2015

Robust Estimates for hp-Adaptive Approximations of Non-Self-Adjoint Eigenvalue Problems

Stefano Giani
Durham University

Luka Grubišić
University of Zagreb

Agnieszka Międlar
Technische Universität Berlin

Jeffrey S. Ovall
Portland State University, jovall@pdx.edu

Noname manuscript No.
(will be inserted by the editor)

Robust estimates for *hp*-adaptive approximations of non-self-adjoint eigenvalue problems

Stefano Giani · Luka Grubišić · Agnieszka Międlar · Jeffrey S. Ovall

Received: date / Accepted: date

Abstract We present new residual estimates based on Kato's square root theorem for spectral approximations of non-self-adjoint differential operators of convection–diffusion–reaction type. These estimates are incorporated as part of an *hp*-adaptive finite element algorithm for practical spectral computations, where it is shown that the resulting *a posteriori* error estimates are reliable. Provided experiments demonstrate the efficiency and reliability of our approach.

Keywords eigenvalue problems · non-self-adjoint operators · convection–diffusion–reaction operators · a posteriori error estimates · *hp*-adaptive finite elements

Mathematics Subject Classification (2000) 65N30 · 65N25 · 65N15

1 Introduction

This paper concerns the direct residual analysis of approximation errors involved in the variational approximation of eigenvalues and eigenvectors of linear convection–diffusion–reaction operators in bounded polygonal domains $\Omega \subset \mathbb{R}^2$, as given by the formal differential expression

$$\mathcal{A}\psi := -\nabla \cdot A\nabla\psi + b \cdot \nabla\psi + c\psi = \lambda\psi \ , \tag{1}$$

where standard assumptions (see Section 2) on the coefficients ensure that the inverse $\mathcal{A}^{-1}$ is a compact operator. Abstract estimates based on Kato's square root theorem [4, 25] provide the basis for constructing a practical *hp*-adaptive finite element method for eigencomputations.

The utilization of the Kato's square root theorem in the context of finite element approximation is one of the main contributions of this paper, and we provide first-principle proofs wherever possible to emphasize its role in our error estimation technique. This allows us to use functional calculus for sectorial operators to derive bounds which are cluster robust, in the sense of [38], and are formulated without requiring Galerkin orthogonality constraints. This feature is particularly convenient if one wants to incorporate the effect of inexact linear algebra computations into the overall error control and balancing framework. In this context, we consider a *cluster* to be any finite collection of eigenvalues which are

Stefano Giani
Durham University, School of Engineering and Computing Sciences, South Road, Durham DH1 3LE, United Kingdom.
Tel.: +44 (0) 191 33 42397, E-mail: stefano.giani@durham.ac.uk

Luka Grubišić
University of Zagreb, Department of Mathematics, Bijenička 30, 10000 Zagreb, Croatia.
Tel.: +385 1 4605 881, E-mail: luka.grubisic@math.hr

Agnieszka Międlar
Technische Universität Berlin, Institut für Mathematik, Strasse des 17. Juni 136, Berlin, Germany.
Tel.: +49 (0)30 314 - 23439, Fax: +49 (0)30 314 - 79706 E-mail: miedlar@math.tu-berlin.de

Jeffrey S. Ovall
Portland State University, Fariborz Maseeh Department of Mathematics and Statistics, 315 Neuberger Hall, Portland, OR 97201, USA.
Tel.: +1 503 725-3610, E-mail: jovall@ms.uky.edu

enclosed by a Jordan curve that does not intersect the spectrum of $\mathcal{A}$, and *cluster robustness* means that our estimates depend on the distance between the Jordan curve and the unwanted component of the spectrum (multiplied by a "local measure of nonormality" of the operator), not on distances between eigenvalues within the cluster. In comparison, standard approaches to a posteriori error analysis from [11,21] use a direct variational analysis of the Galerkin approximation of an eigenvalue/vector problem. The results in these contributions are (mostly) presented for single eigenvalues, and the approximation constants which appear are not cluster robust when extended to multiple eigenvalues or clusters of eigenvalues.

We specialize some of our general estimates in the case of diagonalizable operators. In this context we call an operator diagonalizable (also called a scalar operator in the terminology of Dunford–Schwartz [13]) if it is similar to a normal operator, so we can use spectral calculus (in the place of functional calculus) to further improve the results. Also, the use of spectral analysis allows us to formulate our results in terms of geometric restrictions on the size of the residual and the separation between wanted and unwanted components of the spectrum rather than having to resort to the use of saturation assumptions as in e.g. [21].

Let us denote the adjoint operator to $\mathcal{A}$ as $\mathcal{A}^*$. The compactness of the resolvent of $\mathcal{A}$ implies that its spectrum $\mathrm{Spec}(\mathcal{A})$ is a countable set without any finite accumulation points. Furthermore, for each eigenvalue $\lambda \in \mathrm{Spec}(\mathcal{A})$ the space $\mathrm{Ker}(\lambda - \mathcal{A})$ is finite dimensional and its dimension is called the *geometric multiplicity* of λ. Further, given a sectorial operator $\mathcal{A}$ and Jordan curves $\mathfrak{C} \subset \mathbb{C} \setminus \mathrm{Spec}(\mathcal{A})$ and $\mathfrak{C}_d \subset \mathbb{C} \setminus \mathrm{Spec}(\mathcal{A}^*)$, we define the following bounded operators

$$\mathbf{S}(\mathfrak{C}) = \frac{1}{2\pi \mathrm{i}} \int_{\mathfrak{C}} (z - \mathcal{A})^{-1} dz \quad , \quad \mathbf{S}_d(\mathfrak{C}_d) = \frac{1}{2\pi \mathrm{i}} \int_{\mathfrak{C}_d} (z - \mathcal{A}^*)^{-1} dz \ , \tag{2}$$

which we refer to as *spectral projectors*. If $\lambda \in \mathrm{Spec}(\mathcal{A})$ is the only element of $\mathrm{Spec}(\mathcal{A})$ inside a curve $\mathfrak{C}$ we say that $\mathrm{Ran}(\mathbf{S}(\mathfrak{C}))$ is the *algebraic eigenspace* (or *principal eigenspace*) of λ, and its dimension is the *algebraic multiplicity* of λ. If $\dim \mathrm{Ker}(\lambda - \mathcal{A}) = \dim \mathrm{Ran}(\mathbf{S}(\mathfrak{C}))$ holds, then we call the eigenvalue λ *semisimple*. We note that the algebraic multiplicities of $\lambda \in \mathrm{Spec}(\mathcal{A})$ and $\overline{\lambda} \in \mathrm{Spec}(\mathcal{A}^*)$ are identical. If the curve $\mathfrak{C}$ contains several semisimple eigenvalues (and no others) then $\dim \mathrm{Ran}(\mathbf{S}(\mathfrak{C})) = r$, and we call r the *joint algebraic multiplicity* of the eigenvalues inside $\mathfrak{C}$. If in addition $\mathfrak{C}_d$ encloses only the conjugates of the eigenvalues which were enclosed by $\mathfrak{C}$, then $\dim \mathrm{Ran}(\mathbf{S}_d(\mathfrak{C}_d)) = r$ as well.

Finally, according to [12, Theorem 9.2.19], the norm $\|\mathbf{S}(\mathfrak{C})\|$ is an appropriate measure of the local sensitivity of a semisimple eigenvalue enclosed by $\mathfrak{C}$. It can be characterized as

$$\|\mathbf{S}(\mathfrak{C})\| = \lim_{z \to \lambda} |z - \lambda| \ \|(z - \mathcal{A})^{-1}\| \ . \tag{3}$$

More details on the spectral theory of non-self-adjoint operators can be found in [12,26] and the classical reference [18].

The problem (1)—under standard restrictions on the coefficients—provides an important example of a more general class of non-self-adjoint eigenvalue problems in a Hilbert space for which a Riesz basis can be constructed from associated eigenvectors, see Example 10 and [12,18,34] for further discussion and references. A priori estimation theory for finite element approximations of such problems appeared for the first time in [29,30].

Definition 1 A sequence of vectors (functions) $(f_i)_{i \in \mathbb{N}}$ is called a *Riesz basis* of a Hilbert space H if there exists an orthonormal basis $(e_i)_{i \in \mathbb{N}}$ of H and a bounded operator $\mathcal{X}$ with a bounded inverse $\mathcal{X}^{-1}$ such that

$$f_i = \mathcal{X} e_i, \quad \text{with} \quad i \in \mathbb{N}.$$

Criteria for the existence of the Riesz basis of eigenvectors were given in [18,34]. To keep the paper more self-contained, in Example 10 we provide a first-principle argument for the existence of a Riesz basis for some operators of type (1). Such problems, although particular, will be used to benchmark our more general estimates.

Remark 2 Since an orthonormal basis is a Riesz basis with $\mathcal{X} = I$, it is reasonable to use the quantity $\kappa(\mathcal{X}) := \|\mathcal{X}\| \|\mathcal{X}^{-1}\|$ as a measure of the "non-orthogonality" of a basis $(f_i)_{i \in \mathbb{N}}$. It is important to notice that the size of $\kappa(\mathcal{X})$ has a strong impact on the performance of the numerical linear algebra routines used to solve the discretized (algebraic) eigenvalue problems. In the context of numerical linear algebra $\kappa(\mathcal{X})$ is known as the condition number of the eigenvectors.

The paper is organized as follows. In Section 2 we introduce notation and some basic properties of the model problem (1). Section 3 contains the abstract reliability results based on Kato's square root theorem. Here, we show that our eigenvector results are *cluster robust*, i.e., when approximating a subset of the spectrum from a subspace whose dimension is the same as the dimension of the associated spectral subspace, then the estimates depend only on the distance between the computed Ritz values and the complement of the rest of the spectrum. In Section 4 we describe an *hp*-adaptive algorithm for non-self-adjoint problems and use it to illustrate some of our key results for diagonalizable operators with real spectra.

2 Notation and Basic Results

Let $\Omega \subset \mathbb{R}^d$ be an open, bounded domain, and

$$H_0^1(\Omega) = \{v \in H^1(\Omega) : \ v = 0 \text{ on } \partial\Omega \text{ in the sense of trace}\} \ .$$

We denote the standard $L^2(\Omega)$-norm and $H^1(\Omega)$-norm and seminorm on Ω, respectively, by $\|\cdot\| = \|\cdot\|_0$, $\|\cdot\|_1$ and $|\cdot|_1$. We also use

$$(\varphi, \phi) = \int_\Omega \varphi\overline{\phi}\, dx \tag{4}$$

to denote the standard $L^2(\Omega)$ inner-product.

Definition 3 Given real-valued $A \in \left[L^\infty(\Omega)\right]^{2\times 2}$, $b \in \left[L^\infty(\Omega)\right]^2$ with $\nabla \cdot b \in L^\infty(\Omega)$, and $c \in L^\infty(\Omega)$, we define the bilinear form $B : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{C}$ by

$$B(w, v) = \int_\Omega A\nabla w \cdot \nabla\overline{v} + (b \cdot \nabla w + cw)\overline{v}\, dx \ . \tag{5}$$

We make the following common assumptions on the coefficients:

(A1) A is symmetric and uniformly positive definite a.e. in Ω, i.e., there is a $\sigma_0 > 0$ for which $(A(x)z)\cdot z \geq \sigma_0 z \cdot z$ for all $z \in \mathbb{R}^2$ and a.e. $x \in \Omega$.

(A2) $\sigma_0 + \min(0, pc_\Omega) > 0$, where $p = \operatorname{essinf}\{c(x) - \nabla \cdot b(x)/2 : \ x \in \Omega\}$, and c_Ω is the optimal Poincaré constant for the domain, $\|v\|_0 \leq c_\Omega |v|_1$ for all $v \in H_0^1(\Omega)$.

One could relax these assumptions further, provided the associated operator (6) remains positive and sectorial, but we do not pursue that discussion here. When we consider *hp* finite element discretizations in Section 4, we restrict our attention to polygonal domains in $\mathbb{R}^2$, and make one further practical assumption on A,

(A3) There is a partition $\overline{\Omega} = \cup_{k=1}^p \overline{\Omega}_k$ of Ω into polygons Ω_k with disjoint interiors such that $A|_{\Omega_k} \in W^{1,\infty}(\Omega_k)$ for each k.

Since all of the coefficients are bounded, the bilinear form $B(\cdot,\cdot)$ is clearly bounded (continuous): there is a $\gamma_1 > 0$ such that $|B(w, v)| \leq \gamma_1 \|v\|_1 \|w\|_1$ for all $w, v \in H_0^1(\Omega)$. Assumptions (A1)–(A2) guarantee that $B(\cdot,\cdot)$ is also coercive (cf. [15, Theorem 3.8]): there is a $\gamma_0 > 0$ such that $|B(v, v)| \geq \operatorname{Re}(B(v, v)) \geq \gamma_0 \|v\|_1^2$ for all $v \in H_0^1(\Omega)$. For the convenience of the reader we provide the Hermitian and anti-Hermitian parts of $B(\cdot,\cdot)$,

$$B_H(u, v) = \frac{1}{2}(B(u, v) + \overline{B(v, u)}) = \int_\Omega A\nabla u \cdot \nabla\overline{v} + \frac{b}{2} \cdot (\overline{v}\nabla u + u\nabla\overline{v}) + cu\overline{v}\, dx \ ,$$
$$B_A(u, v) = \frac{1}{2}(B(u, v) - \overline{B(v, u)}) = \int_\Omega \frac{b}{2} \cdot (\overline{v}\nabla u - u\nabla\overline{v})\, dx,$$

and note that $B_H(\cdot,\cdot)$ is an inner-product, with energy norm $|\!|\!| \cdot |\!|\!|$ given by

$$|\!|\!| v |\!|\!|^2 = B_H(v, v) = \operatorname{Re}\Big(B(v, v)\Big) = \int_\Omega A\nabla v \cdot \nabla\overline{v} + (c - \nabla \cdot b/2)|v|^2\, dx \ .$$

Also, recall that operators whose real parts are scalar products are called *accretive*. An operator is *maximal accretive* if it has no proper accretive extension.

