-1/2} f, \mathbf{H}_b^{1/2} v) - (\mathbf{H}_b^{1/2} \mathbf{H}_\infty^\dagger f, \mathbf{H}_b^{1/2} v) = (f, v) - (\mathbf{H}_\infty^{1/2} \mathbf{H}_\infty^\dagger f, \mathbf{H}_\infty^{1/2} v) = 0, \qquad v \in \mathcal{Q}_\infty.$$

The relation (3.4.6) is the perpendicularity of the residual to the test space written in the geometry of space $\mathcal{H}$ instead of it being written in the equivalent space $\mathcal{Q}(h_b)^*$, as is usually done. Note that

$$\mathsf{ker}(\mathbf{H}_e^{1/2}\mathbf{H}_b^{-1/2}) = \mathbf{H}_b^{1/2}(\mathsf{ker}(\mathbf{H}_e^{1/2}) \cap \mathcal{Q}(h_b)) = \mathbf{H}_b^{1/2}\mathcal{Q}_\infty,$$

so

$$\mathcal{H} = \mathsf{ran}((\mathbf{H}_e^{1/2}\mathbf{H}_b^{-1/2})^*) \oplus \mathbf{H}_b^{1/2}\mathcal{Q}_\infty$$

and

$$r_f \in \mathsf{ran}((\mathbf{H}_e^{1/2}\mathbf{H}_b^{-1/2})^*). \tag{3.4.7}$$

Given (3.4.2) and (3.4.7), we define $q_f \in \mathcal{H}$ as

$$q_f = (\mathbf{H}_e^{1/2}\mathbf{H}_b^{-1/2})^{*\dagger} r_f. \tag{3.4.8}$$

Relation (3.4.2) implies that $\|(\mathbf{H}_e^{1/2}\mathbf{H}_b^{-1/2})^{*\dagger}\| < \infty$, so we have the estimate

$$\|q_f\| \leq \|(\mathbf{H}_e^{1/2}\mathbf{H}_b^{-1/2})^\dagger\| \|r_f\|. \tag{3.4.9}$$

We can now rewrite (3.4.4) as

$$h_b(\mathbf{H}_\infty^\dagger f, v) + (q_f, \mathbf{H}_e^{1/2} v) = (f, v), \qquad v \in \mathcal{Q}(h_b), \tag{3.4.10}$$

since for every $v \in \mathcal{Q}(h_b)$

$$\begin{aligned} h_b(\mathbf{H}_\infty^\dagger f, v) + (q_f, \mathbf{H}_e^{1/2} v) &= (\mathbf{H}_b^{1/2}\mathbf{H}_\infty^\dagger f, \mathbf{H}_b^{1/2} v) + ((\mathbf{H}_e^{1/2}\mathbf{H}_b^{-1/2})^{*\dagger} r_f, \mathbf{H}_e^{1/2} v) \\ &= (\mathbf{H}_b^{1/2}\mathbf{H}_\infty^\dagger f, \mathbf{H}_b^{1/2} v) + (r_f, (\mathbf{H}_e^{1/2}\mathbf{H}_b^{-1/2})^\dagger (\mathbf{H}_e^{1/2}\mathbf{H}_b^{-1/2}) \mathbf{H}_b^{1/2} v) \\ &= (\mathbf{H}_b^{1/2}\mathbf{H}_\infty^\dagger f + r_f, \mathbf{H}_b^{1/2} v) \\ &= (\mathbf{H}_b^{1/2}\mathbf{H}_\infty^\dagger f + \mathbf{H}_b^{-1/2} f - \mathbf{H}_b^{1/2}\mathbf{H}_\infty^\dagger f, \mathbf{H}_b^{1/2} v) = (f, v). \end{aligned}$$

Now we are ready to give the proof of the main result[10] from [59].

Lemma 3.4.2 (Tambača). *Take $f \in \mathcal{H}$, then*

$$h_\eta[\mathbf{H}_\eta^{-1} f - \mathbf{H}_\infty^\dagger f] \leq \frac{C_1^2}{\eta^2}(f, f)$$

with

$$C_1 = \|(\mathbf{H}_e^{1/2}\mathbf{H}_b^{-1/2})^\dagger\| (\|\mathbf{H}_b^{-1/2}\| + \|\mathbf{H}_\infty^{1/2\dagger}\|).$$

[10] This result appeared subsequently in [36].

PROOF. *(New proof of Lemma 3.4.2.)*
For any $f \in \mathcal{H}$, we have

$$\begin{aligned} h_b(\mathbf{H}_\infty^\dagger f, v) + (q_f, \mathbf{H}_e^{1/2} v) &= (f, v), \qquad v \in \mathcal{Q}(h_b), \\ h_b(\mathbf{H}_\eta^{-1} f, v) + \eta^2 h_e(\mathbf{H}_\eta^{-1} f, v) &= (f, v), \qquad v \in \mathcal{Q}(h_b). \end{aligned}$$

Now, it follows

$$\begin{aligned} h_\eta(\mathbf{H}_\eta^{-1} f - \mathbf{H}_\infty^\dagger f, v) = h_b(\mathbf{H}_\eta^{-1} f - \mathbf{H}_\infty^\dagger f, v) + \eta^2 h_e(\mathbf{H}_\eta^{-1} f, v) = (q_f, \mathbf{H}_e^{1/2} v) \\ \leq \|q_f\| \|\mathbf{H}_e^{1/2} v\| \end{aligned}$$

and in particular

$$\eta^2 h_e[\mathbf{H}_\eta^{-1} f] \leq \|q_f\| h_e[\mathbf{H}_\eta^{-1} f]^{1/2}.$$

This yields

$$h_\eta[\mathbf{H}_\eta^{-1} f - \mathbf{H}_\infty^\dagger f] \leq \frac{\|q_f\|^2}{\eta^2} \tag{3.4.11}$$

which together with (3.4.9) gives

$$h_\eta[\mathbf{H}_\eta^{-1} f - \mathbf{H}_\infty^\dagger f] \leq \frac{\left[\|(\mathbf{H}_e^{1/2} \mathbf{H}_b^{-1/2})^{*\dagger}\| (\|\mathbf{H}_b^{-1/2}\| + \|\mathbf{H}_\infty^{1/2\dagger}\|)\right]^2}{\eta^2} (f, f),$$

which is the desired estimate. □

Remark 3.4.3. A cruder, but simpler estimate of the constant C is

$$C \leq 2\|(\mathbf{H}_e^{1/2} \mathbf{H}_b^{-1/2})^\dagger\| \|\mathbf{H}_b^{-1/2}\|.$$

So, in order for Lemma 3.4.2 to be useful in obtaining the exactly computable estimates we need to assume that h_b be a known object. This is to say that we assume that we have additional information on the geometry of the space $(\mathcal{Q}(h_b), h_b)$. Since we are from the start assuming that h_∞ is the known object, this is a reasonable assumption.

In the next lemma we shall establish the connection between Lemma 3.4.2 and the estimates of the difference between the harmonic Rayleigh quotients, that are needed in Lemma 3.3.1.

Lemma 3.4.4. *Take $f \in \overline{\mathcal{Q}_\infty}$, then*

$$h_\eta[\mathbf{H}_\eta^{-1} f - \mathbf{H}_\infty^\dagger f] = (f, \mathbf{H}_\eta^{-1} f) - (f, \mathbf{H}_\infty^\dagger f) = \|r_f\|^2.$$

PROOF. The proof is a straight forward computation. Take $f \in \overline{\mathcal{Q}_\infty}$, then

$$h_\eta[\mathbf{H}_\infty^\dagger f] = (f, \mathbf{H}_\infty^\dagger f)$$

and we have

$$\begin{aligned} h_\eta[\mathbf{H}_\eta^{-1} f - \mathbf{H}_\infty^\dagger f] &= (f, \mathbf{H}_\eta^{-1} f) - h_\eta(\mathbf{H}_\eta^{-1} f, \mathbf{H}_\infty^\dagger f) - h_\eta(\mathbf{H}_\infty^\dagger f, \mathbf{H}_\eta^{-1} f) + (f, \mathbf{H}_\infty^\dagger f) \\ &= (f, \mathbf{H}_\eta^{-1} f) - (\mathbf{H}_\eta^{-1/2} f, \mathbf{H}_\eta^{1/2} \mathbf{H}_\infty^\dagger f) - (\mathbf{H}_\eta^{1/2} \mathbf{H}_\infty^\dagger f, \mathbf{H}_\eta^{-1/2} f) \\ &\quad + (f, \mathbf{H}_\infty^\dagger f) \\ &= (f, \mathbf{H}_\eta^{-1} f) - (f, \mathbf{H}_\infty^\dagger f). \end{aligned}$$

The other equality follows analogously from (3.4.5). □

Lemma 3.4.4 also implies the known fact

$$(f, \mathbf{H}_\eta^{-1} f) \geq (f, \mathbf{H}_\infty^\dagger f), \qquad f \in \overline{\mathcal{Q}_\infty}.$$

Since $(f, \mathbf{H}_\infty^\dagger f) = 0$, for $f \in \mathcal{Q}_\infty^\perp$, we have established

$$\mathbf{H}_\eta^{-1} - \mathbf{H}_\infty^\dagger \geq 0.$$

This in turn implies the norm resolvent convergence.

Theorem 3.4.5. *Assume* $\|(\mathbf{H}_e^{1/2} \mathbf{H}_b^{-1/2})^\dagger\| < \infty$, *then* $\mathbf{H}_\eta \to \mathbf{H}_\infty$ *in the norm resolvent sense and*

$$\|\mathbf{H}_\eta^{-1} - \mathbf{H}_\infty^\dagger\| \leq \frac{\|(\mathbf{H}_e^{1/2} \mathbf{H}_b^{-1/2})^\dagger\|^2}{\eta^2} \|\mathbf{H}_b^{-1} - \mathbf{H}_\infty^\dagger\|$$

and more specifically

$$\frac{\|r_f^\eta\|^2}{\|r_f\|^2} \leq \frac{\|(\mathbf{H}_e^{1/2} \mathbf{H}_b^{-1/2})^\dagger\|^2}{\eta^2}$$

for $r_f^\eta = \mathbf{H}_\eta^{-1/2} f - \mathbf{H}_\eta^{1/2} \mathbf{H}_\infty^\dagger f$.

PROOF. The operator $\mathbf{H}_\eta^{-1} - \mathbf{H}_\infty^\dagger$ is a nonnegative operator, hence

$$\|\mathbf{H}_\eta^{-1} - \mathbf{H}_\infty^\dagger\| = \sup_{\|f\|=1} (f, (\mathbf{H}_\eta^{-1} - \mathbf{H}_\infty^\dagger) f), \qquad \eta \geq 0.$$

Combining Lemma 3.4.4 with (3.4.8) and (3.4.11) yields

$$\frac{(f, \mathbf{H}_\eta^{-1} f) - (f, \mathbf{H}_\infty^\dagger f)}{(f, \mathbf{H}_b^{-1} f) - (f, \mathbf{H}_\infty^\dagger f)} = \frac{h_\eta[\mathbf{H}_\eta^{-1} f - \mathbf{H}_\infty^\dagger f]}{h_b[\mathbf{H}_b^{-1} f - \mathbf{H}_\infty^\dagger f]} = \frac{\|r_f^\eta\|^2}{\|r_f\|^2} \leq \frac{\|(\mathbf{H}_e^{1/2} \mathbf{H}_b^{-1/2})^\dagger\|^2}{\eta^2}. \tag{3.4.12}$$

This proves both estimates. Now, $\mathbf{H}_b$ was assumed to be a positive definite operator and obviously

$$\mathbf{H}_b \leq \mathbf{H}_\eta.$$

Hence, the convergence in norm of the resolvent at zero implies the norm resolvent convergence of $\mathbf{H}_\eta$, cf. [41]. □

To assess the accuracy of Corollary 3.4.5 consider the following example. We will show that the estimate (3.4.12) is quite sharp. This also shows that the core of the analysis in Lemma 3.4.4, up to the estimate (3.4.11), is quite sharp.

Example 3.4.6. We will present this example as an abstract variation on Problem 3.2. Let $\mathbf{H}$ be a positive definite operator and let P be a projection. Consider

$$h_\eta(u,v) = ((\mathbf{I}+\eta^2 P)\mathbf{H}^{1/2}u, \mathbf{H}^{1/2}v) = h_b(u,v) + \eta^2 h_e(u,v),$$

then

$$\|(\mathbf{H}_e^{1/2}\mathbf{H}_b^{-1/2})^\dagger\| \leq 1$$

and Corollary 3.4.5 gives

$$\frac{\|r_x^\eta\|^2}{\|r_x\|^2} = \frac{(x,\mathbf{H}_\eta^{-1}x) - (x,\mathbf{H}^{-1/2}P_\perp\mathbf{H}^{-1/2}x)}{(x,\mathbf{H}^{-1}x) - (x,\mathbf{H}^{-1/2}P_\perp\mathbf{H}^{-1/2}x)} = \frac{(x,\mathbf{H}_\eta^{-1}x) - (x,\mathbf{H}^{-1/2}P_\perp\mathbf{H}^{-1/2}x)}{(\mathbf{H}^{-1/2}x, P\mathbf{H}^{-1/2}x)} \leq \frac{1}{\eta^2}. \tag{3.4.13}$$

We compute

$$\mathbf{H}_\eta^{-1} = \mathbf{H}^{-1/2}(P_\perp + \frac{1}{1+\eta^2}P)\mathbf{H}^{-1/2}$$

to establish

$$(x,\mathbf{H}_\eta^{-1}x) - (x,\mathbf{H}^{-1/2}P_\perp\mathbf{H}^{-1/2}x) = \frac{1}{1+\eta^2}(\mathbf{H}^{-1/2}x, P\mathbf{H}^{-1/2}x). \tag{3.4.14}$$

Formulae (3.4.13) and (3.4.14) give

$$\frac{\|r_x^\eta\|^2}{\|r_x\|^2} = \frac{1}{1+\eta^2} \leq \frac{1}{\eta^2},$$

which is very favorable estimate for η large.

3.4.1 A model problem from 1D theory of elasticity

As an illustration of the applicability of Lemma 3.4.2, we consider the small frequency problem for the circular arch as described in [19, Chapter 8.8:3] and [51], cf. Figure 3.5.

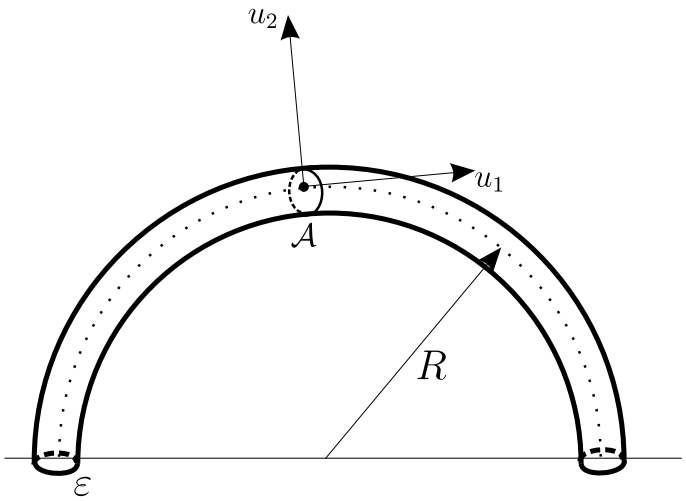

Figure 3.5: The Curved rod model

Let $\phi : [0, l] \to \mathbb{R}^2$ be the middle curve of the arch. We take ϕ to be the upper part of the circle with the radius R. The arch (the model problem we are considering) will be a thin homogeneous, elastic body of the constant cross-section $\mathcal{A}$, whose area is $A > 0$. The arch will be clamped at one end and free at the other, cf. [36]. The strain energy of the arch is given[11] by the positive definite form

$$a(\mathbf{u}, \mathbf{v}) = EI \int_0^l \left(u_2' + \frac{u_1}{R}\right)' \left(v_2' + \frac{v_1}{R}\right)' \, ds + EA \int_0^l \left(u_1' - \frac{u_2}{R}\right) \left(v_1' - \frac{v_2}{R}\right) \, ds, \quad (3.4.15)$$
$$\mathbf{u}, \mathbf{v} \in \mathcal{Q}(a) = \{\mathbf{u} \in H^1[0, l] \times H^2[0, l] \ : \ \mathbf{v}(0) = 0, v_2'(0) = 0\}.$$

Here $\mathbf{u} = (u_1, u_2)$ and $\mathbf{v} = (v_1, v_2)$ are the functions of the curvilinear abscissa $s \in [0, l]$, the constant E is the Young modulus of elasticity, the constant A is the area of the cross-section $\mathcal{A}$ and the constant I is the moment of inertia of the cross-section $\mathcal{A}$.

Let us assume we have the referent arch with the cross-section area A and the cross-section moment I. We consider the family of rods whose cross-section and the moment of inertia of the cross-section behave like

$$A_\eta = \frac{1}{\eta^2} A = \varepsilon^2 A, \qquad I_\eta = \frac{1}{\eta^4} I = \varepsilon^4 I.$$

We want to study the spectral properties of this family of arches as $\varepsilon \to 0$. More general arch models have been examined in [36, 59], cf. [58].

[11] See also [36, 59, 58].

For some given $\eta > 0$, $\eta := 1/\varepsilon$, we write

$$(\mathbf{A}_\eta^{1/2}\mathbf{u}, \mathbf{A}_\eta^{1/2}\mathbf{v}) = a_\eta(\mathbf{u}, \mathbf{v})$$
$$= E\frac{1}{\eta^4}I\int_0^l \left(u_2' + \frac{u_1}{R}\right)'\left(v_2' + \frac{v_1}{R}\right)' \,\mathrm{d}s + E\frac{1}{\eta^2}A\int_0^l \left(u_1' - \frac{u_2}{R}\right)\left(v_1' - \frac{v_2}{R}\right)\, ds$$

and the eigenvalues of the thus defined self adjoint operator will be $\lambda_i(\mathbf{A}_\eta)$, $i = 1, 2, \ldots$, since the operator $\mathbf{A}_\eta$ has only the discrete spectrum. After rescaling

$$\lambda_i(\mathbf{A}_\eta) = \frac{1}{\eta^4}\lambda_i^\eta$$

we see that λ_i^η are the eigenvalues of the operator $\mathbf{H}_\eta$, which is defined by

$$(\mathbf{H}_\eta^{1/2}\mathbf{u}, \mathbf{H}_\eta^{1/2}\mathbf{v}) = h_b(\mathbf{u}, \mathbf{v}) + \eta^2 h_e(\mathbf{u}, \mathbf{v})$$
$$= EI\int_0^l \left(u_2' + \frac{u_1}{R}\right)'\left(v_2' + \frac{v_1}{R}\right)' \, ds + \eta^2 EA\int_0^l \left(u_1' - \frac{u_2}{R}\right)\left(v_1' - \frac{v_2}{R}\right)\, ds$$

for $\mathbf{u}, \mathbf{v} \in \mathcal{Q}(a_A) = \mathcal{Q}(h_\eta)$. Since λ_i^η enable us to describe only the eigenvalues of $\mathbf{A}_\eta$ for which

$$\lim_{\eta\to\infty} \frac{1}{\eta^4}\lambda_i(\mathbf{A}_\eta) < \infty$$

we see where the name "*low frequency problem*", for the eigenvalue problem for $\mathbf{H}_\eta$, comes from. The low frequency problem satisfies the conditions of Theorem 3.2.1, so we conclude that the limiting form is

$$h_\infty(\mathbf{u}, \mathbf{v}) = EI\int_0^l \left(u_2' + \frac{u_1}{R}\right)'\left(v_2' + \frac{v_1}{R}\right)' \,\mathrm{d}s, \quad \mathbf{u}, \mathbf{v} \in \{\mathbf{f} \in \mathcal{Q}(a_\eta), f_1' - \frac{f_2}{R} = 0\}. \quad (3.4.16)$$

In [58] Tambača has shown that (3.4.16) is the strain energy of the Curved rod model and that h_η is positive definite with

$$\mathcal{Q}(h_\eta) = \{\mathbf{u} \in H^1[0, l] \times H^2[0, l] \;:\; \mathbf{v}(0) = 0, v_2'(0) = 0\}.$$

In [36] the estimates of the convergence rate of the "low frequency arch model" eigenvalues to the curved rod eigenvalues have been proved for general middle curves. We shall present the calculation in this specific case only as an illustration of the general theory. However, we refer an interested reader to [36] for more details on Arch Model and Curved Rod Model.

Remark 3.4.7. From (3.4.16) we can see the significance of the condition

$$f_1' - \frac{f_2}{R} = 0. \tag{3.4.17}$$

Assume the rod is locally straight. That is to say, assume $R \to \infty$, then (3.4.17) turns into

$$f_1' = 0,$$

a condition of the inextensibility of the middle curve of the straight rod. The fact that $f_1' - \frac{f_2}{R} = 0$ is an *inextensibility condition* for the middle curve of the curved rod can be established by a rigorous differential geometric argument, see [58]. Continuing this heuristic reasoning, we conclude that Curved rod model describes the transversal vibrations (perpendicular to the middle curve) of the curved rod (as does Straight rod model for the straight rod). Arch model "couples" the longitudinal vibrations of the rod with the transversal vibrations. The asymptotic analysis will show (yet again) a surprising fact, with the sharp growth of the coupling constant it is the decoupling that happens. Longitudinal vibrations correspond to the "*middle frequency problem*", which will not be further considered here.

Based on (3.4.15) and (3.4.16) one concludes that the sequence h_η satisfies the assumptions of Lemma 3.4.2 and Theorem 3.3.4. Here is a word of additional explanation in order. We have formulated all of our results about the forms h_b and h_e based on the representations

$$h_b(u,v) = (\mathbf{H}_b^{1/2}u, \mathbf{H}_b^{1/2}v)_{\mathcal{H}},$$
$$h_e(u,v) = (\mathbf{H}_e^{1/2}u, \mathbf{H}_e^{1/2}v)_{\mathcal{H}}.$$

However, we can represent, as is done in (3.4.16), the forms h_b and h_e with the help of the operators $\mathbf{R}_b : \mathcal{Q}(h_b) \to \mathcal{H}_b$ and $\mathbf{R}_e : \mathcal{Q}(h_e) \to \mathcal{H}_e$. The only assumptions on the operators $\mathbf{R}_b$ (and $\mathbf{R}_e$) is that they have a closed range in the auxiliary Hilbert spaces $\mathcal{H}_b$ (and $\mathcal{H}_e$), cf. [37]. The representation theorem for the nonnegative definite forms implies

$$h_b(u,v) = (\mathbf{H}_b^{1/2}u, \mathbf{H}_b^{1/2}v)_{\mathcal{H}} = (\mathbf{R}_b u, \mathbf{R}_b v)_{\mathcal{H}_b}, \tag{3.4.18}$$
$$h_e(u,v) = (\mathbf{H}_e^{1/2}u, \mathbf{H}_e^{1/2}v)_{\mathcal{H}} = (\mathbf{R}_e u, \mathbf{R}_e v)_{\mathcal{H}_e}. \tag{3.4.19}$$

The relations (3.4.18) and (3.4.19) imply that there exist isometric isomorphisms $Q_b : \mathcal{H}_b \to \mathcal{H}$ and $Q_e : \mathcal{H}_e \to \mathcal{H}$ such that

$$\mathbf{H}_b^{1/2} = Q_b \mathbf{R}_b, \qquad \mathbf{H}_e^{1/2} = Q_e \mathbf{R}_e$$

and in particular

$$(\mathbf{H}_b^{1/2}u, \mathbf{H}_b^{1/2}v)_{\mathcal{H}} = (Q_b\mathbf{R}_b u, Q_b\mathbf{R}_b v)_{\mathcal{H}} = (\mathbf{R}_b u, \mathbf{R}_b v)_{\mathcal{H}_b},$$
$$(\mathbf{H}_e^{1/2}u, \mathbf{H}_e^{1/2}v)_{\mathcal{H}} = (Q_e\mathbf{R}_e u, Q_e\mathbf{R}_e v)_{\mathcal{H}} = (\mathbf{R}_e u, \mathbf{R}_e v)_{\mathcal{H}_e}.$$

We also have for $\mathbf{u} \in \mathcal{Q}(h_b)$

$$Q_b^{-1}\mathbf{H}_b^{1/2}\mathbf{u} = \mathbf{R}_b\mathbf{u} = \Big(u_2' + \frac{u_1}{R}\Big)',$$
$$Q_e^{-1}\mathbf{H}_e^{1/2}\mathbf{u} = \mathbf{R}_e\mathbf{u} = \Big(u_1' - \frac{u_2}{R}\Big)$$

and $\mathbf{R}_b : \mathcal{Q}(h_b) \to \mathcal{H}_b = L^2[0,l]$ and $\mathbf{R}_e : \mathcal{Q}(h_e) \to \mathcal{H}_e = L^2[0,l]$.

