9-18-2017

Filtered Subspace Iteration for Selfadjoint Operators

Jay Gopalakrishnan  
*Portland State University, gjay@pdx.edu*

Luka Grubišić  
*University of Zagreb*

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

Let us know how access to this document benefits you.

Follow this and additional works at: [https://pdxscholar.library.pdx.edu/pics_pub](https://pdxscholar.library.pdx.edu/pics_pub)

Part of the Numerical Analysis and Computation Commons

**Citation Details**

Gopalakrishnan, Jay; Grubišić, Luka; and Ovall, Jeffrey S., "Filtered Subspace Iteration for Selfadjoint Operators" (2017). *Portland Institute for Computational Science Publications*. 3.  
[https://pdxscholar.library.pdx.edu/pics_pub/3](https://pdxscholar.library.pdx.edu/pics_pub/3)

This Pre-Print is brought to you for free and open access. It has been accepted for inclusion in Portland Institute for Computational Science Publications by an authorized administrator of PDXScholar. For more information, please contact pdxscholar@pdx.edu.
FILTERED SUBSPACE ITERATION FOR SELFADJOINT OPERATORS

JAY GOPALAKRISHNAN, LUKA GRubišiĆ, AND JEFFREY OVALL

Abstract. We consider the problem of computing a cluster of eigenvalues (and its associated eigenspace) of a (possibly unbounded) selfadjoint operator in a Hilbert space. A rational function of the operator is constructed such that the eigenspace of interest is its dominant eigenspace, and a subspace iteration procedure is used to approximate this eigenspace. The computed space is then used to obtain approximations of the eigenvalues of interest. An eigenvalue and eigenspace convergence analysis that considers both iteration error and discretization error is provided. A realization of the proposed approach for a model second-order elliptic operator is based on a discontinuous Petrov-Galerkin discretization of the resolvent, and a variety of numerical experiments illustrate its performance.

1. Introduction

Let $A : \text{dom}(A) \subseteq \mathcal{H} \rightarrow \mathcal{H}$ be a closed, selfadjoint operator (not necessarily bounded) in a complex Hilbert space $\mathcal{H}$ (cf. [22]), whose (real) spectrum is denoted by $\Sigma(A)$. We are interested in computationally approximating a subset $\Lambda$ of the spectrum that consists of a finite collection of eigenvalues of finite multiplicity. To approximate $\Lambda$, we discretize and apply filtered subspace iteration, which can be roughly described as follows. First, a “filter” is selected that transforms the eigenspace of the cluster of eigenvalues of interest (namely $\Lambda$) to the dominant eigenspace of another, bounded operator. Next, subspace iteration is applied using the bounded operator. Starting with an initial subspace (usually chosen randomly), the bounded operator is repeatedly applied to it, generating a sequence of subspaces that typically converges to the eigenspace of $\Lambda$.

This eigenspace, namely the span of all the eigenvectors associated with elements of $\Lambda$, is denoted by $E$. Then $m = \dim E$, being the sum of multiplicities of each element of $\Lambda$, is finite, and we assume $m \geq 1$. Throughout this paper, the multiplicity $\ell$ of an eigenvalue $\lambda$ of an operator refers to its algebraic multiplicity, i.e., $\lambda$ is a pole of order $\ell$ of the resolvent of that operator; and we recall that, for a selfadjoint operator $A$, the algebraic multiplicity of $\lambda$ coincides with its geometric multiplicity, $\dim \ker (\lambda - A)$. In the next section, we shall see that one can transform $E$ into the dominant eigenspace of certain filtered operators (precisely defined later) provided $\Lambda$ is strictly separated from the remainder of the spectrum. To quantify the separation, we assume there are $y \in \mathbb{R}$, $\delta > 0$ and $\gamma > 0$ such that

$$\Lambda \subset \{ x \in \mathbb{R} : |x - y| < \gamma \}, \quad \Sigma(A) \setminus \Lambda \subset \{ x \in \mathbb{R} : |x - y| \geq (1 + \delta) \gamma \},$$

The number $\delta$ provides a measure of the relative gap between the eigenvalues of interest and the rest of the spectrum—relative to the radius $\gamma$ of the interval in which we are seeking eigenvalues. The assumption (1) holds tacitly throughout the paper.

Date: September 18, 2017.

This work was partially supported by AFOSR grant FA9550-17-1-0090, 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.
As seen in the next section, the construction of filters can be motivated by approximations of a Dunford-Taylor contour integral. Following [3], we identify two different classes of methods in the existing literature that use such contour integrals for computation of a targeted cluster of eigenvalues of matrices. One class of methods, that often goes by the name SSM [29] (see also [6, 20]), approximates Λ by the eigenvalues of a system of moment matrices based on contour integrals. The matrices are obtained by approximating the integrals by a quadrature, and the approximation error depends on the accuracy of the quadrature.

Another class of methods, more related to our present contribution in that they apply filtered subspace iterations to matrices, goes by the name of “FEAST” [25] (see also [18, 32]). They also use quadratures to approximate a contour integral, but the eigenvalues are approximated by a Rayleigh-Ritz procedure using the original matrix. In our view, the use of quadratures in such algorithms is essentially different from their use in SSM-like approaches. 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 essentially irrelevant in our error analysis.

All the above-mentioned previous works considered matrices on finite-dimensional spaces. Our main point of departure is in the consideration of (possibly unbounded) linear operators $A$ on Hilbert spaces, thus at once generalizing the matrix case, as well as the case of bounded linear operators. This generalization is motivated by our desire to approximate eigenvalues of differential operators that are closed unbounded operators in appropriate Lebesgue spaces. A classical example of such an operator is provided by the Schrödinger operator on an unbounded domain with a potential well, which gives rise to spectra with a few practically interesting eigenvalues separated from its continuous spectrum (see § 6.4).

Of course, in practice, one must work with finite-dimensional approximations obtained by discretizations. Nonetheless, to analyze how the errors in the finally computed eigenvalues depend on the discretization and filter parameters, it is useful to start from the infinite-dimensional case and study the perturbations made. The difficulty in analyzing discretization errors for the unbounded operator eigenproblem is that many of the existing standard tools [4] are not directly applicable.

When $A$ is a differential operator on an infinite-dimensional space, some approximations to bring the computations into finite-dimensional spaces are practically necessary. To quickly outline the approximation approach 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 $S_N^h$. Here $h$ is some discretization parameter inversely related to a computational finite-dimensional space (usually related to grid spacing). By repeated application of $S_N^h$, a subspace iteration produces a sequence of subspaces that potentially approximate $E$.

Does this iteration converge to some space $E_h$? We give a positive answer for certain filters when (1) holds. The convergence of subspaces is measured in the “gap” metric, which may be defined using a possibly stronger norm than the $H$-norm. The example of selfadjoint elliptic differential operators on $H = L^2(\Omega)$ illustrates the need to measure eigenfunction errors in a stronger norm like the $H^1(\Omega)$-norm. To our knowledge, this is first work to address the issue of convergence of FEAST iterations in stronger norms. Among the several other questions we address are these: Does $E_h$ approximate $E$? Do the Ritz values of $E_h$ approximate $\Lambda$? As we shall see, such questions can be satisfactorily answered for general selfadjoint operators by identifying certain abstract conditions. As a worked-out example, we verify these conditions
for a model operator and a discontinuous Petrov Galerkin (DPG) discretization. Our findings indicate that, while increasing \( N \) may affect the speed of convergence of the filtered subspace iteration (to \( E_h \)), it has little effect on the discretization error (the gap between \( E \) and \( E_h \)).

The rest of the paper is organized as follows. In Section 2, we analyze practical rational filters based on quadrature approximation of a contour integral by the trapezoid rule on circles or ellipses associated with an interval containing the eigenvalues of interest. Section 3 provides an analysis of an inexact subspace iteration, which allows for perturbation errors in the application of the filter that might naturally arise in practice, and permits the use of other norms that may be of interest. Discretization errors are considered in Section 4, where the gap between the dominant eigenspaces of the filter and its discrete realization is bounded using the errors in discretizing the resolvent at the quadrature points of the filter. Eigenvalue errors are then bounded using the square of this subspace gap. Analysis of a DPG discretization of the resolvent of a model operator in Section 5 provides an example of how abstract conditions on the resolvent might be verified in practice. The practical performance of the FEAST algorithm with the DPG discretization is reported in Section 6, where we also compute the bound states of a Schrödinger operator with a short range potential. An appendix offers remarks on a Zolotarev rational filter.