By the first representation theorem from Kato [26], the operator $\mathcal{A}$ is related to the bilinear form $B(\cdot,\cdot)$ through

$$B(\varphi,\phi) = (\mathcal{A}\varphi,\phi), \qquad \varphi \in \mathrm{Dom}(\mathcal{A}), \phi \in H_0^1(\Omega). \tag{6}$$

We consider the following primal and dual eigenvalue problems:
Find (λ,ψ) *and* (λ^d,ψ^d) *in* $\mathbb{C}\times H_0^1(\Omega)$ *such that* $\|\psi\| = \|\psi^d\| = 1$ *and*

$$B(\psi,\phi) = \lambda(\psi,\phi) \quad \text{ and } \quad B(\varphi,\psi^d) = \lambda^d(\varphi,\psi^d) \quad \text{ for all } \phi,\varphi \in H_0^1(\Omega)\ . \tag{7}$$

Obviously, on the operator level, $\mathcal{A}\psi = \lambda\psi$ and $\mathcal{A}^*\psi^d = \overline{\lambda^d}\psi^d$. Note that λ is an eigenvalue of $\mathcal{A}$ if and only if $\overline{\lambda}$ is an eigenvalue of $\mathcal{A}^*$. Therefore given an eigenvalue λ there exist vectors ψ and ψ^d, $\|\psi\| = \|\psi^d\| = 1$, such that $\mathcal{A}\psi = \lambda\psi$ and $\mathcal{A}^*\psi^d = \overline{\lambda}\psi^d$. By analogy with the linear algebraic version of this problem, we refer to ψ and ψ^d, respectively, as right and left eigenfunctions for λ. Following [20], and in analogy to the linear algebra counterpart, when analyzing the variational eigenvalue problem we will primarily be considering the approximation quality of (λ,ψ,ψ^d), which will be referred to as an eigentriple.

We will now summarize some basic facts about the spectral theory of operators which are defined by (5). A recent reference is [12], see in particular [12, Example 13.4.4]. For classical references we point the reader to the monographs [18,34] and the references therein. Since Ω is bounded, $H_0^1(\Omega)$ is compactly embedded in $L^2(\Omega)$. Since the domain of the bilinear form $B(\cdot,\cdot)$ is precisely $H_0^1(\Omega)$, we conclude that the solution operator which maps the function $f \in L^2(\Omega)$ to $u(f) \in H_0^1(\Omega) \subset L^2(\Omega)$ is also compact (as a mapping from $L^2(\Omega)$ to $L^2(\Omega)$). The solution operator is defined by

$$B(u(f),\phi) = (f,\phi), \quad \text{ for all } \quad \phi \in H_0^1(\Omega). \tag{8}$$

Therefore the eigenvalue problem (7) is attained by a sequence of eigenpairs $(\lambda_n,\psi_n) \in \mathbb{C}\times(H_0^1(\Omega)\setminus\{0\})$, $n \in \mathbb{N}$ such that $|\lambda_n| \to \infty$ as $n \to \infty$. The eigenvalue of the smallest modulus is real and simple (of multiplicity one), and the corresponding eigenvector may be chosen to be positive almost everywhere. Furthermore, the associated Jordan chains of vectors (i.e. generalized eigenvectors) are of finite length, see [18].

Since the original problem has a compact solution operator, the adjoint problem also has a compact solution operator which maps the function $f \in L^2(\Omega)$ to $u^d(f)$ and is defined by

$$B(\varphi,u^d(f)) = (f,\varphi), \quad \text{ for all } \quad \varphi \in H_0^1(\Omega). \tag{9}$$

The eigenvalues of the dual (adjoint) problem are the complex conjugates of the eigenvalues of the original problem. Furthermore, in (2), assuming λ is the only eigenvalue of $\mathcal{A}$ enclosed by $\mathfrak{C}$ and $\overline{\lambda}$ is the only eigenvalue of $\mathcal{A}^*$ enclosed by $\mathfrak{C}_d$, we have that $\mathrm{Ran}(\mathbf{S}(\mathfrak{C}))$ is the subspace containing all right eigenvectors and $\mathrm{Ran}\,\mathbf{S}_d(\mathfrak{C}_d)$ is the subspace of all left eigenvectors of λ. For further discussion of general basic properties of eigenvalue problems see the classical reference of Babuška and Osborn [6].

For any maximal accretive operator $\mathcal{A}$ which is defined by a regularly accretive sesquilinear form associated with a differential expression (1) in the space $H_0^1(\Omega)$, there exists a unique maximal accretive operator $\mathcal{A}^{1/2}$ which solves the operator equation $\mathcal{Z}^2 = \mathcal{A}$. Such an operator is called *the square root of* $\mathcal{A}$, and for an operator which satisfies the conditions (A1)-(A3) it is given by the Balakrishan formula [9]

$$\mathcal{A}^{1/2} = \frac{1}{\pi}\int_0^\infty t^{-1/2}\mathcal{A}(t+\mathcal{A})^{-1}\ dt\ .$$

For further references on the existence of fractional powers of maximal accretive operators as well as for precise definitions of these terms see [9,25,26]. It was a long-standing open problem, known as Kato's conjecture or Kato's square root problem (see [25,33] for the origin of the problem), to determine the domain of definition of the operator $\mathcal{A}^{1/2}$. The conjecture was that the domain of the operator $\mathcal{A}^{1/2}$ should be the same as the domain of the abstract bilinear form $B(\cdot,\cdot)$ which defines the operator $\mathcal{A}$. This hypothesis turned out to be false for the most general abstract form of the bilinear operator (see [33,35]). However, in the case when $\mathcal{A}$ is a convection–diffusion–reaction operator of the form (1), Kato's conjecture does hold in all dimensions (see [3, Theorem 1.11] and [2,4]). In particular, in [4] it is proven that there exist constants c_K, c_K^*, C_K and C_K^* such that

$$c_K\|\phi\|_1 \le \|\mathcal{A}^{1/2}\phi\| \le C_K\|\phi\|_1, \qquad \phi \in H_0^1(\Omega), \tag{10}$$

$$c_K^*\|\phi\|_1 \le \|\mathcal{A}^{*1/2}\phi\| \le C_K^*\|\phi\|_1, \qquad \phi \in H_0^1(\Omega), \tag{11}$$

$$c_H\|\phi\|_1 \le |\!|\!|\phi|\!|\!| \le C_H\|\phi\|_1, \qquad \phi \in H_0^1(\Omega). \tag{12}$$

3 Eigenvalue and Eigenvector Approximation Estimates

This section contains the main theoretical contributions of this paper, concerning the reliability of residual-based estimates of eigenvalue and eigenvector approximations.

3.1 Operator dependent norms of a residual

As a first step we introduce the right and the left residual as a functional on $H_0^1(\Omega)$. For $\varphi, \phi \in H_0^1(\Omega)$ and $\mu \in \mathbb{C}$ we define the residual form

$$\mathfrak{r}(\mu)[\varphi, \phi] = B(\varphi, \phi) - \mu(\varphi, \phi) \ . \tag{13}$$

Given the vector $\varphi \in H_0^1(\Omega)$ and the scalar $\mu \in \mathbb{C}$ we define the functional $\mathfrak{r}(\mu)[\varphi, \cdot]$ which we call the right residual of $\varphi \in H_0^1(\Omega)$ and $\mu \in \mathbb{C}$. The number

$$\|\mathfrak{r}(\mu)[\varphi, \cdot]\|_{-1} = \sup_{\phi \in H_0^1(\Omega)\setminus\{0\}} \frac{|\mathfrak{r}(\mu)[\varphi, \phi]|}{\|\phi\|_1} \tag{14}$$

is the $H^{-1}(\Omega)$-norm of the right residual. Analogously, the functional $\mathfrak{r}(\mu)[\cdot, \phi]$ is called the left residual of $\phi \in H_0^1(\Omega)$ and $\mu \in \mathbb{C}$, and the number

$$\|\mathfrak{r}(\mu)[\cdot, \phi]\|_{-1} = \sup_{\varphi \in H_0^1(\Omega)\setminus\{0\}} \frac{|\mathfrak{r}(\mu)[\varphi, \phi]|}{\|\varphi\|_1} \tag{15}$$

denotes the $H^{-1}(\Omega)$-norm of the left residual. Let us introduce the notation $\|\cdot\|_{\mathcal{A},1/2} = \|\mathcal{A}^{1/2}\cdot\|$ and $\|\cdot\|_{\mathcal{A}^*,1/2} = \|\mathcal{A}^{*1/2}\cdot\|$. With this notation, (10)–(12) yields the obvious norm equivalences

$$c_K^* \sup_{\phi \in H_0^1(\Omega)\setminus\{0\}} \frac{|\mathfrak{r}(\lambda)[\varphi, \phi]|}{\|\phi\|_{\mathcal{A}^*,1/2}} \le \|\mathfrak{r}(\mu)[\varphi, \cdot]\|_{-1} \le C_K^* \sup_{\phi \in H_0^1(\Omega)\setminus\{0\}} \frac{|\mathfrak{r}(\lambda)[\varphi, \phi]|}{\|\phi\|_{\mathcal{A}^*,1/2}} \ , \forall \varphi \in H_0^1(\Omega) \ , \tag{16}$$

$$c_K \sup_{\varphi \in H_0^1(\Omega)\setminus\{0\}} \frac{|\mathfrak{r}(\lambda)[\varphi, \phi]|}{\|\varphi\|_{\mathcal{A},1/2}} \le \|\mathfrak{r}(\mu)[\cdot, \phi]\|_{-1} \le C_K \sup_{\varphi \in H_0^1(\Omega)\setminus\{0\}} \frac{|\mathfrak{r}(\lambda)[\varphi, \phi]|}{\|\varphi\|_{\mathcal{A},1/2}} \ , \forall \phi \in H_0^1(\Omega) \ , \tag{17}$$

$$c_H \sup_{\varphi \in H_0^1(\Omega)\setminus\{0\}} \frac{|\mathfrak{r}(\lambda)[\varphi, \phi]|}{|\!|\!|\phi|\!|\!|} \le \|\mathfrak{r}(\mu)[\cdot, \phi]\|_{-1} \le C_H \sup_{\varphi \in H_0^1(\Omega)\setminus\{0\}} \frac{|\mathfrak{r}(\lambda)[\varphi, \phi]|}{|\!|\!|\phi|\!|\!|} \ , \forall \phi \in H_0^1(\Omega) \tag{18}$$

Although these inequalities show that any pair of residual measures is equivalent, they are quite different from the perspective of numerical analysis. Estimates based on the norms $\|\cdot\|_{\mathcal{A}^*,1/2}$ and $\|\cdot\|_{\mathcal{A},1/2}$ are ideal for algebraic manipulation, whereas the estimates based on the energy norm $|\!|\!|\cdot|\!|\!|$ and, in particular those based on estimating the first order Sobolev norm $\|\cdot\|_1$, are much more convenient from approximation theory point of view. We define the dual norms to $\|\cdot\|_{\mathcal{A}^*,1/2}$ and $\|\cdot\|_{\mathcal{A},1/2}$ in the natural way and use them to measure $\mathfrak{r}(\mu)[\varphi, \cdot]$ and $\mathfrak{r}(\mu)[\cdot, \psi]$,

$$\|\mathfrak{r}(\mu)[\varphi, \cdot]\|_{\mathcal{A}^*,-1/2} := \sup_{\nu \in H_0^1(\Omega)\setminus\{0\}} \frac{|\mathfrak{r}(\mu)[\varphi, \nu]|}{\|\phi\|_{\mathcal{A}^*,1/2}} = \|\mathcal{A}^{1/2}\varphi - \mu\mathcal{A}^{-1/2}\varphi\| \ , \tag{19}$$

$$\|\mathfrak{r}(\mu)[\cdot, \phi]\|_{\mathcal{A},-1/2} := \sup_{\nu \in H_0^1(\Omega)\setminus\{0\}} \frac{|\mathfrak{r}(\mu)[\nu, \phi]|}{\|\varphi\|_{\mathcal{A},1/2}} = \|\mathcal{A}^{*1/2}\phi - \mu\mathcal{A}^{-1/2*}\phi\| \ . \tag{20}$$

3.2 Cluster robust eigenvector estimates

In this section we will obtain estimates of the eigenvector errors. Eigenvector approximation errors are assessed by measuring the angle between the given approximations $\widehat{\psi}$ and $\widehat{\psi}^d$ and the subspaces $\mathcal{S}$ and $\mathcal{S}^d$ which are spanned by all right and left eigenvectors associated with a chosen eigenvalue λ. The angle

between a subspace $\mathcal{S}$ and a vector $\widehat{\psi}$ in any norm $\|\cdot\|$ such as e.g. $L^2(\Omega)$- or $H^1(\Omega)$-norm, is defined as $\theta \in [0,\pi)$ such that

$$\sin\theta = \frac{1}{\|\widehat{\psi}\|}\inf_{\phi\in\mathcal{S}}\|\phi-\widehat{\psi}\| \ . \tag{21}$$

The proofs of our eigenvector error estimates will follow from the Cauchy integral representation of the spectral projections $\mathbf{S}(\mathfrak{C})$ and $\mathbf{S}_d(\mathfrak{C})$, so we point back to their definitions in (2) and the surrounding discussion.

