2-2019

Spectral Discretization Errors in Filtered Subspace Iteration

Jay Gopalakrishnan
Portland State University, gjay@pdx.edu

Luka Grubišić
University of Zagreb

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

MATHEMATICS OF COMPUTATION
Volume 00, Number 0, Pages 000–000
S 0025-5718(XX)0000-0

SPECTRAL DISCRETIZATION ERRORS IN FILTERED SUBSPACE ITERATION

JAY GOPALAKRISHNAN, LUKA GRUBIŠIĆ, AND JEFFREY OVALL

Abstract. We consider filtered subspace iteration for approximating a cluster of eigenvalues (and its associated eigenspace) of a (possibly unbounded) selfadjoint operator in a Hilbert space. The algorithm is motivated by a quadrature approximation of an operator-valued contour integral of the resolvent. Resolvents on infinite dimensional spaces are discretized in computable finite-dimensional spaces before the algorithm is applied. This study focuses on how such discretizations result in errors in the eigenspace approximations computed by the algorithm. The computed eigenspace is then used to obtain approximations of the eigenvalue cluster. Bounds for the Hausdorff distance between the computed and exact eigenvalue clusters are obtained in terms of the discretization parameters within an abstract framework. A realization of the proposed approach for a model second-order elliptic operator using a standard finite element discretization of the resolvent is described. Some numerical experiments are conducted to gauge the sharpness of the theoretical estimates.

1. Introduction

The goal of this study is to provide an analysis of discretization errors that arise when a popular filtered subspace iteration algorithm is employed to compute eigenvalues of selfadjoint partial differential operators. Instead of a specific differential operator, we consider a general linear, closed, selfadjoint operator $A : \mathrm{dom}(A) \subseteq \mathcal{H} \to \mathcal{H}$ (not necessarily bounded) in a complex Hilbert space $\mathcal{H}$, whose (real) spectrum is denoted by $\Sigma(A)$. We are interested in computationally approximating a subset Λ of the spectrum that consists of a finite collection of eigenvalues of finite multiplicity.

Filtered subspace iteration is a method for approximating Λ and its corresponding eigenspace (invariant subspace) and is a natural generalization of the power method [23, 28]. It can roughly be described as follows. First, the eigenspace of the cluster Λ is transformed to the dominant eigenspace of another, bounded operator called the "filter." Next, a subspace iteration is applied using the bounded filter. Starting with an initial subspace (usually chosen randomly), the bounded operator is repeatedly applied to it, generating a sequence of subspaces that approximates the eigenspace of Λ. Approximations of Λ are obtained from the approximate

Received by the editor April 3, 2019.

This work was partially supported by the AFOSR (through AFRL Cooperative Agreement #18RDCOR018 and grant FA9451-18-2-0031), the Croatian Science Foundation grant HRZZ-9345, bilateral Croatian-USA grant (administered jointly by Croatian-MZO and NSF) and NSF grant DMS-1414365. The numerical studies were facilitated by the equipment acquired using NSF's Major Research Instrumentation grant DMS-1624776.

eigenspaces by a Rayleigh-Ritz procedure. To apply this filtered subspace iteration in practice requires computable finite-rank approximations of the resolvent at a few points, obtained by some *discretization* process. It is the errors incurred by such discretizations that form the subject of this paper.

The exact eigenspace, namely the span of all the eigenvectors associated with elements of Λ, is denoted by E. Then $m = \dim E$, being the sum of multiplicities of each element of Λ, is finite, and we assume $m \geqslant 1$. Throughout this paper, the multiplicity ℓ of an eigenvalue λ of an operator refers to its algebraic multiplicity, i.e., λ is a pole of order ℓ of the resolvent of that operator. Recall that, for a selfadjoint operator A, the algebraic multiplicity of λ coincides with its geometric multiplicity, $\dim \ker(\lambda - A)$.

As mentioned above, the idea behind filtered subspace iteration is to transform E into the dominant eigenspace of certain filter operators. We shall see in the next section that the construction of these filters can be motivated by approximations of a Dunford-Taylor contour integral. There has been a resurgence of interest in contour integral methods for eigenvalues due to their excellent parallelizability [2,4, 14,15,24]. Following [2], we identify two different classes of methods in the existing literature that use contour integrals for computation of a targeted cluster of matrix eigenvalues. One class of methods, that often goes by the name SS-methods [24] (see also [4,16]), approximates Λ by the eigenvalues of a system of moment matrices based on contour integrals. The moment matrices are obtained by approximating the integrals by a quadrature, and the spectral approximation error depends on the accuracy of the quadrature.

The other class of methods are referred to by the name FEAST [21] (see also [14, 28]). They are more related to our present contribution (the difference being that while FEAST is a matrix algorithm, we focus on filtered subspace iterations applied to inifinite-dimensional selfadjoint operators and their discretizations). Like SS-methods, FEAST also uses quadratures to approximate a contour integral. In our view, the use of quadratures in FEAST is essentially different from their use in approaches like the SS-method. Quadratures in FEAST are only used to develop the filter used in a subspace iteration. A consequence of this is that the quadrature error is not as relevant in FEAST as in the SS-method. The analysis in this paper will show this in precise terms. In particular, our findings show that the rate of convergence of the discretization error is unaffected by the quadrature error.

When A is a differential operator on an infinite-dimensional space, some approximations to bring the computations into finite-dimensional spaces are necessary. The central concern in this paper is the study of how these approximations affect the final spectral approximations that the filtered subspace iteration yields asymptotically. The main technical difficulty in analyzing discretization errors for the unbounded operator eigenproblem is that many of the existing standard tools [3] are not directly applicable to our situation. We present an abstract framework that allows one to study approximation of spectral clusters of unbounded selfadjoint operators with compact resolvent. Very general discretizations are allowed through a set of abstract assumptions.

To quickly outline the approximation approach studied in this paper, recall that the spectral projector onto E, which we denote by S, is characterized by a Dunford-Taylor contour integral of the resolvent $R(z)$. Its N-term quadrature approximation

is denoted by S_N. In the expression defining S_N, when $R(z)$ is replaced by a computable finite-rank approximation $R_h(z)$, we obtain S_N^h, a practically computable filter. Here h is some discretization parameter (such as the grid spacing) inversely related to a computational finite-dimensional space. By repeated application of S_N^h, the iteration produces a sequence of subspaces $\{E_h^{(\ell)} : \ell = 1, 2, \ldots\}$, which we study. While sufficient conditions for convergence of the FEAST iteration for matrices can be found in [14,28], it is not immediately clear that the iteration continues to converge when the operator is perturbed, such as the perturbation of S_N to S_N^h. We begin our analysis by showing that the iterates $E_h^{(\ell)}$ do converge under certain sufficient conditions, after which we focus on analyzing the limit.

To summarize the novelty of this work, this is the first work to study the effect of the *discretization parameter* h (in addition to N). The errors in eigenspace approximations often need to be measured in stronger norms than the base $\mathcal{H}$-norm. The example of elliptic differential operators on $\mathcal{H} = L^2(\Omega)$ illustrates the need to measure eigenfunction errors in a stronger norm like the $H^1(\Omega)$-norm. To our knowledge, this is the first work to give bounds for eigenspace discretization errors arising in filtered subspace algorithms in $\mathcal{H}$-norm as well as a stronger $\mathcal{V}$-norm (see Theorem 4.1). We provide the first result showing that the Hausdorff distance between the eigenvalue cluster computed by filtered subspace iteration and Λ converges to zero at predictable rates as the discretization parameter $h \to 0$ (see Theorem 5.7 and Corollary 5.8). In the process of doing so, we develop a general result (Lemma 5.1) bounding the perturbation of Ritz values of an unbounded selfadjoint operator. To highlight one more conclusion from our analysis, increasing N has little effect on the spectral discretization error as measured by the gap between E and E_h (although it may affect the speed of convergence of $E_h^{(\ell)}$ as implied by the results of [14, 28]).

The rest of the paper is organized as follows. In Section 2, we describe precisely the above-mentioned process of double approximation (going from S to S_N^h) and introduce the necessary assumptions for the error analysis. Section 3 introduces the space to which filtered subspace iteration using S_N^h converges. Bounds for the gap between computed and exact eigenspaces are proved in Section 4. Eigenvalue errors are then bounded using the square of this gap. Analysis of a standard finite element discretization of the resolvent of a model operator in Section 6 provides an example of how abstract conditions on the resolvent might be verified in practice. The practical performance of the algorithm with the Lagrange finite element discretization is reported in Section 7.

2. Preliminaries

Let A, Λ and E be as discussed previously. As already mentioned, filters are linear operators on $\mathcal{H}$ having E as their dominant eigenspace, in the sense made precise below.

Suppose that $\Gamma \subset \mathbb{C}\backslash\Sigma(A)$ is a positively oriented, simple, closed contour that encloses Λ and excludes $\Sigma(A)\backslash\Lambda$, and let $G \subset \mathbb{C}$ be the open set whose boundary is Γ. By the Cauchy Integral Formula,

$$r(\xi) = \frac{1}{2\pi \mathrm{i}} \oint_\Gamma (z - \xi)^{-1}\, dz = \begin{cases} 1, & \xi \in G, \\ 0, & \xi \in \mathbb{C}\backslash(G \cup \Gamma). \end{cases} \tag{2.1}$$

Thus $r(\xi)$ equals a.e. the indicator function of G in $\mathbb{C}$. The associated (orthogonal) spectral projection $S : \mathcal{H} \to \mathcal{H}$ is the bounded linear operator given by the Dunford-Taylor integral

$$S = \frac{1}{2\pi \mathrm{i}} \oint_{\Gamma} R(z)\, dz, \tag{2.2}$$

where $R(z) = (z - A)^{-1}$ is the resolvent, a bounded linear operator on $\mathcal{H}$ for each $z \in \Gamma$. Since Γ encloses Λ and no other element of $\Sigma(A)$, its well known that

$$E = \mathrm{ran}(S). \tag{2.3}$$

Furthermore, by functional calculus (see [22, Theorem VIII.5], [26, Theorem 5.9] or [5, Section 6.4]), if $(\lambda, \phi) \in \Sigma(A) \times \mathrm{dom}(A)$ satisfies $A\phi = \lambda\phi$, then $S\phi = r(A)\phi = r(\lambda)\phi$. Since $r(\lambda)$ equals 1 for all $\lambda \in \Lambda$ and equals 0 for all other elements of $\Sigma(A)$, the desired eigenspace E of A is now the dominant eigenspace of $S = r(A)$. In this sense, S is an *ideal filter.*

Motivated by quadrature approximations of (2.1), in the same spirit as [4,13,21, 24,28], we approximate $r(\xi)$ by

$$r_N(\xi) = w_N + \sum_{k=0}^{N-1} w_k (z_k - \xi)^{-1}\ , \tag{2.4}$$

for some $w_k, z_k \in \mathbb{C}$. The corresponding *rational filter* is the operator

$$S_N = r_N(A) = w_N + \sum_{k=0}^{N-1} w_k R(z_k), \tag{2.5}$$

which can be viewed as an approximation of S. It is common to refer to S_N, as well as the rational function $r_N(\xi)$, as the *filter.* As in the case of S, if $(\lambda, \phi) \in \Sigma(A) \times \mathrm{dom}(A)$ satisfies $A\phi = \lambda\phi$, then $S_N\phi = r_N(\lambda)\phi$. In particular, the set Λ of eigenvalues of interest have been mapped to $\{r_N(\lambda) : \lambda \in \Lambda\}$ by the filter.

These mapped eigenvalues are dominant eigenvalues of S_N if

$$\min_{\lambda \in \Lambda} |r_N(\lambda)| > \sup_{\mu \in \Sigma(A) \setminus \Lambda} |r_N(\mu)| \tag{2.6}$$

holds. This dominance can be obtained provided Λ is strictly separated from the remainder of the spectrum. To quantify the separation, we consider the following strictly separated subsets of $\mathbb{R}$ centered around $y \in \mathbb{R}$

$$I_\gamma^y = \{x \in \mathbb{R} : |x - y| \leqslant \gamma\}, \qquad O_{\delta,\gamma}^y = \{x \in \mathbb{R} : |x - y| \geqslant (1 + \delta)\gamma\}.$$

for some positive numbers γ and δ. If the spectral cluster of interest is within I_γ^y, then the number δ provides a measure of the relative gap between it and the rest of the spectrum—relative to the radius γ of the interval wherein we seek eigenvalues. Using the numbers y, γ, and δ, define

$$W = \sum_{k=0}^{N} |w_k|, \qquad \hat{\kappa} = \frac{\sup\limits_{x \in O_{\delta,\gamma}^y} |r_N(x)|}{\inf\limits_{x \in I_\gamma^y} |r_N(x)|}. \tag{2.7}$$

These definitions help us to formulate the following assumption on the filter and cluster separation.