2. Rational Filters

Let \( A, \Lambda \) and \( E \) be as discussed previously. The goal of this section is to provide a few concrete examples of filters. 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 \( \Lambda \) and excludes \( \Sigma(A) \backslash \Lambda \), and let \( G \subset \mathbb{C} \) be the open set whose boundary is \( \Gamma \). By the Cauchy Integral Formula,

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

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

\[
 S = \frac{1}{2\pi i} \oint_{\Gamma} R(z) \, dz,
\]

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

\[
 E = \text{ran}(S).
\]

Furthermore, by functional calculus (see [26, Theorem VIII.5], [30, Theorem 5.9] or [7, Section 6.4]), if \( (\lambda, \phi) \in \Sigma(A) \times \text{dom}(A) \) 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), in the same spirit as [6,17,25,29,32], we approximate \( r(\xi) \) by

\[
 r_N(\xi) = \sum_{k=0}^{N-1} w_k (z_k - \xi)^{-1},
\]
for some \( w_k, z_k \in \mathbb{C} \). The corresponding rational filter is the operator

\[
S_N = r_N(A) = \sum_{k=0}^{N-1} w_k R(z_k),
\]

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 \((\lambda, \phi) \in \Sigma(A) \times \text{dom}(A)\) satisfies \( A \phi = \lambda \phi \), then \( S_N \phi = r_N(\lambda) \phi \). In particular, the set \( \Lambda \) of eigenvalues of interest have been mapped to \( \{ r_N(\lambda) : \lambda \in \Lambda \} \) by the filter.

Below, we provide an analysis of some practical filters \( r_N \) such that these mapped eigenvalues are dominant eigenvalues of \( S_N \) in the following sense:

\[
\min_{\lambda \in \Lambda} |r_N(\lambda)| > \sup_{\mu \in \Sigma(A) \setminus \Lambda} |r_N(\mu)|,
\]

or, equivalently, the ratio

\[
\kappa = \sup_{\mu \in \Sigma(A) \setminus \Lambda} \frac{|r_N(\mu)|}{\min_{\lambda \in \Lambda} |r_N(\lambda)|}
\]

is less than 1. We take the number \( \kappa \) to be a measure of filter quality. We shall see in Section 3 that one wants \( \kappa \ll 1 \) for fast convergence of a subspace iteration. Using the numbers \( y, \gamma, \) and \( \delta \) in (1), we define

\[
W = \sum_{k=0}^{N-1} |w_k|, \quad \hat{\kappa} = \frac{\sup_{x \in O^\delta_{\gamma,\eta}} |r_N(x)|}{\inf_{x \in I^\gamma_\eta} |r_N(x)|},
\]

where \( O^\delta_{\gamma,\eta} = \{ x \in \mathbb{R} : |x - y| \geq (1 + \delta) \gamma \} \) and \( I^\gamma_\eta = \{ x \in \mathbb{R} : |x - y| \leq \gamma \} \). Since \( \kappa \leq \hat{\kappa} \), the quantity \( \hat{\kappa} \) gives a bound for the filter quality. The relevance of \( W \) will be clear in Section 4.

2.1. Filters based on the trapezoid rule on a circle. Let \( \eta > 0, y \in \mathbb{C}, \phi \in \mathbb{R} \). If the \( N \)-subinterval uniform trapezoid rule is used approximate \( r(\xi) \) on the circle \( \Gamma = \{ \eta e^{i(\theta + \phi)} + y : \theta \in [0, 2\pi) \} \), the corresponding quadrature points and weights are

\[
(10) \quad \theta_k = 2\pi k/N, \quad z_k = \eta e^{i(\theta_k + \phi)} + y, \quad w_k = \eta e^{i(\theta_k + \phi)} / N.
\]

An obvious choice of parameters is when the radius of the circular contour \( \eta \) equals \( \gamma \) and the angular shift \( \phi \) is zero, but this does not guarantee \( \hat{\kappa} < 1 \) for all \( \delta \) and \( N \), in contrast to the following two cases:

\[
(11a) \quad \eta = \gamma, \quad \phi = \pm \pi/N.
\]

\[
(11b) \quad \eta = \gamma / 2^{1/N}, \quad \phi = 0.
\]

In the context of digital filters, the case \( \phi = -\pi/N \) is often referred to as the Butterworth filter (see [19, Section 12.6] or [18, (3.8)]). The second case has real poles slightly inside the interval of interest, and is chosen so that, for \( x, y \in \mathbb{R} \), \( |r_N(x)| \geq 1 \) only for \( |x - y| \leq \gamma \). The filters and their corresponding poles (quadrature points) are pictured in Figure 1 for \( N = 8 \).

**Lemma 2.1** (Properties of circle filters). Consider the circle filters given by (10). We have

\[
(12a) \quad W = \eta \leq \gamma, \quad \text{for both filters in (11)}.
\]
If $N$ is even, then

(12b) \[ \hat{k} = \frac{2}{(1 + \delta)^N + 1} < 1 \quad \text{for the filter in (11a)}, \]

(12c) \[ \hat{k} = \frac{1}{2(1 + \delta)^N - 1} < 1 \quad \text{for the filter in (11b)}. \]

**Proof.** It is obvious from the expression for $w_k$ in (10) that $W = \sum_{k=0}^{N-1} \eta/N = \eta$, so (12a) follows immediately.

Next, we claim that any circle filter given by (10) satisfies

(13) \[ r_N(\xi) = \sum_{k=0}^{N-1} w_k(z_k - \xi)^{-1} = \frac{e^{iN\phi}}{e^{iN\phi} - ((\xi - y)/\eta)^N}. \]

For the special case $\eta = 1, y = 0, \phi = 0$, 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 (13) to the real line, it follows by inspection that

(14) \[ \min_{x \in I^y_\eta} |r_N(x)| = \begin{cases} 1/2, & \eta = \gamma, \phi = \pm \pi/N, \\ 1, & \eta = \gamma/2^{1/N}, \phi = 0, \end{cases} \]

(15) \[ \max_{x \in O_{y,\eta}^y} |r_N(x)| = \begin{cases} [(1 + \delta)^N + 1]^{-1}, & \eta = \gamma, \phi = \pm \pi/N, \\ [2(1 + \delta)^N - 1]^{-1}, & \eta = \gamma/2^{1/N}, \phi = 0, \end{cases} \]

for any $y \in \mathbb{R}$. This proves (12b) and (12c). \qed

**Remark 2.2.** It can proved by similar arguments that the case $\eta = \gamma, \phi = 0$ has

$$ W = \eta, \quad \hat{k} = \frac{1}{(1 + \delta)^N - 1}. $$

Thus this circle filter has $\hat{k} < 1$ only when $(1 + \delta)^N > 2$, whereas the two filters we considered in (11) have $\hat{k} < 1$ unconditionally. We also note that, when $(1 + \delta)^N > 3$, the case $\eta = \gamma, \phi = 0$ yields a smaller $\hat{k}$ than the case $\eta = \gamma, \phi = \pm \pi/N$.

### 2.2. Filters Based on the Trapezoid Rule on an Ellipse

Let $\eta > 0, y \in \mathbb{C}, \sigma, \phi \in \mathbb{R}$ and $\rho > 1$. If the $N$-subinterval uniform trapezoid rule is used to approximate $r(\xi)$ on the Bernstein ellipse

$$ \Gamma = \left\{ e^{i\sigma} \eta \frac{\eta}{2} \left( \rho e^{i(\theta + \phi)} + \rho^{-1}e^{-i(\theta + \phi)} \right) + y : \theta \in [0, 2\pi) \right\}, $$

the corresponding quadrature points and weights are

(16) \[ z_k = e^{i\sigma} \eta \frac{\eta}{2} \left( \rho e^{i(\theta_k + \phi)} + \rho^{-1}e^{-i(\theta_k + \phi)} \right) + y, \quad w_k = e^{i\sigma} \frac{\eta}{2N} \left( \rho e^{i(\theta_k + \phi)} - \rho^{-1}e^{-i(\theta_k + \phi)} \right), \]

for $\theta_k = 2\pi k/N$. The semi-major axis of the ellipse is $\eta(\rho + \rho^{-1})/2$, and it is aligned with the line $te^{i\sigma}, t \in \mathbb{R}$; the semi-minor axis is $\eta(\rho + \rho^{-1})/2$, and it is aligned with the line $i\rho e^{i\sigma}, t \in \mathbb{R}$. Because our eigenvalues are real, we restrict our attention to the case $\sigma = 0$ and introduce natural analogues of (11a) and (11b):

(17a) \[ \eta = 2\gamma/(\rho + \rho^{-1}), \quad \phi = \pm \pi/N, \quad \sigma = 0, \]

(17b) \[ \eta = 2\gamma/(a + a^{-1}), \quad \phi = 0, \quad \sigma = 0, \]
where $a$ is given by

$$a = \left( (1 + \rho^N + \rho^{-N}) + \sqrt{(1 + \rho^N + \rho^{-N})^2 - 1} \right)^{1/N}.$$

The case (17b) has its real poles slightly inside the interval of interest, and we will see in Proposition 2.4 that the circular filters (11) are limiting cases of the ellipse filters (17), as $\rho \to \infty$. The filters and their corresponding poles (quadrature points) are pictured in Figure 2 for $N = 8$ and $\rho = 1.5$.

These filters are related to Chebyshev polynomials. This is readily seen in the special case $\eta = 1$, $y = 0$, $\phi = \sigma = 0$, where $z_k = (\rho e^{i\theta_k} + (\rho e^{i\theta_k})^{-1})/2$. Recall the identity

$$T_N((b + b^{-1})/2) = (b^N + b^{-N})/2, \quad b \neq 0,$$

where $T_N$ is the degree $N$ first-kind Chebyshev polynomial, normalized so that $T_N(1) = 1$. Applying (18) with $b = \rho e^{i\theta_k}$, we see that $z_k$ is a root of the polynomial $T_N(\xi) - (\rho^N + \rho^{-N})/2$ in $\xi$. In fact,

$$2^{1-N}(T_N(\xi) - (\rho^N + \rho^{-N})/2) = \prod_{k=0}^{N-1}(\xi - z_k).$$

For a discussion of “Chebyshev filters” in the context of digital filters, we refer interested readers to [19, Section 13.4–13.6].