Proposition 4 *Let $\mathcal{A}$ be given by* (1) *and let $(\widehat{\lambda},\widehat{\psi},\widehat{\psi}^d)\in\mathbb{C}\times H_0^1(\Omega)\times H_0^1(\Omega)$. Assume c_K and c_K^* are the constants defined in* (10)–(11) *and let $\mathfrak{C}\subset\mathbb{C}\setminus\mathrm{Spec}(\mathcal{A})$ be a Jordan curve which encloses both $\widehat{\lambda}$ and an isolated subset $\Lambda=\{\lambda_i:\ i=1,\dots,r\}$ of* $\mathrm{Spec}(\mathcal{A})$ *and no other elements of* $\mathrm{Spec}(\mathcal{A})$*, and define the corresponding spectral projection $\mathbf{S}=\mathbf{S}(\mathfrak{C})$. Then*

$$\inf_{\phi\in\mathrm{Ran}(\mathbf{S})}\|\phi-\widehat{\psi}\|_k \le \frac{\|\mathcal{A}^{(k+1)/2}\,\mathbf{T}_{\widehat{\lambda}}\,\|}{(c_K)^k}\|\mathfrak{r}(\widehat{\lambda})[\widehat{\psi},\cdot]\|_{\mathcal{A}^*,-1/2} \ ,$$

where $\mathbf{T}_{\widehat{\lambda}}=(\widehat{\lambda}-\mathcal{A}')^{-1}(I-\mathbf{S})$, with $\mathcal{A}'$ denoting the restriction of $\mathcal{A}$ on the space $\mathrm{Ran}(I-\mathbf{S})$*. Similarly, if $\mathfrak{C}_d$ encloses both $\overline{\widehat{\lambda}}$ and $\overline{\Lambda}$, but no other elements of* $\mathrm{Spec}(\mathcal{A}^*)$*, and we define $\mathbf{S}_d=\mathbf{S}_d(\mathfrak{C}_d)$, then*

$$\inf_{\phi\in\mathrm{Ran}(\mathbf{S}_d)}\|\phi-\widehat{\psi}_d\|_k \le \frac{\|\mathcal{A}^{*\,(k+1)/2}\,\mathbf{T}^{(\mathrm{d})}_{\widehat{\lambda}}\,\|}{(c_K^*)^k}\|\mathfrak{r}(\widehat{\lambda})[\cdot,\widehat{\psi}^d]\|_{\mathcal{A},-1/2} \ ,$$

where $\mathbf{T}^{(\mathrm{d})}_{\widehat{\lambda}}=(\overline{\widehat{\lambda}}-\mathcal{A}^{*\prime})^{-1}(I-\mathbf{S}_d)$ and $\mathcal{A}^{*\prime}$ denotes the restriction of $\mathcal{A}^*$ on the space $\mathrm{Ran}(I-\mathbf{S}_d)$*. Here, we let $k=0,1$.*

We emphasize that the results presented above do not require that $\widehat{\lambda}$, $\widehat{\psi}$ or $\widehat{\psi}^d$ satisfy any Galerkin orthogonality conditions.

Proof Noting that, for $z\in\mathbb{C}\setminus\mathrm{Spec}(\mathcal{A})$, the bounded operators $(z-\mathcal{A})^{-1}$ and $\mathcal{A}^{-1/2}$ commute, and the operators $\mathbf{S}$ and $I-\mathbf{S}$ commute with powers of $\mathcal{A}$, we obtain the following identities by direct computation:

$$\begin{aligned}\frac{1}{z-\widehat{\lambda}}\widehat{\psi}-(z-\mathcal{A})^{-1}\widehat{\psi} &= \frac{1}{z-\widehat{\lambda}}\mathcal{A}^{1/2}(z-\mathcal{A})^{-1}\Big[(z-\mathcal{A})\mathcal{A}^{-1/2}\widehat{\psi}-(z-\widehat{\lambda})\mathcal{A}^{-1/2}\widehat{\psi}\Big]\\ &= \frac{1}{z-\widehat{\lambda}}\mathcal{A}^{1/2}(z-\mathcal{A})^{-1}\Big[-\mathcal{A}^{1/2}\widehat{\psi}+\widehat{\lambda}\mathcal{A}^{-1/2}\widehat{\psi}\Big] \ .\end{aligned} \tag{22}$$

Since, $(I-\mathbf{S})(I-\mathbf{S})=(I-\mathbf{S})$ it follows that

$$\begin{aligned}\|\widehat{\psi}-\mathbf{S}\widehat{\psi}\| &= \|(I-\mathbf{S})(\widehat{\psi}-\mathbf{S}\widehat{\psi})\|\\ &= \frac{1}{2\pi}\left\|\int_{\mathfrak{C}}(I-\mathbf{S})\Big[\frac{1}{z-\widehat{\lambda}}\widehat{\psi}-(z-\mathcal{A})^{-1}\widehat{\psi}\Big]\,dz\right\|\\ &= \frac{1}{2\pi}\left\|\mathcal{A}^{1/2}\int_{\mathfrak{C}}\frac{1}{z-\widehat{\lambda}}(z-\mathcal{A})^{-1}(I-\mathbf{S})\Big[-\mathcal{A}^{1/2}\widehat{\psi}+\widehat{\lambda}\mathcal{A}^{-1/2}\widehat{\psi}\Big]\,dz\right\|\\ &\le \|\mathcal{A}^{1/2}(\widehat{\lambda}-\mathcal{A}')^{-1}(I-\mathbf{S})\|\ \|-\mathcal{A}^{1/2}\widehat{\psi}+\widehat{\lambda}\mathcal{A}^{-1/2}\widehat{\psi}\| \ .\end{aligned}$$

Note that $(z-\mathcal{A}')^{-1}$ has no singularities inside $\mathfrak{C}$. Applying $\inf_{\phi\in\mathrm{Ran}(\mathbf{S})}\|\phi-\widehat{\psi}\|\le\|\mathbf{S}\widehat{\psi}-\widehat{\psi}\|$ and $c_K\|\widehat{\psi}-\mathbf{S}\widehat{\psi}\|_1\le\|\mathcal{A}^{1/2}(\widehat{\psi}-\mathbf{S}\widehat{\psi})\|$ completes the proof of the the first bound. The proof for the second bound follows analogously. □

Remark 5 By looking at the operator $\mathbf{T}_{\widehat{\lambda}}$ we would like $\widehat{\lambda}$ to be well-separated from the unwanted portion of $\mathrm{Spec}(\mathcal{A})$, which would allow to decrease the size of norms $\|\mathcal{A}\,\mathbf{T}_{\widehat{\lambda}}\,\|$ and $\|\mathcal{A}^*\,\mathbf{T}^{(\mathrm{d})}_{\widehat{\lambda}}\,\|$. We also point out that, in general, the L^2 bounds for the finite element approximations considered in the next sections are not sharp. However, we do not need this property in our further investigations.

Below we establish an estimate on the norm difference of the projections onto algebraic eigenspaces. We start by recalling elementary properties of orthogonal projections in a Hilbert space, which we then specialize to our setting. In particular, let P and Q be orthogonal projections onto two finite dimensional subspaces $\mathcal{P}$ and $\mathcal{Q}$, with $\mathcal{P}$ having the basis $\mathfrak{W} = \{p_1, \ldots, p_r\}$. We define the *subspace separation* between $\mathcal{P}$ and $\mathcal{Q}$ as the maximum of $\|(I-Q)P\|_k$ and $\|(I-P)Q\|_k$, for $k = 0, 1$. Recall that we sometimes abbreviate, $\|\cdot\| = \|\cdot\|_0$. If $\|P - Q\|_k < 1$, it holds that

$$\|P - Q\|_k = \|(I-Q)P\|_k = \sup_{p\in\mathcal{P}} \inf_{q\in\mathcal{Q}} \frac{\|q-p\|_k}{\|p\|_k} \ , \qquad k = 0, 1 \ ,$$

and we can reverse the roles of projectors P and Q and the spaces $\mathcal{P}$ and $\mathcal{Q}$ above. Note that the assumption $\|P - Q\|_k < 1$ is equivalent with saying that $\dim \mathcal{Q} = r = \dim \mathcal{P}$ and no non-zero vectors in $\mathcal{P}$ are orthogonal to $\mathcal{Q}$ and vice versa. Here, the orthogonality is meant with respect to the scalar product $(\cdot,\cdot)_k$, $k = 0, 1$.

Let $G^{(k)} \in \mathbb{C}^{r\times r}$ be the Gramian matrix for $\mathfrak{W}$, $G^{(k)}_{ij} = (p_j, p_i)_k$. We argue below that

$$\|(I-Q)P\|_k \leq \kappa_k(\mathfrak{W}) \sqrt{\sum_{i=1}^{r} \inf_{v_2\in\mathcal{Q}} \frac{\|q-p_i\|_k^2}{\|p_i\|_k^2}} \ , \tag{23}$$

where $\kappa_k(\mathfrak{W}) = (\|G^{(k)}\|\|(G^{(k)})^{-1}\|)^{-1/2}$, is the square root of the spectral condition number of $G^{(k)}$. In the special case that $\mathfrak{W}$ is an orthonormal basis, (23) clearly simplifies to

$$\|(I-Q)P\|_k \leq \sqrt{\sum_{i=1}^{r} \inf_{v_2\in\mathcal{Q}} \frac{\|q-p_i\|_k^2}{\|p_i\|_k^2}} \ . \tag{24}$$

Proof (of equation (23)*)* We show the argument for $k = 0$ and note that the argument is the same for $k = 1$ with the obvious changes in the Hilbert space structure. Let $F : \mathbb{R}^r \to \mathcal{P}$ be the linear operator defined by $Fe_i = p_i$, $i = 1, \ldots, r$, where $e_i \in \mathbb{R}^r$ are the canonical basis vectors. It is straight-forward to see that $\{\widehat{p}_i = F(F^*F)^{-1/2}e_i \,:\, 1 \leq i \leq r\}$ is an orthonormal basis of $\mathcal{P}$, and we have the following sequence of inequalities,

$$\begin{aligned} \|(I-Q)P\|^2 &\leq \sum_{i=1}^{r} \|(I-Q)\widehat{p}_i\|^2 \leq \|G^{-1}\| \sum_{i=1}^{r} \|(I-Q)p_i\|^2 \\ &= \|G^{-1}\| \sum_{i=1}^{r} \frac{\|(I-Q)p_j\|^2}{\|p_j\|^2} \|p_j\|^2 \leq \|G^{-1}\|\|G\| \sum_{i=1}^{r} \frac{\|(I-Q)p_j\|^2}{\|p_j\|^2} \ . \end{aligned} \tag{25}$$

This is clearly equivalent to (23). The first of these inequalities holds for any orthonormal basis of $\mathcal{P}$, and the second is true due to the particular relationship between the two bases for $\mathcal{P}$. □

We will now formulate cluster robust subspace approximation estimates. Let $\widehat{\mathfrak{C}} \subset \mathbb{C} \setminus \mathrm{Spec}(\mathcal{A})$ be given and let

$$\Lambda := \{\lambda_i \ : \ i = 1, \ldots, r\} \tag{26}$$

be a set of all eigenvalues of $\mathcal{A}$ inside $\widehat{\mathfrak{C}}$. We assume that all λ_i are semisimple eigenvalues. Furthermore, let Q and Q_d be the $L^2(\Omega)$ orthogonal projections onto the sets of corresponding left and right eigenvectors $\mathcal{Q} = \mathrm{Ran}(\mathbf{S}(\widehat{\mathfrak{C}}))$ and $\mathcal{Q}_d = \mathrm{Ran}(\mathbf{S}_d(\widehat{\mathfrak{C}}))$, respectively. Let us now assume that we are given two sets of linearly independent vectors $\mathfrak{M} = \{\widehat{\psi}_i \ : \ i = 1, \ldots, r\}$ and $\mathfrak{M}_d = \{\widehat{\psi}_i^d \ : \ i = 1, \ldots, r\}$ and a sequence of scalars $\widehat{\lambda}_i$, $i = 1, \ldots, r$ which are all inside $\widehat{\mathfrak{C}}$.