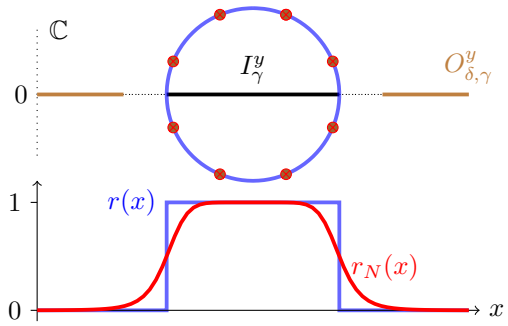


FIGURE 1. The Butterworth filter with $N = 8$ points: the points z_k are plotted in the complex plane (top) and the functions r and r_N are compared on the real line (bottom).

Assumption 2.1. There exist $y \in \mathbb{R}$, $\delta > 0$ and $\gamma > 0$ such that

$$\Lambda \subset I_\gamma^y, \qquad \Sigma(A)\backslash\Lambda \subset O_{\delta,\gamma}^y. \tag{2.8}$$

We assume that r_N is a rational function of the form (2.4) with the property that $z_k \notin \overline{\Sigma(A)}$, $W < \infty$, and $\hat{\kappa} < 1$.

Note that if $\hat{\kappa} < 1$, then (2.6) holds. When an N-point trapezoidal rule is used for quadrature approximation we obtain an r_N as in (2.4) with $w_N = 0$. When the Zolotarev rational approximation of $r(\xi)$ is used to construct r_N, the term w_N is nonzero [14].

Example 2.2 (Butterworth filter)**.** Consider the filter obtained by setting $w_N = 0$, and for $k = 0, \ldots, N-1$,

$$z_k = \gamma e^{\mathrm{i}(\theta_k+\phi)} + y, \qquad w_k = \gamma e^{\mathrm{i}(\theta_k+\phi)}/N. \tag{2.9}$$

with $\theta_k = 2\pi k/N$ and $\phi = \pm\pi/N$. These weights and points are obtained using the N-point uniform trapezoid rule approximation of the contour integral in (2.1) when Γ is set to the circle $\Gamma = \{\gamma e^{\mathrm{i}(\theta+\phi)} + y : \theta \in [0, 2\pi)\}$ enclosing a spectral cluster Λ that satisfies (2.8) – see Figure 1. It is obvious from the expression for w_k that $W = \sum_{k=0}^{N-1} \gamma/N = \gamma$, so the requirement of Assumption 2.1 that $W < \infty$ is satisfied.

An additional important requirement of Assumption 2.1 is that the filter should satisfy $\hat{\kappa} < 1$. Let us now show this holds for the Butterworth filter when N is even. We claim that

$$r_N(\xi) = \sum_{k=0}^{N-1} w_k (z_k - \xi)^{-1} = \frac{e^{\mathrm{i}N\phi}}{e^{\mathrm{i}N\phi} - ((\xi - y)/\gamma)^N}. \tag{2.10}$$

For the special case $\gamma = 1$, $y = 0$, $\phi = 0$, this claim follows from a partial fraction decomposition of $(\xi^N - 1)^{-1}$, recognizing that $\xi^N - 1 = \prod_{k=0}^{N-1}(\xi - z_k)$. Its

extension to the general case readily follows from the obvious change of variable. Restricting (2.10) to the real line, it follows by inspection that

$$\min_{x\in I_\gamma^y} |r_N(x)| = \frac{1}{2} \qquad\qquad \max_{x\in O_{\delta,\gamma}^y} |r_N(x)| = \frac{1}{(1+\delta)^N + 1}$$

for any $y \in \mathbb{R}$. Thus $\hat\kappa = 2[(1+\delta)^N + 1]^{-1} < 1$. □

Next, we introduce a subspace $\mathcal{V} \subseteq \mathcal{H}$, motivated by the need to prove results that bound discretization errors in norms stronger than the $\mathcal{H}$-norm. We place the following assumption and give example classes of operators where the assumption holds.

Assumption 2.3. Suppose there is a Hilbert space $\mathcal{V} \subseteq \mathcal{H}$ such that $E \subseteq \mathcal{V}$, there is a $C_{\mathcal{V}} > 0$ such that for all $u \in \mathcal{V}$, $\|u\|_{\mathcal{H}} \leqslant C_{\mathcal{V}}\|u\|_{\mathcal{V}}$, and $\mathcal{V}$ is an invariant subspace of $R(z)$ for all z in the resolvent set of A.

Example 2.4 ($\mathcal{V}$ is the whole space)**.** Set $\mathcal{V} = \mathcal{H}$, with $(\cdot,\cdot)_{\mathcal{V}} = (\cdot,\cdot)_{\mathcal{H}}$. In this case it is obvious that all statements of Assumption 2.3 hold. □

Example 2.5 ($\mathcal{V}$ is the domain of a positive form)**.** Suppose $a(u,v)$ is a densely defined closed sesquilinear Hermitian form on $\mathcal{H}$ and there is a $\delta > 0$ such that

$$a(v,v) \geqslant \delta\|v\|_{\mathcal{H}}^2, \qquad v \in \operatorname{dom}(a). \tag{2.11}$$

Set

$$\mathcal{V} = \operatorname{dom}(a), \qquad \|v\|_{\mathcal{V}} = a(v,v)^{1/2}.$$

To show that Assumption 2.3 holds in this case, first set the operator A to be the closed selfadjoint *operator associated with the form*, namely it satisfies $a(u,v) = (Au,v)$ for all $u \in \operatorname{dom}(A) \subseteq \operatorname{dom}(a)$ and all $v \in \operatorname{dom}(a)$ (see the first representation theorem [18, TheoremVI.2.1] or [26, Theorem 10.7]). Note that, in this case, A is a positive operator. Hence A has a unique selfadjoint positive square root [18, Theorem V.3.35], denoted by $A^{1/2}$, that commutes with any bounded operator that commutes with A. By the second representation theorem [18, Theorem VI.2.23], the form domain is characterized by $\operatorname{dom}(a) = \operatorname{dom}(A^{1/2})$, and $\|v\|_{\mathcal{V}} = \|A^{1/2}v\|_{\mathcal{H}}$ for $v \in \mathcal{V}$. The strict positivity of a ensures that both A and $A^{1/2}$ are invertible on their respective domains.

Since a is closed, $\mathcal{V}$ is complete. Due to (2.11), $\mathcal{V}$ is continuously embedded in $\mathcal{H}$, with the constant $C_{\mathcal{V}} = \delta^{-1/2}$. The exact eigenspace E is contained in $\operatorname{dom}(A) \subseteq \operatorname{dom}(A^{1/2}) = \mathcal{V}$. Since $A^{1/2}$ and $A^{-1/2}$ commutes with $R(z)$, for any $v, w \in \mathcal{V}$, we have for any $v \in \mathcal{V} = \operatorname{dom}(A^{1/2})$ and z in the resolvent set of A,

$$R(z)v = (z-A)^{-1}v = A^{-1/2}(z-A)^{-1}A^{1/2}v\ .$$

Since $\operatorname{ran}(A^{-1/2}) = \operatorname{dom}(A^{1/2}) = \mathcal{V}$, we see that $R(z)\mathcal{V} \subseteq \mathcal{V}$. Thus Assumption 2.3 is verified. □

Example 2.6 ($\mathcal{V}$ is a graph space)**.** Given A, put $\mathcal{V} = \operatorname{dom}(A) \subseteq \mathcal{H}$ and endow the set $\mathcal{V}$ with the topology of the graph norm

$$\|v\|_{\mathcal{V}} = \left(\|v\|_{\mathcal{H}}^2 + \|Av\|_{\mathcal{H}}^2\right)^{1/2}, \qquad v \in \mathcal{V}.$$

We claim that Assumption 2.3 holds in this case. Indeed, since A is closed, the graph norm makes $\mathcal{V}$ into a Hilbert space. Obviously $E \subset \mathcal{V}$ and $\mathcal{V}$ is continuously embedded into $\mathcal{H}$ with $C_{\mathcal{V}} = 1$. Since A commutes with $R(z)$ for any z in the resolvent set of A, we have $R(z)\operatorname{dom}(A) \subseteq \operatorname{dom}(A)$. □

The next essential ingredient in our study is the approximation of $R(z)$. When A is a differential operator on an infinite-dimensional space, to obtain numerical spectral approximations, we perform a discretization to approximate the resolvent of A in a computable finite-dimensional space. Accordingly, let $\mathcal{V}_h$ be a finite-dimensional subspace of $\mathcal{V}$, where h is a parameter inversely related to the finite dimension, e.g., a mesh size parameter h that goes to 0 as the dimension increases. Let $R_h(z) : \mathcal{H} \to \mathcal{V}_h$ be a finite-rank approximation to the resolvent $R(z)$ satisfying the following assumption.

Assumption 2.7. Assume that the operators $R_h(z_k)$ and $R(z_k)$ are bounded in $\mathcal{V}$ and satisfy

$$\lim_{h\to 0} \|R_h(z_k) - R(z_k)\|_{\mathcal{V}} = 0 \tag{2.12}$$

for all $k = 0, 1, \ldots, N-1$.

Note that this assumption implies that $R(z_k)$, being the limit of finite-rank operators, is compact in $\mathcal{V}$. Its also compact as an operator on $\mathcal{H}$ due to Assumption 2.3. Consequently $R(z)$ is compact for all z in the resolvent set. Relaxing Assumption 2.7 to go beyond operators with compact resolvent is outside the scope of the current work.

Consider the approximation of S_N given by

$$S_N^h = w_N + \sum_{k=0}^{N-1} w_k R_h(z_k). \tag{2.13}$$

In view of Assumption 2.7, we shall from now on view both S_N and S_N^h as bounded operators on $\mathcal{V}$. Note that S_N^h need not be selfadjoint. In Section 6, we shall consider an example of S_N^h, obtained by a standard finite element discretization of $R(z)$ based on symmetrically located z_k, that is selfadjoint. But in general S_N^h may fail to be selfadjoint due to the configuration of $\{z_k\}$ or due to the properties of the discretization (see e.g., [8]).

With the resolvent discretization, filtered subspace iteration can be described mathematically in very simple terms. Namely, starting with a subspace $E_h^{(0)} \subseteq \mathcal{V}_h$, compute

$$E_h^{(\ell)} = S_N^h E_h^{(\ell-1)}, \qquad \text{for } \ell = 1, 2, \ldots. \tag{2.14}$$

Of course, in practice, one must include (implicit or explicit) normalization steps and maintain a basis for the spaces $E_h^{(\ell)}$, but these details are immaterial in our ensuing analysis. The convergence of the FEAST algorithm in Euclidean (ℓ^2 and matrix-based) norms was previously studied in [14,23]. In Section 3, we shall utilize some of their ideas to show that (2.14) converges in $\mathcal{V}$, despite the perturbations caused by the above-mentioned resolvent approximations. Here however, we are solely interested in studying the *discretization errors found in the final asymptotic product of the algorithm*, i.e., the discretization errors in what the algorithm outputs as the "limit space" when (2.14) converges.

3. The limit space

The purpose of this section is to identify to what space convergence of (2.14) might happen. We also briefly examine in what sense $E_h^{(\ell)}$ converges to it.

In view of Assumption 2.3, $\mathcal{V}$ is an invariant subspace of the resolvent. Hence in the remainder of the paper, we will proceed *viewing S_N and S_N^h as operators on $\mathcal{V}$*. To measure the distance between two linear subspaces M and L of $\mathcal{V}$, we use the standard notion of gap [18] defined by

$$\operatorname{gap}_{\mathcal{V}}(M,L) = \max\left[\sup_{m\in U_M^{\mathcal{V}}} \operatorname{dist}_{\mathcal{V}}(m,L),\ \sup_{l\in U_L^{\mathcal{V}}} \operatorname{dist}_{\mathcal{V}}(l,M)\right]. \tag{3.1}$$

Here and throughout, for any linear subspace $M \subseteq \mathcal{V}$, we use $U_M^{\mathcal{V}}$ to denote its unit sphere $\{w \in M : \|w\|_{\mathcal{V}} = 1\}$.

Recall that $E = \operatorname{ran} S$, the exact eigenspace corresponding to eigenvalues $\lambda_1, \dots, \lambda_m$ of A that we wish to approximate. If Assumption 2.1 holds, then the operator $S_N = r_N(A)$ has dominant eigenvalues

$$\mu_i = r_N(\lambda_i), \qquad i = 1, 2, \dots, m,$$

strictly separated in absolute value from the remainder of $\Sigma(S_N)$. In particular, since $\hat{\kappa} < 1$, we have $\mu_i \neq 0$ for $i \leqslant m$, and letting $\mu_* = \sup\{|\mu| : \mu \in \Sigma(S_N)\backslash\{\mu_1, \dots \mu_m\}\}$,

$$\mu_* < |\mu_i|, \quad i = 1, 2, \dots, m. \tag{3.2}$$

In view of these facts, we can find a simple rectifiable curve Θ in the complex plane that encloses $\{\mu_1, \dots, \mu_m\}$ and lies strictly outside the circle of radius μ_*. In particular, Θ encloses no other element of $\Sigma(S_N)$. Define the spectral projector of S_N by

$$P_N = \frac{1}{2\pi \mathrm{i}} \oint_\Theta (z - S_N)^{-1}\, dz.$$

Then $E_N = \operatorname{ran} P_N$ is the eigenspace of S_N corresponding to its eigenvalues $\mu_1, \dots, \mu_m$.