**Lemma 2.3** (Properties of ellipse filters). Consider the ellipse filters given by (16). We have

(20a) \hspace{1cm} W \leq \eta(\rho + \rho^{-1})/2 \leq \gamma, \quad \text{for both filters in (17).}

If $N$ is even, then

(20b) \hspace{1cm} \hat{\kappa} = \frac{2(\rho^N + \rho^{-N})}{2T_N((\rho + \rho^{-1})(1 + \delta)/2) + \rho^N + \rho^{-N}} < 1 \quad \text{for the filter in (17a)},

(20c) \hspace{1cm} \hat{\kappa} = \frac{\rho^N + \rho^{-N} + 2}{2T_N((\rho + \rho^{-1})(1 + \delta)/2) - (\rho^N + \rho^{-N})} < 1 \quad \text{for the filter in (17b)}.$
Proof. It is obvious from the expression for \( w_k \) that \(|w_k| \leq \eta (\rho + \rho^{-1})/(2N)\), so the first inequality in (20a) follows immediately. The choice of \( \eta \) in (17a) yields \( \eta (\rho + \rho^{-1})/2 = \gamma \). Since \( a > \rho \) and \((x + x^{-1})/2\) is an increasing function of \( x \), the choice of \( \eta \) in (17b) yields \( \eta (\rho + \rho^{-1})/2 < \gamma \).

For the general quadrature points and weights as in (16), we have

\begin{equation}
    r_N(\xi) = \sum_{k=0}^{N-1} w_k (z_k - \xi)^{-1} = \frac{((\rho e^{i\phi})^N - (\rho e^{i\phi})^{-N})/2}{((\rho e^{i\phi})^N + (\rho e^{i\phi})^{-N})/2 - T_N(e^{-i\phi}(\xi - y)/\eta)}.
\end{equation}

A proof of the last equality for the special case \( \eta = 1, y = 0, \phi = \sigma = 0 \), follows by viewing the second expression as a partial fraction expansion of the last, recognizing that (19) holds in this case. The general case then follows by change of variable.

Taking \( w = (\xi - y)/\gamma \) as before, we specialize (21) to the two cases (17a) and (17b):

\begin{equation}
    r_N(\xi) = \begin{cases} 
    \frac{\rho^N - \rho^{-N}}{\rho^N + \rho^{-N} + 2T_N((\rho + \rho^{-1})w/2)}, & \text{in case (17a)}, \\
    \frac{-\rho^N - \rho^{-N}}{\rho^N + \rho^{-N} - 2T_N((\rho + \rho^{-1})w/2)}, & \text{in case (17b)}. 
\end{cases}
\end{equation}

Using \((\rho^N - \rho^{-N})/(\rho^N + \rho^{-N} \pm 2) = (\rho^N \pm 1)/(\rho^N \pm 1)\), we determine that

\begin{equation}
    \frac{\rho^N - 1}{\rho^N + 1} \leq r_N(\xi) \leq \frac{\rho^N + 1}{\rho^N - 1} \quad \text{for } \xi \in [y - \eta, y + \eta].
\end{equation}

Note that \( r_N(\xi) \) oscillates between these extreme values on the associated intervals, and has no local extrema outside of these intervals. Restricting (22) to the real line, it follows that

\begin{equation}
    \min_{x \in \mathbb{T}_y} |r_N(x)| = \begin{cases} 
    \frac{\rho^N - \rho^{-N}}{2(\rho^N + \rho^{-N})}, & \text{in case (17a)}, \\
    \frac{\rho^N - 1}{\rho^N + 1}, & \text{in case (17b)}. 
\end{cases}
\end{equation}

These minima in absolute value occur when \( w = \pm 1 \). At this stage, it should be clear that the parameter \( a > \rho \) was chosen to satisfy \( T_N((a + a^{-1})/2) = 1 + \rho^N + \rho^{-N} \), so that

\[ \frac{\rho^N - \rho^{-N}}{\rho^N + \rho^{-N} - 2T_N((a + a^{-1})w/2)} = -\frac{\rho^N - 1}{\rho^N + 1} \quad \text{for } w = \pm 1. \]

It also follows that

\begin{equation}
    \max_{x \in \mathbb{O}_{\delta,\gamma}} |r_N(x)| = \begin{cases} 
    \frac{\rho^N - \rho^{-N}}{2T_N((\rho + \rho^{-1})(1 + \delta)/2 + (\rho^N + \rho^{-N})}, & \text{in case (17a)}, \\
    \frac{\rho^N - \rho^{-N}}{2T_N((a + a^{-1})(1 + \delta)/2 - (\rho^N + \rho^{-N})}, & \text{in case (17b)}. 
\end{cases}
\end{equation}

Now the equalities in (20b) and (20c) follow directly from (24) and (25). That \( \hat{\kappa} < 1 \) in both cases follows from the facts that \( \hat{\kappa} = 1 \) when \( \delta = 0 \), and \( \hat{\kappa} \) is a strictly decreasing function of \( \delta \), because \( T_N \) is strictly increasing outside of \([-1, 1]\). \( \Box \)

The limiting cases \( \rho \to \infty \) and \( \rho \to 1 \) are described in the following result.
Proposition 2.4. Each ellipse filter converges pointwise to its circle filter counterpart as $\rho \to \infty$, except at the poles of the corresponding circle filter. Each ellipse filter converges pointwise to zero as $\rho \to 1$, except at finitely many points as described in the proof.

Proof. The results for $\rho \to \infty$ are perhaps most readily seen by considering the convergence, with respect to $\rho$, of the quadrature points and weights. For $\eta = 2\gamma/(\rho + \rho^{-1})$, we have $\rho\eta/2 \to \gamma$ as $\rho \to \infty$, so the quadrature points and weights (16)-(17a) converge to those of (10)-(11a). For $\eta = 2\gamma/(a + a^{-1})$, we have $\rho\eta/2 \to \gamma/2^{1/N}$ as $\rho \to \infty$, so the quadrature points and weights (16)-(17b) converge to those of (10)-(11b).

We now consider their limits as $\rho \to 1$. It is clear from (22) that the limiting functions will be zero except at the points where the limiting denominators vanish. For (17a), the points at which we do not have convergence to zero are $\xi = \gamma \cos(\theta_k + \pi/N) + y$. Finally, for (17b), the points at which we do not have convergence to zero are $\xi = \eta \cos \theta_k + y$, where $\eta = 2\gamma/(a + a^{-1})$ and $a = (3 + \sqrt{8})^{1/N}$. □
2.3. **Summary.** Because a straightforward application of the trapezoidal rule (as in Remark 2.2) does not always provide a \( \hat{\kappa} < 1 \), we have discussed four other filters.

- **Filter 1:** \( w_k, z_k \) are set by (10) and (11a);
- **Filter 2:** \( w_k, z_k \) are set by (10) and (11b);
- **Filter 3:** \( w_k, z_k \) are set by (16) and (17a);
- **Filter 4:** \( w_k, z_k \) are set by (16) and (17b).

In the remainder of the paper, we shall proceed under these settings:

\[
\begin{align*}
\text{(26a)} & \quad z_k \notin \Sigma(A), \\
\text{(26b)} & \quad N \text{ is even,} \\
\text{(26c)} & \quad r_N \text{ is set to one of Filters 1, 2, 3, or 4.}
\end{align*}
\]

Note that (26a) always holds for Filters 1 and 3. For the remaining two filters, in the event \( \Sigma(A) \) intersects with the two real poles \( z_0, z_{N/2} \), the quantities \( \gamma \) and/or \( N \) can be slightly adjusted to have (26a) hold. Note also that (26b), while not required for key identities such as (13) and (21), is needed for the identities concerning \( \hat{\kappa} \). Both (26a) and (26b) are standard assumptions (cf. [3,18]) and are not restrictive in practice.

### 3. Errors in filtered subspace iteration

In the last section, we saw how the eigenvalues of \( A \) of interest are filtered to become the dominant eigenvalues of \( S_N \). Repeated application of this operator is a common subspace iteration technique to target the dominant eigenspace. Such filtered subspace iterations have been gaining popularity for matrix eigenvalue problems under the name “FEAST iterations” [3, 25, 32]. We now analyze the convergence of these iterations by standard techniques. We begin with an identity that applies generally to a bounded linear operator. We will apply it to \( S_N \) afterward.