Theorem 6 *Let $\mathcal{A}$ be a sectorial operator defined by* (1) *and let P and P_d be $L^2(\Omega)$ orthogonal projections onto subspaces $\mathcal{P} = \operatorname{span}\mathfrak{M}$ and $\mathcal{P}^d = \operatorname{span}\mathfrak{M}_d$. Then*

$$\|(I-Q)P\|_k \leq \frac{\max_{i=1,\ldots,r} \|\mathcal{A}^{(1+k)/2}\mathbf{T}_{\widehat{\lambda}_i}\|\kappa(\mathfrak{M})}{c_K^*} \sqrt{\sum_{i=1}^{r} \frac{\|\mathfrak{r}(\widehat{\lambda}_i)[\widehat{\psi}_i, \cdot]\|^2_{\mathcal{A}^*,-1/2}}{\|\widehat{\psi}_i\|^2}} \ , \tag{27}$$

$$\|(I-Q_d)P_d\|_k \leq \frac{\max_{i=1,\ldots,r} \|\mathcal{A}^{*(1+k)/2}\mathbf{T}^{(d)}_{\widehat{\lambda}_i}\|\kappa(\mathfrak{M}_d)}{c_K} \sqrt{\sum_{i=1}^{r} \frac{\|\mathfrak{r}(\widehat{\lambda}_i)[\cdot, \widehat{\psi}_i^d]\|^2_{\mathcal{A},-1/2}}{\|\widehat{\psi}_d\|^2}} \ . \tag{28}$$

Furthermore, if the right hand side in (27) *or* (28) *is less than one, then*

$$\|P - Q\|_k = \|(I - Q)P\|_k, \quad or \quad \|P_d - Q_d\|_k = \|(I - Q_d)P_d\|_k \ ,$$

respectively, with $k = 0, 1$.

Proof The first result follows directly from (23) and Proposition 4, by summing over all elements of $\mathfrak{M}$, and the second is obtained analogously. The last claim follows from Kato's alternative. Since $\dim \mathcal{P} = \dim \mathcal{Q}$ and by the assumption of the theorem $\|(I - Q)P\| < 1$, then $\|P - Q\| = \|(I - Q)P\|$. The other identity follows analogously. □

3.3 Estimating the separation measure

In general, the measures of separation of a cluster of eigenvalues given by

$$\|\mathcal{A}^{(1+k)/2}\mathbf{T}_{\widehat{\lambda}_i}\| \quad \text{and} \quad \|\mathcal{A}^{*(1+k)/2}\mathbf{T}^{(d)}_{\widehat{\lambda}_i}\|,$$

are not easily accessible to further estimation. However, here, following the argument of Heuveline and Rannacher [22], we will present a general lower estimate of the separation measures. Later on, we will also show how to estimate the separation measures for some special classes of operators used in our numerical experiments. All of these results are stated in the context of (26) and the surrounding discussion.

Proposition 7 *Let* $v \neq 0$ *and* $w \neq 0$ *be eigenvectors associated with the eigenvalue*

$$\lambda_{\rm gap} = \text{arg-min}\{|\chi - \nu| \ : \ \chi \in \text{Spec}(\mathcal{A}) \setminus \Lambda, \ \nu \in \Lambda\} \ ,$$

such that

$$\mathcal{A}v = \lambda_{\rm gap} v, \qquad \mathcal{A}^* w = \overline{\lambda_{\rm gap}}\ w, \qquad (v, w) = 1 \ .$$

Then

$$\|\mathcal{A}^{(1+k)/2}\mathbf{T}_{\widehat{\lambda}_i}\| \geq \frac{|\lambda_{\rm gap}^{(1+k)/2}|}{|\widehat{\lambda}_i - \lambda_{\rm gap}|\|v\|} \quad and \quad \|\mathcal{A}^{*(1+k)/2}\mathbf{T}^{(d)}_{\widehat{\lambda}_i}\| \geq \frac{|\lambda_{\rm gap}^{(1+k)/2}|}{|\widehat{\lambda}_i - \lambda_{\rm gap}|\|w\|}$$

for $k = 0, 1$.

Proof We notice that for $v \neq 0$ the first result is a straight-forward consequence of

$$\|\mathcal{A}^{(1+k)/2}\mathbf{T}_{\widehat{\lambda}_i}\| \geq \frac{1}{\|v\|}\|\mathcal{A}^{(1+k)/2}\mathbf{T}_{\widehat{\lambda}_i} v\| = \frac{1}{\|v\|}\frac{|\lambda_{\rm gap}^{(1+k)/2}|}{|\widehat{\lambda}_i - \lambda_{\rm gap}|} \ .$$

The proof of the second estimate follows analogously. □

3.3.1 Estimates for special classes of operators

The main stability measures in the eigenvector estimates are the condition numbers $\|\mathcal{A}^{1/2}\,\mathbf{T}_{\widehat{\lambda}}\|$ and $\|\mathcal{A}^{*1/2}\,\mathbf{T}^{(\rm d)}_{\widehat{\lambda}}\|$. Since, in general, they are not readily accessible to quantitative estimation, we will now provide their estimates in terms of the separation distance between the wanted and unwanted components of the spectrum of $\mathcal{A}$ and some measure of the non-normality of $\mathcal{A}$. The separation distance is a consequence of using the spectral calculus in establishing the perturbation estimates, whereas a measure of non-normality determines how far the given operator is from an operator which has spectral calculus.

Proposition 8 *Let* $\mathcal{A}$ *be a normal operator and let* $\widehat{\lambda} \in \mathbb{C}$ *be given. Then*

$$\|\mathcal{A}^{1/2}\,\mathbf{T}_{\widehat{\lambda}}\| = \|\mathcal{A}^{*1/2}\,\mathbf{T}^{(\rm d)}_{\widehat{\lambda}}\| = \max_{\xi \in \text{Spec}(\mathcal{A})\setminus\Lambda} \frac{|\xi|}{|\xi - \widehat{\lambda}|}.$$

Moreover, if there exists a bounded operator $\mathcal{X}$*, with a bounded inverse* $\mathcal{X}^{-1}$*, and a normal operator* $\mathcal{H}$ *such that* $\mathcal{A} = \mathcal{X}\mathcal{H}\mathcal{X}^{-1}$*,then*

$$\|\mathcal{A}\,\mathbf{T}_{\widehat{\lambda}}\| \leq \kappa(\mathcal{X}) \max_{\xi \in \text{Spec}(\mathcal{A})\setminus\Lambda} \frac{|\xi|}{|\xi - \widehat{\lambda}|}, \qquad \|\mathcal{A}^{*1/2}\,\mathbf{T}^{(\rm d)}_{\widehat{\lambda}}\| \leq \kappa(\mathcal{X}) \max_{\xi \in \text{Spec}(\mathcal{A})\setminus\Lambda} \frac{|\xi|}{|\xi - \overline{\widehat{\lambda}}|} \ .$$

Proof We will prove the second statement, and the first statement follows by the specialization of the argument. Let $f_i = \mathcal{X}e_i$, $i \in \mathbb{N}$ be eigenvectors of $\mathcal{A}$ which, by the assumption of the theorem, make up the Riesz basis of $\mathcal{H}$. For $\mathcal{A} = \mathcal{X}\mathcal{H}\mathcal{X}^{-1}$, where $\mathcal{H}$ is a normal operator, we compute

$$\mathcal{A}\,\mathbf{T}_{\widehat{\lambda}}\,\mathcal{X}e_i = \mathcal{A}(\widehat{\lambda} - \mathcal{A})^{-1}\mathcal{X}e_i = \frac{\lambda_i}{\widehat{\lambda} - \lambda_i}\mathcal{X}e_i$$

for all i such that $\lambda_i \notin \Lambda$. The vectors $\mathcal{X}e_i$, for all i such that $\lambda_i \notin \Lambda$ make a Riesz basis of the space $\mathrm{Ran}\,\big(I - S(\mathfrak{C})\big)$. Since e_i are the orthonormal eigenvectors of the normal operator $\mathcal{X}^{-1}\mathcal{A}(\widehat{\lambda} - \mathcal{A}')^{-1}\mathcal{X}$ we conclude the statement of the theorem. □

We note that operators which are similar to a normal operator in the sense of $\mathcal{A} = \mathcal{X}\mathcal{H}\mathcal{X}^{-1}$ are common in applications. Let us recall the following definition.

Definition 9 An operator $\mathcal{A}$ in the Hilbert space $L^2(\Omega)$ is said to be *diagonalizable* if there exists a bounded operator $\mathcal{X}$ with a bounded inverse $\mathcal{X}^{-1}$ and a normal (possibly unbounded) operator $\mathcal{H}$ such that

$$\mathcal{A} = \mathcal{X}\mathcal{H}\mathcal{X}^{-1}\ , \tag{29}$$

and $\mathcal{X}^{-1}\,\mathrm{Dom}(\mathcal{A}) \subset \mathrm{Dom}(\mathcal{H})$. For diagonalizable operators, the square root operator is given explicitly by the formula

$$\mathcal{A}^{1/2} = \mathcal{X}\mathcal{H}^{1/2}\mathcal{X}^{-1}. \tag{30}$$

Here $\mathcal{H}^{1/2}$ denotes the positive square root of a normal operator $\mathcal{H}$ defined by the spectral calculus. When $\mathcal{A}$ is diagonalizable, then so is $\mathcal{A}^*$, and we can write $\mathcal{A}^* = \mathcal{X}^{-*}\mathcal{H}\mathcal{X}^*$ and $\mathcal{A}^{*1/2} = \mathcal{X}^{-*}\mathcal{H}^{1/2}\mathcal{X}^*$.

In some cases, all eigenvalues of $\mathcal{A}$ are known to be real, i.e., $\lambda = \overline{\lambda}$, and we can naturally consider eigentriplets $(\lambda, \psi, \psi^d) \in \mathbb{R} \times H_0^1(\Omega) \times H_0^1(\Omega)$ of (7). Diagonalizable operators with real spectra are of the main focus of our numerical experiments, since they allow us to easily compute highly accurate benchmark solutions.

We now provide an example of a family of diagonalizable operators with real spectra which will be studied in the rest of this paper.

Example 10 Let $\mathcal{A}u := -\nabla\cdot(A\nabla u) + b\cdot\nabla u + cu$, where the coefficients A, b and c satisfy the conditions (A1)–(A2) prescribed above. Furthermore, let us define the multiplication operator $\mathcal{X}u := e^{\beta}u$ for some function $\beta \in W^{1,\infty}(\Omega)$. The following identities are obtained by direct computation:

$$\begin{aligned} e^{-\beta}[\nabla\cdot(A\nabla(e^{\beta}u))] &= \nabla\cdot(A\nabla u) + 2A\nabla\beta\cdot\nabla u + (\nabla\cdot(A\nabla\beta) + (A\nabla\beta)\cdot\nabla\beta)u\ , \\ e^{-\beta}[b\cdot\nabla(e^{\beta}u)] &= b\cdot\nabla u + (b\cdot\nabla\beta)u\ . \end{aligned}$$

If $A^{-1}b$ is a conservative vector field, then we choose β such that $\nabla\beta = \frac{1}{2}A^{-1}b$, and determine that $\mathcal{H} := \mathcal{X}^{-1}\mathcal{A}\mathcal{X}$ is self-adjoint and positive definite. In particular,

$$\mathcal{H}u = \mathcal{X}^{-1}\mathcal{A}\mathcal{X}u = -\nabla\cdot(A\nabla u) + \left(c - \frac{1}{2}\nabla\cdot b + \frac{1}{4}b\cdot(A^{-1}b)\right)u\ .$$

From this argument, we see that (λ,ϕ) is an eigenpair of $\mathcal{H}$ if and only if $(\lambda, e^{\beta}\phi, e^{-\beta}\phi)$ is an eigentriple of $\mathcal{A}$. If A and b are constant, then the choice $\beta(x) = \frac{1}{2}A^{-1}b\cdot x$, $x\in\Omega$ is obvious, and we see that the eigenvalues of $\mathcal{A}$ only differ from those of $\mathcal{B}u := -\nabla\cdot(A\nabla u) + cu$ only by an additive constant $\frac{1}{4}b\cdot(A^{-1}b)$.

For such operators we can now prove a first-order residual estimate for eigenvalues. In the field of numerical linear algebra, such a result is known as a *Bauer–Fike* type estimate, see [14]. This estimate is also cluster robust, but it is not of optimal order when estimating approximation errors of eigenvalues.