Note that $\mathbf{H}_b$ is not positive definite but $\mathbf{H}_1$, which is defined by the form $h_1 = h_b + h_e$, is. For the details see [36, 39, 57]. If we were to change the notation we would have to set $\widetilde{h}_b := h_1$. Since this would unnecessarily complicate the exposition we opt not to do so.

We show that

$$\|(\mathbf{H}_e^{1/2}\mathbf{H}_1^{-1/2})^\dagger\| \leq \frac{\sqrt{1+R^2}}{R} \tag{3.4.20}$$

for our model problem. We adapt the procedure from [36, 59] to the new notation. The statement

$$\|(\mathbf{H}_e^{1/2}\mathbf{H}_1^{-1/2})^\dagger\| \leq k$$

is equivalent to the statement

$$\|(\mathbf{H}_e^{1/2}\mathbf{H}_1^{-1/2})^* q_f\| = \sup_{\mathbf{v}\in\mathcal{Q}(h_b)} \frac{|(q_f, \mathbf{H}_e^{1/2}\mathbf{v})|}{\|\mathbf{H}_1^{1/2}\mathbf{v}\|} \geq \frac{1}{k}\,\|P_{\mathcal{Q}_\infty} q_f\|,$$

since

$$\mathsf{ker}((\mathbf{H}_e^{1/2}\mathbf{H}_1^{-1/2})^*) = \mathsf{ker}(\,\overline{\mathbf{H}_1^{-1/2}\mathbf{H}_e^{1/2}}\,) = \overline{\mathsf{ker}(\mathbf{H}_e^{1/2})} = \overline{\mathcal{Q}_\infty}.$$

For $Q_e^{-1}q_f \in L^2[0,l]$ we define[12] $\mathbf{v}_0 = (\int_0^{(\cdot)}(Q_e^{-1}q_f)(s)ds, 0)$ (an element of $\mathcal{Q}(h_\eta)$). For general $\mathbf{v}$ we have

$$\|\mathbf{H}_1^{1/2}\mathbf{v}\| = \Big(\int_0^l \Big(\big[v_2' + \frac{v_1}{R}\big]'\Big)^2\,ds + \int_0^l \Big(v_1' - \frac{v_2}{R}\Big)^2\,ds\Big)^{1/2}.$$

Now, set $\mathbf{v} = \mathbf{v}_0$ and compute

$$\|\mathbf{H}_1^{1/2}\mathbf{v}_0\| = \frac{\sqrt{1+R^2}}{R}\|q_f\|.$$

[12]The idea to use this construction to compute the estimates from Lemma 3.4.2 is taken from [59].

This establishes

$$\sup_{\mathbf{v}\in\mathcal{Q}(h_b)} \frac{|(q_f, \mathbf{H}_e^{1/2}\mathbf{v})|}{\|\mathbf{H}_1^{1/2}\mathbf{v}\|} \geq \frac{|(q_f, \mathbf{H}_e^{1/2}\mathbf{v}_0)|}{\|\mathbf{H}_1^{1/2}\mathbf{v}_0\|} \geq \frac{R|(Q_e^{-1}q_f, \mathbf{R}_e\mathbf{v}_0)_{L^2}|}{\sqrt{1+R^2}\|q_f\|} = \frac{R}{\sqrt{1+R^2}}\|q_f\|,$$

which completes the proof of (3.4.20).

We are now in a position to derive eigenvalue (eigenvector) estimates. Assume

$$\lambda_{m-1}^{\infty} < \lambda_m^{\infty} = \cdots = \lambda_{m+n-1}^{\infty} < \lambda_{m+n}^{\infty},$$

set $\mathsf{sin}\Theta_\eta := \mathsf{sin}\Theta_\eta(E_\infty(\{\lambda_m^\infty\}))$ and then apply Theorem 3.3.8. One obtains

$$\frac{|\widehat{\lambda}_m^\eta - \lambda_m^\infty|}{\lambda_m^\infty} < \sin^2\Theta_\eta \ \frac{3}{\gamma_c(\lambda_m^\infty)} \ \frac{1}{1-\frac{3\sin\Theta_\eta}{\gamma_s(\lambda_m^\infty)}}.$$

Corollary 3.4.5 and (3.4.20) yield

$$\frac{\|r_x^\eta\|^2}{\|r_x^1\|^2} = \frac{(x,\mathbf{H}_\eta^{-1}x) - (x,\mathbf{H}_\infty^\dagger x)}{(x,\mathbf{H}_1^{-1}x) - (x,\mathbf{H}_\infty^\dagger x)} \leq \frac{\|(\mathbf{H}_e^{1/2}\mathbf{H}_1^{-1/2})^\dagger\|^2}{\eta^2} \leq \varepsilon^2\frac{4(1+R^2)}{3R^2}$$

for any $x \in \mathsf{ran}(E_\infty\{\lambda_m^\infty\})$ of norm one and $\varepsilon \leq 1/2$. Corollary 3.3.3 can be used to compute

$$\sin^2\Theta_\eta \leq \frac{\beta_\eta}{1+\beta_\eta} = \varepsilon^2\frac{4(1+R^2)}{3R^2} \ \frac{\max\limits_{x\in E_\infty(\{\lambda_m^\infty\})}\frac{\|r_x^1\|^2}{h[\mathbf{H}_\infty^\dagger x]}}{1+\frac{\varepsilon^2 4(1+R^2)}{3R^2}\max\limits_{x\in E_\infty(\{\lambda_m^\infty\})}\frac{\|r_x^1\|^2}{h[\mathbf{H}_\infty^\dagger x]}} \leq \frac{K_x(1+R^2)}{R^2}\ \varepsilon^2.$$

The constant K_x is a "local" quantity defined on the referent rod.

Remark 3.4.7 has established that the Curved rod model describes the transversal vibrations of a thin elastic body. The Arch model allows the coupling of transversal and longitudinal movements. However, this analysis has shown that as the diameter ε diminishes, the Arch model converges to the Curved rod model with the speed of convergence that is controlled by ε^2. The conclusion is that, provided we are interested in the transversal vibrations of the rod, we can ignore the Arch model for $\varepsilon \to 0$. The advantage of the Curved rod model is that it is better behaved, with respect to the finite element approximations, cf. Remark 3.3.9. For more on the lower dimensional models in the theory of elasticity see [19, 36, 39, 51, 57, 58, 59].

3.5 Conclusion

We have presented a new abstract framework for an asymptotic analysis of the families of positive definite forms that satisfy the assumptions of Theorem 3.2.1. A couple of applications in Mathematical physics were presented to illustrate the abstract convergence results.

A new regularity criterion for the perturbation h_e, equivalent to the Babuška–Brezzi inf–sup condition, was derived. It was established by rewriting the proof of the variational estimates from [59] in the environment Hilbert space. The new criterion appears to be yielding estimates whose components (constants) have sound physical interpretations, cf. (3.4.20). The introduction of the middle space residual, that was facilitated by the new proof, enabled us to discuss the sharpness of the energy norm estimates from [59] in a new light.

It should also be emphasized that our asymptotic estimates, unlike the asymptotic estimates from [17], are directly dependent on relative quantities like (3.3.21). This makes the constants that are appearing in the asymptotic estimates well scaled.

Contributions in this chapter

The results of this chapter represent one of the applications of the theory from Chapter 2. The study of the spectral asymptotics in the regular case has been influenced by the results from the joint work with Josip Tambača, Zagreb, see [36, 59]. We now list the main contributions in this chapter:

- We provide an explicit invariant subspace estimates for the families h_η that satisfy the assumptions of Theorem 3.2.1, see Section 3.3.
- The new "quadratic" relative estimates for eigenvalues have been interpreted (and derived) in the context of the relative perturbation theory for symmetric forms (in a Hilbert space) from [41], see Theorem 3.3.8 and Section 3.3.1.
- We have introduced a notion of the middle space residual into the analysis of [36, 59] and thus reinterpreted and complemented it, see Section 3.4.
- We have reformulated a Babuška–Brezzi inf–sup condition in terms of generalized inverses. This has enabled us to characterize a class of regular perturbations $\eta^2 h_e$—the so called regular case—which allow sharp residual based convergence estimates, cf. Theorem 3.4.5 and Example 3.4.6.

Chapter 4

Finite element spectral approximations

The title of this chapter could have been: "Finite dimensional approximations of eigenvalues of nonnegative operators". This would have been consistent with the abstract framework we have been using so far. For most of the results of this chapter to hold, h need not generate a differential operator. All of our theory can be expressed in terms of an abstract form h.

However, we feel that such abstract presentation of approximation results would not be particularly illustrative. Instead, we will concentrate on an analysis of finite element methods for the approximation of eigenvalues of nonnegative definite operators in divergence form. This will enable us to better place our theory in the context of other available results. To be precise, we will consider operators $\mathbf{H}$ defined by nonnegative forms

$$h(u,v) = \int_{\mathcal{R}} (A(\cdot)\nabla u)^* \, \nabla v \, dx + \int_{\mathcal{R}} b(\cdot) u \, v \, dx = (\mathbf{H}^{1/2}u, \mathbf{H}^{1/2}v), \qquad u, v \in \mathcal{Q}(h) \subset L^2(\mathcal{R}).$$

Here $\mathcal{R} \subset \mathbb{R}^r$, $r = 1, 2$ is assumed to be bounded, *polygonal region*[1] and $\mathcal{Q}(h)$ is assumed to be dense in $L^2(\mathcal{R})$. The precise definition of $\mathcal{Q}(h)$ is deliberately left vague. This should emphasize that any $\mathcal{Q}(h)$ (any set of boundary conditions) such that h is a nonnegative definite form is admissible. The discussion of the computational details will be concentrated on a (re)consideration of several case studies. The conclusions remain unchanged in a general case, but the discussion is (unnecessarily) more technically involved.

[1] All of the results will also hold in the case when $r = 3$. Then we would speak about the polyhedral regions, see [14]. The model problems that will be analyzed in detail are all for $r = 1, 2$, so in order to avoid a possible misunderstandings we have not stressed the case $r = 3$ in text. Our results will be derived under the abstract assumptions from [14], so they hold at least in the generality considered there.

A word about notation. The *Sobolev seminorm* of order $k \in \mathbb{N}$ is defined by

$$|u|_{k,2} = \sqrt{\sum_{|\alpha|=k} \int |D^\alpha u|^2 \, \mathrm{d}x} \ ,$$

where for *multi index* $\alpha = (\alpha_1, \ldots, \alpha_r)$

$$D^\alpha u = \frac{\partial^{|\alpha|} u}{\partial^{\alpha_1} x_1 \cdots \partial^{\alpha_r} x_r}$$

denotes the standard *weak derivative*. By

$$H^k(\mathcal{R}) = \{u \in L^2(\mathcal{R}) : \|u\|_{k,2} = \sqrt{\sum_{j=0}^{k} |u|_{j,2}^2} < \infty\}$$

we denote the standard *Sobolev spaces*. The space $H_0^1(\mathcal{R})$ is the subspace of the Sobolev space $H^1(\mathcal{R})$ consisting of all the functions that vanish on $\partial\mathcal{R}$ (this is meant in the sense of the *trace operator*). The space $H_0^1(\mathcal{R})$ is assumed to be equipped with the norm $\|u\|_{H_0^1} = |u|_{1,2}$.

Assume now that h is positive definite. The general nonnegative case can be reduced to the positive definite case thanks to Theorems 2.3.12 and 2.4.2. To reduce the notational overhead take $A(\cdot) = \mathbf{I}$, $b(x) = b \in \mathbb{R}$ (of course, b must be such that h remains positive definite) and

$$\mathcal{Q}(h) = H_0^1(\mathcal{R}).$$

Since $\mathcal{R}$ is a polygonal (polyhedral) region $\overline{\mathcal{R}} = \bigcup_{K \in \mathcal{T}_d} K$, where $\mathcal{T}_d$ is the set of closed triangles (tetrahedrons) and

$$d = \max_{K \in \mathcal{T}_d} \operatorname{diam}(K) = \max_{K \in \mathcal{T}_d} d_K.$$

The set $\mathcal{T}_d$ is called a *triangulation of the polygonal domain* $\mathcal{R}$ if it consists of the triangles such that union of these triangles is $\mathcal{R}$ and such that the intersection of two such triangles either consists of a common side or of a common vertex of both triangles or is empty. For a given triangulation $\mathcal{T}_d$ we define the finite dimensional function spaces:

$$\begin{aligned}
\mathcal{V}_{\mathcal{T}_d}^1 &= \{u \in \mathcal{Q} : v|_K \text{ is a linear function}, K \in \mathcal{T}_d, \ u \in \ C(\ \overline{\mathcal{R}}\ )\}, \\
\mathcal{V}_{\mathcal{T}_d}^2 &= \{u \in \mathcal{Q} : v|_K \text{ is a quadratic function}, K \in \mathcal{T}_d, \ u \in \ C(\ \overline{\mathcal{R}}\ )\} \ .
\end{aligned}$$

4.1 Estimates of sinΘ for single vector approximations

Given the form h we will consider two problems.

Problem 4.1. *The stationary problem:* For $f \in L^2(\mathcal{R})$ find $u \in \mathcal{D}(\mathbf{H})$ such that

$$h(u, v) = (f, v), \qquad v \in \mathcal{Q}(h).$$

Problem 4.2. *The eigenvalue problem:* Find (u, λ), where $u \neq 0$ and $\lambda \in \mathbb{R}$, such that

$$h(u, v) = \lambda(u, v), \qquad v \in \mathcal{Q}(h).$$

Obviously, the vector

$$u = \mathbf{H}^{-1} f$$

is the solution of Problem 4.1. On the other hand, (u, λ) solves Problem 4.2 if and only if

$$\mathbf{H}u = \lambda u.$$

Although our prime concern lies with the eigenvalue estimates, Lemma 3.4.4 suggests that Problems 4.1 and 4.2 are intimately connected— through the use of sinΘ we reduce the study of the eigenvalue problem to the study of an auxiliary stationary problem with a special right hand side. Let us explore this statement further.

Assume $\mathcal{Y} = \mathsf{ran}(P) \subset \mathcal{Q}(h)$ is a finite dimensional subspace. In this chapter we shall have to simultaneously consider the main perturbation construction of Chapter 2 for several different subspaces $\mathcal{Y} = \mathsf{ran}(P) \subset \mathcal{Q}(h)$. To ease the understanding we write

$$h_{\mathcal{Y}}(u, v) = h(Pu, Pv) + h(P_\perp u, P_\perp v),$$

for $P_\perp = \mathbf{I} - P$ and $\mathcal{Y} = \mathsf{ran}(P)$, cf. Remark 2.4.3. Also, by $\mathbf{H}_{\mathcal{Y}}$ we denote the positive definite operator defined by $h_{\mathcal{Y}}$.

As a first step, we reformulate Lemma 3.4.4 and introduce the middle space residual (cf. (3.4.5))

$$r_{(f,\mathcal{Y},\mathcal{Q}(h))} = \mathbf{H}^{-1/2} f - \mathbf{H}^{1/2} \mathbf{H}_{\mathcal{Y}}^{-1} f,$$

for the finite dimensional space $\mathcal{Y} \subset \mathcal{Q}(h)$ and $f \in \mathcal{Y}$.

Lemma 4.1.1. *Assume $\mathcal{Y} = \mathsf{ran}(P) \subset \mathcal{Q}(h)$ is finite dimensional. Take $f \in \mathcal{Y}$, then $\mathbf{H}_{\mathcal{Y}}^{-1} f \in \mathcal{Y}$ and*

$$h[\mathbf{H}^{-1} f - \mathbf{H}_{\mathcal{Y}}^{-1} f] = (f, \mathbf{H}^{-1} f) - (f, \mathbf{H}_{\mathcal{Y}}^{-1} f) = \|r_{(f,\mathcal{Y},\mathcal{Q}(h))}\|^2.$$

Furthermore

$$h(\mathbf{H}_{\mathcal{Y}}^{-1} f, v) = (f, v), \quad v \in \mathcal{Y},$$

so $\mathbf{H}_{\mathcal{Y}}^{-1} f \in \mathcal{Y}$ is a Galerkin approximation to the solution of Problem 4.1.

PROOF. The proof is a straight forward reformulation of the proof of Lemma 3.4.4. Since the context is new, we repeat the argument.

Take $f \in \mathcal{Y}$, then $\mathbf{H}_{\mathcal{Y}}^{-1} f \in \mathcal{Y}$ and

$$h[\mathbf{H}_{\mathcal{Y}}^{-1} f] = h_{\mathcal{Y}}[\mathbf{H}_{\mathcal{Y}}^{-1} f] = (f, \mathbf{H}_{\mathcal{Y}}^{-1} f).$$

The rest of the proof follows as in Lemma 3.4.4 :

$$\begin{aligned} h[\mathbf{H}^{-1} f - \mathbf{H}_{\mathcal{Y}}^{-1} f] &= h(\mathbf{H}^{-1} f, \mathbf{H}^{-1} f) - h(\mathbf{H}^{-1} f, \mathbf{H}_{\mathcal{Y}}^{-1} f) \\ &\quad - h(\mathbf{H}_{\mathcal{Y}}^{-1} f, \mathbf{H}^{-1} f) + h(\mathbf{H}_{\mathcal{Y}}^{-1} f, \mathbf{H}_{\mathcal{Y}}^{-1} f) \\ &= (f, \mathbf{H}^{-1} f) - (f, \mathbf{H}_{\mathcal{Y}}^{-1} f) - (\mathbf{H}_{\mathcal{Y}}^{-1} f, f) + h_{\mathcal{Y}}(\mathbf{H}_{\mathcal{Y}}^{-1} f, \mathbf{H}_{\mathcal{Y}}^{-1} f) \\ &= (f, \mathbf{H}^{-1} f) - (f, \mathbf{H}_{\mathcal{Y}}^{-1} f) \ . \end{aligned}$$

The other equality follows by an analogous computation. The property $\mathbf{H}_{\mathcal{Y}}^{-1} f \in \mathcal{Y}$, which is a consequence of Lemma 2.3.2, implies

$$h(\mathbf{H}_{\mathcal{Y}}^{-1} f, v) = h_{\mathcal{Y}}(\mathbf{H}_{\mathcal{Y}}^{-1} f, v) = (f, v), \qquad v \in \mathcal{Y},$$

i.e. $\mathbf{H}_{\mathcal{Y}}^{-1} f$ is a Galerkin approximation from the subspace $\mathcal{Y}$ to the $\mathbf{H}^{-1} f$. □

It will become obvious that Lemma 4.1.1 is an alternative way to state that the residual of the Galerkin approximation is perpendicular to the test space (cf. Lemma 4.2.7).

Any positive definite form h defines the norm

$$\|u\|_E = \sqrt{h[u]}, \qquad u \in \mathcal{Q}(h).$$

Traditionally $\|\cdot\|_E$ is called *the energy norm* on $\mathcal{Q}(h)$ and the expression

$$\|\mathbf{H}^{-1} f - \mathbf{H}_{\mathcal{Y}}^{-1} f\|_E = \sqrt{h[\mathbf{H}^{-1} f - \mathbf{H}_{\mathcal{Y}}^{-1} f]} = \|r_{(f,\mathcal{Y},\mathcal{Q}(h))}\| \tag{4.1.1}$$

constitutes the energy norm of the error of the Galerkin approximation $\mathbf{H}_{\mathcal{Y}}^{-1} f$ to the $\mathbf{H}^{-1} f$.

We will show that estimates of $h[\mathbf{H}^{-1} f - \mathbf{H}_{\mathcal{Y}}^{-1} f]$ can be used to establish eigenvalue estimates. To see this assume $x \in \mathcal{Y}$ is of norm one and $\mathbf{H}_{\mathcal{Y}} x = \mu x$. Insert $f = x$ in (4.1.1) to obtain

$$(x, \mathbf{H}^{-1} x) - \mu^{-1}(x, x) = h[\mathbf{H}^{-1} x - \mathbf{H}_{\mathcal{Y}}^{-1} x].$$

Lemma 4.1.1 implies $(x, \mathbf{H}^{-1} x)^{-1} \leq (x, \mathbf{H}_{\mathcal{Y}}^{-1} x)^{-1}$ which together with some elementary trigonometry yields

$$\sin \Theta(\mathbf{H}^{-1/2} x, \mathbf{H}^{1/2} x) = \frac{(x, \mathbf{H}^{-1} x) - \mu^{-1}(x, x)}{(x, \mathbf{H}^{-1} x)} \leq \sqrt{h[\mathbf{H}^{-1} x - \mathbf{H}_{\mathcal{Y}}^{-1} x] h[x]} \ .$$

From Theorem 2.3.17 we conclude: *If* $x \in \mathcal{Q}$, $\|x\| = 1$, *is such that*

$$h[\mathbf{H}^{-1}x - \mathbf{H}_{\mathcal{Y}}^{-1}x]h[x] < 1,$$

then there exists an eigenvalue λ *of the operator* $\mathbf{H}$ *such that*

$$\frac{|\lambda - h[x]|}{h[x]} \leq \sin\Theta(\mathbf{H}^{-1/2}x, \mathbf{H}^{1/2}x) \leq \sqrt{h[\mathbf{H}^{-1}x - \mathbf{H}_{\mathcal{Y}}^{-1}x]h[x]}\ . \tag{4.1.2}$$

For finer results we need additional assumptions, see Figure 1.2. For instance, assume

$$h[x] < D \leq \lambda_2, \qquad \frac{\sin\Theta(\mathbf{H}^{-1/2}x, \mathbf{H}^{1/2}x)}{1 - \sin\Theta(\mathbf{H}^{-1/2}x, \mathbf{H}^{1/2}x)} < \frac{D - h[x]}{D + h[x]}, \tag{4.1.3}$$

then

$$\frac{|\lambda - h[x]|}{h[x]} \leq \frac{D + h[x]}{D - h[x]} h[\mathbf{H}^{-1}x - \mathbf{H}_{\mathcal{Y}}^{-1}x]h[x]. \tag{4.1.4}$$

There is a vast amount of literature that addresses the problem of estimating the energy norm of the error of the Galerkin approximation to the solution of the stationary problem for the form h, see [14, 60]. With the help of (4.1.2) and (4.1.4) we intend to tap into this knowledge base to establish estimates for eigenvalue approximations.