#### 3.1. General subspace iteration

Let \( \mathcal{W} \) be a (complex) Hilbert space and let \( B : \mathcal{W} \rightarrow \mathcal{W} \) be a bounded linear operator. Let \( \Upsilon \) be a finite set of eigenvalues of finite multiplicity of \( B \) that are isolated from the rest of \( \Sigma(B) \). Let \( P \) be any projector onto the algebraic eigenspace of \( B \) associated to \( \Upsilon \). (\( P \) need not be the spectral projection.) We study an inexact subspace iteration to approximate eigenvectors corresponding to eigenvalues in \( \Upsilon \). To account for the inexactness, we consider \( \bar{B}_\ell = B + \Delta_\ell \) where \( \Delta_\ell : \mathcal{W} \rightarrow \mathcal{W} \) is a bounded linear operator representing perturbations at step \( \ell \) of the iteration. The iterations are started using a given initial finite-dimensional subspace \( Q_0 \subset \mathcal{W} \). At step \( \ell \), the inexact subspace iteration computes the subspace

\[
Q_\ell = \bar{B}_\ell Q_{\ell-1}, \quad \ell = 1, 2, \ldots
\]

The following lemma is motivated by the analysis of subspace iteration in [28].

**Lemma 3.1.** Suppose \( \dim(PQ_0) = \dim(\text{ran}(P)) \). Then for each \( 0 \neq \mu \in \Upsilon \) and \( 0 \neq v \in \mathcal{W} \) satisfying \( Bv = \mu v \), there is a sequence \( q^{(\ell)}, \ell = 0, 1, 2, \ldots \), such that \( q^{(\ell)} \in Q_\ell \) and

\[
v - q^{(\ell)} = \frac{1}{\mu^\ell} \bar{B}_\ell (I - P)(v - q^{(0)}) + \frac{1}{\mu^\ell} \left[ B^\ell - \left( \bar{B}_\ell \bar{B}_{\ell-1} \cdots \bar{B}_1 \right) \right] q^{(0)}.
\]

**Proof.** Since \( PQ_0 \subseteq \text{ran} P \), the assumption \( \dim(PQ_0) = \dim(\text{ran} P) \) implies that \( PQ_0 = \text{ran} P \). Hence there is a \( q^{(0)} \in Q_0 \) such that \( Pq^{(0)} = v \). Set \( q^{(\ell)} = \mu^{-\ell} \bar{B}_\ell \cdots \bar{B}_1 q^{(0)} \). Clearly \( q^{(\ell)} \)
lies in $Q_{\ell}$. Moreover,
\[
v - q^{(\ell)} = v - \mu^{-\ell}B^{\ell}q^{(0)} + \mu^{-\ell}\left[B^{\ell} - (\tilde{B}_{\ell} \cdots \tilde{B}_{1})\right]q^{(0)}
\]
\[
= v - \mu^{-\ell}B^{\ell}\left[Pq^{(0)} + (I - P)q^{(0)}\right] + \mu^{-\ell}\left[B^{\ell} - (\tilde{B}_{\ell} \cdots \tilde{B}_{1})\right]q^{(0)}.
\]
Since $Pq^{(0)} = v$ is an eigenfunction of $B^{\ell}$, this simplifies to
\[
v - q^{(\ell)} = -\mu^{-\ell}B^{\ell}(I - P)q^{(0)} + \mu^{-\ell}\left[B^{\ell} - (\tilde{B}_{\ell} \cdots \tilde{B}_{1})\right]q^{(0)}.
\]
The proof is completed by observing that $(I - P)q^{(0)} = (I - P)^{2}q^{(0)} = (I - P)(q^{(0)} - v)$. □

Set $\mu^* = \text{rad}(B(I - P))$, the spectral radius of $B(I - P)$ in $\mathcal{W}$. The set $\mathcal{Y}$ is a set of dominant eigenvalues if $\mu^* < |\mu|$ for all $\mu \in \mathcal{Y}$. In this case, the inequality of the next lemma implies a convergence result.

**Lemma 3.2.** Suppose the assumptions of Lemma 3.1 hold, $v$ and $q^{(\ell)}$ are as in Lemma 3.1, and $\Delta_{\ell} = 0$ for all $\ell$. Then for any $\varepsilon > 0$, there is an integer $\ell_0 \geq 1$ such that for all $\ell \geq \ell_0$,
\[
\|v - q^{(\ell)}\|_{\mathcal{W}} \leq \left|\frac{\varepsilon + \mu^*}{\mu}\right|\|v - q^{(0)}\|_{\mathcal{W}}.
\]
If in addition $B$ is selfadjoint with respect to the inner product of $\mathcal{W}$, then we may choose $\varepsilon = \ell_0 = 0$.

**Proof.** We apply Lemma 3.1 with $P$ set to the spectral projector of $B$ associated to $\mathcal{Y}$. Then $B$ commutes with $P$, so $B^{\ell}(I - P) = [B(I - P)]^{\ell}$ and
\[
\|v - q^{(\ell)}\|_{\mathcal{W}} \leq \frac{1}{|\mu|^\ell}\|B(I - P)\|^\ell_{\mathcal{W}}\|v - q^{(0)}\|_{\mathcal{W}}.
\]
Since $\mu^* = \lim_{l \to \infty} \|B(I - P)\|^{1/\ell}_{\mathcal{W}}$, given any $\varepsilon > 0$, there is an $\ell_0$ such that for all $\ell \geq \ell_0$
\[
\|B(I - P)\|^{1/\ell}_{\mathcal{W}} \leq \varepsilon + \mu^*.
\]
Hence the inequality of the lemma follows. If $B$ is selfadjoint, then $B(I - P)$ is also self adjoint with respect to the scalar product of $\mathcal{W}$ and we can instead use $\mu^* = \|B(I - P)\|_{\mathcal{W}}$ (see e.g. [10, Lemma 5.1.5]) to obtain the stated improvement. □

**3.2. Application to filtered subspace iteration.** We now proceed to analyze the inexact subspace iteration applied to the filter $S_N$, or a perturbation of it. One often prefers to measure these perturbations and convergence rates in a norm different from the norm on $\mathcal{H}$, hence the following assumption.

**Assumption 1.** Suppose there is a Hilbert space $\mathcal{V} \subseteq \mathcal{H}$ that is dense in $\mathcal{H}$ such that $E \subseteq \mathcal{V}$. Suppose $\mathcal{V}$ is continuously embedded in $\mathcal{H}$, i.e., there is a $C_{\mathcal{V}} > 0$ such that for all $u \in \mathcal{V}$, $\|u\|_{\mathcal{H}} \leq C_{\mathcal{V}}\|u\|_{\mathcal{V}}$. Suppose also that, for all $z$ in the resolvent set of $A$, the resolvent $R(z)$ satisfies $(R(z)v, w)_{\mathcal{V}} = (v, R(z)w)_{\mathcal{V}}$ for all $v, w \in \mathcal{V}$, and that $\mathcal{V}$ is an invariant subspace of $R(z)$.

We now give three natural and important examples where this assumption holds.

**Example 3.3 ( $\mathcal{V}$ is the whole space).** Set $\mathcal{V} = \mathcal{H}$, with $\langle \cdot, \cdot \rangle_{\mathcal{V}} = \langle \cdot, \cdot \rangle_{\mathcal{H}}$. In this case it is obvious that all statements of Assumption 1 hold.
Before giving the next example, let us recall that a linear operator \( L \) on \( \mathcal{H} \) is called [22] positive if \( (Lv,v)_{\mathcal{H}} > 0 \) for all \( 0 \neq v \in \text{dom} L \). Recall also that any selfadjoint positive operator \( L \) has a unique selfadjoint positive square root [22, Theorem V.3.35], which we denote by \( L^{1/2} \). Furthermore, \( L^{1/2} \) commutes with any \( \mathcal{H} \)-bounded operator that commutes with \( L \).

**Example 3.4** (\( \mathcal{V} \) 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

\[
(a(v,v) \geq \delta \|v\|_{\mathcal{H}}^2, \quad v \in \text{dom}(a)).
\]

Set

\[
\mathcal{V} = \text{dom}(a), \quad \|v\|_{\mathcal{V}} = a(v,v)^{1/2}.
\]

Set the operator \( A \) to be the closed selfadjoint operator associated with the form, namely it satisfies \( a(u,v) = (Au,v) \) for all \( u \in \text{dom}(A) \subseteq \text{dom}(a) \) and all \( v \in \text{dom}(a) \) (see the first representation theorem [22, Theorem VI.2.1] or [30, Theorem 10.7]). Note that, in this case, \( A \) is a positive operator. It is also important to note that, by the second representation theorem [22, Theorem VI.2.23], the form domain is characterized by \( \text{dom}(a) = \text{dom}(A^{1/2}) \), 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.

All statements of Assumption 1 hold in this case. Since \( a \) is closed, \( \mathcal{V} \) is complete. Since \( a \) is densely defined, \( \mathcal{V} = \text{dom}(a) \) is dense in \( \mathcal{H} \). Due to (28), \( \mathcal{V} \) is continuously embedded in \( \mathcal{H} \), with the constant \( C_{\mathcal{V}} = \delta^{-1/2} \). The exact eigenspace \( \mathcal{E} \) is contained in \( \text{dom}(A) \subseteq \text{dom}(A^{1/2}) = \mathcal{V} \). Since \( A^{1/2} \) commutes with \( R(z) \), for any \( v,w \in \mathcal{V} \),

\[
(R(z)v,w)_{\mathcal{V}} = (A^{1/2}R(z)v,A^{1/2}w)_{\mathcal{H}} = (R(z)A^{1/2}v,A^{1/2}w)_{\mathcal{H}} = (A^{1/2}v,R(z)A^{1/2}w)_{\mathcal{H}} = (v,R(z)w)_{\mathcal{V}}.
\]

Finally, for any \( v \in \mathcal{V} = \text{dom}(A^{1/2}) \) and \( z \) in the resolvent set of \( A \), we have

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

Since \( \text{ran}(A^{-1/2}) = \text{dom}(A^{1/2}) = \mathcal{V} \), we see that \( R(z)\mathcal{V} \subseteq \mathcal{V} \). This completes the verification of all statements in Assumption 1.