Proposition 11 *Let $\widehat{\lambda}\in\mathbb{R}$ and $\widehat{\psi}\in H_0^1(\Omega)$, $\|\widehat{\psi}\| = 1$ be given and let $\mathcal{A}$, defined as in (1), be similar to some positive definite self-adjoint operator (e.g. diagonalizable with real and positive spectrum). Then*

$$\min_{\xi\in\mathrm{Spec}(\mathcal{A})}\frac{|\widehat{\lambda}-\xi|}{\sqrt{|\widehat{\lambda}|\xi}} \le \frac{\kappa(\mathcal{X})}{\sqrt{|\widehat{\lambda}|}}\|\mathfrak{r}(\widehat{\lambda})[\widehat{\psi},\cdot]\|_{\mathcal{A}^*,-1/2}\ .$$

Proof Due to (19) and the diagonalizability of $\mathcal{A}$, i.e., $\mathcal{A} = \mathcal{X}\mathcal{H}\mathcal{X}^{-1}$, it holds that

$$\|\mathfrak{r}(\widehat{\lambda})[\widehat{\psi}, \cdot]\|_{\mathcal{A}^*,-1/2} = \|\mathcal{A}^{1/2}\widehat{\psi} - \widehat{\lambda}\mathcal{A}^{-1/2}\widehat{\psi}\| = \|\mathcal{X}(\mathcal{H}^{1/2} - \widehat{\lambda}\mathcal{H}^{-1/2})\mathcal{X}^{-1}\widehat{\psi}\|.$$

Since the operator $\mathcal{H}^{1/2} - \widehat{\lambda}\mathcal{H}^{-1/2}$ is self-adjoint, using the standard spectral calculus for self-adjoint operators, e.g., [39, Theorems VIII.5 and VIII.6], yields

$$\mathrm{Spec}(\mathcal{H}^{1/2} - \widehat{\lambda}\mathcal{H}^{-1/2}) = \left\{ \sqrt{|\widehat{\lambda}|}\ \frac{\xi - \widehat{\lambda}}{\sqrt{|\widehat{\lambda}|\xi}}\ :\ \xi \in \mathrm{Spec}(\mathcal{A}) \right\},$$

and the smallest in modulus eigenvalue of $\mathcal{H}^{1/2} - \widehat{\lambda}\mathcal{H}^{-1/2}$ is given by $\sqrt{|\widehat{\lambda}|} \min\limits_{\xi \in \mathrm{Spec}(\mathcal{A})} \dfrac{|\xi - \widehat{\lambda}|}{\sqrt{\xi|\widehat{\lambda}|}}$. This, together with some general properties of norm $\|\cdot\|$ and an assumption $\|\widehat{\psi}\| = 1$ yields

$$\begin{aligned}\|\mathfrak{r}(\widehat{\lambda})[\widehat{\psi}, \cdot]\|_{\mathcal{A}^*,-1/2} &= \|\mathcal{X}(\mathcal{H}^{1/2} - \widehat{\lambda}\mathcal{H}^{-1/2})\mathcal{X}^{-1}\widehat{\psi}\| \\ &\geq \|\mathcal{X}^{-1}\|^{-1}\sqrt{|\widehat{\lambda}|} \min_{\xi \in \mathrm{Spec}(\mathcal{A})} \frac{|\xi - \widehat{\lambda}|}{\sqrt{\xi|\widehat{\lambda}|}} \|\mathcal{X}^{-1}\widehat{\psi}\|\ ,\end{aligned}$$

which completes the proof. □

Remark 12 Similar estimate also holds for a complex $\widehat{\lambda}$ and a general diagonalizable operator $\mathcal{A}$, however, limited in space, we leave out the details here. Furthermore, such a result is not relevant in our numerical experiments.

3.4 Eigenvalue estimates

The estimates presented so far give a bound on the distance of the approximated eigenvector to the eigenspace spanned by the exact eigenvectors of interest. In the case of diagonalizable operators, we have obtained a cluster robust error estimates for the eigenvalue closest to a given scalar $\widehat{\lambda}$. However, in both cases, we have neither localized the approximated eigenvalue in the cluster, nor obtained which eigenvector was approximated. Moreover, in order to obtain the eigenvalue estimates for general non-diagonalizable operators, it is necessary to consider the distance to the next nearest eigenvalue in the bounds. As a consequence these estimates are not any more cluster robust. In Section 3.4.1 we will discuss the issue of cluster robustness in the non-self-adjoint case. Note that the self-adjoint case has been resolved in [28] using the majorization principle. For an alternative approach using symmetric gauge functions, see [19].

Up to now, we had no restrictions on the choice of $\widehat{\lambda}$. However, in the following theorem we make a special choice for the scalar $\widehat{\lambda}$, which results in a slightly different notation. Given two non-orthogonal vectors $\widehat{\psi}, \widehat{\psi}^d \in H_0^1(\Omega)$ we define the generalized Rayleigh quotient

$$\widetilde{\lambda} = \frac{B(\widehat{\psi}, \widehat{\psi}^d)}{(\widehat{\psi}, \widehat{\psi}^d)}.$$

Let us now prove a general residual estimate, which can be applied after the first, cluster robust, phase of the convergence is resolved to sufficient accuracy.

Theorem 13 *Let $\mathcal{A}$ be as in Proposition 4 and let $(\widehat{\psi}_i, \widehat{\psi}_i^d) \in H_0^1(\Omega) \times H_0^1(\Omega)$, $(\widehat{\psi}_i, \widehat{\psi}_i^d) \neq 0$, $i = 1, \ldots, r$, be given. If a semisimple eigenvalue $\lambda \in \mathbb{C}$ of multiplicity r is the only eigenvalue inside $\mathfrak{C}$ and the right hand sides of both inequalities* (27) *and* (28) *are less than one, then, for the corresponding generalized Rayleigh quotient*

$$\widetilde{\lambda}_i = B(\widehat{\psi}_i, \widehat{\psi}_i^d)/(\widehat{\psi}_i, \widehat{\psi}_i^d),$$

the following estimate holds

$$\sum_{i=1}^{r} \frac{|\lambda - \widetilde{\lambda}_i|}{|\lambda|} \leq C_{\mathrm{cluster}} \sum_{i=1}^{r} \Big[\frac{1}{|(\widehat{\psi}_i, \widehat{\psi}_i^d)|}\Big] \sqrt{\sum_{i=1}^{r} \Big[\frac{\|\mathfrak{r}(\widetilde{\lambda}_i)[\widehat{\psi}_i, \cdot]\|^2_{\mathcal{A}^*,-1/2}}{\|\widehat{\psi}_i\|^2}\Big] \sum_{i=1}^{r} \Big[\frac{\|\mathfrak{r}(\widetilde{\lambda}_i)[\cdot, \widehat{\psi}_i^d]\|^2_{\mathcal{A},-1/2}}{\|\widehat{\psi}_i^d\|^2}\Big]}\ .$$

Proof For an eigentriple $(\lambda, \psi, \psi^d) \in \mathbb{C} \times H_0^1(\Omega) \times H_0^1(\Omega)$ of (7), using [31, Lemma 3.6] and the assumptions of the theorem, we obtain

$$\widetilde{\lambda} - \lambda = \frac{B(\widehat{\psi}, \widehat{\psi}^d)}{(\widehat{\psi}, \widehat{\psi}^d)} - \lambda = \frac{B(\widehat{\psi} - \psi, \widehat{\psi}^d - \psi^d)}{(\widehat{\psi}, \widehat{\psi}^d)} - \lambda \frac{(\widehat{\psi} - \psi, \widehat{\psi}^d - \psi^d)}{(\widehat{\psi}, \widehat{\psi}^d)} \ .$$

Now, the triangle inequality and the continuity of $B(\cdot, \cdot)$ yield

$$\frac{|\widetilde{\lambda} - \lambda|}{|\lambda|} \leq \frac{2 \max\{\gamma_1/|\lambda|, c_\Omega\}}{|(\widehat{\psi}, \widehat{\psi}^d)|} \|\widehat{\psi} - \psi\|_1 \|\widehat{\psi}^d - \psi^d\|_1 \ . \tag{31}$$

Note that here we have no restriction on the choice of the approximated eigentriple. From now on we will use the assumption on the size of the residual which is not cluster robust.

Let Q and Q_d be the $L^2(\Omega)$ orthogonal projections such that $\mathrm{Ran}(Q)$ and $\mathrm{Ran}(Q_d)$ are spanned by all right and left eigenvectors, respectively. Then, (27) and (28) now read

$$\kappa(\mathfrak{M}) \sqrt{\sum_{i=1}^{r} \inf_{q \in \mathrm{Ran}(Q)} \frac{\|q - \widehat{\psi}_i\|_1^2}{\|\widehat{\psi}_i\|_1^2}} \leq \frac{\max_{i=1,\ldots,r} \|\mathcal{A}\mathbf{T}_{\widetilde{\lambda}_i}\| \kappa(\mathfrak{M})}{c_K^*} \sqrt{\sum_{i=1}^{r} \frac{\|\mathfrak{r}(\widetilde{\lambda}_i)[\widehat{\psi}_i, \cdot]\|_{\mathcal{A}^*,-1/2}^2}{\|\widehat{\psi}_i\|^2}} < 1, \tag{32}$$

$$\kappa(\mathfrak{M}_d) \sqrt{\sum_{i=1}^{r} \inf_{q_d \in \mathrm{Ran}(Q_d)} \frac{\|q_d - \widehat{\psi}_i^d\|_1^2}{\|\widehat{\psi}_i^d\|_1^2}} \leq \frac{\max_{i=1,\ldots,r} \|\mathcal{A}^* \mathbf{T}_{\widetilde{\lambda}_i}^{(d)}\| \kappa(\mathfrak{M}_d)}{c_K} \sqrt{\sum_{i=1}^{r} \frac{\|\mathfrak{r}(\widetilde{\lambda}_i)[\cdot, \widehat{\psi}_i^d]\|_{\mathcal{A},-1/2}^2}{\|\widehat{\psi}_d\|^2}} < 1. \tag{33}$$

Now, for each $i = 1, \ldots r$, there exists a vector q_i and a scalar $0 < \theta_i < \pi/2$ such that

$$\sin\theta_i = \frac{\|q_i - \widehat{\psi}_i\|_1}{\|\widehat{\psi}_i\|_1} = \inf_{q \in \mathrm{Ran}(Q)} \frac{\|q - \widehat{\psi}_i\|_1}{\|\widehat{\psi}_i\|_1} \ .$$

Since $q_i \in \mathrm{Ran}(Q)$, we define the eigenvectors as $\psi_i = (1/\|q_i\|_1) q_i$, and therefore $\|\psi_i\|_1 = 1$. Due to the assumption of semisimplicity of λ, all ψ_i, $i = 1, \ldots, r$, belong to λ, and, since $\sum_{i=1}^{r} \sin^2\theta_i < 1$, they span $\mathrm{Ran}(Q)$. Analogously, we define the left eigenvectors ψ_i^d, $i = 1, \ldots, r$.
Note that

$$\|\psi_i - \widehat{\psi}_i\|_1 = \|\widehat{\psi}_i\|_1 \sin\theta_i / \cos\frac{\theta_i}{2} \leq \sqrt{2} \|\widehat{\psi}_i\|_1 \sin\theta_i \ , \qquad i = 1, \ldots, r$$

and equivalently for $\|\widehat{\psi}_i^d - \psi_i^d\|_1$.

Therefore, applying (31) for this particular choice of vectors, and summing over all $i = 1, \ldots, r$, completes the proof, with the local quantity C_{cluster} given as

$$C_{\mathrm{cluster}} = \frac{2\sqrt{2} \max\{\gamma_1/|\lambda|, c_\Omega\}}{c_K c_K^*} \max\{\|\mathcal{A}\mathbf{T}_{\widetilde{\lambda}_i}\| \|\mathcal{A}^* \mathbf{T}_{\widetilde{\lambda}_i}^{(d)}\| \ : \ i = 1, \ldots, r\} \ .$$

□

Obviously, much more can be said about the sharpness of the eigenvalue estimate when we are willing to accept the lack of cluster robustness.

Remark 14 Using [41, Proposition 2.1], as has also been done in [20, Remark 9], we obtain the efficiency estimate

$$\|\mathfrak{r}(\widehat{\lambda})[\widehat{\psi}, \cdot]\|_{\mathcal{A}^*,-1/2}^2 + \|\mathfrak{r}(\widehat{\lambda})[\cdot, \widehat{\psi}^d]\|_{\mathcal{A},-1/2}^2 \leq c\big[|\lambda - \widehat{\lambda}| + \|\psi - \widehat{\psi}\|_1^2 + \|\psi^d - \widehat{\psi}^d\|_1^2\big] \ , \tag{34}$$

under the assumption that λ is a simple eigenvalue and ψ and ψ^d are the right and left eigenvector, respectively, satisfying $\|\widehat{\psi}\| = (\widehat{\psi}, \widehat{\psi}^d) = 1$. Here, the constant c depends solely on the problem (1) and the equivalence constants c_K and c_K^*.

3.4.1 Concerning cluster robustness

Analogously to Theorem 13 we can provide similar estimates for the case when, instead of λ, we have a cluster of semisimple eigenvalues λ_i, $i = 1, \ldots, r$, counted according to their multiplicity, whose convex hull is inside the contour $\mathfrak{C}$. Accordingly the space $\mathrm{Ran}(Q)$ in (32)–(33) is now a geometric eigenspace belonging to a group of eigenvalues. Consequently, vectors ψ_i and ψ_i^d are, in general, a linear combination of eigenvectors belonging to different eigenvalues in the cluster. Alternatively, as in Kato [26] we can study the stability of $\lambda_{\mathrm{avg}} := 1/r \sum_{i=1}^{r} \lambda_i = 1/r \,\mathrm{tr}(\mathcal{A}\mathbf{S}(\mathfrak{C}))$ when approximated as a function of $\tilde{\lambda}_i$, $i = 1, \cdots, r$. Such quantities (one can also consider a harmonic mean) can be estimated in a cluster robust way since we have cluster robust estimates for $\mathbf{S}(\mathfrak{C})$, however the properties of individual eigenvalues are hidden in an average. We will not present proofs of these estimates (since they are a straight forward computation). Instead we will show results of numerical computation.