4.2 Estimates by discrete residuals measures

The problem of estimating

$$h[\mathbf{H}^{-1}x - \mathbf{H}_{\mathcal{Y}}^{-1}x] = \|r_{(x,\mathcal{Y},\mathcal{Q}(h))}\|^2 \tag{4.2.1}$$

is a challenging problem in the analysis of $\mathcal{Q}(h) = H_0^1(\mathcal{R})$. We intend to point out a plausible assumption which will enable us to substitute (4.2.1) by an equivalent problem in a ("carefully" constructed) finite dimensional (function) subspace $\mathcal{V} \subset \mathcal{Q}(h) = H_0^1(\mathcal{R})$. In order to be definite we can safely assume that $\mathcal{Y} = \mathcal{V}_{\mathcal{T}_d}^1$ and $\mathcal{V} = \mathcal{V}_{\mathcal{T}_d}^2$, where $\mathcal{T}_d$ is triangulation of the polygonal domain $\mathcal{R}$. According to [60, Verfürth] the basic ingredients to establish error estimates for the stationary problem are:

1. The measure of the stability of the infinite dimensional variational problem.

2. An error representation formula.

3. Error estimates for an interpolation operator under minimal regularity assumptions.

An analysis of eigenvalue approximations will be performed according to this agenda. The main tool will be the theory of Chapter 2 and the conclusions will be estimates like (4.1.4). A quick and rough answer to the problem of how to adapt the agenda from [60, Verfürth] to tackle eigenvalue estimates would be (assuming we are approximating λ_1!):

1. $\frac{D+h[x]}{D-h[x]}$ measures the stability of the problem of the approximation of λ_1.
2. (4.1.2) and (4.1.4) are the error representation formulas.
3. The analysis of the Galerkin approximations to the stationary problem for h with the test vector x as the right hand side should take care of all of the remaining details.

This answer does not literally correspond to the given paradigm. This is due to the fact that the eigenvalue problem has more complex structure than the stationary problem. The stationary problem is better understood than the eigenvalue problem. Therefore, our aim is to reduce the analysis of the eigenvalue problem to the analysis of special auxiliary stationary problems. Those auxiliary problems will then be analyzed by known techniques from the literature. The subspace approximation estimates will also be considered.

We will present the analysis for a general positive definite h, attempting at the same time to keep the notational burden to the minimum. To compensate for this abstractness a detailed discussion of the estimates will be performed on several model problems. In Section 4.3.1 we will concentrate on differential operators on regular domains[2] and the numerical results will be presented in full detail.

In the discussion that follows we have, in particular, been influenced by [14, 48, 67]. Let $\mathcal{Y} \subset \mathcal{Q}(h)$ and $\mathcal{Z} \subset \mathcal{Q}(h)$ be finite dimensional spaces. Consider the subspace

$$\mathcal{V} = \mathcal{Y} + \mathcal{Z}$$

as an enlargement of $\mathcal{Y}$ and consider the operators $\mathbf{H}_{\mathcal{Y}}$ and $\mathbf{H}_{\mathcal{V}}$. Take $x \in \mathcal{Y}$, then

$$(\mathbf{H}_{\mathcal{V}})^{-1}_{\mathcal{Y}}\ x = \mathbf{H}^{-1}_{\mathcal{Y}}\ x$$

and Lemma 4.1.1 implies

$$h_{\mathcal{V}}[\mathbf{H}^{-1}_{\mathcal{V}}x - \mathbf{H}^{-1}_{\mathcal{Y}}x] = h_{\mathcal{V}}[\mathbf{H}^{-1}_{\mathcal{V}}x - (\mathbf{H}_{\mathcal{V}})^{-1}_{\mathcal{Y}}x] = (x, \mathbf{H}^{-1}_{\mathcal{V}}x) - (x, \mathbf{H}^{-1}_{\mathcal{Y}}x). \tag{4.2.2}$$

This is an identity in which only the objects that "live" on $\mathcal{V}$ feature. Inspired by this observation we define the *discrete residual* of the vector $x \in \mathcal{Y}$ as

$$r_{(x,\mathcal{Y},\mathcal{V})} = \mathbf{H}^{-1/2}_{\mathcal{V}}x - \mathbf{H}^{1/2}_{\mathcal{V}}\mathbf{H}^{-1}_{\mathcal{Y}}x.$$

[2]The regular domain is the one which is additionally assumed to be convex, or to have a smooth boundary (in the sense of Miranda-Talenti Theorem from [7]).

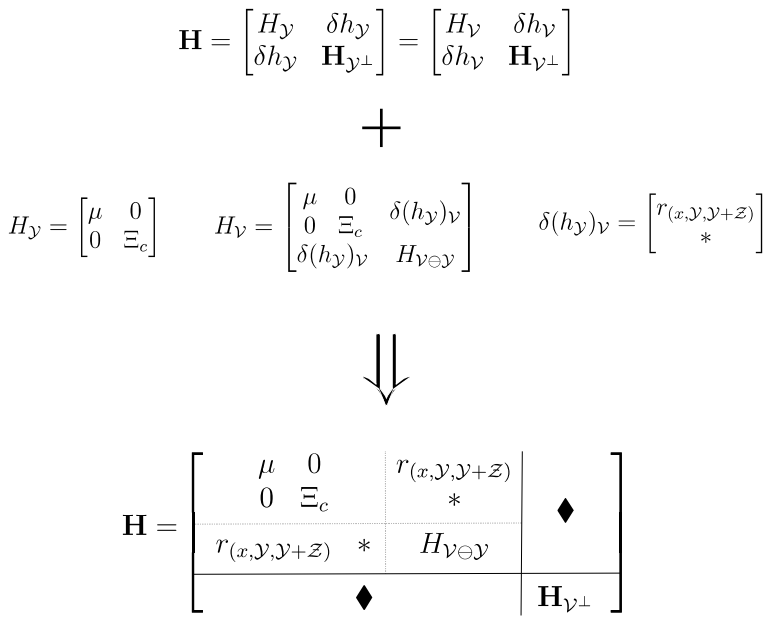

Figure 4.1: Measuring the accuracy of by a discrete residual.

Lemma 4.1.1 is applicable to the discrete residual $r_{(x,\mathcal{Y},\mathcal{V})}$, so we obtain

$$h_{\mathcal{V}}[\mathbf{H}_{\mathcal{V}}^{-1}x - \mathbf{H}_{\mathcal{Y}}^{-1}x] = \|r_{(x,\mathcal{Y},\mathcal{V})}\|^2. \tag{4.2.3}$$

Since $\mathbf{H}_{\mathcal{V}}^{-1}x, \mathbf{H}_{\mathcal{Y}}^{-1}x \in \mathcal{V}$, (4.2.3) can be written as

$$\|\mathbf{H}_{\mathcal{V}}^{-1}x - \mathbf{H}_{\mathcal{Y}}^{-1}x\|_E^2 = h[\mathbf{H}_{\mathcal{V}}^{-1}x - \mathbf{H}_{\mathcal{Y}}^{-1}x] = \|r_{(x,\mathcal{Y},\mathcal{V})}\|^2. \tag{4.2.4}$$

We aim to use (4.2.4) to derive an upper bound on $\|\mathbf{H}^{-1}x - \mathbf{H}_{\mathcal{Y}}^{-1}x\|_E$.

Not every enlargement $\mathcal{V} = \mathcal{Y}+\mathcal{Z}$ will lead to the desired estimate. We need to quantify the situation in which the space $\mathcal{Z}$ "captures" enough information, so that $\|r_{(x,\mathcal{Y},\mathcal{Q})}\|^2$ is bounded in terms of the "finite dimensional" quantity $\|r_{(x,\mathcal{Y},\mathcal{Y}+\mathcal{Z})}\|^2$, cf. Figure 4.1. To get the estimate we need we have to utilize the *saturation assumption*.

Assumption 4.2.1. Take the finite dimensional subspaces $\mathcal{V} \subset \mathcal{Q}(h)$ and $\mathcal{Y} \subset \mathcal{V}$. We say that the subspaces $\mathcal{Y}$ and $\mathcal{V}$ satisfy the saturation assumption with regard to $x \in \mathcal{Y}$ and $\mathbf{H}$ when there exists $0 \leq \beta_{\mathrm{s},x} < 1$ such that

$$\|\mathbf{H}^{-1}x - \mathbf{H}_{\mathcal{V}}^{-1}x\|_E \leq \beta_{\mathrm{s},x}\|\mathbf{H}^{-1}x - \mathbf{H}_{\mathcal{Y}}^{-1}x\|_E.$$

Assumption 4.2.1 states that the larger space $\mathcal{V} \supset \mathcal{Y}$ leads to a better approximation $\mathbf{H}_{\mathcal{V}}^{-1}x \neq \mathbf{H}_{\mathcal{Y}}^{-1}x$ of $\mathbf{H}^{-1}x$. According to [14, Theorem 2.1], Assumption 4.2.1 is equivalent

to any of the following

$$\|\mathbf{H}^{-1}x - \mathbf{H}_{\mathcal{Y}}^{-1}x\|_E \leq (1-\beta_{\mathrm{s},x}^2)^{-1/2}\|\mathbf{H}_{\mathcal{V}}^{-1}x - \mathbf{H}_{\mathcal{Y}}^{-1}x\|_E, \tag{4.2.5}$$

$$\|\mathbf{H}^{-1}x - f\|_E \leq (1-\beta_{\mathrm{s},x}^2)^{-1/2}\|\mathbf{H}_{\mathcal{V}}^{-1}x - f\|_E, \qquad f \in \mathcal{Y}. \tag{4.2.6}$$

On the other hand, [14, Proposition 2.1] gives

$$\|\mathbf{H}_{\mathcal{V}}^{-1}x - \mathbf{H}_{\mathcal{Y}}^{-1}x\|_E \leq \|\mathbf{H}^{-1}x - \mathbf{H}_{\mathcal{Y}}^{-1}x\|_E. \tag{4.2.7}$$

Inequalities (4.2.7) and (4.2.5) jointly imply

$$\|\mathbf{H}_{\mathcal{V}}^{-1}x - \mathbf{H}_{\mathcal{Y}}^{-1}x\|_E \leq \|\mathbf{H}^{-1}x - \mathbf{H}_{\mathcal{Y}}^{-1}x\|_E \leq (1-\beta_{\mathrm{s},x}^2)^{-1/2}\|\mathbf{H}_{\mathcal{V}}^{-1}x - \mathbf{H}_{\mathcal{Y}}^{-1}x\|_E \tag{4.2.8}$$

and this property is, according to [14, Theorem 2.1], equivalent to the saturation assumption.

More on the makeup of $\beta_{\mathrm{s},x}$, as well as on how to chose the spaces $\mathcal{V} = \mathcal{Y} + \mathcal{Z}$, can be found in [14, 25, 60]. The analysis so far can be summed up in a lemma whose proof follows from (4.2.2) and [14, Theorem 2.1].

Lemma 4.2.2. *Let $\mathcal{Y}$ and $\mathcal{V}$ be the subspaces which satisfy the Assumption 4.2.1 with regard to $x \in \mathcal{Y}$ and $\mathbf{H}$, then*

$$h_{\mathcal{V}}[\mathbf{H}_{\mathcal{V}}^{-1}x - \mathbf{H}_{\mathcal{Y}}^{-1}x] \leq h[\mathbf{H}^{-1}x - \mathbf{H}_{\mathcal{Y}}^{-1}x] \leq (1-\beta_{s,x}^2)^{-1}h_{\mathcal{V}}[\mathbf{H}_{\mathcal{V}}^{-1}x - \mathbf{H}_{\mathcal{Y}}^{-1}x] \tag{4.2.9}$$

$$(x, \mathbf{H}_{\mathcal{V}}^{-1}x - \mathbf{H}_{\mathcal{Y}}^{-1}x) \leq (x, \mathbf{H}^{-1}x - \mathbf{H}_{\mathcal{Y}}^{-1}x) \leq (1-\beta_{s,x}^2)^{-1}(x, \mathbf{H}_{\mathcal{V}}^{-1}x - \mathbf{H}_{\mathcal{Y}}^{-1}x) \tag{4.2.10}$$

$$\|r_{(x,\mathcal{Y},\mathcal{V})}\|^2 \leq \|r_{(x,\mathcal{Y},\mathcal{Q})}\|^2 \leq (1-\beta_{s,x}^2)^{-1}\|r_{(x,\mathcal{Y},\mathcal{V})}\|^2. \tag{4.2.11}$$

Furthermore, each of (4.2.9), (4.2.10) and (4.2.11) are equivalent to the saturation assumption.

This lemma, in particular statement (4.2.11), says that we can use the discrete residual $r_{(f,\mathcal{Y},\mathcal{V})}$ to measure the middle space residual $r_{(f,\mathcal{Y},\mathcal{Q})}$ if and only if $\mathcal{Y}$ and $\mathcal{V}$ satisfy the saturation assumption with regard to f and $\mathbf{H}$. Assumption 4.2.1 is the minimal regularity assumption need to establish the estimate (4.2.8) or (4.2.11). We will now formulate an appropriate equivalent of this lemma in the case of eigenvalue estimates.

Theorem 4.2.3. *Let $x \in \mathcal{Y}$ be of norm one and let $\mathcal{Y} \subset \mathcal{V}$ be such that they satisfy Assumption 4.2.1 for $x \in \mathcal{Y}$ and $\mathbf{H}$. Assume $\mathbf{H}_{\mathcal{Y}}\, x = \mu\, x$, then*

$$\sin^2\Theta(\mathbf{H}_{\mathcal{V}}^{-1/2}\, x, \mathbf{H}_{\mathcal{V}}^{1/2}\, x) \leq \sin^2\Theta(\mathbf{H}^{-1/2}x, \mathbf{H}^{1/2}x) \leq (1-\beta_{s,x}^2)^{-1}\sin^2\Theta(\mathbf{H}_{\mathcal{V}}^{-1/2}\, x, \mathbf{H}_{\mathcal{V}}^{1/2}\, x)\ .$$

PROOF. Lemma 4.2.2 and (4.2.2) yield

$$\begin{aligned}\sin^2\Theta(\mathbf{H}^{-1/2}x,\mathbf{H}^{1/2}x) &\le \frac{h[\mathbf{H}^{-1}x-\mathbf{H}_{\mathcal{Y}}^{-1}\,x]}{(x,\mathbf{H}_{\mathcal{V}}^{-1}x)} \le \frac{(1-\beta_{\mathrm{s},x}^2)^{-1}h_{\mathcal{V}}[\mathbf{H}_{\mathcal{V}}^{-1}x-\mathbf{H}_{\mathcal{Y}}^{-1}x]}{(x,\mathbf{H}_{\mathcal{V}}^{-1}x)}\\ &\le \frac{(1-\beta_{\mathrm{s},x}^2)^{-1}((x,\mathbf{H}_{\mathcal{V}}^{-1}x)-(x,\mathbf{H}_{\mathcal{Y}}^{-1}x))}{(x,\mathbf{H}_{\mathcal{V}}^{-1}\,x)}\\ &\le (1-\beta_{\mathrm{s},x}^2)^{-1}\sin^2\Theta(\mathbf{H}_{\mathcal{V}}^{-1/2}\,x,\mathbf{H}_{\mathcal{V}}^{1/2}\,x).\end{aligned}$$

On the other hand,

$$\frac{(x,\mathbf{H}_{\mathcal{Y}}^{-1}x)}{(x,\mathbf{H}_{\mathcal{V}}^{-1}x)} \ge \frac{(x,\mathbf{H}_{\mathcal{Y}}^{-1}x)}{(x,\mathbf{H}^{-1}x)}, \qquad x\in\mathcal{Y},$$

so

$$1-\frac{(x,\mathbf{H}_{\mathcal{Y}}^{-1}x)}{(x,\mathbf{H}^{-1}x)} \ge 1-\frac{(x,\mathbf{H}_{\mathcal{Y}}^{-1}x)}{(x,\mathbf{H}_{\mathcal{V}}^{-1}x)}$$

and the other inequality is proved. □

We will now formulate a sort of a converse to this result.

Proposition 4.2.4. *Assume* $\mathbf{H}_{\mathcal{Y}}\,x=\mu\,x$ *and set*

$$C_s=\frac{\sin^2\Theta(\mathbf{H}^{-1/2}x,\mathbf{H}^{1/2}x)}{\sin^2\Theta(\mathbf{H}_{\mathcal{V}}^{-1/2}x,\mathbf{H}_{\mathcal{V}}^{1/2}x)}.$$

Then Assumption 4.2.1 holds with

$$\beta_{\mathrm{s},x}=\sqrt{1-\frac{(x,\mathbf{H}_{\mathcal{V}}^{-1}x)}{C_s(x,\mathbf{H}^{-1}x)}}.$$

PROOF.

Let $C_s>1$ be such that

$$\sin^2\Theta(\mathbf{H}_{\mathcal{V}}^{-1/2}\,x,\mathbf{H}_{\mathcal{V}}^{1/2}\,x)\le\sin^2\Theta(\mathbf{H}^{-1/2}x,\mathbf{H}^{1/2}x)\le C_s\sin^2\Theta(\mathbf{H}_{\mathcal{V}}^{-1/2}\,x,\mathbf{H}_{\mathcal{V}}^{1/2}\,x),$$

then

$$\frac{h[\mathbf{H}^{-1}x-\mathbf{H}_{\mathcal{Y}}^{-1}\,x]}{(x,\mathbf{H}^{-1}x)}\le C_s\frac{h_{\mathcal{V}}[\mathbf{H}_{\mathcal{V}}^{-1}x-\mathbf{H}_{\mathcal{Y}}^{-1}x]}{(x,\mathbf{H}_{\mathcal{V}}^{-1}x)}.$$

This is to say that

$$h[\mathbf{H}^{-1}x-\mathbf{H}_{\mathcal{Y}}^{-1}\,x]\le C_s\frac{(x,\mathbf{H}^{-1}x)}{(x,\mathbf{H}_{\mathcal{V}}^{-1}x)}h_{\mathcal{V}}[\mathbf{H}_{\mathcal{V}}^{-1}x-\mathbf{H}_{\mathcal{Y}}^{-1}x]$$

and

$$C_s\frac{(x,\mathbf{H}^{-1}x)}{(x,\mathbf{H}_{\mathcal{V}}^{-1}x)}>1.$$

Subsequently, there exists $0 \leq \beta_{\mathrm{s},x} < 1$ such that

$$C_s \frac{(x, \mathbf{H}^{-1}x)}{(x, \mathbf{H}_{\mathcal{V}}^{-1}x)} = (1 - \beta_{\mathrm{s},x}^2)^{-1}.$$

The rest of the proof follows from Lemma 4.2.2. □

The eigenvalue estimates, under the minimal regularity assumptions, can now be obtained as a consequence of Theorem 4.2.3 and the results of Chapter 2.

Corollary 4.2.5. *Let $x \in \mathcal{Y}$ be of norm one and let $\mathcal{Y}$ and $\mathcal{V}$ be such that $\mathcal{Y} \subset \mathcal{V}$ and that they satisfy Assumption 4.2.1 for $x \in \mathcal{Y}$ and $\mathbf{H}$. Also, let $\mathbf{H}_{\mathcal{Y}} x = \mu x$.*

1. *Assume $(1 - \beta_{s,x}^2)^{-1/2} \mathsf{sin}\Theta(\mathbf{H}_{\mathcal{V}}^{-1/2}\ x, \mathbf{H}_{\mathcal{V}}^{1/2}\ x) < 1$, then there exists an eigenvalue λ of the operator $\mathbf{H}$ such that*

$$\frac{|\lambda - \mu|}{\mu} \leq (1 - \beta_{s,x}^2)^{-1/2} \sin\Theta(\mathbf{H}_{\mathcal{V}}^{-1/2}\ x, \mathbf{H}_{\mathcal{V}}^{1/2}\ x).$$

2. *Assume $\lambda_1 < D \leq \lambda_2$ and*

$$\frac{(1 - \beta_{s,x}^2)^{-1/2} \sin\Theta(\mathbf{H}_{\mathcal{V}}^{-1/2}\ x, \mathbf{H}_{\mathcal{V}}^{1/2}\ x)}{1 - (1 - \beta_{s,x}^2)^{-1/2} \sin\Theta(\mathbf{H}_{\mathcal{V}}^{-1/2}\ x, \mathbf{H}_{\mathcal{V}}^{1/2}\ x)} < \frac{D - \mu}{D + \mu}$$

then

$$\frac{|\lambda_1 - \mu|}{\mu} \leq \frac{D + \mu}{D - \mu} (1 - \beta_{s,x}^2)^{-1} \sin^2\Theta(\mathbf{H}_{\mathcal{V}}^{-1/2}\ x, \mathbf{H}_{\mathcal{V}}^{1/2}\ x)$$

and

$$\sin\Theta(E(\lambda_1), x) \leq \frac{(1 - \beta_{s,x}^2)^{-1/2} \sin\Theta(\mathbf{H}_{\mathcal{V}}^{-1/2}\ x, \mathbf{H}_{\mathcal{V}}^{1/2}\ x)}{\sqrt{1 - (1 - \beta_{s,x}^2)^{-1/2} \sin\Theta(\mathbf{H}_{\mathcal{V}}^{-1/2}\ x, \mathbf{H}_{\mathcal{V}}^{1/2}\ x)}} \frac{\sqrt{D\ h[h]}}{D + h[h]}.$$

Similar estimates can be formulated for other eigenvalues λ of multiplicity one.

In order for Theorem 4.2.3 to make a basis to develop a computational procedure, we require a computationally inexpensive way to estimate

$$\sin^2\Theta(\mathbf{H}_{\mathcal{V}}^{-1/2}\ x, \mathbf{H}_{\mathcal{V}}^{1/2}\ x) = \frac{x^* H_{\mathcal{V}}^{-1} x - x^* H_{\mathcal{Y}}^{-1} x}{x^* H_{\mathcal{V}}^{-1} x}.$$

Here we use

$$H_{\mathcal{V}} = \mathbf{H}_{\mathcal{V}}|_{\mathcal{V}}, \qquad H_{\mathcal{Y}} = H_{\mathcal{V}}|_{\mathcal{Y}}$$

and write $x^* H_{\mathcal{V}}^{-1} x$ and $x^* H_{\mathcal{Y}}^{-1} x$, for $x \in \mathcal{X}$, as we did in Section 2.7.3.

There are many ways to tackle this problem, see [14, 48, 60]. We follow [48] in so much as we will also explore the possibility to use (hierarchical basis) preconditioners in the context of efficient evaluation of

$$x^* H_{\mathcal{V}}^{-1} x - x^* H_{\mathcal{Y}}^{-1} x = h[\mathbf{H}_{\mathcal{V}}^{-1} x - \mathbf{H}_{\mathcal{Y}}^{-1} x] = h_{\mathcal{V}}[\mathbf{H}_{\mathcal{V}}^{-1} x - \mathbf{H}_{\mathcal{Y}}^{-1} x].$$

So far we have established a method to compute an estimate to the eigenvalue that is approximated by $x \in \mathcal{Y}$. What remains is to show that Lemmata 4.1.1 and 4.2.2 can also be modified to obtain the subspace approximation estimates for some $\mathcal{X} \subset \mathcal{Y}$.

4.2.1 Bounding sinΘ for subspace approximations

In the last section we have analyzed the eigenvalue approximations by the Ritz value associated with the vector

$$x \in \mathcal{Y} \subset \mathcal{V}.$$

Here it was assumed that the subspaces $\mathcal{V}$ and $\mathcal{Y}$ satisfy the saturation assumption (Assumption 4.2.1). The saturation assumption (Assumption 4.2.1) is dependent upon the vector $x \in \mathcal{Y}$. To emphasize this fact we use $\beta_{s,x}$ to denote the saturation constant in (4.2.1).