**Example 3.5** (\( \mathcal{V} \) is a graph space). Given \( A \), put \( \mathcal{V} = \text{dom}(A) \subseteq \mathcal{H} \) and endow the set \( \mathcal{V} \) with the topology of the graph norm

\[
\|v\|_{\mathcal{V}} = (\|v\|_{\mathcal{H}}^2 + \|Av\|_{\mathcal{H}}^2)^{1/2}, \quad v \in \mathcal{V}.
\]

Assumption 1 holds in this case. Indeed, since \( A \) is closed, the graph norm makes \( \mathcal{V} \) into a Hilbert space. Obviously \( \mathcal{E} \subseteq \mathcal{V} \) and \( \mathcal{V} \) is continuously embedded into \( \mathcal{H} \) with \( C_{\mathcal{V}} = 1 \). Since \( A \) is selfadjoint, \( \text{dom}(A) = \mathcal{V} \) is dense in \( \mathcal{H} \). Since \( A \) commutes with \( R(z) \) for any \( z \) in the resolvent set of \( A \), we have \( R(z)\text{dom}(A) \subseteq \text{dom}(A) \). That \( (R(z)v,w)_{\mathcal{V}} = (v,R(z)w)_{\mathcal{V}} \) follows by a minor modification of the argument in Example 3.4.

**Lemma 3.6.** Suppose Assumption 1 holds. Let \( B \) be a bounded selfadjoint operator on \( \mathcal{H} \) and let \( \mathcal{V} \) be an invariant subspace of \( B \). Suppose \( B|_{\mathcal{V}} \), again denoted by \( B \), is also selfadjoint and bounded in \( \mathcal{V} \). Then \( \|B\|_{\mathcal{H}} = \|B\|_{\mathcal{V}} \).
Proof. By the representation theorem [22, Theorem VI.2.23], there is a unique positive self-
adjoint operator \( G \) whose square root characterizes \((\cdot,\cdot)_{\mathcal{V}}\) as follows:

\[
(29) \quad (u,v)_{\mathcal{V}} = (G^{1/2}u,G^{1/2}v)_{\mathcal{H}}, \quad u,v \in \mathcal{V} = \text{dom}(G^{1/2}).
\]

The operator \( G^{1/2} \) has a bounded inverse \( G^{-1/2} : \mathcal{H} \to \mathcal{H} \) with \( \text{ran}(G^{-1/2}) = \mathcal{V} \) (see e.g., [22, Theorem V.3.35]). Since any \( v \in \mathcal{V} \) can be written as \( v = G^{-1/2}w \) for some \( w \in \mathcal{H} \), we have

\[
(30) \quad \|B\|_{\mathcal{V}} = \sup_{v \in \mathcal{V}} \frac{\|Bv\|_{\mathcal{V}}}{\|v\|_{\mathcal{V}}} = \sup_{w \in \mathcal{H}} \frac{\|G^{1/2}Bv\|_{\mathcal{H}}}{\|w\|_{\mathcal{H}}} = \|G^{1/2}BG^{-1/2}\|_{\mathcal{H}}.
\]

Since \( B \) is bounded in \( \mathcal{V} \), we conclude that \( G^{1/2}BG^{-1/2} \) is bounded in \( \mathcal{H} \). Furthermore, the selfadjointness of \( B \) in \( \mathcal{V} \) implies

\[
(G^{1/2}Bu,G^{1/2}v)_{\mathcal{H}} = (G^{1/2}u,G^{1/2}Bv)_{\mathcal{H}}
\]

for all \( u,v \in \mathcal{V} \). Putting \( u = G^{-1/2}x \) and \( v = G^{-1/2}y \) we obtain \( (G^{1/2}BG^{-1/2}x,y)_{\mathcal{H}} = (x,G^{1/2}BG^{-1/2}y)_{\mathcal{H}} \) for all \( x,y \in \mathcal{H} \). Thus \( G^{1/2}BG^{-1/2} = G^{-1/2}BG^{1/2} \), and so \( BG^{-1} = G^{-1}B \).

Since \( B \) commutes with \( G^{-1} \), it commutes with \( G^{-1/2} \) (again see [22, Theorem V.3.35]). Hence, \( G^{1/2}BG^{-1/2} = B \), so \( \|B\|_{\mathcal{V}} = \|B\|_{\mathcal{H}} \) by (30).

\[\square\]

**Lemma 3.7.** Suppose Assumption 1 holds. Then \( S \) and \( S_N \) are bounded selfadjoint operators in \( \mathcal{V} \) and

\[
(31) \quad \|S_N\|_{\mathcal{V}} = \max\{|r_N(\lambda)| : \lambda \in \Lambda\},
\]

\[
(32) \quad \|(I-S)S_N(I-S)\|_{\mathcal{V}} = \sup\{|r_N(\lambda)| : \lambda \in \Sigma(A)\setminus\Lambda\}.
\]

**Proof.** The filter \( r_N \) is real-valued on \( \mathbb{R} \), and bounded on \( \Sigma(A) \) because of (26a). Therefore, \( S_N = r_N(A) \) is bounded and selfadjoint on \( \mathcal{H} \), with \( \|S_N\|_{\mathcal{H}} = \sup\{|r_N(\lambda)| : \lambda \in \Sigma(A)\} \), as a result of functional calculus (see, e.g., [30, Theorem 5.9]). This supremum is equal to \( \max\{|r_N(\lambda)| : \lambda \in \Lambda\} \) by Lemmas 2.1 and 2.3. Applying a similar argument to \((1-r)r_N(1-r)\) also, we conclude that

\[
(33) \quad \|S_N\|_{\mathcal{H}} = \max\{|r_N(\lambda)| : \lambda \in \Lambda\},
\]

\[
(34) \quad \|(I-S)S_N(I-S)\|_{\mathcal{H}} = \sup\{|r_N(\lambda)| : \lambda \in \Sigma(A)\setminus\Lambda\}.
\]

To show that these expressions also give the respective \( \mathcal{V} \)-norms, we will apply Lemma 3.6.

Since \( \mathcal{V} \) is an invariant subspace of \( R(z) \) due to Assumption 1, \( S_N \) maps \( \mathcal{V} \) into \( \mathcal{V} \). Furthermore, for the filter functions in (26c), the set of quadrature weights and nodes \( \{(w_k,z_k) : k = 0,\ldots,N-1\} \) coincides with \( \{(|w_k,z_k|) : k = 0,\ldots,N-1\} \). Hence,

\[
S_N = r_N(A) = \sum_{k=0}^{N-1} w_k R(z_k) = \sum_{k=0}^{N-1} \overline{w_k} R(\overline{z_k}).
\]

This, combined with the equality \( (R(z)u,v)_{\mathcal{V}} = (R(\overline{z})v)_{\mathcal{V}} \) given by Assumption 1, then yields \( (S_N u,v)_{\mathcal{V}} = (u,S_N v)_{\mathcal{V}} \) for all \( u,v \in \mathcal{V} \). Together with the fact that \( S_N \) is defined everywhere on \( \mathcal{V} \), this implies, by the Hellinger-Toeplitz theorem [26, Corollary of Theorem III.12], that \( S_N \) is a bounded selfadjoint operator on \( \mathcal{V} \). Now (31) follows by Lemma 3.6 and (33).
The symmetry of $S$ in $V$ also follows from the same equality $(R(z)u, v)_V = (u, R(\bar{z})v)_V$ for all $u, v \in V$. Indeed, by (3), we have $(S u, v)_V = (u, S^* v)_V$ where

$$S^* = \frac{1}{2\pi i} \int_{\Gamma} R(\bar{z}) \, dz = \frac{1}{2\pi i} \int_{\bar{\Gamma}} R(z) \, dz = S$$

because $\Gamma$ and $\bar{\Gamma}$ (when positively oriented) are the same for any contour $\Gamma$ that is symmetric with respect to the real line, and $S$ is independent of the choice of $\Gamma$, provided that it encloses $\Lambda$ and excludes $\Sigma(A) \setminus \Lambda$. Using the Hellinger-Toeplitz theorem again, we conclude that $S$ is a bounded selfadjoint operator on $V$. Consequently, $(I - S)S_N(I - S)$ is a bounded selfadjoint operator on $V$, so Lemma 3.6 and (33) together prove (32).