In the self-adjoint case, using monotonicity together with majorization inequalities (which solves the problem of the optimal choice of averaging), we were able to establish cluster robust efficiency estimates for individual eigenvalues, see [19] for more details. Alternatively, the same problem has been solved using the majorization inequalities by Knyazev and Argrentati in [28]. Unfortunately this approach cannot be generalized directly to the non-self-adjoint case, largely due to the absence of a canonical ordering of general complex eigenvalues.

Remark 15 In order to conclude the discussion of cluster robustness, we compare our eigenvector estimates with those obtained by a direct residual analysis. Assume $\lambda \in \mathrm{Spec}(\mathcal{A})$ is given, then by $S(\lambda)$ we denote the orthogonal projection onto the geometric eigenspace $\mathrm{Ker}(\lambda - \mathcal{A})$. According to [11,20], there exists a constant C_{LBB} depending on the distance between the eigenvalue λ and the rest of the spectrum such that

$$\|(I - S(\lambda))\varphi\|_1 \leq C_{LBB} \sup_{\phi \in H_0^1(\Omega)\setminus\{0\}} \frac{|\mathfrak{r}(\lambda)[\varphi, \phi]|}{\|\phi\|_1} \,. \tag{35}$$

Obviously, for highly clustered eigenvalues constant C_{LBB} will be large. Using a perturbation argument, we may now obtain an estimate of (35) for a given $(\widehat{\lambda}, \widehat{\psi}) \in \mathbb{R} \times H_0^1(\Omega)$ which reads

$$\|(I - S(\lambda))\widehat{\psi}\|_1 \leq \widehat{C}_{LBB} \sup_{\phi \in H_0^1(\Omega)\setminus\{0\}} \frac{|\mathfrak{r}(\widehat{\lambda})[\widehat{\psi}, \phi]|}{\|\phi\|_1} \,.$$

Here, the modified constant $\widehat{C}_{LBB}$ depends on the distance between $\widehat{\lambda}$ and the nearest element of $\mathrm{Spec}(\mathcal{A}) \setminus \{\lambda\}$ and therefore is not cluster robust.

On the other hand, we have seen that our eigenvector estimates, based on the Cauchy integral, are cluster robust, and it is the distance between the given eigenvalue and the nearest unwanted eigenvalue which matters (in the case of clustered eigenvalues we assume that all of the eigenvalues in the cluster are wanted).

Proposition 4 nicely illustrates this claim. By choosing an isolated subset Λ of $\mathrm{Spec}(\mathcal{A})$, we see that only the distance to its complement matters in assessing the residual approximation. Subsequently, by summarizing estimates for each $\widehat{\psi}_i$, $i = 1, \ldots, r$, as we did in Theorem 6, we end up with the estimate which is cluster robust (the inter-cluster distances do not appear in the estimate).

4 An *hp*-Adaptive Finite Element Algorithm, and Numerical Validation

We now briefly describe the *hp*-adaptive finite element algorithm used to numerically illustrate some of our key results. We focus on problems for which $\Omega \subset \mathbb{R}^2$ and the eigenvalues are real, so we consider approximations of real eigentriplets (λ, ψ, ψ^d) of (7) in the real space $H_0^1(\Omega)$. Let $\mathcal{T} = \mathcal{T}_h$ be a triangulation of Ω with the piecewise constant mesh function $h : \mathcal{T}_h \to (0, 1)$, $h(T) = \mathrm{diam}(T)$ for $T \in \mathcal{T}_h$. We implicitly assume that $\mathcal{T}_h$ is subordinate to the polygonal partition of Ω discussed in (A3) of Definition 3; in other words, each $T \in \mathcal{T}_h$ is contained in precisely one of the polygons Ω_k. Given a piecewise constant distribution of polynomial degrees, $p : \mathcal{T}_h \to \mathbb{N}$, we define the space

$$V = V_{hp} = \{v \in H_0^1(\Omega) \cap C(\overline{\Omega}) : \; v\big|_T \in \mathbb{P}_{p(T)} \text{ for each } T \in \mathcal{T}_h\} \,,$$

where $\mathbb{P}_{p(T)}$ is the collection of polynomials of total degree not greater than p on a given element $T \in \mathcal{T}_h$. We assume that the family of spaces satisfy the following standard *regularity properties* on $\mathcal{T}$ and p: *There exists a constant* $\gamma > 0$ *for which*

($\mathfrak{C}_1$) $\gamma^{-1} h(T) \le h(T') \le \gamma h(T)$ for adjacent $T, T' \in \mathcal{T}$, $\overline{T} \cap \overline{T'} \neq \emptyset$. In other words, the diameters of adjacent elements are comparable.

($\mathfrak{C}_2$) $\gamma^{-1}(p(T)+1) \le p(T')+1 \le \gamma(p(T)+1)$ for adjacent $T, T' \in \mathcal{T}$, $\overline{T} \cap \overline{T'} \neq \emptyset$. In other words, the polynomial degrees associated with adjacent elements are comparable.

The corresponding discrete version of (7) is:
Find an eigentriple $(\widehat{\lambda}, \widehat{\psi}, \widehat{\psi}^d) \in \mathbb{R} \times V \times V$ *such that*

$$B(\widehat{\psi}, \phi) = \widehat{\lambda}(\widehat{\psi}, \phi) \quad \text{and} \quad B(\phi, \widehat{\psi}^d) = \widehat{\lambda}(\phi, \widehat{\psi}^d) \quad \text{for all } \phi \in V, \tag{36}$$

with $\|\widehat{\psi}\| = \|\widehat{\psi}^d\| = 1$. Choosing a (standard, real) basis $\{v_1, v_2, \ldots, v_N\}$ of V, we obtain the algebraic eigenvalue problems

$$B\mathbf{x} = \widehat{\lambda} M \mathbf{x} \quad \text{and} \quad B^T \mathbf{y} = \widehat{\lambda} M \mathbf{y} \quad \text{with} \quad \mathbf{x}^T M \mathbf{x} = \mathbf{y}^T M \mathbf{y} = 1, \tag{37}$$

where $B_{ij} = B(v_j, v_i)$, $M_{ij} = (v_j, v_i) = (v_i, v_j)$. The vectors $\mathbf{x}$ and $\mathbf{y}$ are the coefficient vectors of $\widehat{\psi}$ and $\widehat{\psi}^d$, respectively, i.e., $\widehat{\psi} = \sum_{i=1}^{N} \mathbf{x}_i v_i$ and $\widehat{\psi}^d = \sum_{i=1}^{N} \mathbf{y}_i v_i$.

We also consider the discrete analogues of (8) and (9). In particular, for $f \in L^2(\Omega)$, we define $\widehat{u}(f), \widehat{u}^d(f) \in V$ as the solutions of

$$B(\widehat{u}(f), v) = (f, v) \quad \text{and} \quad B(v, \widehat{u}^d(f)) = (f, v) \quad \text{for all } v \in V. \tag{38}$$

With these definitions, it is clear that $\widehat{u}(\widehat{\psi}) = \widehat{\lambda}^{-1}\widehat{\psi}$ and $\widehat{u}^d(\widehat{\psi}^d) = \widehat{\lambda}^{-1}\widehat{\psi}^d$ or, equivalently, $\widehat{u}(f) = \widehat{\psi}$ and $\widehat{u}^d(f^d) = \widehat{\psi}^d$, where $f = \widehat{\lambda}\widehat{\psi}$ and $f^d = \widehat{\lambda}\widehat{\psi}^d$. From this we obtain the following expressions for the $H^{-1}(\Omega)$-norms of the right and left residuals:

$$\|\mathfrak{r}(\widehat{\lambda})[\widehat{\psi}, \cdot]\|_{-1} = \sup_{v \in H_0^1(\Omega)\setminus\{0\}} \frac{|B(\widehat{u}(f) - u(f), v)|}{\|v\|_1}\,, \tag{39}$$

$$\|\mathfrak{r}(\widehat{\lambda})[\cdot, \widehat{\psi}^d]\|_{-1} = \sup_{v \in H_0^1(\Omega)\setminus\{0\}} \frac{|B(v, \widehat{u}^d(f^d) - u^d(f^d))|}{\|v\|_1}\,. \tag{40}$$

These expressions make apparent the link between the dual norms of the residual and the discretization errors in associated boundary value problems with source terms f and f^d. One may also consider other dual norms of the right and left residuals, for example by replacing $\|v\|_1$ in (39)–(40) with the energy norm $|||v|||$ associated with the hermitian part B_H of B, or by $\|v\|_{\mathcal{A}^*,1/2}$ for (39) and $\|v\|_{\mathcal{A},1/2}$ for (40). Any of these three variations are to be considered as *idealized* (non-computable) error estimates, and although those involving the dual norms $\|\cdot\|_{\mathcal{A},-1/2}$ and $\|\cdot\|_{\mathcal{A}*,-1/2}$ are preferred from a theoretical point of view, it is more natural to derive a practical algorithm based on computable approximations of the dual norms $\|\cdot\|_{-1}$ or $|||\cdot|||_{-1}$. The most direct way to obtain such approximations of (39) begins with the obvious bounds

$$\gamma_0 \|\widehat{u}(f) - u(f)\|_1 \le \|\mathfrak{r}(\widehat{\lambda})[\widehat{\psi}, \cdot]\|_{-1} \le \gamma_1 \|\widehat{u}(f) - u(f)\|_1\,, \tag{41}$$

$$c_0 |||\widehat{u}(f) - u(f)||| \le |||\mathfrak{r}(\widehat{\lambda})[\widehat{\psi}, \cdot]|||_{-1} \le c_1 |||\widehat{u}(f) - u(f)|||\,, \tag{42}$$

and then uses one of the many available techniques (cf. [10,16,36]) for approximating $\|\widehat{u}(f) - u(f)\|_1$ or $|||\widehat{u}(f) - u(f)|||$, or their analogues for f^d.

Our experiments have been carried out using the APTOFEM package (`www.aptofem.com`) on a single processor desktop machine. In particular, we have used ARPACK [32] to solve the algebraic eigenvalue problems, employing MUMPS [1] to solve the necessary linear systems. Since ARPACK is based on the Arnoldi algorithm, we have to solve the projected eigenvalue problem twice—once for the right- and once for the left-eigenvectors. In contrast, by using the nonsymmetric Lanczos algorithm, such as the one implemented in ABLEPACK [8], one obtains simultaneously both the right and the left eigenvectors approximations. The choice of the most efficient algebraic eigensolver is beyond the scope of this article.

Let us shortly summarize the adaptive algorithm used in our simulations. At first we choose the indices i of the eigenvalues of interest. On the initial coarse mesh we compute the approximations $(\widehat{\lambda}_{i,hp}, \widehat{\psi}_{i,hp}, \widehat{\psi}^d_{i,hp})$, and estimate (39)–(40) via $\|\widehat{u}(f_i) - u(f_i)\|_1$ and $\|\widehat{u}^d(f_i^d) - u(f_i^d)\|_1$, for $f_i = \widehat{\lambda}_{i,hp}\widehat{\psi}_{i,hp}$ and $f_i^d = \widehat{\lambda}_{i,hp}\widehat{\psi}^d_{i,hp}$, using the approach of [36]. This hp-weighted residual method (HPR), though derived and analyzed for the Laplacian, is easily extended to treat operators considered here, and we also use its obvious extension for estimates in the energy norm. We determine the elements $T \in \mathcal{T}$ to be marked for refinement by using a simple fixed-fraction marking strategy based on the values of the corresponding local error indicators, with different percentages for refinement and de-refinement. The choice between h- or p-refinement is based on an estimation of the local analyticity of the exact eigenvectors using their approximations, see [24] for further detail. Finally, a refined space is generated and the process is restarted by taking the previously calculated eigentriples $(\widehat{\lambda}_{i,hp}, \widehat{\psi}_{i,hp}, \widehat{\psi}^d_{i,hp})$ as the initial values for the computations in the refined space. By this process we generate a (nested) family of spaces $\{V_{hp}\}$.

The HPR approach described above is cheap and well-suited for guiding adaptive refinement, but tends to significantly overestimate H^1 and energy norms of $\widehat{u}(f) - u(f)$, so it does not provide sufficient insight into how well our idealized error estimates approximate eigenvalue and eigenvector error in practice. To this end, we also compute a more expensive, but far more accurate, goal oriented dual weighted residual (DWR) error estimator in our experiments for the purposes of effectivity (the ratio of estimated error over true error) analysis. Our DWR approach is described in [17], and employs the family of meshes $\{\mathcal{T}_{hp}\}$ and spaces $\{V_{hp}\}$ generated by the HPR approach.

Remark 16 As noted above, we do not compute an eigentriple directly, but instead use two independent runs of ARPACK to produce two eigenpairs for index i. The two computed eigenvalues will be much closer to each other than they are to the true eigenvalue they are approximation, i.e. the numerical errors coming from ARPACK are smaller than the discretization errors, so it is really immaterial whether or not we have use a single value $\widehat{\lambda}_{i,hp}$ as described above for our computed error estimates. In fact, it is the vectors $\widehat{\psi}_{i,hp},, \widehat{\psi}^d_{i,hp}$ which are of primary significance, and we could use the generalized Rayleigh quotient $B(\widehat{\psi}_{i,hp},, \widehat{\psi}^d_{i,hp})/(\widehat{\psi}_{i,hp},, \widehat{\psi}^d_{i,hp})$ as our approximate eigenvalue at any rate.