Lemma 3.1. *We have $E_N = E$ and $P_N = S$.*

Proof. Since $\dim E_N = \dim E = m$, it suffices to prove that $E \subseteq E_N$. If $e_i \in E$ is an eigenfunction of A corresponding to the eigenvalue $\lambda_i \in \Lambda$, then $S_N e_i = r_N(\lambda_i) e_i$, so $e_i \in E_N$. Since P_N and S are both orthogonal projectors and have the same range, they are the same operator. □

Next, observe that when Assumption 2.7 is used after subtracting the expression for S_N^h in (2.13) from that of S_N, we obtain

$$\|S_N - S_N^h\|_{\mathcal{V}} \leqslant W \max_{k=0,\dots,N-1} \|R_h(z_k) - R(z_k)\|_{\mathcal{V}} \to 0 \tag{3.3}$$

as h goes to 0. Let us recall the standard ramifications of the convergence of operators in norm given by (3.3) (see e.g., [18, Theorem IV.3.16] or [1]). Namely, given an open disc enclosing an isolated eigenvalue of S_N of multiplicity ℓ, (3.3) implies that for sufficiently small h, there are exactly ℓ eigenvalues (counting algebraic multiplicities) of S_N^h in the same disc. In particular, this implies that, for sufficiently small h, the contour Θ is in the resolvent set of S_N^h and encloses only semisimple eigenvalues of joint multiplicity m of S_N^h, which we shall enumerate as $\mu_1^h, \mu_2^h, \dots, \mu_m^h$. Hence, the integral

$$P_h = \frac{1}{2\pi \mathrm{i}} \oint_\Theta \left(z - S_N^h\right)^{-1} dz$$

is well defined. Henceforth we assume that h has been made small enough as mentioned above.

Definition 3.2. Let $E_h = \operatorname{ran} P_h$.

Clearly, P_h is the spectral projector of S_N^h corresponding to the eigenvalues $\mu_1^h, \mu_2^h, \ldots, \mu_m^h$. Hence,

$$\dim E_h = m. \tag{3.4}$$

Note also that by construction of Θ,

$$\mu_i^h \neq 0, \qquad i = 1, 2, \ldots, m. \tag{3.5}$$

Remark 3.3. We may consider approximating A by A_h and then setting $R_h(z_k) = (z_k - A_h)^{-1}$ in the formation of S_N^h, or more generally, approximate the resolvent $R(z_k) = (z_k - A)^{-1}$ directly by some $R_h(z_k)$. In the former case, the results of [14, 28] will show convergence of the FEAST iteration, applied to the matrix A_h, provided their assumptions on A_h can be verified. The latter case allows for different discretizations at different z_k, as well as for discretizations of the resolvent by least-squares approaches, including discontinuous Petrov-Galerkin methods [8]. These approaches are of interest because, even when $R_h(z_k)$ is not selfadjoint, the application of $R_h(z_k)$ reduces to the solution of a Hermitian positive definite system.

We now show that the above-defined E_h is the limit space of subspace iterates $E_h^{(\ell)}$.

Theorem 3.4. *Starting with a subspace $E_h^{(0)} \subseteq \mathcal{V}_h$ satisfying $\dim(E_h^{(0)}) = \dim(P_h E_h^{(0)}) = m$, we compute $E_h^{(\ell)}$ by (2.14). Suppose Assumptions 2.1–2.7 hold. Then there is an $h_0 > 0$ such that, for all $h < h_0$,*

$$\lim_{\ell \to \infty} \operatorname{gap}_{\mathcal{V}}(E_h^{(\ell)}, E_h) = 0.$$

Proof. Step 1: Recall from (3.4) that $\dim(E_h) = m$ for sufficiently small h. Together with $P_h E_h^{(0)} \subseteq E_h$ and the assumption $\dim(P_h E_h^{(0)}) = m$, this leads to the equality

$$P_h E_h^{(0)} = E_h. \tag{3.6}$$

Thus

$$P_h E_h^{(\ell)} = P_h (S_N^h)^\ell E_h^{(0)} = (S_N^h)^\ell P_h E_h^{(0)} = (S_N^h)^\ell E_h = E_h.$$

In particular, this implies that $\dim(E_h^{(\ell)}) \geqslant \dim(P_h E_h^{(\ell)}) = \dim(E_h)$. Hence,

$$\dim(E_h^{(\ell)}) = \dim(E_h) = m, \qquad \ell = 0, 1, 2, \ldots. \tag{3.7}$$

Step 2: Let v_i be an eigenvector of S_N^h corresponding to eigenvalue μ_i^h. We shall now find an approximant of v_i in $E_h^{(\ell)}$. Due to (3.6), there is a $q_i^{(0)}$ in $E_h^{(0)}$ such that $P_h q_i^{(0)} = v_i$. Set

$$q_i^{(\ell)} = \left(\frac{1}{\mu_i^h}\right)^\ell (S_N^h)^\ell \, q_i^{(0)}.$$

Clearly $q_i^{(\ell)}$ is well defined due to (3.5) and is in $E_h^{(\ell)}$. Moreover,

$$\begin{aligned} v_i - q_i^{(\ell)} &= v_i - (\mu_i^h)^{-\ell} (S_N^h)^\ell \big[P_h q_i^{(0)} + (I - P_h) q_i^{(0)}\big] \\ &= -(\mu_i^h)^{-\ell} (S_N^h)^\ell (I - P_h) q_i^{(0)}, \end{aligned}$$

Since $(I-P_h)q_i^{(0)} = (I-P_h)^2 q_i^{(0)} = (I-P_h)(q_i^{(0)} - v_i)$, we conclude that

$$v_i - q_i^{(\ell)} = -(\mu_i^h)^{-\ell}(S_N^h)^\ell (I-P_h)(q_i^{(0)} - v_i). \tag{3.8}$$

Step 3: Since S_N^h commutes with P_h, equation (3.8) implies

$$\|v - q_i^{(\ell)}\|_{\mathcal{V}} \leqslant \frac{1}{|\mu_i^h|^\ell} \left\| \left[S_N^h (I-P_h)\right]^\ell \right\|_{\mathcal{V}} \|v - q_i^{(0)}\|_{\mathcal{V}}.$$

Let μ_*^h denote the supremum of $|\mu|$ over all μ in $\Sigma(S_N^h)\backslash\{\mu_1^h, \mu_2^h, \ldots, \mu_m^h\}$, i.e., μ_*^h is the spectral radius of $S_N^h(I-P_h)$, so

$$\mu_*^h = \lim_{\ell\to\infty} \|[S_N^h(I-P_h)]^\ell\|_{\mathcal{V}}^{1/\ell}.$$

Hence, for any given $\varepsilon > 0$, there is an $\ell_0 \geqslant 1$ such that $\|[S_N^h(I-P_h)]^\ell\|_{\mathcal{V}} \leqslant (\mu_h^* + \varepsilon)^\ell$ holds for all $\ell > \ell_0$ and consequently

$$\|v - q_i^{(\ell)}\|_{\mathcal{V}} \leqslant \frac{(\mu_*^h + \varepsilon)^\ell}{|\mu_i^h|^\ell} \|v - q_i^{(0)}\|_{\mathcal{V}}. \tag{3.9}$$

Step 4: As already seen, a consequence of Assumptions 2.1 and 2.7, is that by making h sufficiently small, we ensure that the eigenvalues $\mu_1^h, \mu_2^h, \ldots, \mu_m^h$ of S_N^h are strictly separated in magnitude from the remaining eigenvalues – cf. (3.2). Hence we may choose an $\varepsilon > 0$ so small that $\delta_i = (\mu_*^h + \varepsilon)/|\mu_i^h| < 1$. Then, with $\alpha_i = \|v_i - q_i^{(0)}\|_{\mathcal{V}}$, the estimate (3.9) implies

$$\|v_i - q_i^{(\ell)}\|_{\mathcal{V}} \leqslant \alpha_i \delta_i^\ell, \qquad \ell > \ell_0. \tag{3.10}$$

Note that v_i, $i = 1, \ldots, m$ form a basis for E_h. Hence, we may expand an arbitrary $v_h \in U_{E_h}^{\mathcal{V}}$ in this basis and construct an approximation of v_h using the same coefficients:

$$v_h = \sum_{i=1}^m c_i v_i, \qquad q_\ell = \sum_{i=1}^m c_i q_i^{(\ell)}.$$

Then, by (3.10),

$$\operatorname{dist}_{\mathcal{V}}(v_h, E_h^{(\ell)}) \leqslant \|v_h - q_\ell\|_{\mathcal{V}} \leqslant \sum_{i=1}^m |c_i \alpha_i| \delta_i^\ell \leqslant \alpha \left(\sum_{i=1}^m |c_i|^2\right)^{1/2} \left(\sum_{i=1}^m \delta_i^{2\ell}\right)^{1/2}. \tag{3.11}$$

where $\alpha = \max_i \alpha_i$.

Step 5: Denote one of the two suprema in the definition of $\operatorname{gap}_{\mathcal{V}}(E_h, E_h^{(\ell)})$ by

$$\delta_{h,\ell} = \sup_{v_h \in U_{E_h}^{\mathcal{V}}} \operatorname{dist}_{\mathcal{V}}\left(v_h, E_h^{(\ell)}\right).$$

Let g denotes the minimal eigenvalue of the $m \times m$ Gram matrix of the v_i-basis (whose (i,j)th entry is $(v_i, v_j)_{\mathcal{V}}$). Then $g\sum_{i=1}^m |c_i|^2 \leqslant \|v_h\|_{\mathcal{V}}^2 = 1$. (Note that g may depend on h, but is independent of ℓ.) Hence (3.11) implies $\delta_{h,\ell}^2 \leqslant (\alpha^2/g)\sum_{i=1}^m \delta_i^{2\ell}$ which converges to 0 as $\ell \to \infty$ since $\delta_i < 1$.

In particular, for large enough ℓ, we have $\delta_{h,\ell} < 1$. Hence, by [18, Theorem I.6.34] there is a subspace $\tilde{E}_h^{(\ell)} \subseteq E_h^{(\ell)}$ such that $\operatorname{gap}_{\mathcal{V}}(E_h, \tilde{E}_h^{(\ell)}) = \delta_{h,\ell} < 1$. Hence, $\dim(E_h) = \dim(\tilde{E}_h^{(\ell)}) = m$. But by (3.7), the only subspace $\tilde{E}_h^{(\ell)} \subseteq E_h^{(\ell)}$ of dimension m is $\tilde{E}_h^{(\ell)} = E_h^{(\ell)}$. Thus, for sufficiently large ℓ,

$$\operatorname{gap}_{\mathcal{V}}(E_h, E_h^{(\ell)}) = \delta_{h,\ell},$$

and the proof is complete since $\delta_{h,\ell} \to 0$ as $\ell \to \infty$. □

To summarize this section, we have defined a space E_h (in Definition 3.2) using S_N^h, but independently of the filtered subspace iteration (2.14), and have shown (in Theorem 3.4) that under certain conditions the iteration converges to it. The convergence of FEAST iterations for matrices (disregarding any discretization errors) were previously studied in [14] when $\mathcal{H} = \mathbb{R}^n$ and $\|\cdot\|_{\mathcal{V}} = \|\cdot\|_{\mathcal{H}}$ using the theory of subspace iterations [23]. In fact the identity obtained in Step 2 of the above proof was motivated by a standard argument in the analysis of subspace iterations [23]. Our proof of Theorem 3.4 gives a rigorous justification for the intuition that if the discretization is good, then despite the errors in the resolvent approximations, filtered subspace iteration should converge for well-separated eigenvalue clusters.

4. Discretization errors in eigenspace

In this section we study how the discrete eigenspace E_h approaches the exact eigenspace E as the discretization parameter h goes to 0.