With these lemmas, we are now ready to analyze the inexact filtered subspace iteration. Consider the subspace iteration (27) with $\tilde{B}_\ell$ set to $\tilde{S}_N = S_N + \Delta_\ell$, a perturbation of the filter $B = S_N$. Let

$$\tilde{S}_N^{(\ell)} = S_N + \Delta_\ell, \quad \|\Delta_\ell\|_V \leq \epsilon \|S_N\|_V.$$  

Recall that $\Lambda$ is a finite set (see Section 1) consisting of $m$ isolated points, where we count the eigenvalues according to multiplicity, and $E$ is the ($m$-dimensional) span of the corresponding eigenvectors,

$$\Lambda = \{\lambda_1, \ldots, \lambda_m\}, \quad E = \text{span}\{e_1, \ldots, e_m\}, \quad Ae_i = \lambda_i e_i.$$

The following are key quantities for Theorem 3.8 below:

$$r_N^* = \sup_{\lambda \in \Sigma(A) \setminus \Lambda} |r_N(\lambda)|, \quad \kappa_i = \frac{r_N^*}{|r_N(\lambda_i)|}, \quad \beta_i = \frac{\max_{1 \leq j \leq m} |r_N(\lambda_j)|}{|r_N(\lambda_i)|}.$$  

Note that $r_N(\lambda_i)$ is never zero (see (14) and (24)) and $r_N^*$ is finite (see (15) and (25)).

**Theorem 3.8.** Suppose Assumption 1 holds and let $Q_\ell$ be given by (27) with $\tilde{B}_\ell = \tilde{S}_N^{(\ell)}$ starting from a $Q_0 \subseteq V$ satisfying $\dim(SQ_0) = m$. Then for each $e_i$, there is a sequence $q_i^{(\ell)}$ in $Q_\ell$ satisfying

$$\|e_i - q_i^{(\ell)}\|_V \leq \kappa_i \|e_i - q_i^{(0)}\|_V + [(1 + \epsilon)\ell - 1] \beta_i \|q_i^{(0)}\|_V$$

for all $\ell \geq 0$ and $i = 1, \ldots, m$. Moreover, $\kappa_i \leq \bar{\kappa} < 1$ by Lemmas 2.1 and 2.3.

**Proof.** Applying Lemma 3.1 with $B = S_N$, $\tilde{B}_\ell = \tilde{S}_N^{(\ell)}$, $\mu_i = r_N(\lambda_i)$, and $P = S$,

$$e_i - q_i^{(\ell)} = \frac{1}{r_N(\lambda_i)^\ell} S_N^{\ell}(I - S)(e_i - q_i^{(0)}) + \frac{1}{r_N(\lambda_i)^\ell} \left[ S_N^{\ell} - \left( \tilde{S}_N^{\ell}(\tilde{S}_N^{(\ell-1)} - S_N^{(\ell-1)} \cdots \tilde{S}_N^{(1)}) \right) q_i^{(0)} \right].$$

Hence

$$\|e_i - q_i^{(\ell)}\|_V \leq \frac{s_\ell \|e_i - q_i^{(0)}\|_V + d_\ell \|q_i^{(0)}\|_V}{|r_N(\lambda_i)|^\ell}$$

where $s_\ell = \|S_N^{\ell}(I - S)\|_V$, $d_\ell = \|D_\ell\|_V$ and $D_\ell$ is given by

$$D_\ell = S_N^{\ell} - \left( \tilde{S}_N^{\ell} \tilde{S}_N^{(\ell-1)} \cdots \tilde{S}_N^{(1)} \right) = (S_N + \Delta_\ell - \Delta_\ell) S_N^{\ell-1} - (S_N + \Delta_\ell) (\tilde{S}_N^{(\ell-1)} \cdots \tilde{S}_N^{(1)}) = (S_N + \Delta_\ell) D_{\ell-1} - S_N^{\ell-1} \Delta_\ell.$$
The last equality, together with (35) and \(d_0 = 0\), implies that 
\[ d_\ell \leq (1 + \epsilon) d_{\ell - 1} \|S_N\|_\mathcal{V} + \epsilon \|S_N\|_\mathcal{V}. \]
This immediately allows us to use an induction argument to conclude that 
\[ d_\ell \leq [(1 + \epsilon)\ell - 1] \|S_N\|_\mathcal{V}. \]
Using this estimate for \(d_\ell\) in (37), we obtain

\[
\|e_i - q_i^{(0)}\|_\mathcal{V} \leq \frac{s_\ell}{|r_N(\lambda_i)|^\ell} \|e_i - q_i^{(0)}\|_\mathcal{V} + \frac{(1 + \epsilon)^\ell - 1}{|r_N(\lambda_i)|^\ell} \|S_N\|_\mathcal{V} \|q_i^{(0)}\|_\mathcal{V}.
\]

Lemma 3.7 yields \(\|S_N\|_\mathcal{V}/|r_N(\lambda_i)| = \beta_i\).

Now it only remains to bound \(s_\ell\). Since \(I - S\) is a projection and \(S\) commutes with \(A\), \((I - S)\) commutes with \(S_N\) as well, and we have \(S_N^\ell (I - S) = [(I - S)S_N(I - S)]^\ell\). By Lemma 3.7, we have \(s_\ell \leq (\sup\{|r_N(\lambda)| : \lambda \in \Sigma(A) \setminus \Lambda\})^\ell = (r_N^*)^\ell\), which completes the proof.

We conclude this section with a few remarks on Theorem 3.8. Consider the situation when \(\epsilon = 0\) and \(\mathcal{V} = \mathcal{H}\). Then, the estimate of the theorem for Case (a) implies

\[
\min_{z \in Q_\ell} \|e_i - z\|_\mathcal{H} \leq \alpha_i \kappa_i^\ell
\]

where \(\alpha_i = \|e_i - q_i^{(0)}\|_\mathcal{H}\). For finite-dimensional \(\mathcal{H}\), this result was stated in [18, Theorem 2.2] and in [28]. Our generalization covers cases when \(A\) has a continuous spectrum outside \(\Gamma\), and also covers a variety of norms, as Examples 3.4 and 3.5 illustrate. The perturbations \(\Delta_\ell\) give one way to incorporate the approximations due to discretizations or other inaccuracies. Note that in the estimate of the theorem, as the iteration number \(\ell\) increases, the term \(\kappa_i^\ell\) decreases, while \((1 + \epsilon)^\ell - 1\) increases and \(\beta_i^\ell\) does not decrease. (Note that \(\beta_i \geq 1\).) This suggests that there might be an optimal value of \(\ell\) for a given choice of filter (which dictates \(\kappa_i\) and \(\beta_i\)) and a given discretization scheme (which governs the size of \(\epsilon\)). Although reliably estimating \(\kappa_i\) and \(\beta_i\) may be straightforward for many filters, sharp estimates for \(\epsilon\) will typically be more difficult to obtain.

4. Discretization errors

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 \(R_h(z) : \mathcal{H} \rightarrow \mathcal{V}_h\) be a finite-rank approximation to the resolvent \(R(z)\). Consider the approximation of \(S_N\) given by

\[
S_N^h = \sum_{k=1}^N w_k R_h(z_k),
\]

which need not be selfadjoint. In this section, we shall study the error in the eigenspace approximations due to this discretization.

4.1. Convergence of eigenspace discretizations. Recall that \(E = \text{ran} \ S\), the exact eigenspace corresponding to eigenvalues \(\lambda_1, \ldots, \lambda_m\) of \(A\) that we wish to approximate. As we saw previously, the operator \(S_N = r_N(A)\) was designed to have dominant eigenvalues \(r_N(\lambda_1), \ldots, r_N(\lambda_m)\) strictly separated in absolute value from the remainder of \(\Sigma(S_N)\). Let \(\Theta\) be a fixed simple closed rectifiable curve in the complex plane that encloses \(\{r_N(\lambda_1), \ldots, r_N(\lambda_m)\}\)
and no other element of \( \Sigma(S_N) \). Define the spectral projector of \( S_N \) by
\[
P_N = \frac{1}{2\pi i} \oint_{\Theta} (z - S_N)^{-1} \, dz.
\]
Then \( E_N = \text{ran} \, P_N \) is the eigenspace of \( S_N \) corresponding to its eigenvalues \( r_N(\lambda_1), \ldots, r_N(\lambda_m) \). Before we proceed to make the analogous definitions for \( S_h^N \), let us note a simple fact.

**Lemma 4.1.** It holds that \( E_N = E \) 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. \( \square \)

In the remainder of this section, we proceed under the following assumption.

**Assumption 2.** Assume that the operators \( R_h(z_1) \) and \( R(z_1) \) are bounded in \( \mathcal{V} \) and satisfy
\[
\lim_{h \to 0} \| R_h(z_1) - R(z_1) \|_{\mathcal{V}} = 0
\]
for all \( k = 1, 2, \ldots, N \).