The new setting is

$$\mathcal{X} \subset \mathcal{Y} \subset \mathcal{V} \tag{4.2.12}$$

and we will modify the notation from Chapter 2 accordingly. Let $P = XX^*$ and $\mathsf{ran}(P) = \mathcal{X}$ be the usual representation of some n-dimensional subspace $\mathcal{X}$. We are using the Rayleigh quotient[3]

$$\Xi = X^* \mathbf{H}_{\mathcal{X}} X (\simeq \mathbf{H}_{\mathcal{X}}|_{\mathcal{X}} : \mathcal{X} \to \mathcal{X})$$

to compute the Ritz values. We want to assess the quality of these Ritz values, when considered as approximate eigenvalues. We will be using the formula

$$\sin^2 \Theta(\mathbf{H}^{1/2} X, \mathbf{H}^{-1/2} X) = \max_{x \in \mathbb{R}^n \setminus \{0\}} \frac{|x^*(\Omega - \Xi^{-1})x|}{x^* \Omega x} = \max_{x \in \mathcal{X} \setminus \{0\}} \frac{(x, \mathbf{H}^{-1} x) - (x, \mathbf{H}_{\mathcal{X}}^{-1}\, x)}{(x, \mathbf{H}^{-1} x)}, \tag{4.2.13}$$

where $\Omega = X^* \mathbf{H}^{-1} X (\simeq P \mathbf{H}^{-1} P|_{\mathcal{X}})$. Set $\Omega_{\mathcal{V}} = X^* \mathbf{H}_{\mathcal{V}}^{-1} X$, then

$$\sin^2 \Theta(\mathbf{H}_{\mathcal{V}}^{-1/2} X, \mathbf{H}_{\mathcal{V}}^{1/2} X) = \max_{x \in \mathbb{R}^n \setminus \{0\}} \frac{|x^*(\Omega_{\mathcal{V}} - \Xi^{-1})x|}{x^* \Omega_{\mathcal{V}} x} = \max_{x \in \mathcal{X} \setminus \{0\}} \frac{(x, \mathbf{H}_{\mathcal{V}}^{-1} x) - (x, \mathbf{H}_{\mathcal{X}}^{-1} x)}{(x, \mathbf{H}_{\mathcal{V}}^{-1}\, x)}. \tag{4.2.14}$$

[3] In this Chapter we are deliberately vague about wether $\Xi \in \mathbb{C}^{n \times n}$ or $\Xi : \mathcal{X} \to \mathcal{X}$.

As in the stationary and single vector case, we aim to relate $\mathsf{sin}\Theta(\mathbf{H}_{\mathcal{V}}^{-1/2}X, \mathbf{H}_{\mathcal{V}}^{1/2}X)$ and $\mathsf{sin}\Theta := \mathsf{sin}\Theta(\mathbf{H}^{-1/2}X, \mathbf{H}^{1/2}X)$ and thus reduce the problem of estimating $\mathsf{sin}\Theta$ to the matrix problem. In the structure of these estimates the subspaces $\mathcal{V}$ and $\mathcal{Y}$ will feature only through the saturation property that will enable us to perform the reduction. In this case, however, we will need a subspace saturation property.

Assumption 4.2.6. Take the finite dimensional subspaces $\mathcal{V} \subset \mathcal{Q}(h)$ and $\mathcal{Y} \subset \mathcal{V}$. We say that the subspaces $\mathcal{Y}$ and $\mathcal{V}$ satisfy the *subspace saturation assumption* with regard to $\mathcal{X} \subset \mathcal{Y}$ and $\mathbf{H}$ when there exists $0 \leq \beta_{\mathrm{s},\mathcal{X}} < 1$ such that

$$\|\mathbf{H}^{-1}x - \mathbf{H}_{\mathcal{V}}^{-1}x\|_E \leq \beta_{\mathrm{s},\mathcal{X}}\|\mathbf{H}^{-1}x - \mathbf{H}_{\mathcal{Y}}^{-1}x\|_E, \qquad x \in \mathcal{X}.$$

Let the subspace $\mathcal{X}$ satisfy the Assumption 4.2.6, then each of the vectors $x \in \mathcal{X}$ satisfies the Assumption 4.2.1 with the saturation constant $\beta_{\mathrm{s},x} = \beta_{\mathrm{s},\mathcal{X}}$. Now we will modify Lemma 4.1.1 for subspace approximations.

Lemma 4.2.7. *Let $\mathcal{X} := \mathsf{ran}(P) \subset \mathcal{Q}(h)$ be any n-dimensional subspace. Take $x \in \mathcal{X}$, then*

$$h(\mathbf{H}^{-1}x - \mathbf{H}_{\mathcal{X}}^{-1}x, v) = 0, \qquad v \in \mathcal{X}.$$

Also, let $x, y \in \mathcal{X}$, then

$$(x, \mathbf{H}^{-1}y) - (x, \mathbf{H}_{\mathcal{X}}^{-1}y) = h(\mathbf{H}^{-1}x - \mathbf{H}_{\mathcal{X}}^{-1}x, \mathbf{H}^{-1}y - \mathbf{H}_{\mathcal{X}}^{-1}y).$$

Proof. Take $x, y \in \mathcal{X}$, then

$$\begin{aligned} h(\mathbf{H}^{-1}x - \mathbf{H}_{\mathcal{X}}^{-1}x, y) &= h(\mathbf{H}^{-1}x, y) - h(\mathbf{H}_{\mathcal{X}}^{-1}x, y) \\ &= (x, y) - h_{\mathcal{X}}(\mathbf{H}_{\mathcal{X}}^{-1}x, y) = (x, y) - (x, y) = 0. \end{aligned} \tag{4.2.15}$$

This is also known as the orthogonality property of the Galerkin approximation. With the help of (4.2.15) we prove

$$\begin{aligned} h(\mathbf{H}^{-1}x - \mathbf{H}_{\mathcal{X}}^{-1}x, \mathbf{H}^{-1}y - \mathbf{H}_{\mathcal{X}}^{-1}y) &= h(\mathbf{H}^{-1}x - \mathbf{H}_{\mathcal{X}}^{-1}x, \mathbf{H}^{-1}y) \\ &= (\mathbf{H}^{-1}x - \mathbf{H}_{\mathcal{X}}^{-1}x, y). \end{aligned}$$

□

Lemma 4.2.7 and (4.2.14) can now be used to obtain an approximation theorem for $\mathsf{sin}\Theta$.

Theorem 4.2.8. *Take $\mathcal{X} := \mathsf{ran}(X) \subset \mathcal{Y}$, $P = XX^*$ and let $\mathcal{Y}$ and $\mathcal{V}$ be such that $\mathcal{Y} \subset \mathcal{V}$ and that they satisfy Assumption 4.2.6 for $\mathcal{X}$ and $\mathbf{H}$. If $\mathbf{H}_{\mathcal{Y}}X = X\Xi$ then*

$$\sin^2\Theta(\mathbf{H}_{\mathcal{V}}^{-1/2}X, \mathbf{H}_{\mathcal{V}}^{1/2}X) \leq \sin^2\Theta \leq (1 - \beta_{\mathsf{s},\mathcal{X}}^2)^{-1}\sin^2\Theta(\mathbf{H}_{\mathcal{V}}^{-1/2}X, \mathbf{H}_{\mathcal{V}}^{1/2}X).$$

PROOF. The proof is based on Lemma 4.2.7, (4.2.13) and (4.2.14). We compute

$$\begin{aligned}\sin^2\Theta(\mathbf{H}^{-1/2}X,\mathbf{H}^{1/2}X) &= \max_{x\in\mathcal{X}\setminus\{0\}}\frac{(x,\mathbf{H}^{-1}x)-(x,\mathbf{H}_{\mathcal{X}}^{-1}\,x)}{(x,\mathbf{H}^{-1}x)}\\ &= \max_{x\in\mathcal{X}\setminus\{0\}}\frac{(x,\mathbf{H}^{-1}x)-(x,\mathbf{H}_{\mathcal{Y}}^{-1}\,x)}{(x,\mathbf{H}^{-1}x)}\\ &\le \frac{(1-\beta_{\mathrm{s},\mathcal{X}}^2)^{-1}((x,\mathbf{H}_{\mathcal{V}}^{-1}x)-(x,\mathbf{H}_{\mathcal{Y}}^{-1}x))}{(x,\mathbf{H}_{\mathcal{V}}^{-1}\,x)}\\ &\le (1-\beta_{\mathrm{s},\mathcal{X}}^2)^{-1}\sin^2\Theta(\mathbf{H}_{\mathcal{V}}^{-1/2}\,X,\mathbf{H}_{\mathcal{V}}^{1/2}\,X).\end{aligned}$$

As before we have used the fact $\Omega \ge \Omega_{\mathcal{V}}$. This can be proved from

$$(x,\mathbf{H}^{-1}x)-(x,\mathbf{H}_{\mathcal{V}}^{-1}\,x) = h[\mathbf{H}^{-1}x-\mathbf{H}_{\mathcal{X}}^{-1}x]\ge 0,\qquad x\in\mathcal{X}.$$

Lemma 4.2.7 reveals the structure of the matrix $\Omega-\Omega_{\mathcal{V}}$ in full detail. An argument similar to the one that led to the proof of Theorem 4.2.3, implies the other inequality. □

Analogously as before we establish that Assumption 4.2.6 was the minimal regularity requirement necessary to perform the analysis we wanted to perform. Since $\beta_{\mathrm{s},x}=\beta_{\mathrm{s},\mathcal{X}}$, can be used as the single vector saturation constant for every $x\in\mathcal{X}$, the following result can be established.

Proposition 4.2.9. *Assume* $\mathbf{H}_{\mathcal{Y}}X = X\Xi$ *and define*

$$C_s = \frac{\sin^2\Theta(\mathbf{H}^{-1/2}X,\mathbf{H}^{1/2}X)}{\sin^2\Theta(\mathbf{H}_{\mathcal{V}}^{-1/2}X,\mathbf{H}_{\mathcal{V}}^{1/2}X)}.$$

Then Assumption 4.2.6 holds with

$$\beta_{\mathrm{s},\mathcal{X}} = \sqrt{1-\frac{1}{C_s}\min_{x\in\mathcal{X}}\frac{(x,\mathbf{H}_{\mathcal{V}}^{-1}x)}{(x,\mathbf{H}^{-1}x)}}.$$

The proof follows by an analogous argument as was used to prove Theorem 4.2.4 and will, therefore, be omitted.

The subspace approximation theorem has a similar form as Corollary 4.2.5. For simplicity we assume we are given $\mathcal{X} := \mathsf{ran}(X) \subset \mathcal{Y} \subset \mathcal{V}$ and we want to approximate the n-lowest eigenvalues of operator $\mathbf{H}$.

Theorem 4.2.10. *Take* $\mathsf{ran}(X) =: \mathcal{X} \subset \mathcal{Y}$ *such that* $\mathbf{H}_{\mathcal{Y}}\,X = X\Xi$ *and let* $\mathcal{Y}$ *and* $\mathcal{V}$ *be such that* $\mathcal{Y}\subset\mathcal{V}$ *and that they satisfy Assumption 4.2.6 for* $\mathcal{X}\subset\mathcal{Y}$ *and* $\mathbf{H}$. *By* $\mu_1\le\cdots\le\mu_n$ *denote the eigenvalues of the matrix* Ξ.

1. *Assume* $(1-\beta^2_{s,\mathcal{X}})^{-1/2}\mathsf{sin}\Theta(\mathbf{H}_{\mathcal{V}}^{-1/2}\ X, \mathbf{H}_{\mathcal{V}}^{1/2}\ X) < 1$, *then there exist eigenvalues* λ_{i_j}, $j = 1, ..., n$ *of the operator* $\mathbf{H}$ *such that*

$$\frac{|\lambda_{i_j} - \mu_j|}{\mu_j} \leq (1-\beta^2_{s,\mathcal{X}})^{-1/2} \sin\Theta(\mathbf{H}_{\mathcal{V}}^{-1/2}\ X, \mathbf{H}_{\mathcal{V}}^{1/2}\ X), \quad j = 1, ..., n.$$

2. *Assume* $\lambda_n < D \leq \lambda_{n+1}$ *and*

$$\frac{(1-\beta^2_{s,\mathcal{X}})^{-1/2} \sin\Theta(\mathbf{H}_{\mathcal{V}}^{-1/2}\ X, \mathbf{H}_{\mathcal{V}}^{1/2}\ X)}{1-(1-\beta^2_{s,\mathcal{X}})^{-1/2} \sin\Theta(\mathbf{H}_{\mathcal{V}}^{-1/2}\ X, \mathbf{H}_{\mathcal{V}}^{1/2}\ X)} < \frac{D-\mu_n}{D+\mu_n}$$

then

$$\sin\Theta(E(\lambda_n), \mathcal{X}) \leq \frac{(1-\beta^2_{s,\mathcal{X}})^{-1/2} \sin\Theta(\mathbf{H}_{\mathcal{V}}^{-1/2}\ x, \mathbf{H}_{\mathcal{V}}^{1/2}\ x)}{\sqrt{1-(1-\beta^2_{s,\mathcal{X}})^{-1/2} \sin\Theta(\mathbf{H}_{\mathcal{V}}^{-1/2}\ x, \mathbf{H}_{\mathcal{V}}^{1/2}\ x)}} \frac{\sqrt{D\ \mu_n}}{D+\mu_n}.$$

Similar estimates can be formulated for the approximation of other contiguous groups of eigenvalues, cf. Theorem 2.4.2 and Theorem 2.5.6.

3. *If* $\lambda_1 = \cdots = \lambda_n < D \leq \lambda_{n+1}$ *and*

$$\frac{(1-\beta^2_{s,\mathcal{X}})^{-1/2} \sin\Theta(\mathbf{H}_{\mathcal{V}}^{-1/2}\ X, \mathbf{H}_{\mathcal{V}}^{1/2}\ X)}{1-(1-\beta^2_{s,\mathcal{X}})^{-1/2} \sin\Theta(\mathbf{H}_{\mathcal{V}}^{-1/2}\ X, \mathbf{H}_{\mathcal{V}}^{1/2}\ X)} < \frac{D-\mu_n}{D+\mu_n}$$

then

$$\frac{|\lambda_1 - \mu_j|}{\mu_j} \leq \frac{D+\mu_n}{D-\mu_n}(1-\beta^2_{s,\mathcal{X}})^{-1} \sin^2\Theta(\mathbf{H}_{\mathcal{V}}^{-1/2}\ X, \mathbf{H}_{\mathcal{V}}^{1/2}\ X), \quad j = 1, ..., n.$$

Similar estimates can be formulated for other eigenvalues λ *of multiplicity* n.

The proof is a direct combination of the theory from Chapter 2 and the lemmata from this section.

4.2.2 Saturation assumptions

The term "saturation assumption" is borrowed from the approximation theory, see [25]. Assume $\mathcal{T}_d$ is a triangulation of the bounded polygonal domain $\mathcal{R}$ and take $\mathcal{Y} = \mathcal{V}^1_{\mathcal{T}_d}$, $\mathcal{V} = \mathcal{V}^2_{\mathcal{T}_d}$. For a given function $u \in H^1_0(\mathcal{R})$, let the functions u_1 and u_2 be such that

$$\|\nabla(u-u_1)\| = \min_{v \in \mathcal{V}^1_{\mathcal{T}_d}} \|\nabla(u-v)\|$$

$$\|\nabla(u-u_2)\| = \min_{v \in \mathcal{V}^2_{\mathcal{T}_d}} \|\nabla(u-v)\|.$$

We say that the triangulation $\mathcal{T}_d$ satisfies the saturation assumption with regard to $u \in H_0^1(\mathcal{R})$ if

$$\alpha = \frac{\|\nabla(u-u_1)\|}{\|\nabla(u-u_2)\|} < 1.$$

Let $f \in L^2(\mathcal{R})$ be given. We consider $u \in H_0^1(\mathcal{R})$ which is defined as the solution of the problem

$$-\triangle u = f.$$

Let $\mathcal{N}_{\mathcal{T}_d} = \{\xi_1, \ldots, \xi_{n_{\mathcal{T}_d}}\}$ be the set of all the vertices of the triangles from $\mathcal{T}_d$. The functions from $\mathcal{V}^1_{\mathcal{T}_d}$ are uniquely defined by their values on the elements of the finite set $\mathcal{N}_{\mathcal{T}_d}$. The canonical basis functions $\phi_{\xi_i} \in \mathcal{V}^1_{\mathcal{T}_d}$, $1 \le i \le n_{\mathcal{T}_d}$ are defined by requiring that $\phi_{\xi_i}(\xi_j) = \delta_{i,j}$, $1 \le j \le n_{\mathcal{T}_d}$, where $\delta_{i,j}$ is the Kronecker δ-function.

Dörfler and Nochetto have proved in [25] that there exists a constant $C_{\mathcal{T}_d} > 1$, solely depending on the shape regularity of $\mathcal{T}_d$, such that if

$$\nu_f = \frac{\operatorname{osc}(f, \mathcal{T}_d)}{\|\nabla(u-u_1)\|} < \frac{1}{C_{\mathcal{T}_d}}$$

then

$$\frac{\|\nabla(u-u_1)\|^2}{\|\nabla(u-u_2)\|^2} \le 1 - \frac{1}{C_{\mathcal{T}_d}} + \nu^2.$$

$\mathsf{osc}(f, \mathcal{T}_d)$ measures the *oscillation of the function* f on $\mathcal{T}_d$ and is defined by

$$\operatorname{osc}^2(f, \mathcal{T}_d) = \sum_{\xi_i \in \mathcal{N}_d} \sum_{T \in \operatorname{supp}(\phi_{\xi_i})} d_T^2 \|f - f_{\xi_i}\|^2_{L^2(T)}, \qquad f_{\xi_i} = \frac{\int \chi_{\operatorname{supp}(\phi_{\xi_i})} f}{\int \chi_{\operatorname{supp}(\phi_{\xi_i})}}.$$

Let us now go back to Assumption 4.2.1. Set $\mathbf{H} = -\triangle$ and take $x \in \mathcal{V}^1_{\mathcal{T}_d} \subset H^1(\mathcal{R})$ such that $\mathbf{H}_{\mathcal{Y}} x = \mu x$. Setting $u = \mathbf{H}^{-1}x$, one obtains $u_1 = \mathbf{H}_{\mathcal{Y}}^{-1}x$ and $u_2 = \mathbf{H}_{\mathcal{V}}^{-1}x$, so

$$\beta^2_{\mathrm{s},x} = \frac{h[\mathbf{H}^{-1}x - \mathbf{H}_{\mathcal{Y}}^{-1}x]}{h[\mathbf{H}^{-1}x - \mathbf{H}_{\mathcal{V}}^{-1}x]} = \frac{\|\nabla(u-u_1)\|^2}{\|\nabla(u-u_2)\|^2} \le 1 - \frac{1}{C_{\mathcal{T}_d}} + \nu_x^2,$$

where

$$\nu_x = \frac{\operatorname{osc}(x, \mathcal{T}_d)}{h[\mathbf{H}^{-1}x - \mathbf{H}_{\mathcal{Y}}^{-1}x]^{1/2}}. \tag{4.2.16}$$

It was also proved in [25] that $x \in H^1(\mathcal{R})$ implies

$$\operatorname{osc}(x, \mathcal{T}_d) \le C d^2.$$

Therefore, small enough ν_x can be obtained— for our special $x \in H^1(\mathcal{R})$— with realistic triangulations.

Remark 4.2.11. Dörfler and Nochetto also showed that (4.2.16) can be substituted by a similar measure which only involves quantities —estimates of so called jump residuals— that are directly computable from the vector x, cf. [25, Remark 3.4] and Theorem 4.3.2.

Remark 4.2.12. There are other ways to formulate a saturation assumption. The saturation assumption in terms of eigenvalues has been used in [48, Neymeyr] to compute the eigenvalue estimates. The saturation assumption in terms of eigenvalues can be stated as

$$\lambda_i(\mathbf{H}_{\mathcal{V}}|_{\mathcal{V}}) - \lambda_i(\mathbf{H}) \leq \beta_{\mathrm{e}}(\lambda_i(\mathbf{H}_{\mathcal{Y}}|_{\mathcal{Y}}) - \lambda_i(\mathbf{H})), \qquad i = 1, \ldots, \mathsf{dim}(\mathcal{Y}).$$

Here $\beta_{\mathrm{e}} < 1$ is responsible to control two things:

1. It ensures that we are approximating the correct eigenvalues.
2. It measures the quality of the extension $\mathcal{V} = \mathcal{Y} + \mathcal{Z}$.

About β_{e} one can say that it is bounded away from one, cf. [48]. In our analysis $\beta_{\mathrm{s},\mathcal{X}}$ only measures the quality of the space $\mathcal{X}$ with respect to the extension $\mathcal{V} = \mathcal{Y} + \mathcal{Z}$. The localization of the approximated eigenvalues is left to (4.1.3) and Theorem 2.4.1. The behavior of $\beta_{\mathrm{s},\mathcal{X}}$ is better understood, see [14, 25, 60] and Theorem 2.5.5 gives eigenvector approximation error estimates as a bonus. Due to the influence of large eigenvalues β_e should always be a bit more "pessimistic" than $\beta_{\mathrm{s},\mathcal{X}}$.

4.2.3 A case for the use of $\mathsf{sin}\Theta_p$

To illustrate the advantages of the use of $\mathsf{sin}\Theta_p$, as a measure of the quality of Ritz approximations, we go back to Theorem 2.4.2 and Corollary 2.3.15. Theorem 2.4.2 traces the path to a successful approximation method for the approximation of nonnegative eigenvalue problems. Let us remind ourselves what does relatively accurate approximation of an eigenvalue mean in the context of nonnegative operators. It means that the zero eigenvalues are approximated exactly and that we have relative error estimates for nonzero eigenvalues bounded away from one. Subsequently, in order to obtain relatively accurate approximations of the eigenvalues we must chose a test space such that either $\mathsf{ker}(\mathbf{H}) \subset \mathcal{X}$ or $\mathsf{ker}(\mathbf{H}) \perp \mathcal{X}$. We assume, without affecting the level of generality, that $\mathsf{ker}(\mathbf{H}) \perp \mathcal{X}$. Now we see in which way to generalize Assumption 4.2.1. Let $\mathcal{Y}$ and $\mathcal{V}$ be two subspaces such that $\mathcal{X} \subset \mathcal{Y}$, $\mathcal{Y} \subset \mathcal{V}$ and $\mathcal{Y} \perp \mathsf{ker}(\mathbf{H})$ and $\mathcal{V} \perp \mathsf{ker}(\mathbf{H})$. If there exists $0 \leq \beta_{\mathrm{s},x} < 1$ such that

$$h[\mathbf{H}^{\dagger}x - \mathbf{H}_{\mathcal{V}}^{\dagger}x] \leq \beta_{\mathrm{s},x} h[\mathbf{H}^{\dagger}x - \mathbf{H}_{\mathcal{Y}}^{\dagger}x],$$

then the subspaces $\mathcal{Y}$ and $\mathcal{V}$ satisfy the saturation assumption with regard to the vector $x \in \mathcal{Y}$ and the operator $\mathbf{H}$. The equivalence of the saturation assumption and the ability to use the discrete residual can be established by a modification of the arguments used in the positive definite case.

This approach in dealing with the kernel of the nonnegative operator is consistent with the discussion from [1]. The ability to consider nonnegative operators in the same framework with the positive definite operators is one of the main advantages brought in by the use of estimates based on $\mathsf{sin}\Theta_p$. The examples, which will be present in the sequel, only concern positive definite operators. Therefore, we will not further pursue the discussion of the general nonnegative definite case.

In the standard analysis one starts off with a residual of the test vector. Such residual is a functional on $\mathcal{Q}(h)$, and we require positive definite h in order to be able to measure it. Theorem 2.4.2 spells that if we are to expect relative accuracy from our approximation method, then we can safely assume that we are working with the restriction of the form h to the space $\mathsf{ran}(\mathbf{H})$ only—cf. Corollary 2.3.13. Naturally, the form h is positive definite on $\mathsf{ran}(\mathbf{H})$, but our estimates are based on the quantity $\mathsf{sin}\Theta_p$ which is independent of $\mathsf{ran}(\mathbf{H})$ (we do not have to explicitly reduce the form h on $\mathsf{ran}(\mathbf{H})$ in order to analyze error estimates). Also, we get the eigenvector estimates in the same go. Let us describe some of the alternative methods for obtaining eigenvalue estimates.