Theorem 4.1. *Suppose Assumptions 2.1–2.7 hold. Then there is a $C_N > 0$ and an $h_0 > 0$ such that for all $h < h_0$,*

$$\operatorname{gap}_{\mathcal{V}}(E, E_h) \leqslant C_N W \max_{k=1,\dots,N} \left\| \left[R(z_k) - R_h(z_k) \right]\big|_E \right\|_{\mathcal{V}}, \tag{4.1}$$

so, in particular,

$$\lim_{h\to 0} \operatorname{gap}_{\mathcal{V}}(E, E_h) = 0.$$

Proof. Consider one of the two suprema in the definition of $\operatorname{gap}_{\mathcal{V}}(E_N, E_h)$, namely

$$\delta_h = \sup_{e \in U^{\mathcal{V}}_{E_N}} \operatorname{dist}_{\mathcal{V}}(e, E_h). \tag{4.2}$$

Then,

$$\delta_h \leqslant \sup_{e \in U^{\mathcal{V}}_{E_N}} \|e - P_h e\|_{\mathcal{V}} \leqslant \sup_{e \in U^{\mathcal{V}}_{E_N}} \|(P_N - P_h)e\|_{\mathcal{V}}. \tag{4.3}$$

Note that

$$\begin{aligned} P_N - P_h &= \frac{1}{2\pi \mathrm{i}} \oint_\Theta \left[(z - S_N)^{-1} - (z - S_N^h)^{-1} \right] dz \\ &= \frac{1}{2\pi \mathrm{i}} \oint_\Theta (z - S_N^h)^{-1} (S_N - S_N^h)(z - S_N)^{-1}\, dz. \end{aligned}$$

Since E_N is an invariant subspace of $(z - S_N)^{-1}$, the above identity gives the estimate

$$\|P_N e - P_h e\|_{\mathcal{V}} \leqslant \left[\frac{1}{2\pi} \oint_\Theta \|(z - S_N)^{-1}\|_{\mathcal{V}} \|(z - S_N^h)^{-1}\|_{\mathcal{V}}\, dz \right] \|(S_N - S_N^h)|_{E_N}\|_{\mathcal{V}} \|e\|_{\mathcal{V}}.$$

Returning to (4.3), we conclude that $\delta_h \leqslant C_N \|(S_N - S_N^h)|_{E_N}\|_{\mathcal{V}}$, where C_N is a bound for the quantity in square brackets above. Clearly, C_N can be bounded independently of h, since $\|(z - S_N^h)^{-1}\|_{\mathcal{V}} \to \|(z - S_N)^{-1}\|_{\mathcal{V}}$.

Thus, by virtue of (3.3), $\delta_h \to 0$ as $h \to 0$. In particular, for sufficiently small h, we have $\delta_h < 1$. Then, by [18, Theorem I.6.34], there is a closed subspace $\tilde{E}_h \subseteq E_h$

such that $\operatorname{gap}_{\mathcal{V}}(E_N, \tilde{E}_h) = \delta_h < 1$ and $\dim \tilde{E}_h = \dim E_N = m$. Because of (3.4), this implies that $\tilde{E}_h = E_h$. Since $E_N = E$ by Lemma 3.1, we finish the proof of (4.1) by noting that $\operatorname{gap}_{\mathcal{V}}(E, E_h) = \operatorname{gap}_{\mathcal{V}}(E_N, \tilde{E}_h) = \delta_h$. □

Remark 4.2. If $\mathcal{V}$ is replaced by $\mathcal{H}$ in (3.1), we obtain $\operatorname{gap}_{\mathcal{H}}(M, L)$, so

$$\operatorname{gap}_{\mathcal{H}}(E, E_h) = \max \left[\sup_{e \in U_E^{\mathcal{H}}} \operatorname{dist}_{\mathcal{H}}(e, E_h), \sup_{m \in U_{E_h}^{\mathcal{H}}} \operatorname{dist}_{\mathcal{H}}(m, E) \right].$$

Its natural to ask if $\operatorname{gap}_{\mathcal{V}}(E, E_h) \to 0$ implies $\operatorname{gap}_{\mathcal{H}}(E, E_h) \to 0$ as $h \to 0$. Let $\delta_h^{\mathcal{H}}$ denote the first of the two suprema above. Since E is finite-dimensional, there is a $C_m > 0$ such that $\|e\|_{\mathcal{H}} \geqslant C_m \|e\|_{\mathcal{V}}$ for all e in E. Using also Assumption 2.3,

$$\delta_h^{\mathcal{H}} = \sup_{0 \neq e \in E} \frac{\operatorname{dist}_{\mathcal{H}}(e, E_h)}{\|e\|_{\mathcal{H}}} \leqslant \frac{C_{\mathcal{V}}}{C_m} \sup_{0 \neq e \in E} \frac{\operatorname{dist}_{\mathcal{V}}(e, E_h)}{\|e\|_{\mathcal{V}}} \leqslant \frac{C_{\mathcal{V}}}{C_m} \operatorname{gap}_{\mathcal{V}}(E, E_h).$$

Thus, if $\operatorname{gap}_{\mathcal{V}}(E, E_h) \to 0$, taking h sufficiently small, $\dim(E_h) = \dim(E) = m$ and $\delta_h^{\mathcal{H}} < 1$, so using [18, Theorem I.6.34] as in the previous proof, we may conclude that $\operatorname{gap}_{\mathcal{H}}(E, E_h) = \delta_h^{\mathcal{H}}$. This implies that, under the same assumptions as in Theorem 4.1, there is an $h_1 > 0$ such that

$$\operatorname{gap}_{\mathcal{H}}(E, E_h) \leqslant \frac{C_{\mathcal{V}}}{C_m} \operatorname{gap}_{\mathcal{V}}(E, E_h) \tag{4.4}$$

for all $h < h_1$. Note C_m depends only on E and is independent of h.

5. Discretization errors in eigenvalues

In this section, we analyze the eigenvalue approximations that are generated as Ritz values (defined below) of eigenspace approximations obtained from the filtered subspace iteration. To define the Ritz values maintaining the same level of generality as we have so far, we need to consider the (possibly unbounded) sesquilinear form generated by A.

Recall that any selfadjoint operator A admits the polar decomposition $A = U_A|A| = |A|U_A$ (see [18, p. 335]), where U_A is selfadjoint and partially isometric, and $|A|$ is selfadjoint and positive semidefinite. As described in [26, §10.2], the polar decomposition can be used to define the following symmetric sesquilinear *form associated to the operator* A:

$$a(x, y) = (U_A|A|^{1/2}x, |A|^{1/2}y)_{\mathcal{H}} \tag{5.1}$$

for any x, y in $\operatorname{dom}(a) = \operatorname{dom}(|A|^{1/2})$. Let $|u|_a = |a(u, u)|^{1/2}$ for any $u \in \operatorname{dom}(a)$. By the properties of U_A,

$$|u|_a \leqslant (|A|^{1/2}u, |A|^{1/2}u)^{1/2} = \||A|^{1/2}u\|_{\mathcal{H}}, \qquad u \in \operatorname{dom}(a). \tag{5.2}$$

Let $F \subset \operatorname{dom}(a)$ be a closed finite-dimensional subspace of $\mathcal{H}$. We define $A_F : F \to F$ by the relation

$$(A_F x, y)_{\mathcal{H}} = a(x, y) \qquad \text{for all } x, y \in F. \tag{5.3}$$

The spectrum of the linear operator A_F on F, namely $\Sigma(A_F)$, is called the set of *Ritz values of* A *on* F. The operator A_E is defined by (5.3) with E in place of F. Note that the exact eigenspace we wish to approximate, namely E, is contained in $\operatorname{dom}(A) \subset \operatorname{dom}(a)$, and the exact eigenvalue cluster Λ that we wish to approximate is the set of Ritz values of A on E.

Central to the discussion of this section is how the Ritz values change when F is a perturbation of E. To formulate a result on sensitivity of Ritz values, we need more notation. Recall that S is the $\mathcal{H}$-orthogonal projection onto E. Let Q denote the $\mathcal{H}$-orthogonal projection onto F. Using S, we may express A_E as

$$A_E = S|A|^{1/2}U_A|A|^{1/2}S\big|_E$$

and A_F may be similarly expressed using Q. Let

$$|(S-I)Q|_{a,F} = \sup_{0\neq v\in F} \frac{|(S-I)Qv|_a}{\|v\|_{\mathcal{H}}} = \sup_{0\neq v\in F} \frac{|(S-I)v|_a}{\|v\|_{\mathcal{H}}}. \tag{5.4}$$

Note that there is a finite positive constant $\|A_E\|$ such that $\|A_E e\|_{\mathcal{H}} \leqslant \|A_E\|\|e\|_{\mathcal{H}}$ for all $e \in E$ (since $A_E : E \to E$ and E is finite-dimensional). Define the Hausdorff distance between two subsets $\Upsilon_1, \Upsilon_2 \subset \mathbb{R}$ by

$$\operatorname{dist}(\Upsilon_1, \Upsilon_2) = \max\left[\sup_{\mu_1\in\Upsilon_1} \operatorname{dist}(\mu_1, \Upsilon_2), \sup_{\mu_2\in\Upsilon_2} \operatorname{dist}(\mu_2, \Upsilon_1)\right]$$

where $\operatorname{dist}(\mu, \Upsilon) = \inf_{\nu\in\Upsilon} |\mu-\nu|$ for any $\Upsilon \subset \mathbb{R}$. The following lemma is a perturbation result that can be understood independently of the filtered subspace iteration. In particular, we have no need for Assumptions 2.1–2.7 in the lemma.

Lemma 5.1. *Suppose* $\operatorname{gap}_{\mathcal{H}}(E,F) < 1$. *Then there is a* $C_0 > 0$ *such that*

$$\operatorname{dist}(\Sigma(A_E), \Sigma(A_F)) \leqslant |(S-I)Q|_{a,F}^2 + C_0\|A_E\| \operatorname{gap}_{\mathcal{H}}(E,F)^2.$$

Proof. Step 1: Let $R = (S-Q)^2$ and let $\delta < 1$ be any number satisfying $\operatorname{gap}_{\mathcal{H}}(E,F) \leqslant \delta < 1$. Since $\|R\|_{\mathcal{H}} \leqslant \operatorname{gap}_{\mathcal{H}}(E,F)^2 \leqslant \delta^2 < 1$, the binomial series $\sum_{n=0}^{\infty}\binom{-1/2}{n}(-R)^n$ converges and defines $(I-R)^{-1/2}$. Subtracting the first term from this series, define $T = (I-R)^{-1/2} - I$. Since $(1-x)^{-1/2} - 1 = x[\sqrt{1-x} + (1-x)]^{-1}$, we obtain that

$$\|T\|_{\mathcal{H}} \leqslant \sum_{n=1}^{\infty}\binom{-1/2}{n}\|R\|_{\mathcal{H}}^n = (1-\|R\|_{\mathcal{H}})^{-1/2} - 1 = \|R\|_{\mathcal{H}}[\sqrt{1-\|R\|_{\mathcal{H}}} + (1-\|R\|_{\mathcal{H}})]^{-1},$$

which implies

$$\|T\|_{\mathcal{H}} \leqslant \|R\|_{\mathcal{H}}\left[\sqrt{1-\delta^2} + (1-\delta^2)\right]^{-1}. \tag{5.5}$$

We use R to define an isometry $J = (I-R)^{-1/2}[QS + (I-Q)(I-S)]$ on $\mathcal{H}$ (cf. [18, p. 33]) which maps E one-to-one onto F, and whose inverse is

$$J^{-1} = J^* = [SQ + (I-S)(I-Q)]\,(I-R)^{-1/2}. \tag{5.6}$$

Note that the spectra of A_E and the unitarily equivalent $JA_EJ^*|_F$ are identical.

Step 2: Let $D = JA_EJ^*|_F - A_F$, a selfadjoint operator on F. By [18, Theorem V.4.10],

$$\operatorname{dist}(\Lambda, \Lambda_h) \leqslant \|D|_F\|_{\mathcal{H}} = \sup_{0\neq f\in F} \frac{|(Df,f)_{\mathcal{H}}|}{(f,f)_{\mathcal{H}}}. \tag{5.7}$$

For $f \in F$, we have

$$\begin{aligned}(Df,f)_{\mathcal{H}} &= a(J^*f, J^*f) - a(f,f) = a(SJ^*f, SJ^*f) - a(Qf,Qf)\\ &= \operatorname{Re} a(SJ^*f + Qf, SJ^*f - Qf).\end{aligned}$$

Observing that (5.6) implies $SJ^* = SQ(I-R)^{-1/2}$, we split

$$(Df,f)_{\mathcal{H}} = \operatorname{Re} a((SJ^*+Q)f, SQ\left[(I-R)^{-1/2}-I\right]f) + \operatorname{Re} a((SJ^*+Q)f,(S-I)Qf).$$

Labelling the two terms on the right as t_1 and t_2, we proceed to estimate them.