Following [5], for our convergence plots we use error models of the forms

$$|\lambda - \widehat{\lambda}| = Ce^{-2\alpha(\mathrm{DOFs})^r} \quad , \quad \|\psi - \widehat{\psi}\|_1 = Ce^{-\alpha(\mathrm{DOFs})^r} \ , \tag{43}$$

for eigenvalue and eigenvector error in hp-adaptive approximations, with $r = 1/2$ for eigenfunctions which are expected to be smooth, and $r = 1/3$ for eigenfunctions which are expected to have singularities. Here, DOFs is the dimension of the space V_{hp}. We refer to the number α as the *convergence rate*, and we compute it from our data using a least-squares fit of the appropriate error model. These convergence plots not only track the actual error, but also our estimates based on both the HPR and DWR approaches. In the effectivity plots we use only the ratio of the error under consideration and our DWR estimate of it, EFF=(estimated error)/(true error). In all cases we observe that the effectivities are very near 1, which suggests that our ideal error estimates may be asymptotically exact in some some cases, though analysis supporting this conjecture is beyond the scope of this paper.

4.1 Dumbbell Problem

For this example we consider the operator $\mathcal{A}v = -\Delta v + b \cdot \nabla v$, where $b = (0, 1)$, on the Dumbbell domain formed by two $\pi \times \pi$ squares joined by a $\pi/4 \times \pi/4$ "bridge", see Figure 1.

Highly accurate eigenvalues (and eigenvectors) for this problem were computed in the case $b = (0, 0)$ in [40], and we include the first six eigenvalues from that paper here to demonstrate the clustering of eigenvalues which the small bridge induces:

$$\begin{array}{lll} \lambda_1 = 1.9557938 & \lambda_3 = 4.8007611 & \lambda_5 = 4.9968371 \\ \lambda_2 = 1.9606830 & \lambda_4 = 4.8298953 & \lambda_6 = 4.9968509 \end{array}$$

As seen in Example 10, any *constant* convection b just shifts the eigenvalues of $-\Delta$ to those of $\mathcal{A}$ by the constant $|b|^2/4$, so the clustering is not affected by such a shift. In Figure 2(a) we present the convergence history for the error $\sum_{i=3}^{6} |\lambda_i - \widehat{\lambda}_i|/\widehat{\lambda}_i$. In Figure 3, we also present the final computed mesh which was generated by marking with regard to all 4 residuals which are associated to the cluster.

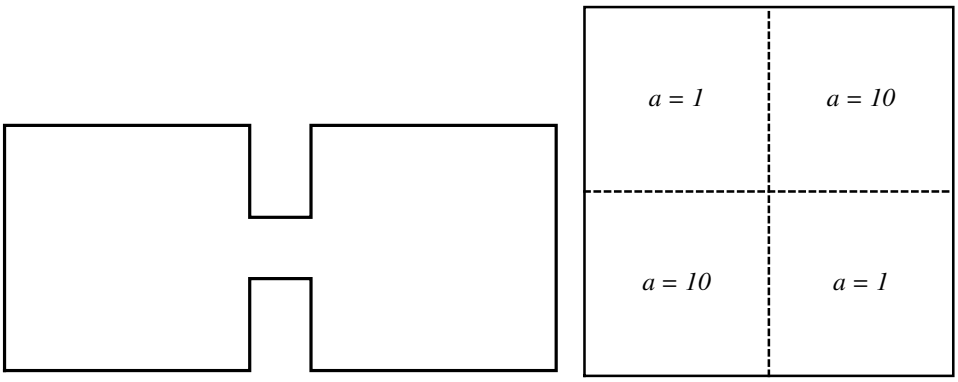
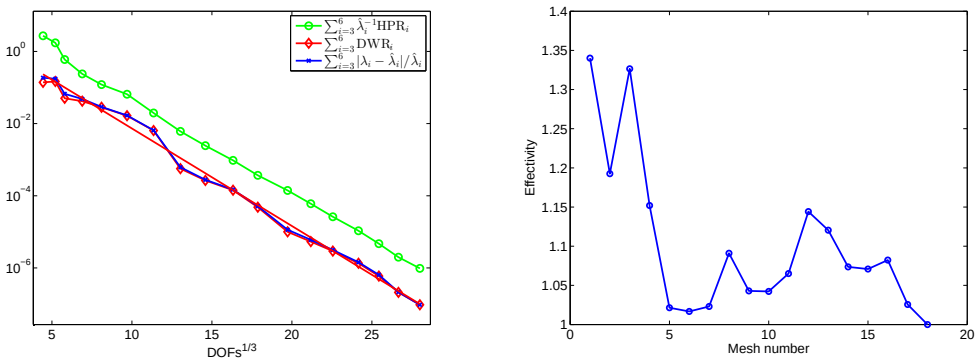


Fig. 1 The dumbbell domain (left) and the square domain with diffusion coefficients for the Kellogg problem.



(a) Convergence history for the cluster λ_i, $i = 3, 4, 5, 6$. (b) Effectivity index for the estimate of λ_{avg}.

Fig. 2 Convergence history and DWR effectivity for the dumbbell problem. The estimated convergence rate is 0.31122.

4.2 Kellogg Problem

For this example we consider the eigenvalue version of a Kellogg problem [27]: Ω is the square domain $\Omega = (-1, 1) \times (-1, 1)$, and $\mathcal{A}v = -\nabla \cdot (a\nabla v) + b \cdot \nabla v$, where $b = (2, 2)$ and $a = 10$ in quadrants I and III, and $a = 1$ in quadrants II and IV, see Figure 1.

For this problem the lowermost eigenvalue is simple and Theorem 13 with $r = 1$ reads

$$\frac{|\lambda - \tilde{\lambda}|}{|\tilde{\lambda}|} \leq \frac{C_{\mathrm{cluster}}}{|(\widehat{\psi}, \widehat{\psi}^d)|} \|\mathfrak{r}(\widehat{\lambda})[\widehat{\psi}, \cdot]\|_{-1} \|\mathfrak{r}(\widehat{\lambda})[\cdot, \widehat{\psi}^d]\|_{-1} \tag{44}$$

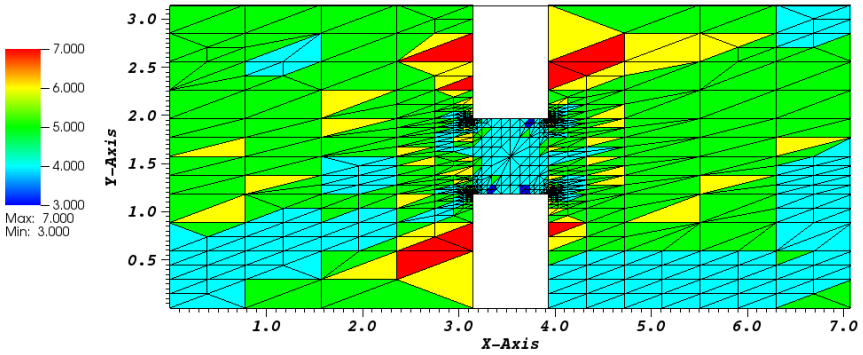


Fig. 3 Final mesh adapted to the cluster of eigenvalues for the dumbbell problem.

for given $\widehat{\psi}$, $\widehat{\psi}^d$, such that $\|\widehat{\psi}\| = \|\widehat{\psi}^d\| = 1$, and the Rayleigh quotient $\tilde{\lambda} = B(\widehat{\psi}, \widehat{\psi}^d)/(\widehat{\psi}, \widehat{\psi}^d)$. Furthermore, according to Proposition 4 there exist eigenvectors ψ and ψ^d such that

$$\|\psi - \widehat{\psi}\|_1 \leq C_l \|\mathfrak{r}(\widehat{\lambda})[\widehat{\psi}, \cdot]\|_{-1}, \qquad \text{and} \qquad \|\psi^d - \widehat{\psi}^d\|_1 \leq C_l \|\mathfrak{r}(\widehat{\lambda})[\cdot, \widehat{\psi}^d]\|_{-1} \ . \tag{45}$$

We can replace the norms $\|\cdot\|_1$ and $\|\cdot\|_{-1}$ with $|||\cdot|||_1$ and $|||\cdot|||_{-1}$ in (44) and (45) by adjusting the constants, and we do so here because the jump discontinuity in a makes an energy norm estimate more appropriate. Convergence histories and effectivities for the lowermost eigenvalue and left and right eigenvectors are provided in Figures 4 and 5.

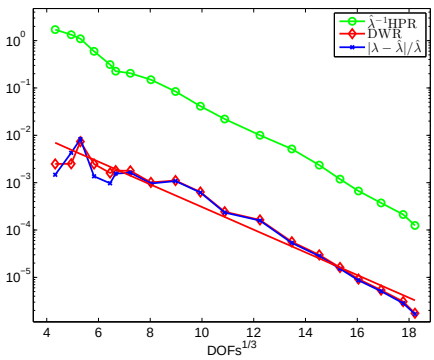
For the lowermost eigenvalue we also study the quotient $\|\widehat{\psi}\|\|\widehat{\psi}^d\|/|(\widehat{\psi}, \widehat{\psi}^d)|$ in relation to the dimension of V_{hp}. Note that, since the lowermost eigenvalue λ_1 is simple we have, by the a priori convergence analysis in [6] and references therein, that

$$\frac{\|\widehat{\psi}\|\|\widehat{\psi}^d\|}{|(\widehat{\psi}, \widehat{\psi}^d)|} \to \lim_{z \to \lambda_1} |z - \lambda_1| \ \|(z - \mathcal{A})^{-1}\|. \tag{46}$$

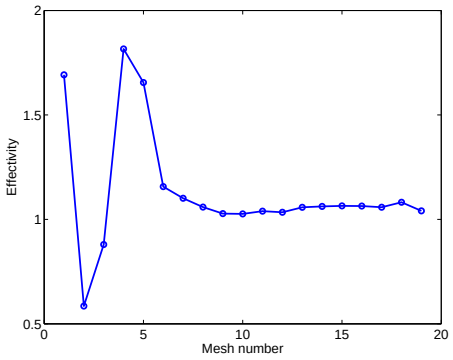
We cannot guarantee that, for a given V_{hp} and computed $\widehat{\psi}$ and $\widehat{\psi}^d$, the scalar product $(\widehat{\psi}, \widehat{\psi}^d)$ will be nonzero; but we can guarantee that $\|\widehat{\psi}\|\|\widehat{\psi}^d\|/|(\widehat{\psi}, \widehat{\psi}^d)|$ must converge to the local condition number, as given in (46) (cf. (3) for broader context), as h decreases and/or p increases. We will not discuss the convergence of adaptive finite element procedures here. We also emphasize that $(\widehat{\psi}, \widehat{\psi}^d)$ is a computable quantity, so the applicability of Theorem 13 can always be checked. We note that the residual norm together with geometric results like Theorem 6 can give a criterion for relating the size of the residual norm to the measure of spectral separation, which would force the computed condition number $\|\widehat{\psi}\|\|\widehat{\psi}^d\|/|(\widehat{\psi}, \widehat{\psi}^d)|$ for a Galerkin approximation to be near the local condition number and certainly away from zero. We leave out the technical details, but instead present the dependence of $\|\widehat{\psi}\|\|\widehat{\psi}^d\|/|(\widehat{\psi}, \widehat{\psi}^d)|$ on the convection term for the approximation of the lowermost eigentriple for the Kellogg problem in Figure 6(b).

The relative eigenvalue error and the error estimate for the first eigenvalue obtained using our *hp*-adaptive scheme with 15% for refinement and 4% for de-refinement are presented in Figure 4(a). The value of the convergence rate for the eigenvalue estimated with the least-squares fitting is $\alpha = 0.2757$ and the reference value for the eigenvalue is 17.714316 with an accuracy of 10^{-6}. In Figure 4(b) the corresponding effectivity indices are displayed. The right and left eigenvector errors together with the associated error estimates, and the effectivity indices are given in Figure 5(a) and Figure 5(b), respectively. The convergence rate for the right and left eigenvectors estimated with the least-squares fitting are $\alpha = 0.1834, 0.1813$. The final *hp*-adapted mesh is presented in Figure 6(a). As expected, the *h*-adaptivity has concentrated around the singularity in the center of the domain.

In Figure 6(b) we see that the eigenproblem of computing the lowermost eigentriplet is well conditioned. This indicates that the measured convergence rate is mainly influenced by the lack of regularity in the eigenfunction due to the type of discontinuity in the diffusion coefficient. This convergence claim is



(a) Convergence of the first eigenvalue.



(b) Effectivity index of the the goal oriented dual weighted residual estimator.

Fig. 4 Convergence and effectivity histories for the lowest eigenvalue for the Kellogg problem. The estimated convergence rate is 0.27567.