For now let h be a positive definite form in $\mathcal{H}$ and let $\mathcal{Q}$ be its domain. By $\mathcal{Q}^*$ we denote the space of continuous functionals on $\mathcal{Q}$. Since $\mathcal{Q}$ is continuously imbedded in $\mathcal{H}$, we have

$$\mathcal{Q} \subset \mathcal{H} = \mathcal{H}^* \subset \mathcal{Q}^*. \tag{4.2.17}$$

Let $\langle \cdot, \cdot \rangle : \mathcal{Q}^* \times \mathcal{Q} \to \mathbb{C}$ denote the usual dual product, antilinear in the first argument and linear in the second. By

$$h(u, v) = \langle \mathbf{H}_{\mathcal{Q}} u, v \rangle$$

we introduce, as in [32], the operator

$$\mathbf{H}_{\mathcal{Q}} : \mathcal{Q} \to \mathcal{Q}^*$$

as an extension of $\mathbf{H}$ to $\mathcal{Q}$. Here we consider $\mathcal{Q}$ as the Hilbert space with the scalar product

$$(u, v)_{\mathcal{Q}} = (\mathbf{H}^{1/2} u, \mathbf{H}^{1/2} v), \quad u, v \in \mathcal{Q}.$$

By $\mathbf{R} : \mathcal{Q} \to \mathcal{H}$ we denote the restriction of the operator $\mathbf{H}^{1/2}$ to its natural domain. It is obvious that $\mathbf{R} : \mathcal{Q} \to \mathcal{H}$ is an isometric isomorphism such that

$$\mathbf{H}_{\mathcal{Q}} = \mathbf{R}^* \mathbf{R}.$$

The space $\mathcal{Q}^*$ can be organized as a Hilbert space with the help of the scalar product

$$(u,v)_{\mathbf{H}^{-1}} = (\mathbf{R}^{-*}u, \mathbf{R}^{-*}v), \quad u,v \in \mathcal{Q}^*. \tag{4.2.18}$$

We write the scalar product (4.2.18) as (see [32])

$$(u,v)_{\mathbf{H}^{-1}} = \left\langle u, \mathbf{H}_{\mathcal{Q}}^{-1} v \right\rangle, \quad u,v \in \mathcal{Q}^*. \tag{4.2.19}$$

The norm $\|\cdot\|_{\mathbf{H}^{-1}}$, induced by the scalar product $(\cdot,\cdot)_{\mathbf{H}^{-1}}$, has the property

$$\|u\|_{\mathbf{H}^{-1}} = \sqrt{(u, \mathbf{H}_{\mathcal{Q}}^{-1}\, u)_{\mathbf{H}^{-1}}} = \sqrt{\left\langle u, \mathbf{H}_{\mathcal{Q}}^{-1}\, u \right\rangle} = \max_{v\in\mathcal{Q}} \frac{|\left\langle u,v\right\rangle|}{\|\mathbf{H}^{1/2}v\|}, \qquad u \in \mathcal{Q}^*.$$

If we assume this Hilbert space structures on $\mathcal{Q}$ and $\mathcal{Q}^*$, then $\mathbf{H}_{\mathcal{Q}}$ is an isometry from $\mathcal{Q}$ into $\mathcal{Q}^*$. Note that for $f \in \mathcal{H}$ we have $\mathbf{H}^{-1}f = \mathbf{H}_{\mathcal{Q}}^{-1}f$.

Let $\widetilde{u} \in \mathcal{Q}(\mathbf{H})$ and $b \in \mathcal{H}$, then the *residual*

$$r_{\widetilde{u}}^b = \mathbf{H}_{\mathcal{Q}}\, \widetilde{u} - b \tag{4.2.20}$$

is an element of $\mathcal{Q}^*$ and its $\mathbf{H}^{-1}$–norm measures the error with which $\widetilde{u}$ approximates the solution of the problem

$$\mathbf{H}\, u = b, \quad u \in \mathcal{D}. \tag{4.2.21}$$

To demonstrate this statement we rewrite the definition of the residual as

$$\left\langle r_{\widetilde{u}}^b, v\right\rangle = h(\widetilde{u},v) - (b,v), \quad v \in \mathcal{Q}$$

and then proceed

$$\begin{aligned}
\|r_{\widetilde{u}}^b\|_{\mathbf{H}^{-1}} &= \max_{v\in\mathcal{Q}} \frac{|\left\langle r_{\widetilde{u}}^b, v\right\rangle|}{\sqrt{(\mathbf{H}^{1/2}v, \mathbf{H}^{1/2}v)}} = \max_{y\in\mathcal{H}} \frac{|(\mathbf{H}^{1/2}\widetilde{u}, y) - (\mathbf{H}u, \mathbf{H}^{-1/2}y)|}{\|y\|} \\
&= \max_{y\in\mathcal{H}} \frac{|(\mathbf{H}^{1/2}\widetilde{u}, y) - (\mathbf{H}^{1/2}u, y)|}{\|y\|} \\
&= h[\widetilde{u}-u]^{1/2} = \|\widetilde{u}-u\|_E,
\end{aligned}$$

i.e. the dual norm of the residual equals the energy norm of the error. Assume now the finite dimensional subspace $\mathcal{V} \subset \mathcal{Q}$ is given, then

$$\|r_{\mathbf{H}_{\mathcal{V}}^{-1}b}^b\|_{\mathbf{H}^{-1}} = h[\mathbf{H}^{-1}b - \mathbf{H}_{\mathcal{V}}^{-1}b]^{1/2} = \|r_{(b,\mathcal{V},\mathcal{Q})}\|$$

illustrates the connection between the middle space residual $r_{(b,\mathcal{V},\mathcal{Q})} \in \mathcal{H}$ and the residual $r_{\mathbf{H}_{\mathcal{V}}^{-1}b}^b \in \mathcal{Q}^*$ as well as the motivation for the terminology.

Let $\|\widetilde{u}\| = 1$ and $\widetilde{\mu} = h[\widetilde{u}]$, by

$$r_{\widetilde{u}} = \mathbf{H}_{\mathcal{Q}}\widetilde{u} - \widetilde{\mu}\widetilde{u} \in \mathcal{Q}^* \tag{4.2.22}$$

we define the residual for the eigenvalue problem

$$\mathbf{H}u = \lambda u, \quad u \in \mathcal{D}, \tag{4.2.23}$$

with respect to the Ritz vector $\widetilde{u}$. As in the case of the stationary problem we prove, cf. (3.4.5) and (3.4.6),

$$\begin{aligned}
\|r_{\widetilde{u}}\|_{\mathbf{H}^{-1}} &= \max_{v\in\mathcal{Q}} \frac{|h(\widetilde{u},v) - \widetilde{\mu}(\widetilde{u},v)|}{\sqrt{(\mathbf{H}^{1/2}v, \mathbf{H}^{1/2}v)}} = \max_{v\in\mathcal{Q}} \frac{|(\mathbf{H}^{1/2}\widetilde{u}, \mathbf{H}^{1/2}v) - \widetilde{\mu}(\widetilde{u},v)|}{\sqrt{(\mathbf{H}^{1/2}v, \mathbf{H}^{1/2}v)}} \\
&= \max_{y\in\mathcal{H}} \frac{|(\mathbf{H}^{1/2}\widetilde{u}, y) - \widetilde{\mu}(\widetilde{u}, \mathbf{H}^{-1/2}y)|}{\|y\|} = \max_{y\in\mathcal{H}} \frac{|(\mathbf{H}^{1/2}\widetilde{u}, y) - \widetilde{\mu}(\mathbf{H}^{-1/2}\widetilde{u}, y)|}{\|y\|} \\
&= \|\mathbf{H}^{1/2}\widetilde{u} - \widetilde{\mu}\mathbf{H}^{-1/2}\widetilde{u}\|.
\end{aligned} \tag{4.2.24}$$

The norm of the residual is tightly connected to $\mathsf{sin}\Theta(\mathbf{H}^{-1/2}\widetilde{u}, \mathbf{H}^{1/2}\widetilde{u})$, namely we will show

$$\sin\angle(\mathbf{H}^{1/2}\widetilde{u}, \mathbf{H}^{-1/2}\widetilde{u}) \le \frac{\|\mathbf{H}^{1/2}\widetilde{u} - \widetilde{\mu}\mathbf{H}^{-1/2}\widetilde{u}\|}{\|\mathbf{H}^{1/2}\widetilde{u}\|} \le \frac{\sin\angle(\mathbf{H}^{1/2}\widetilde{u}, \mathbf{H}^{-1/2}\widetilde{u})}{1 - \sin\angle(\mathbf{H}^{1/2}\widetilde{u}, \mathbf{H}^{-1/2}\widetilde{u})}. \tag{4.2.25}$$

Theorem 4.2.13. *Let $\widetilde{u} \in \mathcal{Q}$ be of norm one. Let $r_{\widetilde{u}}$ and the forms h and $h_{\widetilde{u}}$ be as before, we have*

$$\sin\angle(\mathbf{H}^{1/2}\widetilde{u}, \mathbf{H}^{-1/2}\widetilde{u}) \le \frac{\|\mathbf{H}^{1/2}\widetilde{u} - \widetilde{\mu}\mathbf{H}^{-1/2}\widetilde{u}\|}{\|\mathbf{H}^{1/2}\widetilde{u}\|} \le \frac{\sin\angle(\mathbf{H}^{1/2}\widetilde{u}, \mathbf{H}^{-1/2}\widetilde{u})}{1 - \sin\angle(\mathbf{H}^{1/2}\widetilde{u}, \mathbf{H}^{-1/2}\widetilde{u})}. \tag{4.2.26}$$

Proof. Theorem 2.3.5 gives the right-hand estimate

$$\begin{aligned}
\|\mathbf{H}_{\mathcal{Q}}\widetilde{u} - \widetilde{\mu}\widetilde{u}\|_{\mathbf{H}^{-1}} &= \max_{v} \frac{|h(\widetilde{u},v) - h_{\widetilde{u}}(\widetilde{u},v)|}{\|\mathbf{H}^{1/2}v\|} = \max_{v} \frac{|\delta h(\widetilde{u},v)|}{\|\mathbf{H}^{1/2}v\|} \\
&\le \|\mathbf{H}^{1/2}\widetilde{u}\| \ \frac{\sin\angle(\mathbf{H}^{1/2}\widetilde{u}, \mathbf{H}^{-1/2}\widetilde{u})}{1 - \sin\angle(\mathbf{H}^{1/2}\widetilde{u}, \mathbf{H}^{-1/2}\widetilde{u})}.
\end{aligned}$$

The left-hand side of the inequality follows from (4.2.24),

$$\begin{aligned}
\|r_{\widetilde{u}}\|_{\mathbf{H}^{-1}} &= \|\mathbf{H}^{1/2}\widetilde{u} - \widetilde{\mu}\mathbf{H}^{-1/2}\widetilde{u}\| \\
&= \|\mathbf{H}^{1/2}\widetilde{u}\| \ \|\frac{1}{\|\mathbf{H}^{1/2}\widetilde{u}\|}\mathbf{H}^{1/2}\widetilde{u} - \|\mathbf{H}^{1/2}\widetilde{u}\| \ \mathbf{H}^{-1/2}\widetilde{u}\| \\
&\ge \sqrt{\widetilde{\mu}} \ \sin\angle(\mathbf{H}^{1/2}\widetilde{u}, \mathbf{H}^{-1/2}\widetilde{u}).
\end{aligned}$$

□

In the case in which the dimension of the test space is larger than one we have the following generalization of Theorem 4.2.13.

Theorem 4.2.14. *Let $X\mathbb{C}^n \subset \mathcal{Q}$. Take $\widetilde{u} \in \mathcal{X} = X\mathbb{C}^n$ of norm one and let $r_{\widetilde{u}}$ and the forms h and $h_{\mathcal{X}}$ be as before. Then*

$$\max_{\substack{\widetilde{u}\in\mathcal{X}\\ \|\widetilde{u}\|=1}} \sin\angle(\mathbf{H}^{1/2}\widetilde{u}, \mathbf{H}^{-1/2}\widetilde{u}) \leq \max_{\substack{\widetilde{u}\in\mathcal{X}\\ \|\widetilde{u}\|=1}} \frac{\|\mathbf{H}^{1/2}\widetilde{u} - \widetilde{\mu}\mathbf{H}^{-1/2}\widetilde{u}\|}{\|\mathbf{H}^{1/2}\widetilde{u}\|} \leq \frac{\sin\angle(\mathbf{H}^{1/2}X, \mathbf{H}^{-1/2}X)}{1-\sin\angle(\mathbf{H}^{1/2}X, \mathbf{H}^{-1/2}X)}. \tag{4.2.27}$$

Proof. As in the proof of Theorem 4.2.13, we have

$$\begin{aligned}\|\mathbf{H}_{\mathcal{Q}}\widetilde{u} - \widetilde{\mu}\widetilde{u}\|_{\mathbf{H}^{-1}} &= \max_{v} \frac{|h(\widetilde{u}, v) - h_{\mathcal{X}}(\widetilde{u}, v)|}{\|\mathbf{H}^{1/2}v\|} = \max_{v} \frac{|\delta h(\widetilde{u}, v)|}{\|\mathbf{H}^{1/2}v\|} \\ &\leq \|\mathbf{H}^{1/2}\widetilde{u}\| \, \frac{\sin\angle(\mathbf{H}^{1/2}X, \mathbf{H}^{-1/2}X)}{1-\sin\angle(\mathbf{H}^{1/2}X, \mathbf{H}^{-1/2}X)},\end{aligned}$$

for every $\widetilde{u} \in \mathcal{Q}$, $\|\widetilde{u}\| = 1$. Hence, we have established

$$\max_{\substack{\widetilde{u}\in\mathcal{X}\\ \|\widetilde{u}\|=1}} \frac{\|\mathbf{H}^{1/2}\widetilde{u} - \widetilde{\mu}\mathbf{H}^{-1/2}\widetilde{u}\|}{\|\mathbf{H}^{1/2}\widetilde{u}\|} \leq \frac{\sin\angle(\mathbf{H}^{1/2}X, \mathbf{H}^{-1/2}X)}{1-\sin\angle(\mathbf{H}^{1/2}X, \mathbf{H}^{-1/2}X)}.$$

Theorem 4.2.13 gives

$$\max_{\substack{\widetilde{u}\in\mathcal{X}\\ \|\widetilde{u}\|=1}} \sin\angle(\mathbf{H}^{1/2}\widetilde{u}, \mathbf{H}^{-1/2}\widetilde{u}) \leq \max_{\substack{\widetilde{u}\in\mathcal{X}\\ \|\widetilde{u}\|=1}} \frac{\|\mathbf{H}^{1/2}\widetilde{u} - \widetilde{\mu}\mathbf{H}^{-1/2}\widetilde{u}\|}{\|\mathbf{H}^{1/2}\widetilde{u}\|} \leq \frac{\sin\angle(\mathbf{H}^{1/2}X, \mathbf{H}^{-1/2}X)}{1-\sin\angle(\mathbf{H}^{1/2}X, \mathbf{H}^{-1/2}X)},$$

as was required. □

The measure $\frac{\|\mathbf{H}^{1/2}\widetilde{u} - \widetilde{\mu}\mathbf{H}^{-1/2}\widetilde{u}\|}{\|\mathbf{H}^{1/2}\widetilde{u}\|}$ from (4.2.25) is sometimes called the norm of the scaled residual. It appears in the matrix Temple–Kato inequality from [48]. This inequality was generalized to the operator setting in Theorem 2.6.1. In the original form, implicit in the proof of Theorem 2.6.1, it reads

$$\frac{\lambda_m - \widetilde{\mu}}{\lambda_m} \leq \frac{\lambda_{m+n}}{\lambda_{m+n} - \widetilde{\mu}} \left(\frac{\|\mathbf{H}^{1/2}\widetilde{u} - \widetilde{\mu}\mathbf{H}^{-1/2}\widetilde{u}\|}{\|\mathbf{H}^{1/2}\widetilde{u}\|} \right)^2. \tag{4.2.28}$$

The modification in the Theorem 2.6.1 was made in order to be able to use the symmetric function to measure the relative gap, i.e.

$$\begin{aligned}\frac{|\lambda_m - \widetilde{\mu}|}{\lambda_m} &\leq \max\left\{ \frac{\lambda_{m+n} + \widetilde{\mu}}{\lambda_{m+n} - \widetilde{\mu}}, \frac{\lambda_{m-1} + \widetilde{\mu}}{\widetilde{\mu} - \lambda_{m-1}} \right\} \left(\frac{\|\mathbf{H}^{1/2}\widetilde{u} - \widetilde{\mu}\mathbf{H}^{-1/2}\widetilde{u}\|}{\|\mathbf{H}^{1/2}\widetilde{u}\|} \right)^2 \\ &\leq \max\left\{ \frac{\lambda_{m+n} + \widetilde{\mu}}{\lambda_{m+n} - \widetilde{\mu}}, \frac{\lambda_{m-1} + \widetilde{\mu}}{\widetilde{\mu} - \lambda_{m-1}} \right\} \frac{\sin^2\Theta}{(1-\sin\Theta)^2}.\end{aligned}$$

The main problem with the estimate (4.2.28) is that it measures the approximation error relative to the unknown quantity λ_m. The estimate we prefer is

$$\frac{|\lambda_m - \widetilde{\mu}|}{\widetilde{\mu}} \leq \max\left\{\frac{\lambda_{m+n} + \widetilde{\mu}}{\lambda_{m+n} - \widetilde{\mu}}, \frac{\lambda_{m-1} + \widetilde{\mu}}{\widetilde{\mu} - \lambda_{m-1}}\right\} \sin^2\Theta, \tag{4.2.29}$$

from Theorem 2.6.4. Comparing the estimates (4.2.28) and (4.2.29) is not easy. The sharpness gained on the measure of the scaled residual is lost in the slack allowed in the measure of the relative gap. Furthermore, we are not measuring the error relative to the same quantity. Important feature to note is that the composition of both estimates is the same. Namely, error is bounded by the product of a measure of the gap and a measure of the scaled residual. The computational procedure that is being used to estimate $\frac{\|\mathbf{H}^{1/2}\widetilde{u} - \widetilde{\mu}\mathbf{H}^{-1/2}\widetilde{u}\|}{\|\mathbf{H}^{1/2}\widetilde{u}\|}$ applies equally well in both settings, cf. Theorem 4.2.13 and [48].

To illustrate this statement, as well as to justify our announcement from the introduction to Section 4.2, we bring forward the following theorem which is a mixture of the considerations from Section 4.2.2, Corollary 4.2.5 and Reference [48]. We will concentrate on single vector approximation, but we note that generalization to subspace approximations presents primarily technical difficulties, cf. Theorem 4.2.10.

Theorem 4.2.15. *Let $\mathbf{H} = -\triangle$ be defined in the bounded polygonal domain $\mathcal{R}$ by imposing the Dirichlet boundary condition. Let the triangulation $\mathcal{T}_d$ and the spaces $\mathcal{Y} = \mathcal{V}^1_{\mathcal{T}_d}$ and $\mathcal{V} = \mathcal{V}^2_{\mathcal{T}_d}$ be as in Section 4.2.2. Let $x \in \mathcal{Y}$ be as in Corollary 4.2.5–2. Then there exists a constant $C_{\mathcal{T}_d} > 1$, solely depending on the shape regularity of $\mathcal{T}_d$, such that*

$$\frac{|\lambda_1 - \mu|}{\mu} \leq \frac{D + \mu}{D - \mu}(C^{-1}_{\mathcal{T}_d} - \nu_x^2)^{-1}\frac{\|H_{\mathcal{V}}x - \mu x\|^2_{H^{-1}_{\mathcal{V}}}}{\|H^{1/2}_{\mathcal{V}}x\|^2}.$$

(Here we have freely used the notation from Section 4.2.2 and Corollary 4.2.5.)

Proof. The proof is a combination of Corollary 4.2.5–2 and equations (4.2.16) and (4.2.26). It follows the agenda from the introduction to Section 4.2 to the letter. □

Remark 4.2.16. Analogous theorems accompany the subspace approximation results from Theorem 4.2.10, as well as other variants of Corollary 4.2.5. We leave out the details.

We emphasize, again, that in this case

$$\nu_x = \frac{\mathrm{osc}(x, \mathcal{T}_d)}{|\mathbf{H}^{-1}x - \mu^{-1}x|_{1,2}}$$

essentially depends on an estimate of the "jump residual", cf. Remark 4.2.11 and Theorem 4.3.2. Since this is a computable quantity it can be used to define a refinement procedure for the triangulation $\mathcal{T}_d$. In [43, 48] it has been shown how to reduce

$$\frac{\|H_{\mathcal{V}}x - \mu x\|^2_{H_{\mathcal{V}}^{-1}}}{\|H_{\mathcal{V}}^{1/2}x\|^2}$$

by a refinement of the triangulation $\mathcal{T}_d$, see [43, Section 4] and [48, Section 5.]. We now have two target functions which, according to Theorem 4.2.15, should (could) be reduced by a refinement of the mesh $\mathcal{T}_d$. It remains an interesting task for the future to determine what additional information is brought into the general picture by ν_x.

Notably, both ν_x and $\|H_{\mathcal{V}}x - \mu x\|^2_{H_{\mathcal{V}}^{-1}}/\|H_{\mathcal{V}}^{1/2}x\|^2$ essentially depend on a measure of the "jump residual". This, to some extent, corroborates the numerical evidence from [48, Figure 4.], which showed that $\|H_{\mathcal{V}}x - \mu x\|^2_{H_{\mathcal{V}}^{-1}}/\|H_{\mathcal{V}}^{1/2}x\|^2$ captures most of the error.

Experiment: Estimating $\mathsf{sin}\Theta$ with a use of preconditioners

Assume that we have finite dimensional spaces $\mathcal{Y}$ and $\mathcal{V}$ such that

$$\mathcal{X} \subset \mathcal{Y} \subset \mathcal{V}, \tag{4.2.30}$$

where $\dim \mathcal{X} = n$. Let the matrices $\Xi = \mathbf{H}_{\mathcal{X}}|_{\mathcal{X}}$, $\Omega = P_{\mathcal{X}}\mathbf{H}^{-1}P_{\mathcal{X}}|_{\mathcal{X}}$, $\Omega_{\mathcal{V}} = P_{\mathcal{X}}\mathbf{H}_{\mathcal{V}}^{-1}P_{\mathcal{X}}\big|_{\mathcal{X}}$ and $H_{\mathcal{V}} = \mathbf{H}_{\mathcal{V}}|_{\mathcal{V}}$ be as before. The saturation assumption (Assumption 4.2.6) enabled us to show that

$$\sin^2\Theta_{\mathcal{V}} = \sin^2\Theta(\mathbf{H}_{\mathcal{V}}^{-1/2}\mathcal{X}, \mathbf{H}_{\mathcal{V}}^{1/2}\mathcal{X}) = \max_{x\in\mathcal{X}} \frac{x^*(\Omega_{\mathcal{V}} - \Xi^{-1})x}{x^*\Omega_{\mathcal{V}}x}$$

can be used to estimate

$$\sin^2\Theta = \sin^2\Theta(\mathbf{H}^{-1/2}\mathcal{X}, \mathbf{H}^{1/2}\mathcal{X}) = \max_{x\in\mathcal{X}} \frac{x^*(\Omega - \Xi^{-1})x}{x^*\Omega x}.$$

However, even estimating—yet alone computing—$\Omega_{\mathcal{V}}$ will often be a computationally expensive task.

Assumption 4.2.6 effectively quantified the equivalence of norms $h[\cdot]^{1/2}$ and $h_{\mathcal{V}}[\cdot]^{1/2}$ for the task of computing $\mathsf{sin}\Theta$. Also, we have seen that a "better behaved" quantity

$$\delta_{\mathcal{V}} = \max_{x\in\mathcal{X}} \frac{|x^*(\Omega_{\mathcal{V}} - \Omega)x|}{x^*\Omega_{\mathcal{V}}x} \tag{4.2.31}$$

can be used as a measure of this equivalence in the place of the saturation constant $\beta_{\mathrm{s},\mathcal{X}}$.