Step 3: The first term $t_1 = \operatorname{Re} a((SJ^*+Q)f, SQTf) = \operatorname{Re} a((SJ^*+SQ)f, SQTf)$. Here we have used $S^2 = S$ and $a(Sx,y) = a(x,Sy)$ for all $x, y \in \operatorname{dom}(a)$. The latter follows from (5.1) because S commutes with A, so it commutes with $|A|$ and U [18, p.335ff], and moreover, it commutes with $|A|^{1/2}$ (see e.g. [18, Theorem V.3.35]). Continuing,

$$|t_1| = |\operatorname{Re} a((SJ^*+Q)f, SQTf)| = |\operatorname{Re}(A_E S(J^*+Q)f, SQTf)_{\mathcal{H}}|$$
$$\leqslant \|A_E\|\|S(J^*+Q)f\|_{\mathcal{H}}\|SQTf\|_{\mathcal{H}} \leqslant \frac{2\|A_E\|\|R\|_{\mathcal{H}}}{\sqrt{1-\delta^2}+(1-\delta^2)}\|f\|_{\mathcal{H}}^2$$

where we have used the fact that orthogonal projectors have unit norm as well as the isometry of J^*. Thus

$$|t_1| \leqslant \frac{2\|A_E\|}{\sqrt{1-\delta^2}+(1-\delta^2)}\operatorname{gap}_{\mathcal{H}}(E,F)^2\|f\|_{\mathcal{H}}^2.$$

Step 4: Next, we estimate t_2. Since $a((S-I)x,y) = a(x,(S-I)y)$ for all $x, y \in \operatorname{dom}(a)$,

$$\begin{aligned}|t_2| &= |\operatorname{Re} a((S-I)(SJ^*+Q)f,(S-I)Qf)| \\ &= |\operatorname{Re} a((S-I)Qf,(S-I)Qf)| \\ &\leqslant |(S-I)Q|_{a,h}^2\|f\|_{\mathcal{H}}^2.\end{aligned}$$

Adding the estimates for $|t_1|$ and $|t_2|$ and using it in (5.7), the proof is finished. □

Before applying this lemma to filtered subspace iteration, a few remarks are in order. *(i)* Its clear from the proof that the result of the lemma holds even when dimension of E (and F) is infinite, as long as $\|A_E\| < \infty$. Its also clear from the proof that the constant $C_0 = 2/(\sqrt{1-\delta^2}+1-\delta^2)$ is independent of the location of eigenvalue cluster Λ. *(ii)* The quantity $|(S-I)Q|_{a,F}^2$ is related to the square of the gap (like the other term in the bound of Lemma 5.1). Indeed, if $c_{a,F}$ is any constant that satisfies $|v|_a^2 \leqslant c_{a,F}\|v\|_{\mathcal{H}}^2$ for all $v \in E+F$, then

$$|(S-I)Q|_{a,F}^2 \leqslant c_{a,F}\|(S-I)Q\|_{\mathcal{H}}^2 \leqslant c_{a,F}\operatorname{gap}_{\mathcal{H}}(E,F)^2. \tag{5.8}$$

However, in applications, we usually need to make the dependence of $c_{a,F}$ more explicit (say, on discretization parameters). One technique for this is developed in the proof of Corollary 5.8 below. *(iii)* We highlight that Lemma 5.1 applies to *general unbounded* selfadjoint operators, even those whose spectra extends throughout the real line. *(iv)* Bounds for the Hausdorff distance between Ritz values under space perturbations have been previously studied for *bounded* operators [19, Theorem 5.3] and a part of the above proof above is inspired by their arguments. However we are not able to use their result directly because it holds only for Ritz values located at the extremes (top or bottom) of the spectrum of the bounded operator. Nonetheless, an approach to bring [19, Theorem 5.3] to bear on unbounded operators is to apply it to $R(\mu) = (\mu - A)^{-1}$, which is bounded (even if A is unbounded) provided μ is in the resolvent set of A. To quickly sketch this approach, one choses

a μ such that E becomes the eigenspace of $R(\mu)$ corresponding to the top of its spectrum, then apply [19, Theorem 5.3] to obtain an estimate that bounds the distance between Ritz values of $R(\mu)$, from which one then concludes estimates on the distance between Ritz values of A on E and on F. This technique would yield bounds involving $\mathrm{gap}_{\mathcal{H}}(E,F)^2$ like that of Lemma 5.1 but with other μ-dependent constants. *(v)* For finite-dimensional E, F, perturbations in eigenvalues of bounded operators corresponding to the top or bottom of the spectrum have also been studied in [20, Theorem 2.7] using majorization techniques. Their estimates can also be used to study spectral perturbations of unbounded operators by the technique mentioned in item *(iv)* above. In cases where one can bound specific angles between E and F, then [20, Theorem 2.7] may provide bounds for individual eigenvalue errors that are sharper than what can be concluded from bounds of the Hausdorff distance.

We now turn to the issue of approximating the eigenvalue cluster Λ using the subspaces of $\mathcal{V}_h$ generated by the filtered subspace iteration using S_N^h. Our analysis of this approximation is under the next assumption. Example 5.4 below illustrates the reason to consider forms and place this assumption.

Assumption 5.2. Assume that $\mathcal{V}_h$ is contained in $\mathrm{dom}(a)$.

Example 5.3 (Positive operators)**.** Consider the operator A and the form a in Example 2.5. Here, since A is positive, the factors of the polar decomposition of A are $U_A = I$ and $|A| = A$. Thus $\mathrm{dom}(a) = \mathrm{dom}(|A|^{1/2}) = \mathrm{dom}(A^{1/2})$. Moreover, $\mathcal{V} = \mathrm{dom}(A^{1/2})$ in Example 2.5. Since $\mathcal{V}_h \subset \mathcal{V}$ by definition, we conclude that Assumption 5.2 holds. □

Example 5.4 (A differential operator)**.** To give an example of a partial differential operator fitting the scenario of Example 5.3, suppose Ω is an open subset of $\mathbb{R}^d$, $\beta : \Omega \to \mathbb{R}$ is a bounded positive function, and $\alpha : \Omega \to \mathbb{C}^{d\times d}$ is a bounded Hermitian positive definite matrix function. Suppose the smallest eigenvalue of $\alpha(x)$ and $\beta(x)$ are greater than some $\delta > 0$ for a.e. $x \in \Omega$. Put $\mathcal{H} = L^2(\Omega)$ and set a by

$$a(u,v) = \int_\Omega \alpha \,\mathrm{grad}\, u \cdot \mathrm{grad}\, \overline{v}\; dx + \int_\Omega \beta u\overline{v}\; dx \tag{5.9}$$

for all u, v in $\mathrm{dom}(a) = H^1(\Omega)$. This is a densely defined closed form. Set A to be the closed selfadjoint *operator associated to the form* a, obtained by a representation theorem [26, Theorem 10.7].

When α and β equal the identity and Ω has Lipschitz boundary, the operator A is a Neumann operator whose domain satisfies $\mathrm{dom}(A) \subseteq H^{3/2}(\Omega)$ by a result of [17]. Thus $\mathrm{dom}(A)$ is strictly smaller than $\mathrm{dom}(a) = \mathrm{dom}(A^{1/2}) = H^1(\Omega)$ in this case. Therefore, if $\mathcal{V}_h$ is set to the Lagrange finite element subspace of $H^1(\Omega)$, then Assumption 5.2 holds. Note that it is easier to build finite element subspaces of $H^1(\Omega)$ than $H^{3/2}(\Omega)$, which is why we did *not* require $\mathcal{V}_h$ to be contained in $\mathrm{dom}(A)$ in Assumption 5.2. □

Example 5.5 (Semibounded operators)**.** Suppose A is *lower semibounded*, i.e., there is a $\mu \in \mathbb{R}$ such that $(Ax,x)_{\mathcal{H}} \geqslant \mu\,(x,x)_{\mathcal{H}}$ for all x in $\mathrm{dom}(A)$. Then, by [26, Proposition 10.5],

$$\mathrm{dom}(|A|^{1/2}) = \mathrm{dom}((A-\mu)^{1/2}). \tag{5.10}$$

An example of such an operator is the operator associated to the form a in (5.9) when β no longer satisfies $\beta > 0$, but instead changes sign while remaining bounded on Ω. Then fixing some $\mu < -\|\beta\|_{L^\infty(\Omega)}$, we note that the operator $A-\mu$ is positive and is the operator associated with the positive form $a_\mu(u,v) = a(u,v) - \mu(u,v)_{\mathcal{H}}$. Thus, by Example 5.3, $\mathrm{dom}(a_\mu) = \mathrm{dom}((A-\mu)^{1/2}) = H^1(\Omega)$. Hence by (5.10) we conclude that $\mathrm{dom}(a) = H^1(\Omega)$. □

Remark 5.6. Above we have encountered two related, but distinct concepts, of the *form associated to an unbounded operator* (via the polar decomposition as in (5.1)) and the *operator associated to a unbounded form* (by the first representation theorem [18, TheoremVI.2.1]). If one begins with a form a and then considers the operator A associated to it, we can define another form $\tilde{a}$ that is the form associated to A. The form $\tilde{a}$ need not equal a for a general selfadjoint operator as shown in [12, Example 2.11]). However, a and $\tilde{a}$ are equal if a is a densely defined lower semibounded closed form by [26, Theorem 10.7].

With the above background in mind, we now return to the analysis of eigenvalue approximations. Recall $E_h \subset \mathcal{V}_h \subset \mathrm{dom}(|A|^{1/2})$, the space we studied in Section 3). Using E_h, we compute the spectrum of the finite-dimensional operator A_{E_h},

$$\Lambda_h = \Sigma(A_{E_h}).$$

This set forms our approximation to Λ. In practice, the elements of Λ_h are computed by solving a small dense generalized eigenproblem arising from an equivalent equation of forms: find $\lambda_h \in \mathbb{R}$ and $0 \neq u_h \in E_h$ satisfying

$$a(u_h, v_h) = \lambda_h (u_h, v_h)_{\mathcal{H}}$$

for all $v_h \in E_h$. The collection of all such λ_h forms Λ_h. In the next theorem, we use Q_h to denote the $\mathcal{H}$-orthogonal projection onto E_h.

Theorem 5.7. *Suppose Assumptions 2.1–5.2 hold. Then there are positive constants C_0 and h_0 such that for all $h < h_0$,*

$$\mathrm{dist}(\Lambda, \Lambda_h) \leqslant |(S-I)Q_h|^2_{a,E_h} + C_0 \|A_E\| \, \mathrm{gap}_{\mathcal{H}}(E, E_h)^2.$$

Proof. By Theorem 4.1 and (4.4) we may choose h so small that $\mathrm{gap}_{\mathcal{H}}(E, E_h) \leqslant \delta < 1$. Hence, applying Lemma 5.1, the result follows. □

Corollary 5.8. *In addition to Assumptions 2.1–5.2, suppose $\|u\|_{\mathcal{V}} = \||A|^{1/2}u\|_{\mathcal{H}}$. Then there are positive constants C_0 and h_0 such that for all $h < h_0$, we have $\mathrm{gap}_{\mathcal{V}}(E, E_h) < 1$ and*

$$\mathrm{dist}(\Lambda, \Lambda_h) \leqslant (\Lambda_h^{\max})^2 \, \mathrm{gap}_{\mathcal{V}}(E, E_h)^2 + C_0 \|A_E\| \, \mathrm{gap}_{\mathcal{H}}(E, E_h)^2,$$

where $\Lambda_h^{\max} = \sup_{e_h \in E_h} \||A|^{1/2} e_h\|_{\mathcal{H}} / \|e_h\|_{\mathcal{H}}$ satisfies

$$(\Lambda_h^{\max})^2 \leqslant \left(\frac{1}{1 - \mathrm{gap}_{\mathcal{V}}(E, E_h)} \right)^2 \|A_E\|.$$

Proof. The first inequality will follow from Theorem 5.7 by establishing that

$$|(S-I)Q|_{a,E_h} \leqslant \Lambda_h^{\max} \, \mathrm{gap}_{\mathcal{V}}(E, E_h).$$

Since S is a $\mathcal{H}$-orthogonal projection, it is selfadjoint in $\mathcal{H}$-inner product. Moreover, since S commutes with A, it commutes with $|A|$ and hence with $|A|^{1/2}$. Therefore,

$$(Su, v)_{\mathcal{V}} = (|A|^{1/2} Su, |A|^{1/2} v)_{\mathcal{H}} = (S|A|^{1/2} u, |A|^{1/2} v)_{\mathcal{H}} = (|A|^{1/2} u, S|A|^{1/2} v)_{\mathcal{H}} = (u, Sv)_{\mathcal{V}},$$

for all $u, v \in \mathcal{V}$, i.e., S is selfadjoint in the $\mathcal{V}$-inner product too. This implies that S is also the $\mathcal{V}$-orthogonal projector onto E. Hence, using (5.2),

$$|Se_h - e_h|_a \leqslant \|Se_h - e_h\|_{\mathcal{V}} = \operatorname{dist}_{\mathcal{V}}(e_h, E) \tag{5.11}$$

for any $e_h \in E_h$. Combining (5.4) with (5.11), we conclude that

$$\begin{aligned} |(S-I)Q|_{a,E_h} &\leqslant \left(\sup_{0\neq e_h\in E_h} \frac{\|e_h\|_{\mathcal{V}}}{\|e_h\|_{\mathcal{H}}}\right)\left(\sup_{0\neq e_h\in E_h} \frac{|(S-I)e_h|_a}{\|e_h\|_{\mathcal{V}}}\right) \\ &\leqslant \Lambda_h^{\max}\left(\sup_{0\neq e_h\in E_h} \frac{\operatorname{dist}_{\mathcal{V}}(e_h, E)}{\|e_h\|_{\mathcal{V}}}\right) \end{aligned}$$

The first inequality of the corollary now follows from (3.1).