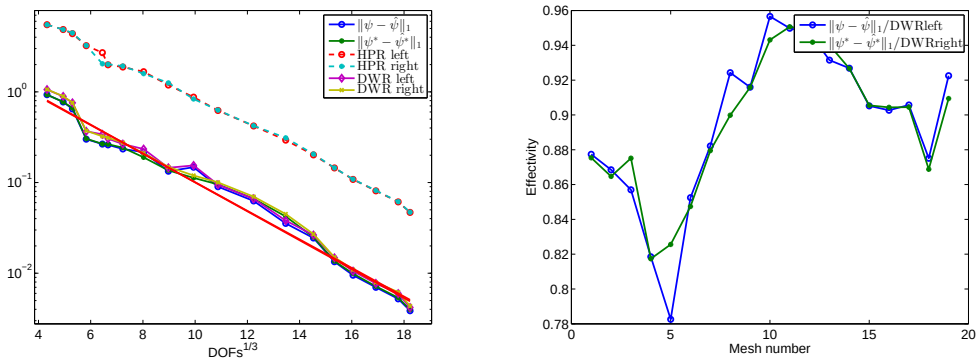
further corroborated by the fact that the convergence rate for the problem with $b = (0, 0)$ (a self-adjoint problem) is is essentially the same as when $b = (2, 2)$. More specifically, for the self-adjoint problem we have $r = 1/3$ and $\alpha = 0.37451$—recall the error model (43).

5 Conclusions

In this paper we have presented new relative estimates for the eigenvalue/function approximation error for a class of convection–diffusion–reaction operators. The main ingredients of our analysis have been Kato's square root theorem, which holds for the whole class of convection–diffusion–reaction type operators with bounded coefficients, and a generalization of the Bauer-Fike type theorem (cf. discussion on [42, p. 95]), which holds only in the case when the eigenfunctions of the operator constitute a Riesz basis of the entire Hilbert space where the problem is posed. The condition number of the Riesz basis of eigenvectors measures the global sensitivity of all eigenvalues and appears in our upper estimates of approximation errors. In the case of convection–diffusion–reaction operators which satisfy the conditions from Example 10, this global quantity is a good measure of the sensitivity of individual eigenvalues as well. We have also presented estimates which hold when no Riesz basis assumption is imposed. Most importantly, our estimation technique does not require imposition of any Galerkin orthogonality constraints. This feature allows us to directly include the treatment of the approximation errors caused by finite precision arithmetic and inexact solvers in our theoretical framework.

Acknowledgement

The authors would like to thank Prof. Dr. V. Mehrmann, Technical University Berlin, for very helpful comments on the manuscript and Dr. C. Engström, University of Umea for bringing the early and



(a) Relative errors

(b) Effectivity indices

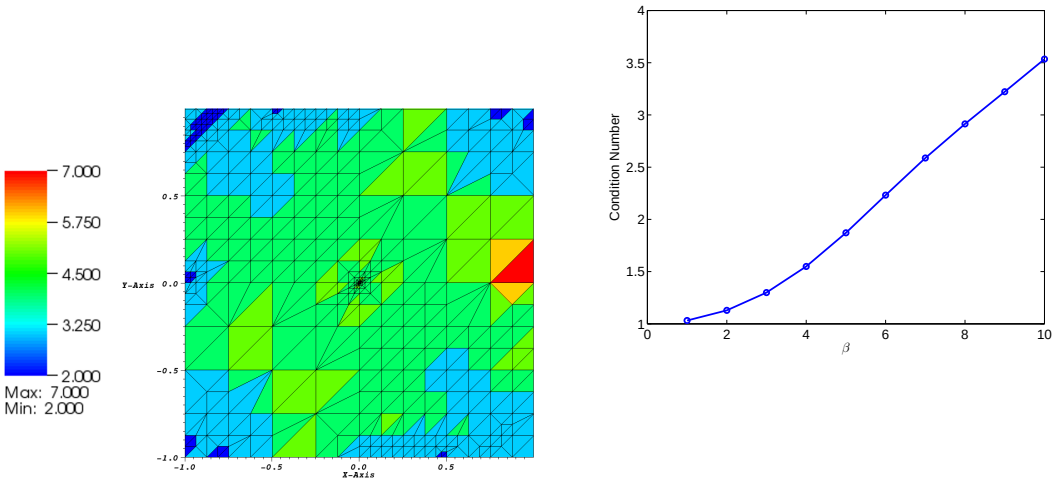
Fig. 5 Convergence of the left and right eigenvectors (eigenfunctions) corresponding to the first eigenvalue for the Kellogg problem. Estimated convergence rates for the left and right eigenvectors (eigenfunctions) are respectively: 0.18338 and 0.18134.

important references [29,30] to our attention. We also thank the referees and editor for very helpful suggestions on refining the manuscript.

L. Grubišić was supported by the Croatian MZOS Grant Nr. $037-0372783-2750$ "Spectral decompositions – numerical methods and applications". A. Międlar was supported by the DFG Research Center Matheon. J. Ovall was supported by the National Science Foundation under contract DMS-1414365.

References

1. P. Amestoy, I. Duff, and J.-Y. L'Excellent. Multifrontal parallel distributed symmetric and unsymmetric solvers. *Comput. Methods in Appl. Mech. Eng.*, 184(2–4):501–520, 2000.
2. P. Auscher, S. Hofmann, M. Lacey, A. McIntosh, and P. Tchamitchian. The solution of the Kato square root problem for second order elliptic operators on $\mathbb{R}^n$. *Ann. of Math. (2)*, 156(2):633–654, 2002.
3. P. Auscher, S. Hofmann, A. McIntosh, and P. Tchamitchian. The Kato square root problem for higher order elliptic operators and systems on $\mathbb{R}^n$. *J. Evol. Equ.*, 1(4):361–385, 2001. Dedicated to the memory of Tosio Kato.
4. P. Auscher and P. Tchamitchian. Square roots of elliptic second order divergence operators on strongly Lipschitz domains: L^2 theory. *J. Anal. Math.*, 90(1):1–12, 2003.
5. I. Babuška and B. Q. Guo. The h-p version of the finite element method for domains with curved boundaries. *SIAM J. Numer. Analysis*, 25 (1988), pp. 837âĂŞ861.
6. I. Babuška and J. Osborn. Eigenvalue problems. In P. G. Ciarlet and J.-L. Lions, editors, *Handbook of numerical analysis. Vol. II*, Handbook of Numerical Analysis, II, pages 641–787. North-Holland, Amsterdam, 1991. Finite element methods. Part 1.
7. I. Babuška and M. Suri. The p and h-p versions of the finite element method, basic principles and properties. *SIAM Rev.*, 36(4):578–632, 1994.
8. Z. Bai and D. Day. Lanczos method. In Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, editors, *Templates for the Solution of Algebraic Eigenvalue Problems: a Practical Guide*. SIAM Philadelphia, 2000.
9. A. V. Balakrishnan. Fractional powers of closed operators and the semigroups generated by them. *Pacific J. Math.*, 10:419–437, 1960.
10. R. E. Bank, J. Xu and B. Zheng. Superconvergent derivative recovery for Lagrange triangular elements of degree p on unstructured grids. *SIAM J. Numer. Anal.*, 45:2032–2046, 2007.



(a) The order of polynomials in each element is expressed in the color scheme.

(b) Dependence of the condition number for the Galerkin approximation of the lowermost eigenvalue of the Kellog's problem on the norm of convection term. The coefficient b of the convection term is set as $b = \beta(1, 1)$

Fig. 6 Final *hp*-adapted mesh for the Kellogg problem. We use this mesh and a varying convection to study the dependence of the local condition number on the convection term.

11. C. Carstensen and J. Gedicke. A posteriori error estimators for convection-diffusion eigenvalue problems. *Comput. Methods Appl. Mech. Engrg.*, 268:160–177, 2014.
12. E. B. Davies. *Linear operators and their spectra*, volume 106 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2007.
13. N. Dunford and J. T. Schwartz. *Linear operators. Part I.* Wiley Classics Library. John Wiley & Sons Inc., New York, 1988. General theory, With the assistance of William G. Bade and Robert G. Bartle, Reprint of the 1958 original, A Wiley-Interscience Publication.
14. S. C. Eisenstat and I. C. F. Ipsen. Three absolute perturbation bounds for matrix eigenvalues imply relative bounds. *SIAM J. Matrix Anal. Appl.*, 20(1):149–158 (electronic), 1999.
15. A. Ern and J.-L. Guermond. *Theory and practice of finite elements*, volume 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2004.
16. A. Ern and M. Vohralík. Polynomial-degree-robust a posteriori estimates in a unified setting for conforming, nonconforming, discontinuous Galerkin, and mixed discretizations. *HAL (Inria)* Preprint 00921583, Dec. 2013.
17. S. Giani, L. Grubisic and J. Ovall. Benchmark results for testing adaptive finite element eigenvalue procedures. *Applied Numerical Mathematics*, 62(2):121–140, 2012.
18. I. C. Gohberg and M. G. Kreĭn. *Introduction to the theory of linear nonselfadjoint operators*. Translated from the Russian by A. Feinstein. Translations of Mathematical Monographs, Vol. 18. American Mathematical Society, Providence, R.I., 1969.
19. L. Grubišić and J. S. Ovall. *On estimators for eigenvalue/eigenvector approximations*, *Math. Comp.*, 78:739–770, 2009.
20. V. Heuveline and R. Rannacher. A posteriori error control for finite approximations of elliptic eigenvalue problems. *Adv. Comput. Math.*, 15(1-4):107–138 (2002), 2001.
21. V. Heuveline and R. Rannacher. Duality-based adaptivity in the *hp*-finite element method. *J. Numer. Math.*, 11(2):95–113, 2003.
22. V. Heuveline and R. Rannacher. Adaptive FEM for eigenvalue problems with application in hydrodynamic stability analysis. In *"Advances in Numerical Mathematics", Proc. Int. Conf. , Sept. 16-17, 2005, Moscow*, Moscow: Institute of Numerical Mathematics RAS, 2006.
23. P. Houston and E. Süli. Adaptive finite element approximation of hyperbolic problems. In T. Barth and H. Deconinck, editors, *Error Estimation and Adaptive Discretization Methods in Computational Fluid Dynamics. Lect. Notes Comput. Sci. Engrg.*, volume 25, pages 269–344. 2002.

24. P. Houston and E. Süli. A note on the design of hp-adaptive finite element methods for elliptic partial differential equations. *Comp. Methods in Appl. Mech. Eng.*, 194(2–5):229–243, Feb. 2005.
25. T. Kato. Fractional powers of dissipative operators. *J. Math. Soc. Japan*, 13:246–274, 1961.
26. T. Kato. *Perturbation theory for linear operators.* Classics in Mathematics. Springer-Verlag, Berlin, 1995. Reprint of the 1980 edition.
27. R. B. Kellogg. On the Poisson equation with intersecting interfaces. *Applicable Anal.*, 4:101–129, 1974/75. Collection of articles dedicated to Nikolai Ivanovich Muskhelishvili.
28. A. V. Knyazev and M. E. Argentati. Rayleigh-Ritz majorization error bounds with applications to FEM. *SIAM J. Matrix Anal. Appl.*, 31(3):1521–1537, 2009.
29. W. G. Kolata. *Spectral approximation and spectral properties of variationally posed nonselfadjoint problems.* ProQuest LLC, Ann Arbor, MI, 1976. Thesis (Ph.D.)–University of Maryland, College Park.
30. W. G. Kolata. Approximation in variationally posed eigenvalue problems. *Numer. Math.*, 29(2):159–171, 1977/78.
31. K. Kolman. A two-level method for nonsymmetric eigenvalue problems. *Acta Math. Appl. Sin. Engl. Ser.*, 21(1):1–12, 2005.
32. R. B. Lehoucq, D. C. Sorensen, and C. Yang. *ARPACK Users' Guide*, volume 6 of *Software, Environments, and Tools.* Solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods. SIAM Philadelphia, 1998.
33. J.-L. Lions. Espaces d'interpolation et domaines de puissances fractionnaires d'opérateurs. *J. Math. Soc. Japan*, 14:233–241, 1962.
34. A. S. Markus. *Introduction to the spectral theory of polynomial operator pencils*, volume 71 of *Translations of Mathematical Monographs.* Translated from the Russian by H. H. McFaden, Translation edited by Ben Silver, With an appendix by M. V. Keldysh. Amer. Math. Soc., Providence, RI, 1988.
35. A. McIntosh. The square root problem for elliptic operators: a survey. In *Functional-analytic methods for partial differential equations (Tokyo, 1989)*, volume 1450 of *Lecture Notes in Math.*, pages 122–140. Springer, Berlin, 1990.
36. J. M. Melenk and B. I. Wohlmuth. On residual-based a posteriori error estimation in *hp*-FEM. *Adv. Comput. Math.*, 15(1-4):311–331 (2002), 2001.
37. J. C. Miller and J. N. Miller. *Statistics for Analytical Chemistry.* Ellis Horwood Ltd, 3 sub (April 1993) edition, July 1993.
38. E. Ovtchinnikov. Cluster robust error estimates for the Rayleigh-Ritz approximation. I. Estimates for invariant subspaces. *Linear Algebra Appl.*, 415(1):167–187, 2006.
39. M. Reed and B. Simon. *Methods of modern mathematical physics. I.* Second Edition. Functional analysis. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York, 1980.
40. L. N. Trefethen and T. Betcke. Computed eigenmodes of planar regions. In *Recent advances in differential equations and mathematical physics*, volume 412 of *Contemp. Math.*, pages 297–314. Amer. Math. Soc., Providence, RI, 2006.
41. R. Verfürth. A posteriori error estimates for nonlinear problems. Finite element discretizations of elliptic equations. *Math. Comp.*, 62(206):445–475, 1994.
42. D. S. Watkins. *The matrix eigenvalue problem.* SIAM Philadelphia, PA, 2007.