Assume now we are given a positive definite matrix $H_{\rm p} : \mathcal{V} \to \mathcal{V}$, such that

$$\delta_{\rm p} = \max_{x \in \mathcal{X}} \frac{|x^*(H_{\rm p}^{-1} - \Omega_{\mathcal{V}})x|}{x^* H_{\rm p}^{-1} x} \tag{4.2.32}$$

is sufficiently small. If we could define what sufficiently small means, assuming that forming the matrix $\Omega_p = P H_{\rm p}^{-1} P\big|_{\mathcal{X}}$ is computationally cheaper than the evaluation of the matrix $\Omega_{\mathcal{V}}$, then

$$s_p^2 = \max_{x \in \mathcal{X}} \frac{|x^*(\Omega_p - \Xi^{-1})x|}{x^* \Omega_p x}$$

should be a good substitute for $\sin^2\Theta_{\mathcal{V}}$ and thus also for $\sin^2\Theta$. Let us state the result first, and then we shall comment on the possible origin of such $H_{\rm p}$.

Proposition 4.2.17. *Let the matrices Ξ, Ω, $\Omega_{\mathcal{V}}$, Ω_p be as before. If $\kappa = \delta_p + \frac{\delta_{\mathcal{V}}}{1-\delta_p} < 1$ then*

$$\sin^2\Theta \le \frac{s_p^2 + \kappa}{1 + \kappa}.$$

Proof. As in the proof of (4.2.46) we compute

$$\begin{aligned} \frac{|x^*(\Omega_{\rm p} - \Omega)x|}{x^*\Omega_{\rm p}x} &\le \frac{|x^*(\Omega_{\rm p} - \Omega_{\mathcal{V}})x|}{x^*\Omega_{\rm p}x} + \frac{|x^*(\Omega_{\mathcal{V}} - \Omega)x|}{x^*\Omega_{\rm p}x} \\ &\le \frac{|x^*(\Omega_{\rm p} - \Omega_{\mathcal{V}})x|}{x^*\Omega_{\rm p}x} + \frac{1}{1-\delta_{\rm p}} \frac{|x^*(\Omega_{\mathcal{V}} - \Omega)x|}{x^*\Omega_{\mathcal{V}}x}. \end{aligned}$$

The conclusion is a consequence of the right hand inequality from Lemma 4.2.19 □

Let us now return to equation (4.2.30). The subspace $\mathcal{V}$ can be written as

$$\mathcal{V} = \mathcal{Y} + \mathcal{Z}.$$

Assume there exist subspaces $\mathcal{V}_k$, $k = 0, 2, \cdots, j$, such that

$$\mathcal{Y} = \mathcal{V}_0 \subset \mathcal{V}_1 \subset \mathcal{V}_2 \subset \cdots \subset \mathcal{V}_j = \mathcal{V}. \tag{4.2.33}$$

Let $I_{\mathcal{V}_k}$ be the interpolation (or some other "projection") operators on $\mathcal{V}_k$. According to [12]

$$\| u \|_{\mathcal{V}}^2 = \|\mathbf{H}_{\mathcal{Y}}^{1/2} I_{\mathcal{Y}} u\|^2 + \sum_{k=1}^{j} 4^{-k} \|(I_{\mathcal{V}_j} - I_{\mathcal{V}_{j-1}})u\|^2, \qquad u \in \mathcal{V}, \tag{4.2.34}$$

is a norm equivalent to the energy norm $h[\cdot]^{1/2}$ on $\mathcal{V}$, cf. [43].

Energy norms on finite dimensional spaces are represented by positive definite matrices. Let H_{p} be the matrix of the "properly scaled" norm $|\!|\!| \cdot |\!|\!|_{\mathcal{V}}$. This means that there exist $0 < \gamma_p < 1$, such that the equivalence of the norms $|\!|\!| \cdot |\!|\!|_{\mathcal{V}}$ and $h_{\mathcal{V}}[\cdot]^{1/2}$ can be expressed as

$$(1-\gamma_p)x^* H_{\mathrm{p}} x \leq x^* H_{\mathcal{V}} x \leq (1+\gamma_p)\ x^* H_{\mathrm{p}} x, \quad x \in \mathcal{V}$$

or equivalently

$$1-\gamma_p \leq \frac{x^* H_{\mathrm{p}}^{-1} x}{x^* H_{\mathcal{V}}^{-1} x} \leq 1+\gamma_p, \qquad x \in \mathcal{V}.$$

The matrix H_{p} can be used as a *preconditioner* for the iterative methods to compute the solution of the equation $H_{\mathcal{V}} x = b$, since $H_{\mathcal{V}} x = b$ is equivalent to the equation

$$H_{\mathrm{p}}^{-1/2} H_{\mathcal{V}} H_{\mathrm{p}}^{-1/2} y = H_{\mathrm{p}}^{-1/2} b, \qquad y = H_{\mathrm{p}}^{1/2} x,$$

which has a well conditioned coefficient matrix $H_{\mathrm{p}}^{-1/2} H_{\mathcal{V}} H_{\mathrm{p}}^{-1/2}$— cf. [2, 13, 66]. We will consistently use the term preconditioner when we refer to H_{p}.

In the view of Corollary 4.2.17 we might get overly pessimistic eigenvalue estimates based on the global bound γ_p, since it is the quality of the preconditioner on the much smaller space $\mathcal{X}$, that matters in the computation of the eigenvalue estimate $\mathsf{sin}\Theta$.

Looking at the formula (4.2.34) one might suspect that $\delta_p < \gamma_p$ would not be an implausible expectation. This is the reason why we have presented Proposition 4.2.17. Further consideration of the optimal use of preconditioners will be subject of the future research.

Remark 4.2.18. Let us illustrate this reasoning on Problem (2.7.7). In what follows we will freely use the notation from Section 2.7.3. The finite element approximation $\mathsf{sin}\Theta_{\mathcal{V}_{4N}^1}$ of $\mathsf{sin}\Theta$ is very accurate since $\delta_{\mathcal{V}_{4N}^1}$ is of the order 10^{-3}, see Figure 4.2. Note that

$$\delta_{\mathcal{V}_{4N}^1} = \|(\widetilde{X}^* H_{4N}^{-1} \widetilde{X})^{-1/2} (I - \widetilde{X}^* T_{4N} \widetilde{X})(\widetilde{X}^* H_{4N}^{-1} \widetilde{X})^{-1/2}\| < 4 \cdot 10^{-3}, \quad N = 80, \ldots, 120$$

is small without H_{4N}^{-1} being a good approximation of T_{4N} on the whole of $\mathcal{V}_{4N}$, since

$$\|T_{4N}^{-1/2}(I - H_{4N}) T_{4N}^{-1/2}\| > 1, \quad N = 80, \ldots, 120. \tag{4.2.35}$$

Here $H_{4N} = (\mathbf{H}^{1/2}\mathbf{V})^* \mathbf{H}^{1/2}\mathbf{V}$ and we compute $T_{4N} = \mathbf{V}^* \mathbf{H}_{\mathcal{V}}^{-1} \mathbf{V}$ using the formula (2.7.19). This suggest that we could use, if available, a nearby matrix H_{p} and compute $x^* \widetilde{X}^* H_{\mathrm{p}}^{-1} \widetilde{X} x$ instead of computing $x^* \widetilde{X}^* H_{\mathcal{V}}^{-1} \widetilde{X} x$, see Corollary 4.2.17. We only need to have available H_{p}^{-1} that is a good approximation of T on the subspace $\mathcal{X}$ and that is cheaper to invert than H.

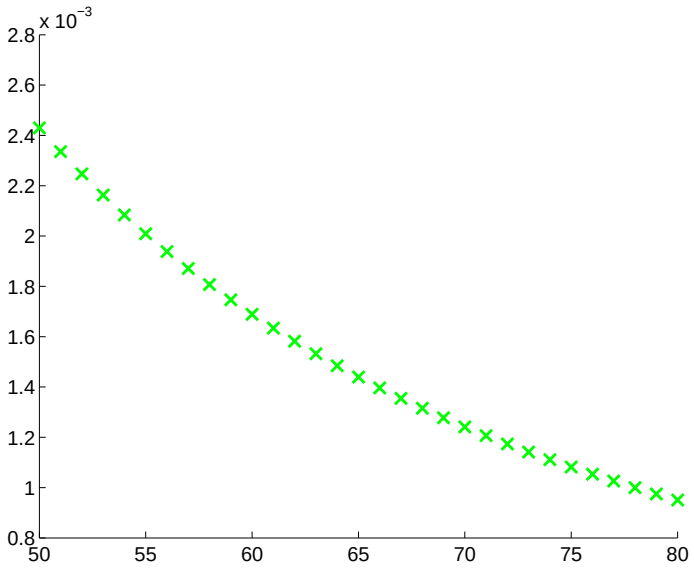

Figure 4.2: An experiment with preconditioning:
The true δ computed using the formula (2.7.18).

Such behavior of H_{4N} is not surprising, since only the low frequency eigenmodes are well approximated by the finite element methods. This example is brought out to corroborate the conclusion of the Corollary 4.2.17. Results of experiments with realistic preconditioners will be reported elsewhere. Precisely (4.2.35) is the reason why the analysis based on Assumption 4.2.1 (or Assumption 4.2.6) should be preferred to the analysis based on the assumption from Remark 4.2.12.

4.2.4 Example: Laplace eigenvalue problem in the square $[-1,1]^2$

The saturation assumption (4.2.6) has enabled us to reduce the problem of estimating the $\|\mathbf{H}^{-1}x - \mathbf{H}_{\mathcal{X}}^{-1}x\|_E$ to the problem of bounding the discrete residual

$$\|\mathbf{H}_{\mathcal{V}}^{-1}x - \mathbf{H}_{\mathcal{Y}}^{-1}x\|_E = h_{\mathcal{V}}[\mathbf{H}_{\mathcal{V}}^{-1}x - \mathbf{H}_{\mathcal{Y}}^{-1}x]^{1/2} = \|r_{(x,\mathcal{Y},\mathcal{V})}\|.$$

To be definite, let us assume that we have subspaces $\mathcal{X} = X\mathbb{R}^n$, $\mathcal{Y}$ and $\mathcal{V}$ satisfying (4.2.12). The subspace saturation constant $\beta_{\mathrm{s},\mathcal{X}}$, which measures the quality of the estimates by the discrete residual, can be evaluated from

$$\beta_{\mathrm{s},\mathcal{X}} = \max_{x\in\mathcal{X}\setminus\{0\}} \frac{(x, \mathbf{H}^{-1}x - \mathbf{H}_{\mathcal{V}}^{-1}x)}{(x, \mathbf{H}^{-1}x - \mathbf{H}_{\mathcal{Y}}^{-1}x)}. \tag{4.2.36}$$

Equivalently, in matrix formulation (4.2.36) is written as

$$\beta_{\mathrm{s},\mathcal{X}} = \max_{x\in\mathbb{R}^n\setminus\{0\}} \frac{x^*(\Omega-\Omega_{\mathcal{V}})x}{x^*(\Omega-\Omega_{\mathcal{Y}})x}. \tag{4.2.37}$$

Although the matrices $\Omega-\Omega_{\mathcal{V}}$ and $\Omega-\Omega_{\mathcal{Y}}$ are both positive definite matrices, computing $\beta_{\mathrm{s},\mathcal{X}}$ from (4.2.37) is at best a difficult problem. In what follows, the saturation measure $\beta_{\mathrm{s},\mathcal{X}}$ will be substituted by (asymptotically) weaker quantity

$$\delta_{\mathcal{V}} = \max_{x\in\mathbb{R}^n\setminus\{0\}} \frac{x^*(\Omega-\Omega_{\mathcal{V}})x}{x^*\Omega_{\mathcal{V}}x}.$$

The nonappearance of $\mathcal{Y}$ is only superficial, since

$$\delta_{\mathcal{V}} = \max_{x\in\mathbb{R}^n\setminus\{0\}} \frac{x^*(\Omega-\Omega_{\mathcal{V}})x}{x^*\Omega_{\mathcal{V}}x} \leq \max_{x\in\mathbb{R}^n\setminus\{0\}} \frac{x^*(\Omega-\Omega_{\mathcal{V}})x}{x^*\Omega_{\mathcal{Y}}x} = \widehat{\delta}_{\mathcal{V}}.$$

In comparison with $\beta_{\mathrm{s},\mathcal{X}}$ the measure $\delta_{\mathcal{V}}$ is not optimally scaled, but will be easier to compute. Our aim in this section is to compute an estimate of $\mathsf{sin}\Theta$ for a 2D model problem to complement the study performed on the 1D model problem in Section 2.7.3.

The main technical result will be Lemma 4.2.7 and the following inequality (which will be formulated as a lemma to ease the reference). It is based upon the estimate from [3].

Lemma 4.2.19. *Let $A_1, A_2, M \in \mathbb{R}^{n\times n}$ be positive definite matrices and let $A_2 \leq A_1$,*

$$s_1 = \max_{x\neq 0} \frac{|x^*(A_1-M)x|}{x^*A_1x}, \quad s_2 = \max_{x\neq 0} \frac{|x^*(A_2-M)x|}{x^*A_2x}, \quad \delta = \max_{x\neq 0} \frac{|x^*(A_1-A_2)x|}{x^*A_2x}$$

then

$$s_2 \leq s_1 \leq \frac{s_2+\delta}{1+\delta}.$$

To compute $\delta_{\mathcal{V}}$ one has to carefully consider the relation between the subspaces $\mathcal{V}, \mathcal{Y} \subset \mathcal{Q}(h) = H_0^1(\mathcal{R})$. Let $\mathcal{R} = [-1,1]^2$ and $\mathcal{V} = \mathcal{V}_{\mathcal{T}_d}^1 = \mathcal{V}_d^1$, where $\mathcal{T}_d$ is the standard triangulation of $\mathcal{R}$. In this case we can directly estimate

$$\|\mathbf{H}^{-1}x - \mathbf{H}_{\mathcal{V}}^{-1}x\|_E, \quad x \in \mathcal{X} \subset \mathcal{V}_d^1,$$

by the norm of the vector x, see [7]. Therefore, we may take $\mathcal{X} = \mathcal{Y}$, which identifies this example as a special case of the theory from the previous chapter.

These estimates are based on the special regularity property of the space $H_0^1(\mathcal{R})$. Let $\mathbf{H} = -\triangle$ and $\mathcal{D}(\mathbf{H}) = H_0^2([-1,1]^2)$, then

$$\max_{v\in\mathcal{D}(\mathbf{H})} \frac{|v|_{2,2}}{\|\mathbf{H}v\|} = 1.$$

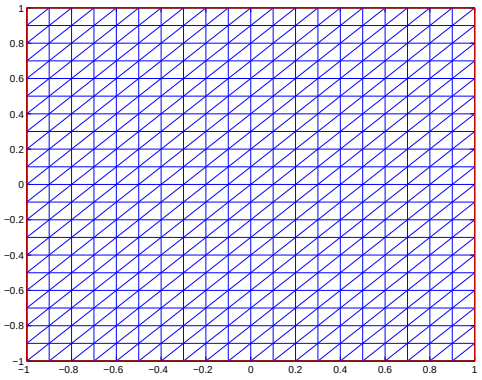

Figure 4.3: The *standard triangulation* on $\mathcal{R}$ with $N = 20$.

This is customarily written as $|v|_{2,2} \leq \|\mathbf{H}v\|$. It implies that $\mathbf{H}^{-1}x \in H^2(\mathcal{R})$ when $x \in L^2(\mathcal{R})$. This procedure can be adapted for a region for which

$$S_2 = \max_{v \in \mathcal{D}(\mathbf{H})} \frac{|v|_{2,2}}{\|\mathbf{H}v\|} < \infty.$$

In particular, when $\mathcal{R}$ is convex then $S_2 = 1$, see [7]. A variant of Miranda-Talenti theorem from [7] describes a class of regions for which $S_2 < \infty$. Although such energy norm estimates can be established for a broader class of regions (cf. [31, 42] and Section 4.3.1) the sharpness the bound is heavily dependent upon the regularity properties of the space $H^2(\mathcal{R})$.

Rather than to pursue the most abstract case, we opt to concentrate on a model problem where we can compute all the relevant constants exactly. We will also show that the assumption $\mathbf{H}_{\mathcal{Y}}^{-1}\ x = \mu\ x$ was of technical nature, only. Allowing x that is not an eigenvector of $\mathbf{H}_{\mathcal{Y}}$ is important when one has finite precision computations in mind.

Now, let us concentrate on the problem of estimating the accuracy of the Rayleigh-Ritz approximations to the solutions of the eigenvalue problem

$$\begin{aligned} -\triangle u &= \omega\ u, \quad \text{in } \mathcal{R} \\ u|_{\partial\mathcal{R}} &= 0. \end{aligned} \tag{4.2.38}$$

Here, $\mathcal{R}$ is a square $[-1,1] \times [-1,1] \subset \mathbb{R}^2$. The eigenvalues of Problem (4.2.38) are known to be

$$\omega_{k,l} = \left(\frac{k^2}{4} + \frac{l^2}{4}\right)\pi^2, \quad k,l \in \mathbb{N}.$$

We take $\mathcal{H} = L^2(\mathcal{R})$ and $\mathbf{H}$ to be the positive definite operator defined by the form

$$h(u,v) = \int (\nabla u)^* \nabla v \, \mathrm{d}x, \quad u, v \in \mathcal{Q} = H_0^1(\mathcal{R}). \tag{4.2.39}$$

In the ordering assumed in Chapter 2 we have

$$\lambda_1 = \frac{1}{2}\pi^2, \ \lambda_2 = \lambda_3 = \frac{5}{4}\pi^2, \ \lambda_4 = 2\pi^2. \tag{4.2.40}$$

We aim to compute the relative estimates of the error of the Rayleigh-Ritz approximations to $\lambda_1, \lambda_2, \lambda_3$. The problem (4.2.38) is discretized on the space $\mathcal{V}_d^1$, the space of piecewise linear functions on the standard triangulation of $\mathcal{R}$ with step size d being $\frac{2}{N}$, see Figure 4.3. We take $X\mathbb{R}^3 \subset \mathcal{V}_d^1$ to be the space spanned by the Ritz vectors constructed by the SPTARN procedure from MATLAB®.

Note that any space $X\mathbb{R}^3 \subset \mathcal{V}_d^1$ will do as the test space, so all the errors incurred in the construction of $X\mathbb{R}^3$ as the Ritz space from $\mathcal{V}_d^1$ are not important. We only need an estimate

$$\frac{\sin\Theta}{1-\sin\Theta} < \frac{2\pi^2 - \mu_3}{2\pi^2 + \mu_3}, \quad \{\mu_1, \mu_2, \mu_3\} = \sigma(\Xi), \tag{4.2.41}$$

in order to be able to apply Theorem 2.4.1. The estimate (4.2.41) is to be interpreted so that μ_3 is taken as the best possible upper estimate of $\sigma(\Xi)$ and $2\pi^2$ is the best possible lower estimate of λ_4. Naturally, less sharp estimates, e.g. computed from the error analysis of a finite precision procedure, will deliver just as rigorous a conclusion. Given (4.2.41), Theorem 2.4.1 guarantees that μ_1, μ_2, μ_3 approximate $\lambda_1, \lambda_2, \lambda_3$. We will now be able to apply Theorem 4.2.10 directly (assuming $X\mathbb{R}^3 = \mathcal{X} = \mathcal{Y}$).

We have used the subspace $\mathcal{V}_d^1$ to generate the test space $X\mathbb{R}^3$. Let now $d_2 \geq d$ be such that $\mathcal{V}_d^1 \subset \mathcal{V}_{d_2}^1$. Take $x \in X\mathbb{R}^3$, then Lemma 4.2.7 applied to the vector x and the subspaces $\mathcal{V}_{d_2}^1$ and $X\mathbb{R}^3 \subset \mathcal{V}_{d_2}^1$ yields

$$\begin{aligned} (x, \Omega x) - (x, \Omega_{\mathcal{V}_{d_2}^1} x) &= \|r_{(x,\mathcal{V}_{d_2}^1, H_0^1(\mathcal{R}))}\|^2 = h[\mathbf{H}^{-1}x - \mathbf{H}_{\mathcal{V}_{d_2}^1}^{-1} x], \\ (x, \Omega_{\mathcal{V}_{d_2}^1} x) - (x, \Xi^{-1}x) &= \|r_{(x,X\mathbb{R}^3,\mathcal{V}_d^1)}\|^2 = h[\mathbf{H}_{\mathcal{V}_{d_2}^1}^{-1} x - \mathbf{H}_{X\mathbb{R}^3}^{-1} x]. \end{aligned} \tag{4.2.42}$$

An a priori estimate (dependent only on the vector $x \in \mathcal{V}_{d_2}^1$) of the energy norm of the error in the Galerkin approximation $\mathbf{H}_{\mathcal{V}_{d_2}^1}^{-1} x \in \mathcal{V}_{d_2}^1$ to the solution of the Poisson's equation

$$\begin{aligned} -\triangle u &= x, \quad \text{in } \mathcal{R}, \\ u|_{\partial\mathcal{R}} &= 0, \end{aligned} \tag{4.2.43}$$

was derived in [4, section 4.]. We have

$$\sqrt{h[\mathbf{H}^{-1}x - \mathbf{H}^{-1}_{\mathcal{V}^1_{d_2}}\, x]} \le 1.207\ d_2\ \|x\|$$

or equivalently

$$(x, \Omega x) - (x, \Omega_{\mathcal{V}^1_{d_2}}\, x) = \|r_{(x,\mathcal{V}^1_{d_2},H^1_0(\mathcal{R}))}\|^2 \le (1.207)^2 d_2^2 \|x\|^2. \tag{4.2.44}$$

We will now briefly illustrate what happened here. Let $\pi_{d_2} : H^2 \cap \mathcal{Q} \to \mathcal{V}^1_{d_2}$ be the interpolation operator and let $\mathbf{H}^{-1}_{\mathcal{V}^1_{d_2}}\, x$ be the Galerkin approximation to $\mathbf{H}^{-1}x$, then

$$h[\mathbf{H}^{-1}x - \mathbf{H}^{-1}_{\mathcal{V}^1_{d_2}}\, x]^{1/2} = \min_{v \in \mathcal{V}^1_{d_2}} h[\mathbf{H}^{-1}x - v]^{1/2} \le h[\mathbf{H}^{-1}x - \pi_{d_2}\mathbf{H}^{-1}x]^{1/2}.$$

By a lengthy but straightforward computation, which can be found in [55], one establishes

$$h[\mathbf{H}^{-1}x - \pi_{d_2}\mathbf{H}^{-1}x]^{1/2} \le 2\ d_2\ |\mathbf{H}^{-1}x|_{2,2}.$$

With the help of the advanced techniques (the Sard kernel theory) this estimate can been improved, see [4], to obtain

$$h[\mathbf{H}^{-1}x - \pi_{d}\mathbf{H}^{-1}x]^{1/2} \le 1.207\ d_2\ |\mathbf{H}^{-1}x|_{2,2}.$$

The convexity of $\mathcal{R}$ implies $|\mathbf{H}^{-1}x|_{2,2} \le \|\mathbf{H}\mathbf{H}^{-1}x\| = \|x\|$, so

$$h[\mathbf{H}^{-1}x - \mathbf{H}^{-1}_{\mathcal{V}^1_{d_2}}\, x]^{1/2} \le h[\mathbf{H}^{-1}x - \pi_{d}\mathbf{H}^{-1}x]^{1/2} \le 1.207\ d_2\ \|x\|.$$

Now, (4.2.42) and (4.2.44) give an estimate

$$\begin{aligned}
\delta_{\mathcal{V}^1_{d_2}} &= \max_{x \in X\mathbb{R}^3} \frac{|(x, \Omega x) - (x, \Omega_{\mathcal{V}^1_{d_2}}\, x)|}{x^*\ \Omega_{\mathcal{V}^1_{d_2}}\ x} \\
&\le \max_{x \in X\mathbb{R}^3} \frac{|(x, \Omega x) - (x, \Omega_{\mathcal{V}^1_{d_2}}\, x)|}{x^*\ \Xi^{-1}\ x} \\
&\le (1.207)^2 d^2 \max_{x \in X\mathbb{R}^3} \frac{\|x\|^2}{x^* \Xi^{-1} x} = \widehat{\delta}_{\mathcal{V}^1_{d_2}}\,.
\end{aligned} \tag{4.2.45}$$

Lemma 4.2.7 implies $\Omega \le \Omega_{\mathcal{V}^1_{d_2}}$ and

$$\sin^2 \Theta_{\mathcal{V}^1_{d_2}} = \max_{x \in \mathbb{R}^3} \frac{x^*\Omega_{\mathcal{V}^1_{d_2}} x - x^*\Xi^{-1}x}{x^*\Omega_{\mathcal{V}^1_{d_2}} x}, \qquad \sin^2 \Theta = \max_{x \in \mathbb{R}^3} \frac{x^*\Omega x - x^*\Xi^{-1}x}{x^*\Omega x}.$$

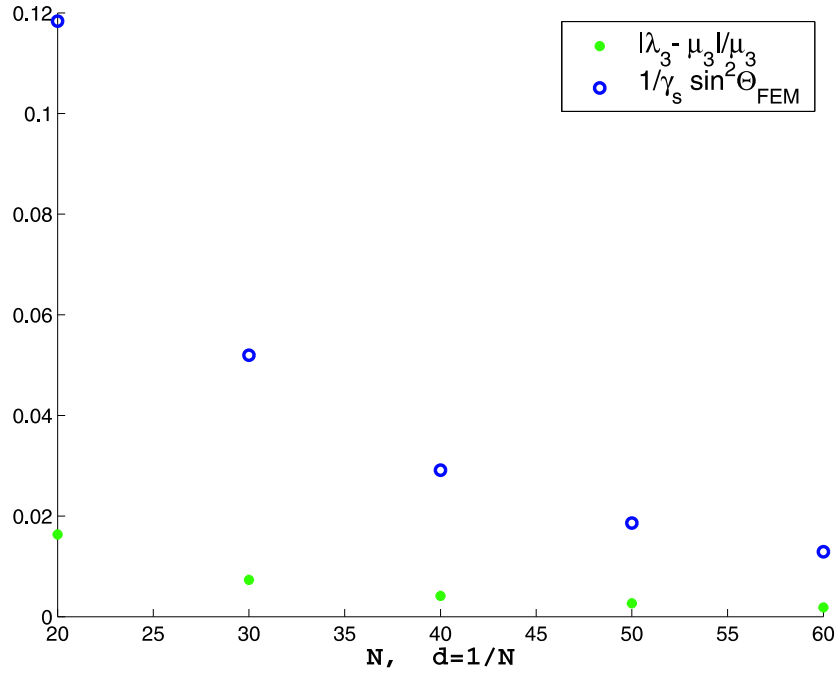

Figure 4.4: The error in the approximation of $\lambda_2 = \lambda_3$ ($\sin^2\Theta_{FEM} = \sin^2\Theta_{FEM}(\mathcal{V}^1_{d/4})$).