Let $g = \operatorname{gap}_{\mathcal{V}}(E, E_h)$ and $e_h \in E_h$. Then (5.11) implies

$$\|Se_h\|_{\mathcal{V}} \geqslant \|e_h\|_{\mathcal{V}} - \|e_h - Se_h\|_{\mathcal{V}} = \|e_h\|_{\mathcal{V}}\left(1 - \frac{\operatorname{dist}_{\mathcal{V}}(e_h, E)}{\|e_h\|_{\mathcal{V}}}\right) \geqslant \|e_h\|_{\mathcal{V}}\,(1-g)\,. \tag{5.12}$$

Therefore,

$$\frac{\|e_h\|_{\mathcal{V}}}{\|e_h\|_{\mathcal{H}}} = \frac{\|e_h\|_{\mathcal{V}}}{\|Se_h\|_{\mathcal{V}}}\,\frac{\|Se_h\|_{\mathcal{V}}}{\|e_h\|_{\mathcal{H}}} \leqslant \frac{1}{1-g}\,\frac{\|Se_h\|_{\mathcal{V}}}{\|e_h\|_{\mathcal{H}}}.$$

Since $\|Se_h\|_{\mathcal{V}}^2 = |(|A|Se_h, Se_h)_{\mathcal{H}}| = |(U_A A Se_h, Se_h)_{\mathcal{H}}| \leqslant \|A_E\|\|Se_h\|_{\mathcal{H}}^2 \leqslant \|A_E\|\|e_h\|_{\mathcal{H}}^2$,

$$\Lambda_h^{\max} = \sup_{0\neq e_h\in E_h} \frac{\|e_h\|_{\mathcal{V}}}{\|e_h\|_{\mathcal{H}}} \leqslant \frac{1}{1-g}\sup_{0\neq e_h\in E_h} \frac{\|Se_h\|_{\mathcal{V}}}{\|e_h\|_{\mathcal{H}}} \leqslant \frac{1}{1-g}\|A_E\|^{1/2}.$$

This completes the proof. □

Note that the second inequality of Corollary 5.8 allows one to bound $\Lambda_h^{\max}$ independently of h when $\operatorname{gap}_{\mathcal{V}}(E, E_h) \to 0$. A class of examples where Corollary 5.8 immediately applies is given by the positive forms of Example 5.3. For such operators, we have $|A|^{1/2} = A^{1/2}$, so $(u, v)_{\mathcal{V}} = a(u, v) = (A^{1/2}u, A^{1/2}v)$ holds for all $u, v \in \operatorname{dom}(a)$ and Corollary 5.8 applies. As a final remark, we also note that in the case in which $\mathcal{V}$ is normed with a norm equivalent to $\||A|^{1/2}\cdot\|_{\mathcal{H}}$, the above argument provides the estimate

$$\operatorname{dist}(\Lambda, \Lambda_h) \leqslant (C_1\,\Lambda_h^{\max})^2\operatorname{gap}_{\mathcal{V}}(E, E_h)^2 + C_0\|A_E\|\operatorname{gap}_{\mathcal{H}}(E, E_h)^2. \tag{5.13}$$

where C_1 depends on the equivalence constants for norms $\|\cdot\|_{\mathcal{V}}$ and $\||A|^{1/2}\cdot\|_{\mathcal{H}}$.

6. Application to the discretization of a model operator

The purpose of this section is to provide an example for application and illustration of the theoretical framework developed in the previous sections. A simple model problem is obtained by setting

$$\mathcal{H} = L^2(\Omega), \quad A = -\Delta, \quad \operatorname{dom}(A) = \{\psi \in H_0^1(\Omega) :\ \Delta\psi \in L^2(\Omega)\}, \quad \mathcal{V} = H_0^1(\Omega),$$

where $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ is a bounded polyhedral domain with Lipschitz boundary. Note that here $\|u\|_{\mathcal{V}}$ is set to $\|A^{1/2}u\|_{\mathcal{H}} = \|\operatorname{grad} u\|_{L^2(\Omega)} = |u|_{H^1(\Omega)}$, which is equivalent to $H^1(\Omega)$-norm due to the boundary condition. Throughout, we use standard notations for norms $(\|\cdot\|_X)$ and seminorms $(|\cdot|_X)$ on Sobolev spaces (X). The above set operator A is the operator associated to the form

$$a(u, v) = \int_\Omega \operatorname{grad} u \cdot \operatorname{grad}\overline{v}\; dx, \quad u, v \in \operatorname{dom}(a) = \mathcal{V} = H_0^1(\Omega).$$

Hence this model problem fits into Example 2.5 and Assumption 2.3 holds.

To calculate the application of the resolvent $u = R(z)v$, we need to solve the operator equation $(z - A)u = v$ for any z in the resolvent set of A. Using the form $b_z(u, v) = z(u, v)_{\mathcal{H}} - a(u, v)$, this equation may be stated in weak form as the problem of finding $R(z)v \in H^1_0(\Omega) \subset \mathcal{H}$ satisfying

$$b_z(R(z)v, w) = (v, w)_{\mathcal{H}} \qquad \text{for all } w \in H^1_0(\Omega). \tag{6.1}$$

We provide a bound on the biliear form b_z that will be used in subsequent analysis.

Lemma 6.1. *It holds that*

$$|b_z(u, v)| \leqslant \alpha(z)|u|_{H^1_0(\Omega)}|v|_{H^1_0(\Omega)} \qquad \textit{for all } u, v \in H^1_0(\Omega)\ ,$$

where $\alpha(z) = \|zA^{-1} - I\|_{\mathcal{H}} = \sup\{|\lambda - z|/|\lambda| : \lambda \in \Sigma(A)\}$.

Proof. Recognizing that $b_z(u, v) = ((zA^{-1} - I)A^{1/2}u, A^{1/2}v)_{\mathcal{H}}$, we see that

$$|b_z(u, v)| \leqslant \|zA^{-1} - I\|_{\mathcal{H}}\|A^{1/2}u\|_{\mathcal{H}}\|A^{1/2}v\|_{\mathcal{H}} = \alpha(z)|u|_{H^1_0(\Omega)}|v|_{H^1_0(\Omega)}\ .$$

Since $zA^{-1} - I$ is normal, its $\mathcal{H}$-norm, $\alpha(z)$, is equal to its spectral radius. Therefore, $\alpha(z) = \sup\{|\lambda - z|/|\lambda| : \lambda \in \Sigma(A)\}$, as claimed. □

The quantity

$$\beta(z) = \|AR(z)\|_{\mathcal{H}} = \sup\{|\lambda|/|\lambda - z| : \lambda \in \Sigma(A)\} \tag{6.2}$$

also figures in our analysis. The last equality of the $\mathcal{H}$-norm and the spectral radius again follows from the normality of $AR(z)$. It can also be found in [18, p. 273, Equation (3.17)]. Since $\Sigma(A)$ is real, we see that $\alpha(\bar{z}) = \alpha(z)$ and $\beta(\bar{z}) = \beta(z)$.

Next, suppose Ω is partitioned by a conforming simplicial finite element mesh Ω_h, where h equals the maximum of the diameters of all mesh elements. We shall assume that Ω_h is shape regular and quasiuniform (see e.g., [6] for standard finite element definitions). Set $\mathcal{V}_h$ equal to the Lagrange finite element subspace of $H^1_0(\Omega)$ consisting of continuous functions, which when restricted to any mesh element K in Ω_h, is in $P_p(K)$ for some $p \geqslant 1$. Here and throughout $P_\ell(K)$ denote the set of polynomials of total degree at most ℓ restricted to K. It is well known [6] that there is a $C_{\text{app}} > 0$ independent of h such that

$$\inf_{\nu_h \in \mathcal{V}_h} |\nu - \nu_h|_{H^1(\Omega)} \leqslant C_{\text{app}} h^r |\nu|_{H^{1+r}(\Omega)} \tag{6.3}$$

for any $0 \leqslant r \leqslant p$ and any $\nu \in H^{1+r}(\Omega)$.

Consider any $v \in \mathcal{H} = L^2(\Omega)$. The approximation of the resolvent by the finite element method, namely $R_h(z)v$, is a function in $\mathcal{V}_h$ satisfying

$$b_z(R_h(z)v, w) = (v, w) \qquad \text{for all } w \in \mathcal{V}_h. \tag{6.4}$$

It will follow from the ensuing analysis that (6.4) uniquely defines $R_h(z)v$ in $\mathcal{V}_h$ provided h is sufficiently small. The analysis is under the following regularity assumption.

Assumption 6.2. Suppose there are positive constants C_{reg} and s such that the solution $u^f \in \mathcal{V}$ of the Dirichlet problem $-\Delta u^f = f$ admits the regularity estimate

$$\|u^f\|_{H^{1+s}(\Omega)} \leqslant C_{\text{reg}}\|f\|_{\mathcal{H}} \quad \text{for any } f \in \mathcal{V}. \tag{6.5}$$

Suppose also that there is a number $s_E \geqslant s$ such that

$$\|u^f\|_{H^{1+s_E}(\Omega)} \leqslant C_{\text{reg}}\|f\|_{\mathcal{H}} \quad \text{for any } f \in E. \tag{6.6}$$

Standard regularity results for elliptic operators (see, e.g. [10, 11]) yield that $\mathrm{dom}(A) \supset H^{1+s}(\Omega)$ for some $s > 0$ depending on the geometry of Ω. For example, if Ω is convex, we may take $s = 1$ in (6.5); and if $\Omega \subset \mathbb{R}^2$ is non-convex, with its largest interior angle at a corner being π/α for some $1/2 < \alpha < 1$, we may take any positive $s < \alpha$. One can often show higher regularity when f is restricted to the eigenspace E, which is why we additionally assume (6.6). For example, if $\Omega = (0,1) \times (0,1)$, all eigenfunctions are analytic, having the form $\sin(m\pi x)\sin(n\pi y)$, for any positive integers m, n. These expressions, when viewed as functions on the L-shaped hexagon $\Omega_{\mathrm{L}} = (0,2)\times(0,2)\backslash[1,2]\times[1,2]$, also yield smooth eigenfunctions of Ω_{L}. But not all eigenfunctions of Ω_{L} are so regular.

The proof of the next result is modeled after a classical argument of [25], and employs the quantities $\alpha(z)$ and $\beta(z)$ introduced above.

Lemma 6.3. *Suppose Assumption 6.2 holds. Let z be in the resolvent set of A. Then there are positive constants C and h_0 (depending on z) such that for all $h < h_0$*

$$\|R(z) - R_h(z)\|_{\mathcal{V}} \leqslant Ch^r, \qquad \left\|\big[R(z) - R_h(z)\big]\big|_E\right\|_{\mathcal{V}} \leqslant Ch^{r_E}, \tag{6.7}$$

$$\|R(z) - R_h(z)\|_{\mathcal{H}} \leqslant Ch^{2r}, \qquad \left\|\big[R(z) - R_h(z)\big]\big|_E\right\|_{\mathcal{H}} \leqslant Ch^{r_E+r}, \tag{6.8}$$

where $r = \min(s,p)$ and $r_E = \min(s_E, p)$. We may choose $h_0 = [2\alpha(z)\beta(z)|z|C_{\mathrm{app}}C_{\mathrm{reg}}]^{-1/r}$ and $C = 2\alpha(z)\beta(z)C_{\mathrm{app}}C_{\mathrm{reg}}$.