To examine a consequence of this assumption, subtract the expression for \( S_h^N \) in (39) from that of \( S_N \). Then
\[
\| S_N - S_h^N \|_{\mathcal{V}} \leq W \max_{k=1,\ldots,N} \| R_h(z_1) - R(z_1) \|_{\mathcal{V}} \to 0
\]
as \( h \to 0 \) due to Assumption 2. Note that \( W < \infty \) by Lemmas 2.1 and 2.3. Let us recall the standard ramifications of the convergence of operators in norm given by (41) (see e.g., [22, Theorem IV.3.16] or [2]). Namely, given an open disc enclosing an isolated eigenvalue of \( S_N \) of multiplicity \( \ell \), (41) implies that for sufficiently small \( h \), there are exactly \( \ell \) eigenvalues (counting multiplicities) of \( S_h^N \) in the same disc. In particular, this implies that, for sufficiently small \( h \), the contour \( \Theta \) is in the resolvent set of \( S_h^N \) and encloses exactly \( m \) eigenvalues (counting multiplicities) of \( S_h^N \), which we shall enumerate as \( \mu_1^h, \mu_2^h, \ldots, \mu_m^h \). Hence, the integral
\[
P_h = \frac{1}{2\pi i} \oint_{\Theta} (z - S_N^h)^{-1} \, dz
\]
is well defined. Let \( E_h = \text{ran} \, P_h \). Clearly, \( P_h \) is the spectral projector of \( S_h^N \) corresponding to the eigenvalues \( \mu_1^h, \mu_2^h, \ldots, \mu_m^h \). Hence,
\[
\dim E_h = m.
\]

Our next result examines the difference between \( E \) and \( E_h \). To compare two linear subspaces \( M \) and \( L \) of \( \mathcal{V} \), it is standard to use the “gap” [22] between \( M \) and \( L \), defined by
\[
\text{gap}_{\mathcal{V}}(M, L) = \max \left[ \sup_{m \in U_M^\mathcal{V}} \text{dist}_{\mathcal{V}}(m, L), \sup_{l \in U_L^\mathcal{V}} \text{dist}_{\mathcal{V}}(l, M) \right].
\]

Here and throughout, for any linear subspace \( M \subseteq \mathcal{V} \), we use \( U_M^\mathcal{V} \) to denote its unit ball \( \{ w \in M : \| w \|_{\mathcal{V}} = 1 \} \).
Theorem 4.2. Suppose Assumptions 1 and 2 hold. Then there is a $C_N > 0$ and an $h_0 > 0$ such that for all $h < h_0$,

$$\text{gap}_\mathcal{V}(E, E_h) \leq C_N \max_{k=1,\ldots,N} \| (R(z_k) - R_h(z_k)) \|_E \|_{\mathcal{V}},$$

so, in particular,

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

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

$$\delta_h = \sup_{e \in U_{E_N}^\mathcal{V}} \text{dist}_\mathcal{V}(e, E_h).$$

Then,

$$\delta_h \leq \sup_{e \in U_{E_N}^\mathcal{V}} \| e - P_h e \|_{\mathcal{V}} \leq \sup_{e \in U_{E_N}^\mathcal{V}} \| (P_N - P_h) e \|_{\mathcal{V}}.$$

Note that

$$P_N - P_h = \frac{1}{2\pi i} \int_{\Theta} \left[ (z - S_N)^{-1} - (z - S_N^h)^{-1} \right] dz$$

$$= \frac{1}{2\pi i} \int_{\Theta} (z - S_N^h)^{-1} (S_N - S_N^h)(z - S_N)^{-1} dz.$$

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

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

Returning to (46), we conclude that $\delta_h \leq C_N \| (S_N - S_N^h) \|_{E_N} \| \mathcal{V} \|$, 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 (41), $\delta_h \to 0$ as $h \to 0$. In particular, for sufficiently small $h$, we have $\delta_h < 1$. Then, by [22, Theorem I.6.34], there is a closed subspace $\tilde{E}_h \subseteq E_h$ such that $\text{gap}_\mathcal{V}(E_N, \tilde{E}_h) = \delta_h < 1$ and $\dim \tilde{E}_h = \dim E_N = m$. Because of (42), this implies that $\tilde{E}_h = E_h$. Since $E_N = E$ by Lemma 4.1, we finish the proof of (44) by noting that $\text{gap}_\mathcal{V}(E, E_h) = \text{gap}_\mathcal{V}(E_N, \tilde{E}_h) = \delta_h$. \qed

Remark 4.3. If $\mathcal{V}$ is replaced by $\mathcal{H}$ in (43), we obtain $\text{gap}_{\mathcal{H}}(M, L)$, so

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

It’s natural to ask if $\text{gap}_\mathcal{V}(E, E_h) \to 0$ implies $\text{gap}_{\mathcal{H}}(E, E_h) \to 0$ as $h \to 0$. Let $\delta^\mathcal{H}_h$ denote the first of the two suprema above. Since $E$ is finite-dimensional, there is a $C_m > 0$ such that $\| e \|_{\mathcal{H}} \geq C_m \| e \|_{\mathcal{V}}$. Using also Assumption 1,

$$\delta^\mathcal{H}_h = \sup_{0 \neq e \in E} \frac{\text{dist}_{\mathcal{H}}(e, E_h)}{\| e \|_{\mathcal{H}}} \leq \frac{C_V}{C_m} \frac{\text{dist}_{\mathcal{V}}(e, E_h)}{\| e \|_{\mathcal{V}}} \leq \frac{C_V}{C_m} \text{gap}_{\mathcal{V}}(E, E_h).$$
Thus, if $\text{gap}_\nu(E, E_h) \to 0$, taking $h$ sufficiently small, $\dim(E_h) = \dim(E) = m$ and $\delta_h^R < 1$, so using [22, Theorem 1.6.34] as in the previous proof, we may conclude that $\text{gap}_{\nu}(E, E_h) = \delta_h^R$. This implies that, under the same assumptions as in Theorem 4.2, there is an $h_1 > 0$ such that

$$\text{gap}_{\nu}(E, E_h) \leq \frac{C_\nu}{C_m} \text{gap}_\nu(E, E_h)$$

for all $h < h_1$.

4.2. Convergence of discrete subspace iterates. In practical computations, we use a finite-dimensional subspace iteration by repeatedly applying $S_N^h$. We now show that the iterates so obtained converge to the above-defined $E_h$.

**Theorem 4.4.** Suppose Assumptions 1 and 2 hold. Starting with a subspace $E_h^{(0)} \subseteq \mathcal{V}_h$ satisfying $\dim(P_hE_h^{(0)}) = m$, compute $E_h^{(\ell)} = S_N^h E_h^{(\ell-1)}$ for $\ell = 1, 2, \ldots$. Then there is an $h_0 > 0$ such that, for all $h < h_0$,

$$\lim_{\ell \to \infty} \text{gap}_\nu(E_h^{(\ell)}, E_h) = 0.$$ 

*Proof.* As we have already seen, 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. Hence we may choose an $\varepsilon > 0$ so small that

$$\varepsilon + \max_{j>m} |\mu_j^h| < \min_{i=1,\ldots,m} |\mu_i^h|$$

Now, since $\text{rad}(S_N^h(I - P_h)) = \max_{j>m} |\mu_j^h|$, applying Lemma 3.2, we find that, for each eigenpair $(\mu_i^h, v_i)$ of $S_N^h$, there is a $q_i^{(\ell)} \in E_h^{(\ell)}$ such that

$$\|v_i - q_i^{(\ell)}\|_\nu \leq \delta_i^\ell, \quad \delta_i = \frac{\varepsilon + \max_{j>m} |\mu_j^h|}{|\mu_i^h|} < 1,$$

for all $\ell$ larger than some $\ell_0 \geq 1$.

Recall from (42) that $\dim(E_h) = m$ for sufficiently small $h$. Together with $P_hE_h^{(0)} \subseteq E_h$ and the assumption $\dim(P_hE_h^{(0)}) = m$, this leads to the equality $P_hE_h^{(0)} = E_h$. Thus $P_hE_h^{(\ell)} = P_h(S_N^h)^\ell E_h^{(0)} = (S_N^h)^\ell P_hE_h^{(0)} = (S_N^h)^\ell E_h = E_h$. In particular, this implies that $\dim(E_h^{(\ell)}) \geq \dim(P_hE_h^{(\ell)}) = \dim(E_h)$. Hence,

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

Next, we identify one of the two suprema in the definition of $\text{gap}_{\nu}(E_h, E_h^{(\ell)})$ and set

$$\delta_{h,\ell} = \sup_{v_h \in E_h^{(\ell)}} \text{dist}_{\nu}(\nu_h, E_h^{(\ell)}).$$

The $m$ eigenvectors $v_i$ of $S_N^h$ form a basis for $E_h$, each of which is approximated by a $q_i^{(\ell)} \in E_h^{(\ell)}$. Hence, we may expand an arbitrary $v_h \in U^\nu_{E_h}$ in this basis and construct an approximation of $v_h$ using the same coefficients:

$$v_h = \sum_{i=1}^m c_i v_i, \quad q = \sum_{i=1}^m c_i q_i^{(\ell)}.$$
By (48),

$$\text{dist}_V(v_h, E_h^{(\ell)}) \leq \left\| v_h - q_{\ell} \right\|_V \leq \sum_{i=1}^{m} |c_i| \delta_{i}^{\ell} \leq \left( \sum_{i=1}^{m} |c_i|^2 \right)^{1/2} \left( \sum_{i=1}^{m} \delta_{i}^{2\ell} \right)^{1/2}.$$  

If $g$ denotes the minimal eigenvalue of the $m \times m$ Gram matrix of the $v_i$-basis (whose $(i,j)$th entry is $(v_i, v_j)_V$), then $g \sum_{i=1}^{m} |c_i|^2 \leq \left\| v_h \right\|_V^2 = 1$. Hence,

$$\delta_{h,\ell} \leq \sup_{v_n \in \tilde{E}_h^{(\ell)}} \left\| v_h - q_{\ell} \right\|_V \leq g^{-1/2} \left( \sum_{i=1}^{m} \delta_{i}^{2\ell} \right)^{1/2}$$

which converges to 0 as $\ell \to \infty$ since $\delta_{i} < 1$.