Lemma 4.2.19 yields

$$\sin^2\Theta_{\mathcal{V}^1_{d_2}} \leq \sin^2\Theta \leq \frac{\sin^2\Theta_{\mathcal{V}^1_{d_2}} + \widehat{\delta}_{\mathcal{V}^1_{d_2}}}{1 + \widehat{\delta}_{\mathcal{V}^1_{d_2}}} =: \sin^2\Theta_{FEM}(\mathcal{V}^1_{d_2}). \tag{4.2.46}$$

The quantity $\sin\Theta_{\mathcal{V}^1_{d_2}}$ measures the defect of the space $X\mathbb{R}^3$ when considered as an invariant subspace of the matrix $H_{\mathcal{V}^1_{d_2}}$ (or of the operator $\mathbf{H}_{\mathcal{V}^1_{d_2}}$, cf. Lemma 2.3.2). Not surprisingly, one obtains $\sin\Theta_{\mathcal{V}^1_d} = 0$, since $X\mathbb{R}^3$ is spanned by the Ritz vectors of the operator $\mathbf{H}$ from the subspace $\mathcal{V}^1_d$.

Remark 4.2.20. Since $X\mathbb{R}^n = \mathcal{X} = \mathcal{Y}$, this shows a way to use $\delta_{\mathcal{V}}$ to (retroactively) remove the assumption $\mathbf{H}_{\mathcal{Y}}x = \mu_x x$ from Theorem 4.2.8. In the theoretical considerations we can always assume $\sin\Theta_{\mathcal{Y}} = 0$. Theorem 4.2.10 now reveals the nature of the approximation to the eigenvalues of the operator $\mathbf{H}$ by the Ritz values from the finite element space $\mathcal{V}^1_d$. For instance we have

$$\frac{\mu_1 - \lambda_1}{\mu_1} = \mathrm{O}(d^2) = \mathrm{O}(\frac{1}{N^2}).$$

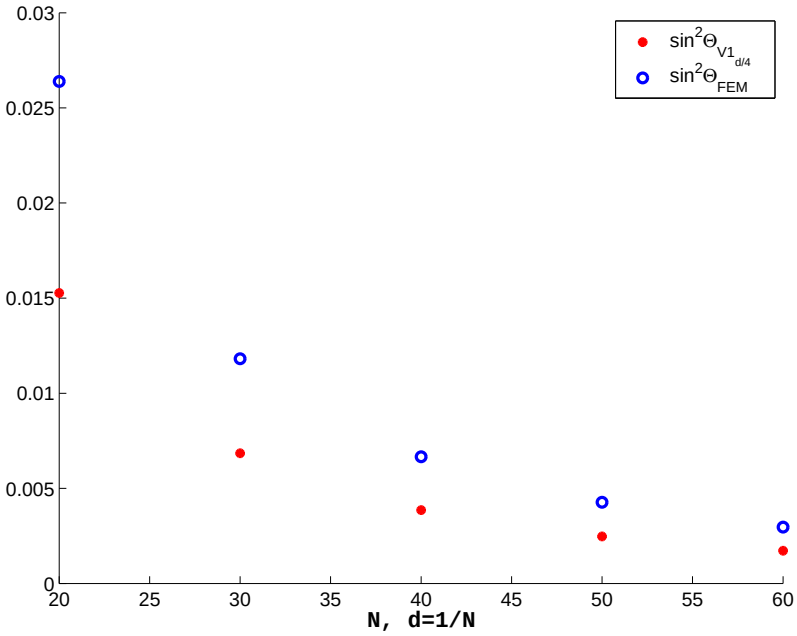

Figure 4.5: Testing the "sandwich" estimate (4.2.46)— $\sin^2\Theta_{FEM}(\mathcal{V}^1_{d/4}) - \sin^2\Theta_{\mathcal{V}^1_{d/4}} \approx 10^{-3}$

The role of $\sin^2\Theta_{\mathcal{V}^1_{d_2}}$ in (4.2.46) is to control the influence of the method that was used to generate the test subspace $X\mathbb{R}^3$. Here it can be used to absorb the influence of SPTARN procedure. In the next section we will compare $\sin\Theta_{\mathcal{V}^1_{d_2}}$ with another measure of the "approximation defect".

4.3 Alternative measures of the residual—direct estimates

So far we have used the $\mathbf{H}^{-1}$–norm of the residual

$$\|r_{\widetilde{u}}\|_{\mathbf{H}^{-1}} = \max_{v\in\mathcal{Q}} \frac{|\langle r_{\widetilde{u}}, v\rangle|}{\|\mathbf{H}^{1/2}v\|} = \|\mathbf{H}^{-1/2}\widetilde{u} - \widetilde{\mu}\mathbf{H}^{-1/2}\widetilde{u}\|$$

to measure the defect of the vector $u \in \mathcal{Q}$ when considered as the solution to the eigenvalue problem (4.2.23). In this section we introduce several alternative measures of the residual, cf. [42]. We will firstly decouple the analysis of the residual measures from the problem of obtaining eigenvalue estimates. The reason is that the problem of how to estimate the residual is a problem in the theory of Sobolev spaces, and the problem of

obtaining the eigenvalue estimates given a measure of the residual is a problem in the perturbation theory. We firmly believe in separating these two features of finite element spectral estimates.

Theorem 4.3.1. *Let* $\mathbf{H}$ *be a positive definite operator. Take* $\widetilde{u} \in \mathcal{Q}$ *of norm one. If*

$$\|r_{\widetilde{u}}\|_{\mathbf{H}^{-2}} = \max_{v \in \mathcal{D}} \frac{|\langle r_{\widetilde{u}}, v\rangle|}{\|\mathbf{H}v\|} < \frac{\lambda_e - \widetilde{\mu}}{\lambda_e} \tag{4.3.1}$$

then

$$\min_{\lambda \in \sigma_d(\mathbf{H})} \frac{|\lambda - \widetilde{\mu}|}{\lambda} \leq \|r_{\widetilde{u}}\|_{\mathbf{H}^{-2}} = \max_{v \in \mathcal{D}(\mathbf{H})} \frac{|\langle r_{\widetilde{u}}, v\rangle|}{\|\mathbf{H}v\|}.$$

Proof. Let $\widetilde{u} \in \mathcal{Q}$, $\|\widetilde{u}\| = 1$, we compute

$$\begin{aligned}
\|r_{\widetilde{u}}\|^2_{\mathbf{H}^{-2}} &= (\mathbf{H}\widetilde{u} - \widetilde{\mu}\widetilde{u}, \mathbf{H}^{-2}\ (\mathbf{H}\widetilde{u} - \widetilde{\mu}\ \widetilde{u})) \\
&= \|\widetilde{u} - \widetilde{\mu}\ \mathbf{H}^{-1}\widetilde{u}\|^2 \\
&= \int \left(\frac{\lambda - \widetilde{\mu}}{\lambda}\right)^2 \mathrm{d}(E(\lambda)\widetilde{u}, \widetilde{u}) \\
&\geq \min_{\lambda \in \sigma(\mathbf{H})} \left(\frac{\lambda - \widetilde{\mu}}{\lambda}\right)^2.
\end{aligned}$$

The assumptions of the theorem assure us that the minimum is achieved on the point in the discrete spectrum of the operator. Finally, we obtain

$$\min_{\lambda \in \sigma_d(\mathbf{H})} \frac{|\lambda - \widetilde{\mu}|}{\lambda} \leq \|r_{\widetilde{u}}\|_{\mathbf{H}^{-2}}.$$

□

There are a lot of examples where one can efficiently estimate $\|r_{\widetilde{u}}\|_{\mathbf{H}^{-2}}$ by a direct analysis, see [42]. Under the assumption (2.6.9) Theorems 2.6.4 and 4.3.1 imply that for every $\widetilde{u} \in \mathsf{ran}(X)$ we have

$$\frac{|\lambda_m - \widetilde{\mu}|}{\lambda_m} \leq \|r_{\widetilde{u}}\|_{\mathbf{H}^{-2}}.$$

We can now use a technique from [42] to estimate

$$\|r_{\widetilde{u}}\|_{\mathbf{H}^{-2}} = \max_{v \in \mathcal{D}(\mathbf{H})} \frac{|\langle r_{\widetilde{u}}, v\rangle|}{\|\mathbf{H}v\|},$$

for $\widetilde{u}$ generated by finite element procedures. The eigenvalue estimate from [42] can be seen as a consequence of our Theorem 4.3.1, but our result is sharper and more general. The technique of obtaining the direct estimates of $\|r_{\widetilde{u}}\|_{\mathbf{H}^{-2}}$ is taken completely from [42]. In order to be able to compare the estimates based on $\|r_{\widetilde{u}}\|_{\mathbf{H}^{-2}}$ with estimates based on $\mathsf{sin}\Theta$ (on several model problems), we compute all the approximation constants explicitly (see Section 4.3.1).

4.3.1 Finite element residual estimates in the regular case

In this section we will demonstrate how to obtain a computational procedure from the theoretical results of the preceding sections. We will present the discussion on two illustrative examples in 1D and 2D.

Assuming the region $\mathcal{R}$ be more regular (e.g. convex) the dual norm of the residual $\|r_{\widetilde{u}}\|_{\mathbf{H}^{-2}}$ can be analyzed directly. The following direct analysis of the residual in the finite element procedures is taken from [31, 42]. It will be applied to estimate $\|r_{\widetilde{u}}\|_{\mathbf{H}^{-2}}$ in Theorem 4.3.1. The estimates will be based upon the computed approximate eigenvector.

We consider various means to compute eigenvalue estimates for a given triangulation (mesh) regardless of how it was constructed. With this task in mind, we compare various eigenvalue estimates on a set of case studies.

Let $\mathcal{R} \subset \mathbb{R}^r$, $r = 1, 2$ be a bounded polygonal region[4] and let $\mathbf{H}$ be a self-adjoint positive definite differential operator defined by the form

$$h(u,v) = \int_{\mathcal{R}} (\nabla u)^* \, \nabla v \, dx + \int_{\mathcal{R}} b(\cdot) u \, v \, dx = (\mathbf{H}^{1/2}u, \mathbf{H}^{1/2}v), \quad u, v \in \mathcal{Q} = H_0^1(\mathcal{R}) \subset L^2(\mathcal{R}). \tag{4.3.2}$$

Other boundary condition that lead to the positive definite form h are also possible. We pick the Dirichlet boundary condition to ease the technical side of the presentation. This, as before, does not lower the level of generality.

For definiteness, we use the finite element space

$$\mathcal{V}_d^1 = \{u \in C(\,\overline{\mathcal{R}}\,) : v|_K \text{ is a linear function}\} \subset \mathcal{Q}$$

and remark that other more general finite element spaces could have also been considered, see [42].

An unfortunate restriction on the region $\mathcal{R}$ is that it must be such that

$$S_2 = \max_{v \in \mathcal{D}(\mathbf{H})} \frac{|v|_{2,2}}{\|\mathbf{H}v\|} < \infty. \tag{4.3.3}$$

Note that estimates from Section 4.2 did not suffer from this deficit.

A class of efficient a posteriori and a priori estimates for eigenvalue problem proposed in [31, 42] is based on this *Stability property*, customarily written as

$$|v|_{2,2} \le S_2 \|\mathbf{H}v\|, \quad v \in \mathcal{D}(\mathbf{H}). \tag{4.3.4}$$

[4] The procedure we are about to describe can be applied on higher dimensional problems, too (See [42]).

In the case of the boundary value problem $\mathbf{H}u = f$, (4.3.4) amounts to

$$|u|_{2,2} \leq S_2 \|f\|. \tag{4.3.5}$$

For boundary value problems (4.2.21)

$$S_1 = \max_{v \in \mathcal{Q}(\mathbf{H})} \frac{\|D^1 v\|}{\|\mathbf{H}^{1/2} v\|} \tag{4.3.6}$$

is used instead. Error estimates based on *Stability property* can deliver overly pessimistic estimates if S_2 turns out to be too large.

A posteriori estimates are also known in the literature as local residual estimates, see [31, 42, 60]. For a priori estimates see [4, 5, 6].

To derive local residual estimates for $\widetilde{u} \in \mathcal{V}_d^1$ we need:
A *trace inequality*

$$\|u\|_{L^2(\partial K)} \leq C_T(d_K^{-1/2}\|u\|_{L^2(K)} + d_K^{1/2}\|Du\|_{L^2(K)}), \quad u \in H^1(\mathcal{R}) \cap \mathcal{Q}$$

and *approximation estimates*

$$\|u - \pi_d u\| \leq C_0 \; d^{\,2}|u|_{2,2}, \quad u \in H^2(\mathcal{R}) \cap \mathcal{Q}, \tag{4.3.7}$$

$$\|u - \pi_d u\|_{H_0^1} \leq C_1 \; d|u|_{2,2}, \quad u \in H^2(\mathcal{R}) \cap \mathcal{Q}. \tag{4.3.8}$$

Here, we have taken $\pi_d : H^2 \cap \mathcal{Q} \to \mathcal{V}_n^1$ to be the interpolation operator. A similar analysis can be performed for the residual of a boundary value problem.

We use several results from [31, 42] which are summed up in the following theorem.

Theorem 4.3.2. *Let* $\mathbf{H}$ *be the operator of the type (4.3.2). We assume that* $\mathcal{R}$ *is a polygonal domain and that the Dirichlet boundary condition is imposed. Let* $r_{\widetilde{u}}$ *and* $r_{\widetilde{u}}^b$ *be the residuals for Problems (4.2.23) and (4.2.21) and* $\widetilde{u} \in \mathcal{V}_d^1$*, then we have the estimates*

$$|\langle r_{\widetilde{u}}, v\rangle| \leq C_{\mathcal{T}_d}\|\widetilde{d}^{\,2} R_{\widetilde{u}}\||v|_{2,2}, \quad v \in H^2 \cap \mathcal{Q} \tag{4.3.9}$$

$$|\langle r_{\widetilde{u}}, v\rangle| \leq C_{\mathcal{T}_d}^1\|\widetilde{d}\; R_{\widetilde{u}}\||v|_{1,2}, \quad v \in H^1 \cap \mathcal{Q} \tag{4.3.10}$$

$$|\langle r_{\widetilde{u}}^b, v\rangle| \leq C_{\mathcal{T}_d}^1\|\widetilde{d}\; R_{\widetilde{u}}^b\||v|_{1,2}, \quad v \in H^1 \cap \mathcal{Q}. \tag{4.3.11}$$

The functions $\widetilde{h}, R_{\widetilde{u}}, R_{\widetilde{u}}^b \in L^2(\mathcal{R})$ *are defined element vise by the formulas*

$$\widetilde{d}\Big|_K = d_K = diam(K),$$
$$R_{\widetilde{u}}^b\big|_K = |\mathbf{H}\widetilde{u} - b|_K + d_K^{-1/2}|vol(K)|^{-1/2}\|[(A\nabla\widetilde{u})^* n]\|_{L^2(\partial K)},$$
$$R_{\widetilde{u}}|_K = |\mathbf{H}\widetilde{u} - \widetilde{\mu}\widetilde{u}|_K + d_K^{-1/2}|vol(K)|^{-1/2}\|[(A\nabla\widetilde{u})^* n]\|_{L^2(\partial K)}.$$

Here, $[(A\nabla\widetilde{u})^* n]$ *denotes the jump across* ∂K $(K \in \mathcal{T}_d)$ *of the exterior normal derivative* $(A\nabla\widetilde{u})^* n$

In general situations assumptions (4.3.7) and (4.3.8) are too restrictive. To prove (4.3.10) and (4.3.11) we need interpolation error estimates that are valid for all $v \in H^1 \cap \mathcal{Q}$, but functions in $H^1(\mathcal{R})$ need not be continuous. To establish estimates like (4.3.7) and (4.3.8), we need a notion of a more general interpolation operator. We do not go into more details here, but point the reader to [31].

Theorem 4.3.2 and (4.3.5) imply

$$\|r_{\widetilde{u}}\|_{\mathbf{H}^{-2}} = \max_{v \in \mathcal{D}(\mathbf{H})} \frac{|\langle r_{\widetilde{u}}, v\rangle|}{\|\mathbf{H}v\|} \leq S_2 \; C_{\mathcal{T}_d} \|d^2 R_{\widetilde{u}}\| \tag{4.3.12}$$

$$\max_{v \in \mathcal{Q}(\mathbf{H})} \frac{|\langle r_{\widetilde{u}}, v\rangle|}{\|\mathbf{H}^{1/2}v\|} \leq S_1 \; C^1_{\mathcal{T}_d} \|d^2 R_{\widetilde{u}}\| \tag{4.3.13}$$

$$\|r^b_{\widetilde{u}}\|_{\mathbf{H}^{-1}} = \max_{v \in \mathcal{Q}(\mathbf{H})} \frac{|\langle r^b_{\widetilde{u}}, v\rangle|}{\|\mathbf{H}^{1/2}v\|} \leq S_1 \; C^1_{\mathcal{T}_d} \|d \; R^b_{\widetilde{u}}\|. \tag{4.3.14}$$

The local residual estimates of this type deliver realistic estimates when we have a small stability constant S_2. Note that in a convex polygonal region $\mathcal{R} \subset \mathbb{R}^2$ we have (for the proof see [7])

$$|u|_{2,2} \leq \|\triangle u\|, \quad u \in H^1_0(\mathcal{R}).$$

Remark 4.3.3. An estimate of the type (4.3.14) is envisaged to be used, in conjunction with the estimate from Theorem 4.2.14, to deliver rigorous eigenvalue estimates for a wide class of eigenvalue problems. In what follows we have a good a priori estimate available, so we use those instead.

4.3.2 The finite element case studies — revisited

Theorem 4.3.1 is temptingly simple. In order to make use of it one is prepared to ignore the additional regularity constraints. However, as the case studies will show, succumbing to this temptation, without further consideration, does more than just restrict the generality of the method. The most important feature of numerical analysis is to correctly identify the measures of the stability of the problem. Our suggestion, based on the paradigm from [60, Verfürth], is to keep the stability analysis of the eigenvalue problem fully separate from the regularity requirements necessary to measure the residual. Furthermore, it is to be attempted to measure the residual under the minimal regularity constraints. The use of the more regular residual measure $\|r_{\widetilde{u}}\|_{\mathbf{H}^{-2}}$ can force the appearance of unpleasantly large "regularity" constants in (4.3.7).

As an illustration we take an eigenvalue bound from Theorem 4.3.1 and compare it to the bound obtained from the "sandwich" inequality (4.2.46) and Theorem 2.6.1. In order to get a fair comparison we will limit ourselves to the case $\mathsf{sin}\Theta_{\mathcal{V}^1_d} = 0$.

Let $\mathbf{H}$ be the operator defined by the form (4.2.39). For a regular triangulation and the interpolation operator $\pi_d : H^2 \cap \mathcal{Q} \to \mathcal{V}_d^1$ we have the following approximation result, see [4, 30]

$$\|v - \pi_d v\| \leq \frac{8\sqrt{6}}{\sqrt{\pi}(2-\sqrt{2})^2}\, d^2 |v|_{2,2} \leq \frac{8\sqrt{6}}{\sqrt{\pi}(2-\sqrt{2})^2}\, d^2 \|\triangle v\|, \tag{4.3.15}$$

$$\|v - \pi_d v\|_{H_0^1} \leq 1.207 d |v|_{2,2} \leq 1.207 d \|\triangle v\|. \tag{4.3.16}$$

Trace inequality takes the form of

$$\|v\|_{L^2(\partial K)} \leq \sqrt{8+4\sqrt{2}}\, (d^{-1/2}\|v\|_{L^2(K_S)} + d^{1/2}\|Dv\|_{L^2(K_S)})$$

for any $v \in H^1 \cap \mathcal{Q}$ and $K \in \mathcal{T}_d$.

The bound from Theorem 4.3.2 reads, assuming $\|\widetilde{u}\| = 1$,

$$\begin{aligned} |\langle r_{\widetilde{u}}, v\rangle| &\leq [\frac{8\sqrt{6}}{\sqrt{\pi}(2-\sqrt{2})^2}\, \widetilde{\mu}\, d^2 \\ &\quad + \frac{\sqrt{8+4\sqrt{2}}}{2}(1.207 + \frac{8\sqrt{6}}{\sqrt{\pi}(2-\sqrt{2})^2})\sqrt{\sum_K \|[(\nabla \widetilde{u})^* n]\|^2}\, d^{3/2}\,]|v|_{2,2} \\ \rightsquigarrow \rho_{FEM} &:= [\frac{8\sqrt{6}}{\sqrt{\pi}(2-\sqrt{2})^2}\, \widetilde{\mu}\, d^2 \\ &\quad + \frac{\sqrt{8+4\sqrt{2}}}{2}(1.207 + \frac{8\sqrt{6}}{\sqrt{\pi}(2-\sqrt{2})^2})\sqrt{\sum_K \|[(\nabla \widetilde{u})^* n]\|^2}\, d^{3/2}\,]. \end{aligned}$$

On Figure 4.6 we see a comparison of the bounds from Theorems 2.6.1 and 4.3.1. The gap γ is roughly 0.2 so Theorem 2.4.1 guarantees that $\widetilde{\mu}$ matches λ_1 in all of the theorems. A bound based on Theorem 4.3.1 and on local residual estimate for $\|r_{\widetilde{u}}\|_{\mathbf{H}^{-2}}$ yields an estimate of the error that is superlinear in d. Yet, it underperforms, even when compared with the bound based on $\mathsf{sin}\Theta = O(d)$ (see Section 4.2.4), due to the large size of the approximation constant in (4.3.15).