Proof. Let $f \in \mathcal{H}$, $e = R(z)f - R_h(z)f$, and $w = R(\overline{z})e$. Then $-\Delta w = Aw = AR(\overline{z})e$. Hence it follows by Assumption 6.2 and (6.2) that

$$\|w\|_{H^{1+r}(\Omega)} \leqslant C_{\mathrm{reg}}\|AR(\overline{z})e\|_{\mathcal{H}} \leqslant C_{\mathrm{reg}}\beta(\overline{z})\|e\|_{\mathcal{H}} = C_{\mathrm{reg}}\beta(z)\|e\|_{\mathcal{H}}. \tag{6.9}$$

Subtracting (6.4) from (6.1), we have $b_z(e, w_h) = 0$ for any $w_h \in \mathcal{V}_h$. Note also that $b_z(v,w) = (v,e)_{\mathcal{H}}$ for any $v \in \mathcal{V}$. Choosing $v = e$ and applying Lemma 6.1,

$$\|e\|_{\mathcal{H}}^2 = (e,e)_{\mathcal{H}} = b_z(e,w) = b_z(e, w - w_h) \leqslant \alpha(z)|e|_{H^1(\Omega)}|w - w_h|_{H^1(\Omega)} .$$

Using (6.3) with $r = \min(s,p) > 0$, together with (6.9), we deduce that

$$\|e\|_{\mathcal{H}}^2 \leqslant \alpha(z)C_{\mathrm{app}}h^r|w|_{H^{1+r}(\Omega)}|e|_{H^1(\Omega)} \leqslant \alpha(z)\beta(z)C_{\mathrm{app}}C_{\mathrm{reg}}h^r\|e\|_{\mathcal{H}}|e|_{H^1(\Omega)},$$

i.e.,

$$\|e\|_{\mathcal{H}} \leqslant \alpha(z)\beta(z)C_{\mathrm{app}}C_{\mathrm{reg}}h^r\|e\|_{\mathcal{V}}. \tag{6.10}$$

Next, setting $u = R(z)f$, observe that for any $v_h \in \mathcal{V}_h$ we have

$$\Big|z\|e\|_{\mathcal{H}}^2 - \|e\|_{\mathcal{V}}^2\Big| = |b_z(e,e)| = |b_z(e, u - v_h)|.$$

Hence (6.10) and Lemma 6.1 imply

$$\big[1 - \alpha(z)\beta(z)C_{\mathrm{app}}C_{\mathrm{reg}}|z|h^r\big]\,\|e\|_{\mathcal{V}}^2 \leqslant \alpha(z)\|e\|_{\mathcal{V}} \inf_{v_h\in\mathcal{V}_h}\|u - v_h\|_{\mathcal{V}}.$$

When h is so small that $1 - \alpha(z)\beta(z)C_{\mathrm{app}}C_{\mathrm{reg}}|z|h^r > 1/2$, using (6.3), we obtain $\|e\|_{\mathcal{V}} \leqslant 2\alpha(z)C_{\mathrm{app}}h^r|u|_{H^{1+r}(\Omega)}$. Moreover, since $-\Delta u = Au = R(z)f$, using Assumption 6.2 and arguing as in (6.9), we have $\|u\|_{H^{1+r}(\Omega)} \leqslant C_{\mathrm{reg}}\beta(z)\|f\|_{\mathcal{H}}$. Thus,

$$\|R(z)f - R_h(z)f\|_{\mathcal{V}} \leqslant 2C_{\mathrm{app}}C_{\mathrm{reg}}\alpha(z)\beta(z)h^r\|f\|_{\mathcal{H}} \tag{6.11}$$

for any $f \in \mathcal{H}$. Now, if f is in $\mathcal{V}$, then since $\|f\|_{\mathcal{H}} \leqslant C\|f\|_{\mathcal{V}}$, the bound (6.11) proves the first inequality in (6.7). Combining with (6.10), we obtain the first inequality of (6.8) as well.

To prove the remaining inequalities, we let $f \in E$ and repeat the above argument leading to (6.11) with $r_E = \min(s_E, p)$ in place of r. This proves the second inequality of (6.7). Then using (6.10) (where we cannot, in general, replace r by r_E), the second inequality of (6.8) also follows. □

With the help of the lemma, we can now prove the main result of this section.

Theorem 6.4. *Suppose Assumption 2.1 (spectral separation) and Assumption 6.2 (elliptic regularity) holds, and that* $\dim E_h^{(0)} = \dim(P_h E_h^{(0)}) = \dim E$. *Then, there are positive constants* C *and* h_0 *such that for all* $h < h_0$, *the subspace iterates* $E_h^{(\ell)}$ *converge (in* $\operatorname{gap}_{\mathcal{V}}$*) to a space* E_h *satisfying*

$$\operatorname{gap}_{\mathcal{V}}(E, E_h) \leqslant C\, h^{r_E}, \tag{6.12}$$

$$\operatorname{gap}_{\mathcal{H}}(E, E_h) \leqslant C\, h^{r_E + r}, \tag{6.13}$$

$$\operatorname{dist}(\Lambda, \Lambda_h) \leqslant C\, h^{2r_E}, \tag{6.14}$$

where $r = \min(s, p)$ *and* $r_E = \min(s_E, p)$.

Proof. The proof proceeds by applying the previous theorems after verifying their assumptions. We have already verified Assumption 2.3 with above set $\mathcal{V} = H_0^1(\Omega)$. In view of (6.7) of Lemma 6.3, since $r > 0$, Assumption 2.7 holds with the same $\mathcal{V}$. We may now apply Theorem 3.4, which yields $\operatorname{gap}_{\mathcal{V}}(E_h^{(\ell)}, E_h) \to 0$. Now the proof of (6.12) reduces to an application of Theorem 4.1. Next observe that Assumptions 2.3 and 2.7 also hold when $\mathcal{V}$ is set to $\mathcal{H}$ (see Example 2.4 and (6.8) of Lemma 6.3). Applying Theorem 4.1 with this $\mathcal{V} = \mathcal{H}$ setting, we obtain the estimate (6.13). Finally, (6.14) follows by combining Corollary 5.8 with (6.12) and (6.13). □

7. Numerical Experiments

We illustrate the convergence results of Theorem 6.4 on the model problem

$$-\Delta e = \lambda e \text{ in } \Omega \quad , \quad e = 0 \text{ on } \partial\Omega \ ,$$

on three different domains $\Omega \subset \mathbb{R}^2$. More specifically, we consider eigenvalue errors and (6.14). The experiments were conducted using [9], which builds a hierarchy of Python classes representing standard Lagrange finite element approximations of the filter S_N based on the resolvent approximations $R_h(z)$, as described in Section 6. We do not write out the details of the algorithm because they can be found in our public code [9] or in many previous papers (see e.g., [23, Algorithm 1.1] and [14]). As in these references, we perform the implicit orthogonalization through a small Rayleigh-Ritz eigenproblem at each iteration. In general, it is not necessary to perform this orthogonalization at every step, but in the experiments reported below, we do so. For all experiments, filtered subspace iteration is applied using the Butterworth filter (2.9) with $N = 8$. The symmetry (about the real axis) of our filter weights and nodes are exploited so that only $N/2$ boundary value problems (rather than N) need to be solved for each right hand side per iteration. The subspace iterations are started with a random subspace of dimension at least as large as the dimension of E, and the algorithm truncates basis vectors that generate Ritz values that are deemed too far outside the search interval; in all cases, we choose this initial subspace dimension to be six. We stop the iterations when successive Ritz values differ by less than 10^{-9}. Though changing N does, in

some cases, change the number of subspace iterations used to achieve a prescribed error tolerance (e.g. three iterations for $N = 8$ versus two iterations for $N = 16$ for the Dumbbell problem with $p = 3$ and $h = 2^{-6}$), it had no effect on the discretization errors reported here, so we do not discuss this parameter further. The finite element discretizations are implemented using a Python interface into the C++ finite element library NGSolve [27]. Two parameters govern the discretization: h is the maximum edge length in a quasi-uniform triangulation of Ω, and p is the polynomial degree in each element.

7.1. **Unit Square.** For the unit square $\Omega = (0,1) \times (0,1)$, the eigenvalues and eigenvectors may be doubly-indexed by

$$\lambda_{m,n} = (m^2 + n^2)\pi^2 \quad , \quad e_{m,n} = \sin(m\pi x)\sin(n\pi y) \quad , \quad m, n \in \mathbb{N} \, .$$

For any subset of the spectrum, the corresponding eigenspaces are analytic ($s_E = \infty$), and the convexity of Ω ensures that $s = 1$. Therefore, Theorem 6.4 indicates that the eigenvalue convergence should behave like $\mathcal{O}(h^{2p})$. This is precisely what is observed in Figure 2 both at the low end of the spectrum, $\Lambda = \{2\pi^2, 5\pi^2\}$, and higher in the spectrum, $\Lambda = \{128\pi^2, 130\pi^2\}$. We note that both $2\pi^2$ and $128\pi^2$ are simple eigenvalues, $5\pi^2$ is a double eigenvalue, and $130\pi^2$ is a quadruple eigenvalue.

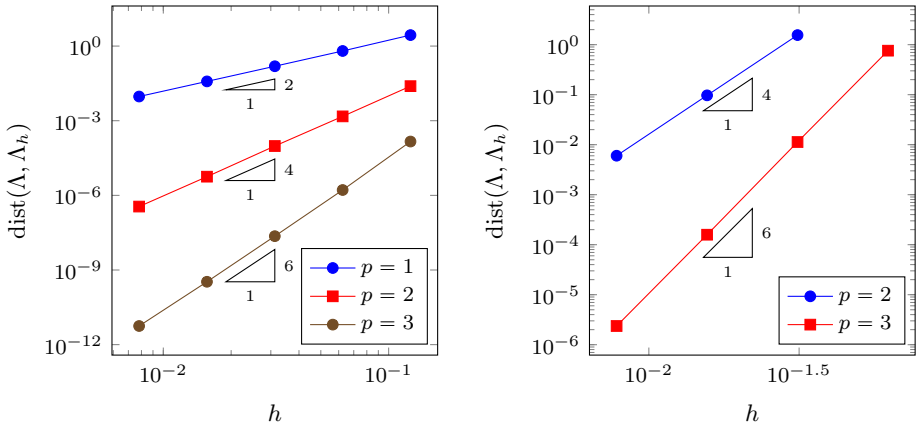
For the first set of experiments, the search interval $(0, 60)$ was chosen, so $y = \gamma = 30$. In Figure 2a, the eigenvalue error, $\mathrm{dist}(\Lambda, \Lambda_h)$, is given with respect to h for (fixed) $p = 1, 2, 3$ and decreasing $h = 2^{-3}, 2^{-4}, \ldots, 2^{-7}$. For the second set of experiments, the search interval was $(1260, 1290)$, so $y = 1275$ and $\gamma = 15$. In order to provide convergence graphs within the same plot for these more highly oscillatory eigenvectors, we use $h = 2^{-5}, 2^{-6}, 2^{-7}$ for $p = 2$, and $h = 2^{-4}, 2^{-5}, 2^{-6}, 2^{-7}$ for $p = 3$, see Figure 2b. For coarser spaces, the approximations of $130\pi^2$ were far enough outside the search interval to be rejected by the algorithm, and an approximation for only $128\pi^2$ was obtained. A plot of the computed basis for the five-dimensional eigenspace corresponding to $\Lambda = \{128\pi^2, 130\pi^2\}$ is given in Figure 3. If we label the computed eigenvalues $\lambda_1^h < \lambda_2^h < \ldots < \lambda_5^h$, with corresponding eigenvectors e_j^h, $1 \leqslant j \leqslant 5$, contour plots of these eigenvectors are given, from left to right, in Figure 3. One sees that $\mathrm{span}\{e_1^h\}$ approximates $\mathrm{span}\{e_{8,8}\}$, and it appears that $\mathrm{span}\{e_2^h, e_5^h\}$ approximates $\mathrm{span}\{e_{3,11}, e_{11,3}\}$, and that $\mathrm{span}\{e_3^h, e_4^h\}$ approximates $\mathrm{span}\{e_{7,9}, e_{9,7}\}$.

7.2. **L-Shape.** Let Ω be the L-shaped domain that is the concatenation of three unit squares; see Figure 4d. In [29], the authors provide very precise approximations of several eigenvalues for this domain (and other planar domains). Their approximations of the first three eigenvalues (accurate to eight digits) are

$$\lambda_1 \approx 9.6397238 \quad , \quad \lambda_2 \approx 15.197252 \quad , \quad \lambda_3 = 2\pi^2 \approx 19.739209 \, ,$$

and we take the first two of these approximations to be the "truth" for purposes of our convergence studies. We use the search interval $(0, 20)$.

These eigenvalues correspond to eigenvectors having very different regularities, and the convergence plots in Figure 4 illustrate that (6.14) can be pessimistic in the sense that it ascribes a single convergence rate to an entire eigenvalue cluster, and this convergence rate is dictated by the worst-case regularity of eigenvectors associated with the cluster. What we see in practice is that individual eigenvalues within a cluster converge at rates determined by the regularity of their corresponding eigenvectors. Since Ω has a re-entrant corner with interior angle $3\pi/2$, we have



(A) Convergence rates for $\Lambda = \{2\pi^2, 5\pi^2\}$. (B) Convergence rates for $\Lambda = \{128\pi^2, 130\pi^2\}$.

FIGURE 2. Square: Convergence rates for clusters located at the bottom and higher up in the spectrum.