To see that this is not the fault of Theorem 4.3.1, but rather of our inability to accurately estimate the "-2"-norm of the residual, we will reconsider Problem (2.7.7). Theorem 4.3.1 yields strikingly accurate estimates when applied to the operator in Problem (2.7.7), even in the presence of the large stability constant S_2. In $1D$ we have $D^2 u = \partial_{xx} u$, so (4.3.3) reads

$$S_2(\alpha) = \max_{u \in \mathcal{D}(\mathbf{H})} \frac{\|\partial_{xx} u\|}{\|\mathbf{H}u\|} = \max_{u \in \mathcal{D}(\mathbf{H})} \frac{\|\partial_{xx} u\|}{\|-\partial_{xx} u - \alpha u\|} = \frac{\theta^2}{-4\,\pi^2\,\alpha + \theta^2}.$$

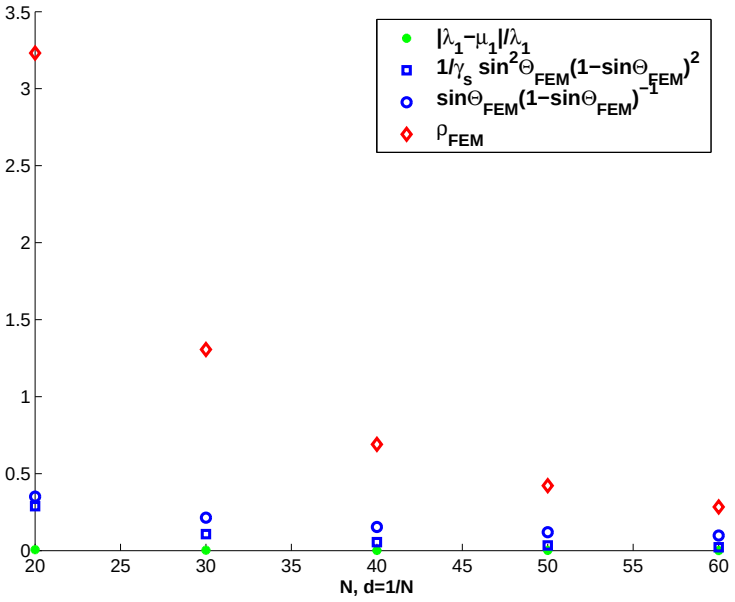

Figure 4.6: Direct estimates for the 2D-model problem.

As $\alpha \to \theta^2/(2\pi)^2$ the stability constant $S_2(\alpha)$ goes to infinity. For this example the Ritz values and the Ritz vectors (assuming $\alpha = 0$) are given by the formula

$$\mu_{(d,k)} = 6d^{-2}\frac{1-\cos(d(-\theta(2\pi)^{-1}+k))}{2+\cos(d(-\theta(2\pi)^{-1}+k))}, \tag{4.3.17}$$

$$u_{(d,k)} = [1 \quad e^{\mathrm{i}\left(\frac{\theta}{2\pi}+(-1)^{k+1}\lfloor\frac{k}{2}\rfloor\right)d^{-1}} \quad \dots \quad e^{\mathrm{i}\left(\frac{\theta}{2\pi}+(-1)^{k+1}\lfloor\frac{k}{2}\rfloor\right)d^{-1}(N-1)}]^* \tag{4.3.18}$$

for $k = 0, \dots, N-1$. In the usual notation we have

$$\mu_1 = \mu_{(d,0)} - \alpha, \qquad u_1 = u_{(d,0)}.$$

The residual estimates in Theorem 4.3.2 are extremely simple in 1D since the boundary terms in (4.3.9) disappear. More importantly, the interpolation estimates (4.3.7) and (4.3.8) can be computed with optimal constants, see [31],

$$\|v - \pi_d v\| \le \frac{1}{\pi^2}\, d^2 |v|_{2,2} = \frac{1}{\pi^2}\, d^2 \|\partial_{xx} v\| \tag{4.3.19}$$

$$|v - \pi_d v|_{1,2} \le \frac{1}{\pi} d |v|_{2,2} = \frac{1}{\pi}\, d \|\partial_{xx} v\|. \tag{4.3.20}$$

The estimate for the eigenvalue problem reads

$$\frac{|\lambda_1 - \mu_1|}{\lambda_1} \le |\alpha + \mu_1| \frac{1}{\pi^2} S_2(\alpha) d^2 = \mu_{(d,0)} \frac{1}{\pi^2} S_2(\alpha) d^2. \tag{4.3.21}$$

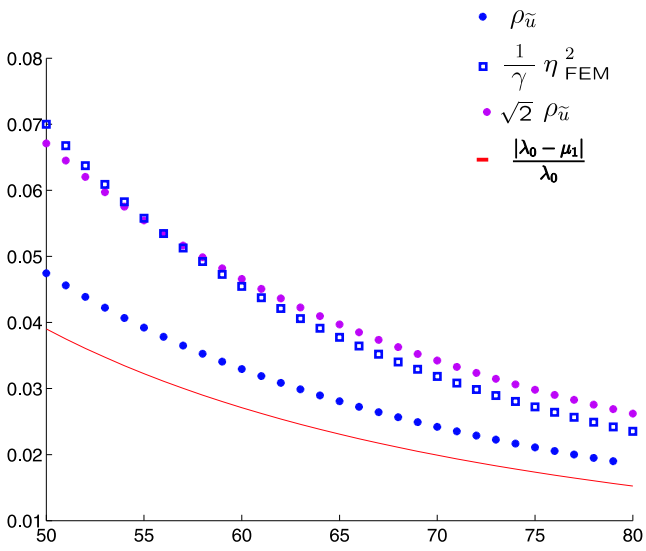

Figure 4.7: Direct estimates for the 1D-model problem

Now, observe that

$$\mu_{(d,0)} = \frac{\theta^2}{4\,\pi^2} + \frac{d^2\,\theta^4}{192\,\pi^4} + \frac{d^4\,\theta^6}{23040\,\pi^6} + O(d^8)$$

and our bound allows us to conclude that $\frac{|\lambda_1-\mu_1|}{\lambda_1} = O(d^2)$. On the other hand,

$$\frac{\mu_1-\lambda_1}{\lambda_1} = \frac{\theta^4\,d^2}{48\,\pi^2\,(-4\,\pi^2\,\alpha+\theta^2)} - \frac{\theta^6\,d^4}{5760\,(4\,\pi^6\,\alpha-\pi^4\,\theta^2)} + O(d)^6$$

and we see that estimate (4.3.21) is of the optimal order. We have established that $\mathsf{sin}\Theta = O(d)$, so we cannot expect (at least asymptotically) to have an estimate that is of higher order than quadratic in $\mathsf{sin}\Theta$. The large value of $S_2(\alpha)$ does not hurt the estimate (4.3.21) since d^2 decays fast enough. We plot all of the bounds on Figure 4.7.

For this example we compute all the estimates exactly using formulas (2.7.19), (4.3.18) and (4.3.21), as well as using the formula (4.3.17) to compute the actual errors.

A deficiency of a result like Theorem 4.3.1 lies in a difficulty to incorporate a mechanism to localize the approximated eigenvalues. Some natural set of a priori assumptions has to be specified to tell us when we are approximating only the desired eigenvalues with the measured accuracy. The character of Theorem 4.3.1 is that it gives an estimate of the distance to the nearest eigenvalue regardless of how many other eigenvalues are clustered together in its neighborhood. On the other hand, the a priori assumptions on the relation

between the spectral gap and the $\sin\Theta$ appear to be a natural requirement. This can best be observed in the proof of Theorem 3.3.8.

4.4 Conclusion

We have shown that the ability to compute relative eigenvalue approximation estimates, by an application of a theorem from Section 2, is equivalent to the assumption that we can construct a subspace $\mathcal{V}$ such that $\mathcal{Y}$ and $\mathcal{V}$ satisfy a saturation assumption for the computed Ritz vector x. An efficient method for obtaining eigenvalue estimates should be based on the construction of equivalent finite dimensional energy norms. Such norms can be constructed under very weak regularity assumptions on the domain $\mathcal{R}$. Estimating the residual directly often leads to estimate that can suffer from unnecessarily high regularity requirements. Furthermore, successful application of the higher order estimates requires a careful localization of the approximated (unknown) eigenvalues. From the discussion of the splitting methods of approximating the solution of the stationary problem (from [2, 14, 66]) we see that an approach through the construction of alternative energy norms leads to the construction of efficient and asymptotically sharp eigenvalue estimates under minimal regularity constraints.

Contributions in this chapter

Here we have presented an application of the results from Chapter 2 to the problem of estimating the quality of the finite element spectral approximations. We now list the main contributions in this chapter:

- We have reduced the analysis of the spectral problem to the analysis of the auxiliary stationary problem (Problem 4.1) with the computed Ritz vector x as the right hand side. Furthermore, an abstract condition (the saturation assumption) necessary to define a finite dimensional procedure to assess the residual measures has been introduced, see Section 4.2.

- We have provided a joint abstract framework to obtain eigenvalue, eigenvector and invariant subspace estimates for finite element spectral approximations, according to Corollary 4.2.5, Theorem 2.5.2 and Theorem 4.2.10.

- A "sandwich" inequality, relating the $\mathbf{H}^{-1}$–norm of the Ritz vector residual and $\sin\Theta$ has been established, see Section 4.2.3 and Theorem 4.2.13.

- We have introduced a new target function, based on the measure of the oscillation of the computed Ritz vector, which can be used as an indicator for mesh refinement. This is envisaged as a companion to the residual refinement techniques of [48], see Section 4.2.2 and Theorem 4.2.15.

- A method to derive computable estimates for the eigenvalues and the eigenvectors of operators defined on the regular domains, as well as detailed comparison with the direct residual measures, has been provided, see Section 4.2.4 and Section 4.3.

For other minor contributions we refer the reader to local references.

Appendices

Bibliography

[1] P. Arbenz and Z. Drmač. On positive semidefinite matrices with known null space. *SIAM J. Matrix Anal. Appl.*, 24(1):132–149 (electronic), 2002.

[2] R. E. Bank, T. F. Dupont, and H. Yserentant. The hierarchical basis multigrid method. *Numer. Math.*, 52(4):427–458, 1988.

[3] J. Barlow and J. Demmel. Computing accurate eigensystems of scaled diagonally dominant matrices. *SIAM J. Numer. Anal.*, 27(3):762–791, 1990.

[4] R. E. Barnhill, J. H. Brown, and A. R. Mitchell. A comparison of finite element error bounds for Poisson's equation. *IMA J. Numer. Anal.*, 1(1):95–103, 1981.

[5] R. E. Barnhill and J. A. Gregory. Interpolation remainder theory from Taylor expansions on triangles. *Numer. Math.*, 25(4):401–408, 1975/76.

[6] R. E. Barnhill and J. A. Gregory. Sard kernel theorems on triangular domains with application to finite element error bounds. *Numer. Math.*, 25(3):215–229, 1975/76.

[7] R. E. Barnhill and C. H. Wilcox. Computable error bounds for finite element approximations to the Dirichlet problem. *Rocky Mountain J. Math.*, 12(3):459–470, 1982.

[8] H. Baumgärtel. *Analytic perturbation theory for matrices and operators*, volume 15 of *Operator Theory: Advances and Applications*. Birkhäuser Verlag, Basel, 1985.

[9] H. Baumgärtel and M. Demuth. Decoupling by a projection. *Rep. Math. Phys.*, 15(2):173–186, 1979.

[10] C. Beattie and F. Goerisch. Methods for computing lower bounds to eigenvalues of self-adjoint operators. *Numer. Math.*, 72(2):143–172, 1995.

[11] R. Bhatia, C. Davis, and A. McIntosh. Perturbation of spectral subspaces and solution of linear operator equations. *Linear Algebra Appl.*, 52/53:45–67, 1983.

[12] F. Bornemann and H. Yserentant. A basic norm equivalence for the theory of multilevel methods. *Numer. Math.*, 64(4):455–476, 1993.

[13] F. A. Bornemann and P. Deuflhard. The cascadic multigrid method for elliptic problems. *Numer. Math.*, 75(2):135–152, 1996.

[14] F. A. Bornemann, B. Erdmann, and R. Kornhuber. A posteriori error estimates for elliptic problems in two and three space dimensions. *SIAM J. Numer. Anal.*, 33(3):1188–1204, 1996.

[15] J. Brasche and M. Demuth. Dynkin's formula and large coupling convergence. *J. Funct. Anal.*, 219(1):34–69, 2005.

[16] F. Brezzi and M. Fortin. *Mixed and hybrid finite element methods*, volume 15 of *Springer Series in Computational Mathematics*. Springer-Verlag, New York, 1991.

[17] V. Bruneau and G. Carbou. Spectral asymptotic in the large coupling limit. *Asymptot. Anal.*, 29(2):91–113, 2002.

[18] F. Chatelin. *Spectral approximation of linear operators*. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York, 1983. With a foreword by P. Henrici, With solutions to exercises by Mario Ahués.

[19] P. G. Ciarlet. *The finite element method for elliptic problems*. North-Holland Publishing Co., Amsterdam, 1978. Studies in Mathematics and its Applications, Vol. 4.

[20] E. B. Davies. ICMS lecture notes on computational spectral theory. In *Spectral theory and geometry (Edinburgh, 1998)*, pages 76–94. Cambridge Univ. Press, Cambridge, 1999.

[21] C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.*, 7:1–46, 1970.

[22] C. Davis, W. M. Kahan, and H. F. Weinberger. Norm-preserving dilations and their applications to optimal error bounds. *SIAM J. Numer. Anal.*, 19(3):445–469, 1982.

[23] M. Demuth, F. Jeske, and W. Kirsch. Rate of convergence for large coupling limits by Brownian motion. *Ann. Inst. H. Poincaré Phys. Théor.*, 59(3):327–355, 1993.

[24] E. G. D′jakonov and M. J. Orehov. Minimization of the computational work in eigenvalue problems. *Dokl. Akad. Nauk SSSR*, 235(5):1005–1008, 1977.

[25] W. Dörfler and R. H. Nochetto. Small data oscillation implies the saturation assumption. *Numer. Math.*, 91(1):1–12, 2002.

[26] Z. Drmač. On relative residual bounds for the eigenvalues of a Hermitian matrix. *Linear Algebra Appl.*, 244:155–163, 1996.

[27] Z. Drmač. On principal angles between subspaces of Euclidean space. *SIAM J. Matrix Anal. Appl.*, 22(1):173–194 (electronic), 2000.

[28] Z. Drmač and V. Hari. Relative residual bounds for the eigenvalues of a Hermitian semidefinite matrix. *SIAM J. Matrix Anal. Appl.*, 18(1):21–29, 1997.

[29] N. Dunford and J. T. Schwartz. *Linear operators. Part II.* Wiley Classics Library. John Wiley & Sons Inc., New York, 1988. Spectral theory. Selfadjoint operators in Hilbert space, With the assistance of William G. Bade and Robert G. Bartle, Reprint of the 1963 original, A Wiley-Interscience Publication.

[30] T. Dupont and R. Scott. Constructive polynomial approximation in Sobolev spaces. In *Recent advances in numerical analysis (Proc. Sympos., Math. Res. Center, Univ. Wisconsin, Madison, Wis., 1978)*, volume 41 of *Publ. Math. Res. Center Univ. Wisconsin*, pages 31–44. Academic Press, New York, 1978.

[31] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson. *Computational differential equations.* Cambridge University Press, Cambridge, 1996.

[32] W. G. Faris. *Self-adjoint operators.* Springer-Verlag, Berlin, 1975. Lecture Notes in Mathematics, Vol. 433.

[33] T. Friese. *Eine Mehrgitter-Methode zur Lösung des Eigenwertproblems der komplexen Helmholtzgleichung.* PhD thesis, Freie Universität Berlin, 1998.

[34] F. R. Gantmacher. *The theory of matrices. Vols. 1, 2.* Chelsea Publishing Co., New York, 1959.

[35] L. Grubišić. Computing the partial eigenvalue problem for symmetric matrices (in Croatian: Numeričko računanje parcijalnog problema vlastitih vrijednosti za simetrične matrice). Master's thesis, PMF-matematički odjel, Sveučilište u Zagrebu, September 2001.

[36] L. Grubišić and J. Tambača. Comparison of the arch model and the curved rod model. *Fernuniversitaet Hagen Preprint*, 2003.

[37] L. Grubišić and K. Veselić. On Ritz approximations for positive definite operators I (theory). *To appear in LAA*, 2002.

[38] R. Hempel and O. Post. Spectral gaps for periodic elliptic operators with high contrast: an overview. In *Progress in analysis, Vol. I, II (Berlin, 2001)*, pages 577–587. World Sci. Publishing, River Edge, NJ, 2003.

[39] M. Jurak and J. Tambača. Linear curved rod model. General curve. *Math. Models Methods Appl. Sci.*, 11(7):1237–1252, 2001.

[40] W. Kahan. Inclusion Theorems for Clusters of Eigenvalues of Hermitian Matrices. *Tech. Report*, 1967. Computer Science Department, University of Toronto.

[41] T. Kato. *Perturbation theory for linear operators*. Springer-Verlag, Berlin, second edition, 1976. Grundlehren der Mathematischen Wissenschaften, Band 132.

[42] M. G. Larson. A posteriori and a priori error analysis for finite element approximations of self-adjoint elliptic eigenvalue problems. *SIAM J. Numer. Anal.*, 38(2):608–625 (electronic), 2000.

[43] P. Leinen, W. Lembach, and K. Neymeyr. An adaptive Subspace Method for Elliptic Eigenproblems with Hierarchical Basis Preconditioning. Sonderforshungsbereich 382 Report Nr. 68, Universität Tübingen, 1997.

[44] S. Levendorskiĭ. *Asymptotic distribution of eigenvalues of differential operators*, volume 53 of *Mathematics and its Applications (Soviet Series)*. Kluwer Academic Publishers Group, Dordrecht, 1990. Translated from the Russian.

[45] R. Mathias and K. Veselić. A relative perturbation bound for positive definite matrices. *Linear Algebra Appl.*, 270:315–321, 1998.

[46] Z. M. Nashed. Perturbations and approximations for generalized inverses and linear operator equations. In *Generalized inverses and applications (Proc. Sem., Math. Res. Center, Univ. Wisconsin, Madison, Wis., 1973)*, pages 325–396. Publ. Math. Res. Center Univ. Wisconsin, No. 32. Academic Press, New York, 1976.

[47] W. Neuschwenger. *Einige Konvergenzaussagen für den Schrödinger-Operator mit tiefen Potentialen*. Diplom Arbeit (under the supervision of K. Veselić), Universität Dortmund, 1979.

[48] K. Neymeyr. A posteriori error estimation for elliptic eigenproblems. *Numer. Linear Algebra Appl.*, 9(4):263–279, 2002.

[49] B. N. Parlett. *The symmetric eigenvalue problem*. Prentice-Hall Inc., Englewood Cliffs, N.J., 1980. Prentice-Hall Series in Computational Mathematics.

[50] M. Reed and B. Simon. *Methods of modern mathematical physics. I–IV*. Academic Press [Harcourt Brace Jovanovich Publishers], New York, 1978.

[51] E. Sánchez-Palencia. Asymptotic and spectral properties of a class of singular-stiff problems. *J. Math. Pures Appl. (9)*, 71(5):379–406, 1992.

[52] B. Simon. A canonical decomposition for quadratic forms with applications to monotone convergence theorems. *J. Funct. Anal.*, 28(3):377–385, 1978.

[53] G. W. Stewart and J. G. Sun. *Matrix perturbation theory*. Academic Press Inc., Boston, MA, 1990.

[54] P. Stollmann. A convergence theorem for Dirichlet forms with applications to boundary value problems with varying domains. *Math. Z.*, 219(2):275–287, 1995.

[55] E. Süli. Finite element methods for partial differential equations (Handouts, University of Oxford). `http://web.comlab.ox.ac.uk/oucl/work/endre.suli/`.

[56] J. Tambača. Vibrations of mechanical systems, (in Croatian: Vibracije mehaničkih sustava). `http://student.math.hr/~tambaca/vms/vms.html`.

[57] J. Tambača. One-dimensional approximations of the eigenvalue problem of curved rods. *Math. Methods Appl. Sci.*, 24(12):927–948, 2001.

[58] J. Tambača. *The evolution model of the curved rod (in Croatian: Evolucijski model zakrivljenog štapa)*. PhD. Thesis, University of Zagreb, 2000.

[59] J. Tambača. Comparison of the arch model and the curved rod model–the stationary case. *University of Zagreb Preprint*, 2003.

[60] R. Verfürth. *A review of a posteriori error estimation and adaptive mesh refinement techniques*. Wiley Teubner, Chichester [u.a.], 1996. Wiley Teubner series advances in Numerical Mathematics.

[61] K. Veselić and I. Slapničar. Floating-point perturbations of Hermitian matrices. *Linear Algebra Appl.*, 195:81–116, 1993.

[62] P.-Å. Wedin. On angles between subspaces of a finite dimensional inner product space. In *Matrix pencils; Proceedings of the Conference held at Pite Havsbad, March 22–24, 1982*. Springer Verlag, 1983.

[63] J. Weidmann. Continuity of the eigenvalues of selfadjoint operators with respect to the strong operator topology. *Integral Equations Operator Theory*, 3(1):138–142, 1980.

[64] J. Weidmann. Stetige Abhängigkeit der Eigenwerte und Eigenfunktionen elliptischer Differentialoperatoren vom Gebiet. *Math. Scand.*, 54(1):51–69, 1984.

[65] J. Weidmann. *Spectral theory of ordinary differential operators*, volume 1258 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1987.

[66] H. Yserentant. On the multilevel splitting of finite element spaces. *Numer. Math.*, 49(4):379–412, 1986.

[67] H. Yserentant. Hierarchical bases. In *ICIAM 91 (Washington, DC, 1991)*, pages 256–276. SIAM, Philadelphia, PA, 1992.

List of Figures

Index

Lebenslauf

19.04.1975: geboren in Zagreb, Kroatien

1981–1989: Grundschule "Ivo Andrić" in Zagreb

1989–1990: XV Gymnasium in Zagreb

1990–1991: Cherwell School Oxford, UK

1991–1993: XV Gymnasium in Zagreb

1993: Abitur

1993–1998: Studium der Mathematik an der Universität Zagreb

1998: Diplom in Mathematik
Titel der Diplomarbeit: A posteriori Berechnung der Singulärvektoren auf einem PVM Rechner (Računanje singularnih vektora a posteriori na PVM računalu)
Betreuer: Prof. Dr. Vjeran Hari

1998–2001: Nachdiplomstudium der Mathematik an der Universität Zagreb

1998–2001: Junior Assistent an der Universität in Zagreb, Lehrgebiet Numerische Mathematik und Wissenschaftliches Rechnen

2001: Magisterdiplom in Mathematik
Titel der Magisterarbeit: Numerische Berechnung des partiellen Eigenwertproblems für symmetrische Matrizen (Numeričko računanje parcijalnog problema vlastitih vrijednosti za simetrične matrice)
Betreuer: Prof. Dr. Zlatko Drmač

ab 2001: Wissenschaftlicher Mitarbeiter an der Fernuniversität in Hagen, Lehrgebiet Mathematische Physik