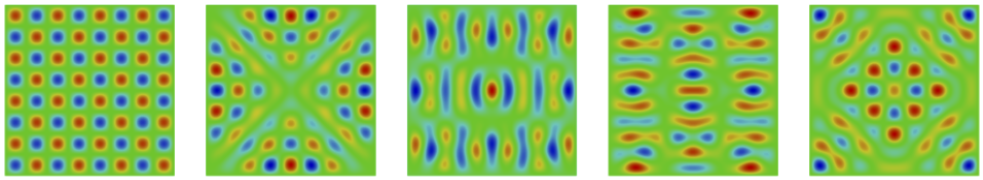
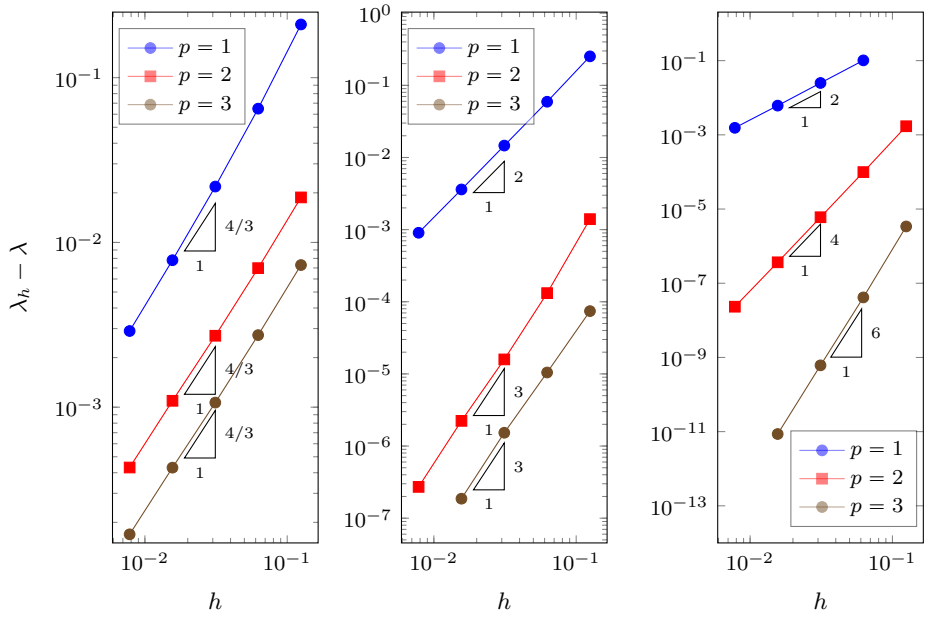


FIGURE 3. Square: Eigenvectors corresponding to the eigenvalues $128\pi^2$ (left) and $130\pi^2$.

$r = \min(s, p) = s$ for any $s < 2/3$, and the first eigenvector actually has this regularity. As such, Theorem 6.4 indicates essentially $\mathcal{O}(h^{4/3})$ convergence for the cluster. While this is true for the cluster as a whole, it is only the first eigenvalue that converges this slowly. The convergence order for the second eigenvalue $\mathcal{O}(h^{\min(2p,3)})$, is consistent with a regularity index $s \leqslant 3/2$; and the convergence order for the third eigenvalue, $\mathcal{O}(h^{2p})$, is precisely what is expected from an analytic eigenvector.

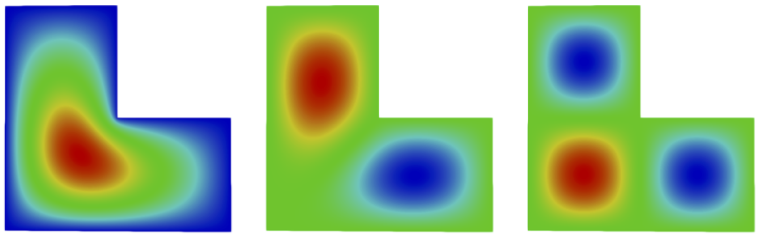
7.3. **Dumbbell.** Let Ω be the dumbbell-shaped domain that is a concatenation of two unit-squares joined by a $1/4 \times 1/4$ square "bridge"; see Figure 5. By tiling the dumbbell with $(1/8) \times (1/8)$ squares, we see that $\lambda = 128\pi^2 \approx 1263.309$ is an eigenvalue, with corresponding eigenvector $e = \sin(8\pi x)\sin(8\pi y)$. In order to determine whether there are other eigenvalues near $128\pi^2$, we choose the search interval $(1262, 1264)$. Because of the highly oscillatory nature of the eigenvector e, we employ relatively fine discretizations, taking $p = 3$ and $h = 2^{-4}, 2^{-5}, 2^{-6}, 2^{-7}$. We have determined that there is one other eigenvalue in the search interval, and it is approximately 1262.41. Labeling these eigenvalues $\lambda_1 \leqslant \lambda_2$, their computed approximations are given in Table 1. For the coarsest of these discretizations, the computed approximation of $128\pi^2$, 1264.02, lies slightly outside the search interval, but is accepted by the algorithm. Since $\lambda_2 = 128\pi^2$ is known, we underline the



(A) Convergence rates for $\lambda_1 \approx 9.6397238$.

(B) Convergence rates for $\lambda_2 \approx 15.197252$.

(C) Convergence rates for $\lambda_3 = 2\pi^2$.



(D) L-Shape: Computed approximations of the first three eigenvectors.

FIGURE 4. L-Shape: Convergence rates for λ_1, λ_2 and λ_3.

TABLE 1. Dumbbell: Computed eigenvalues for the interval $(1262, 1264)$, $p = 3$ and mesh parameters $h = 2^{-4}, 2^{-5}, 2^{-6}, 2^{-7}$.

h	λ_1	λ_2
2^{-4}	1263.178867	1264.020566
2^{-5}	1262.447629	1263.319956
2^{-6}	1262.418298	1263.309521
2^{-7}	1262.410062	1263.309366

number of correct digits in our approximations of it. The error in this approximation is consistent with $\mathcal{O}(h^6)$ eigenvalue error, in agreement with Theorem 6.4.

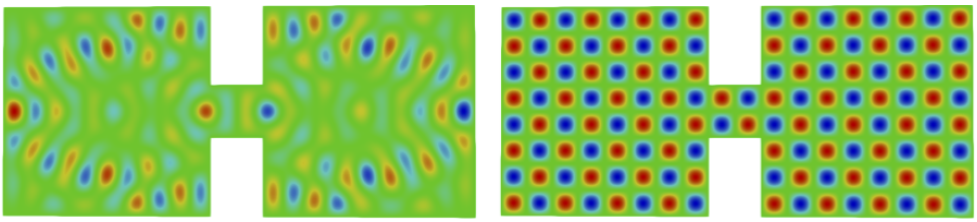


FIGURE 5. Dumbbell: Eigenvectors corresponding to the eigenvalues 1262.41 (left) and $128\pi^2$.

REFERENCES

[1] P. M. ANSELONE AND T. W. PALMER, *Spectral analysis of collectively compact, strongly convergent operator sequences*, Pacific J. Math., 25 (1968), pp. 423–431.

[2] A. P. AUSTIN AND L. N. TREFETHEN, *Computing eigenvalues of real symmetric matrices with rational filters in real arithmetic*, SIAM J. Sci. Comput., 37 (2015), pp. A1365–A1387.

[3] I. BABUŠKA AND J. OSBORN, *Eigenvalue problems*, in Handbook of numerical analysis, Vol. II, Handb. Numer. Anal., II, North-Holland, Amsterdam, 1991, pp. 641–787.

[4] W.-J. BEYN, *An integral method for solving nonlinear eigenvalue problems*, Linear Algebra Appl., 436 (2012), pp. 3839–3863.

[5] T. BÜHLER AND D. A. SALAMON, *Functional Analysis*, vol. 191 of Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, 2018.

[6] A. ERN AND J.-L. GUERMOND, *Theory and practice of finite elements*, vol. 159 of Applied Mathematical Sciences, Springer-Verlag, New York, 2004.

[7] J. GOPALAKRISHNAN, L. GRUBIŠIĆ, AND J. OVALL, *Iterative convergence of filtered subspace iteration for unbounded selfadjoint operators*, In Preparation, (2018).

[8] J. GOPALAKRISHNAN, L. GRUBIŠIĆ, J. OVALL, AND B. Q. PARKER, *Analysis of the FEAST iteration with DPG discretization*, In Preparation, (2018).

[9] J. GOPALAKRISHNAN AND B. Q. PARKER, *Pythonic FEAST*. Software hosted at Bitbucket: `https://bitbucket.org/jayggg/pyeigfeast`, 2017.

[10] P. GRISVARD, *Singularities in boundary value problems*, vol. 22 of Recherches en Mathématiques Appliquées [Research in Applied Mathematics], Masson, Paris; Springer-Verlag, Berlin, 1992.

[11] P. GRISVARD, *Elliptic problems in nonsmooth domains*, vol. 69 of Classics in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011. Reprint of the 1985 original.

[12] L. GRUBIŠIĆ, V. KOSTRYKIN, K. A. MAKAROV, AND K. VESELIĆ, *Representation theorems for indefinite quadratic forms revisited*, Mathematika, 59 (2013), pp. 169–189.

[13] S. GÜTTEL, *Rational krylov methods for operator functions*, Dissertation Thesis, Bergakademie Freiberg, (2010).

[14] S. GÜTTEL, E. POLIZZI, P. T. P. TANG, AND G. VIAUD, *Zolotarev quadrature rules and load balancing for the FEAST eigensolver*, SIAM J. Sci. Comput., 37 (2015), pp. A2100–A2122.

[15] R. HUANG, A. A. STRUTHERS, J. SUN, AND R. ZHANG, *Recursive integral method for transmission eigenvalues*, J. Comput. Phys., 327 (2016), pp. 830–840.

[16] A. IMAKURA, L. DU, AND T. SAKURAI, *Relationships among contour integral-based methods for solving generalized eigenvalue problems*, Jpn. J. Ind. Appl. Math., 33 (2016), pp. 721–750.

[17] D. S. JERISON AND C. E. KENIG, *The Neumann problem on Lipschitz domains*, Bull. Amer. Math. Soc. (N.S.), 4 (1981), pp. 203–207.

[18] T. KATO, *Perturbation theory for linear operators*, Classics in Mathematics, Springer-Verlag, Berlin, 1995. Reprint of the 1980 edition.

[19] A. KNYAZEV, A. JUJUNASHVILI, AND M. ARGENTATI, *Angles between infinite dimensional subspaces with applications to the Rayleigh-Ritz and alternating projectors methods*, J. Funct. Anal., 259 (2010), pp. 1323–1345.

[20] A. V. KNYAZEV AND M. E. ARGENTATI, *Rayleigh-Ritz majorization error bounds with applications to FEM*, SIAM J. Matrix Anal. Appl., 31 (2009), pp. 1521–1537.

[21] E. Polizzi, *Density-matrix-based algorithms for solving eigenvalue problems*, Phys. Rev. B., 79 (2009), p. 115112.
[22] M. Reed and B. Simon, *Methods of modern mathematical physics. I. Functional analysis.*, Academic Press [Harcourt Brace Jovanovich Publishers], New York, 1972.
[23] Y. Saad, *Analysis of subspace iteration for eigenvalue problems with evolving matrices*, SIAM J. Matrix Anal. Appl., 37 (2016), pp. 103–122.
[24] T. Sakurai and H. Sugiura, *A projection method for generalized eigenvalue problems using numerical integration*, in Proceedings of the 6th Japan-China Joint Seminar on Numerical Mathematics (Tsukuba, 2002), vol. 159:1, 2003, pp. 119–128.
[25] A. H. Schatz, *An observation concerning Ritz-Galerkin methods with indefinite bilinear forms*, Math. Comp., 28 (1974), pp. 959–962.
[26] K. Schmüdgen, *Unbounded self-adjoint operators on Hilbert space*, vol. 265 of Graduate Texts in Mathematics, Springer, Dordrecht, 2012.
[27] J. Schöberl, *NGSolve*. `http://ngsolve.org`, 2017.
[28] P. T. P. Tang and E. Polizzi, *FEAST as a subspace iteration eigensolver accelerated by approximate spectral projection*, SIAM J. Matrix Anal. Appl., 35 (2014), pp. 354–390.
[29] L. N. Trefethen and T. Betcke, *Computed eigenmodes of planar regions*, in Recent advances in differential equations and mathematical physics, vol. 412 of Contemp. Math., Amer. Math. Soc., Providence, RI, 2006, pp. 297–314.

Portland State University, PO Box 751, Portland, OR 97207-0751, USA
Email address: `gjay@pdx.edu`

University of Zagreb, Bijenička 30, 10000 Zagreb, Croatia
Email address: `luka.grubisic@math.hr`

Portland State University, PO Box 751, Portland, OR 97207-0751, USA
Email address: `jovall@pdx.edu`