In particular, for large enough $\ell$, we have $\delta_{h,\ell} < 1$. Hence, by [22, Theorem I.6.34] there is a subspace $\tilde{E}_h^{(\ell)} \subseteq E_h^{(\ell)}$ such that $\text{gap}_V(E_h, \tilde{E}_h^{(\ell)}) = \delta_{h,\ell} < 1$. Hence, $\dim(E_h) = \dim(\tilde{E}_h^{(\ell)}) = m$. But by (49), 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 $\ell$,

$$\text{gap}(E_h, E_h^{(\ell)}) = \delta_{h,\ell},$$

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

Remark 4.5. The rate of convergence of the subspace iterates can be quantified in two ways. One of these uses the numbers $\delta_{i}$ in (48). However, this rate is asymptotic since it only holds for sufficiently large iteration numbers $\ell$. In contrast, Theorem 3.8 gives another convergence rate estimate which holds for all $\ell$, namely $\|e_i - q_{i}^{(\ell)}\|_V \leq \kappa_i^{\ell} \|e_i - q_{i}^{(0)}\|_V + [(1 + \epsilon)^{\ell} - 1] \beta^{\ell} \|q_{i}^{(0)}\|_V$ with

$$\epsilon = \frac{\|\Delta_{\ell}\|_V}{\|S_N - S_N^{h}\|_V} \leq \frac{W}{\|S_N\|_V} \max_{k=1,\ldots,N} \|R(z_k) - R_h(z_k)\|_V.$$  

However, since the upper bound given by Theorem 3.8 does not go to zero as $\ell \to \infty$ when $\epsilon$ is positive, we were unable to use it in the previous proof.

4.3. Convergence of eigenvalues. Eigenvalue approximations are obtained as Ritz values of eigenspace approximations. We first recall that any selfadjoint operator $A$ admits the polar decomposition $A = U_A |A| = |A| U_A$ (see [22, p. 335]), where $U_A$ is selfadjoint and partially isometric, and $|A|$ is selfadjoint and positive semidefinite. As described in [30, §10.2], the polar decomposition can be used to define the following (possibly unbounded) symmetric sesquilinear form associated to the operator $A$:

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

for any $x, y$ in $\text{dom}(a) = \text{dom}(|A|^{1/2})$.

Let $F \subseteq \text{dom}(a)$ be a closed finite-dimensional subspace of $\mathcal{H}$ and let $Q$ be the $\mathcal{H}$-orthogonal projector onto $F$. We define $A_F : F \to F$ by the relation $(A_F x, y) = a(x, y)$ for all $x, y \in F$. We refer to $\Sigma(A_F)$, the spectrum of the linear operator $A_F$, on $F$, as the set of Ritz values of $A$ on $F$. Note that when $F = E \subseteq \text{dom}(A) \subseteq \text{dom}(a)$, the exact eigenspace we wish to approximate, the corresponding projector coincides with $S$, so we have $A_E = S |A|^{1/2} U_A |A|^{1/2} S |S| E$. The Ritz values of $A$ on $E$ are precisely the elements of $\Lambda$. Its approximation is computed using the subspaces of $V_h$ generated by the filtered subspace iteration using $S_N^{h}$. Next, we assume that these subspaces are in the form domain. Example 4.7 below illustrates the reason to consider forms and place this assumption.
Assumption 3. Assume that $\mathcal{V}_h$ is contained in $\text{dom}(a)$.

Example 4.6 (Positive operators). Consider the operator $A$ and the form $a$ in Example 3.4. Here, since $A$ is positive, the factors of the polar decomposition of $A$ are $U_A = I$ and $|A| = A$. Thus $\text{dom}(a) = \text{dom}(|A|^{1/2}) = \text{dom}(A^{1/2})$. Moreover, $\mathcal{V} = \text{dom}(A^{1/2})$ in Example 3.4. Since $\mathcal{V}_h \subset \mathcal{V}$ by definition, we conclude that Assumption 3 holds.

Example 4.7 (A differential operator). To give an example of a partial differential operator fitting the scenario of Example 4.6, suppose $\Omega$ 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

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

for all $u, v$ in $\text{dom}(a) = H^1(\Omega)$. 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 [30, Theorem 10.7].

When $\alpha$ and $\beta$ equal the identity and $\Omega$ has Lipschitz boundary, the operator $A$ is a Neumann operator whose domain satisfies $\text{dom}(A) \subseteq H^{3/2}(\Omega)$ by a result of [21]. Thus $\text{dom}(A)$ is strictly smaller than $\text{dom}(a) = \text{dom}(A^{1/2}) = H^1(\Omega)$ in this case. Therefore, if $\mathcal{V}_h$ is set to be the Lagrange finite element subspace of $H^1(\Omega)$, then Assumption 3 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 $\text{dom}(A)$ in Assumption 3.

Example 4.8 (Semibounded operators). Suppose $A$ is lower semibounded, i.e., there is a $\mu \in \mathbb{R}$ such that $(Ax, x)_\mathcal{H} \geq \mu (x, x)_\mathcal{H}$ for all $x$ in $\text{dom}(A)$. Then, by [30, Proposition 10.5],

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

An example of such an operator is the operator associated to the form $a$ in (52) when $\beta$ no longer satisfies $\beta > 0$, but instead changes sign while remaining bounded on $\Omega$. 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 4.6, $\text{dom}(a_\mu) = \text{dom}((A - \mu)^{1/2}) = H^1(\Omega)$. Hence by (53) we conclude that $\text{dom}(a) = H^1(\Omega)$.

Remark 4.9. We have already seen that there are two related, but distinct concepts, of the form associated to an operator (via the polar decomposition as in (51)) and the operator associated to a form (by the first representation theorem [22, Theorem VI.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 [16, Example 2.11]). However, $a$ and $\tilde{a}$ are equal if $a$ is a densely defined lower semibounded closed form by [30, Theorem 10.7].

Returning to the analysis of eigenvalue approximations for a general selfadjoint $A$, after having computed the approximate eigenspace $E_h \subset \mathcal{V}_h \subset \text{dom}(|A|^{1/2})$, we compute the spectrum of the finite-dimensional operator $A_{E_h}$,

$$A_h = \Sigma(A_{E_h}),$$

via the following variational eigenvalue problem: 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 $v_h \in E_h$. Thus, in practice, $\Lambda_h$ is computed by solving a small dense generalized eigenproblem arising from the above equation of forms.

We now examine how the numbers in $\Lambda_h$ differ from the exact eigenvalues in $\Lambda$. Define the Hausdorff distance between these sets by

$$\text{dist}(\Lambda, \Lambda_h) = \max \left[ \sup_{\mu \in \Lambda} \text{dist}(\mu, \Lambda_h), \sup_{\mu_h \in \Lambda_h} \text{dist}(\mu_h, \Lambda) \right]$$

where $\text{dist}(u, L) = \inf_{v \in L} |u - v|$ for any set $L \subset \mathbb{R}$. Bounds for the Hausdorff distance between Ritz values under perturbations in the space have been studied for bounded operators [23, Theorem 5.3] when Ritz values correspond to the top or bottom of the spectrum. Although we are not able to directly use this result, the following proof is inspired by their arguments. Unlike [23], we are only interested in an asymptotic estimate, allowing us to be cavalier with the constants in the estimates for brevity.

**Theorem 4.10.** Suppose Assumptions 1, 2, and 3 hold. Then there are positive constants $C_a$ and $h_1$ such that for all $h < h_1$,

$$\text{dist}(\Lambda, \Lambda_h) \leq C_a \text{ gap}_H(E, E_h)^2.$$  

**Proof.** By Theorem 4.2 and (47) we may choose $h$ so small that $\text{gap}_H(E, E_h) \leq \delta < 1$. Let $P = S$ and $Q$ denote the $\mathcal{H}$-orthogonal projectors onto $E$ and $E_h$, respectively, and put $R = (P - Q)^2$. Since $\|R\|_H \leq \text{gap}_H(E, E_h)^2 \leq \delta^2 < 1$, the binomial series $\sum_{n=0}^{\infty} (-1)^n (R)^n$ converges and defines $(I - R)^{-1/2}$. Subtracting the first term from this series, we get $T = (I - R)^{-1/2} - I$. Since $(1 - x)^{-1/2} - 1 = x [\sqrt{1 - x} + (1 - x)]^{-1}$, we obtain that

$$\|T\|_H \leq \sum_{n=1}^{\infty} \frac{(-1/2)^n \|R\|_H^n}{n} = (1 - \|R\|_H^{-1/2} - 1 = \|R\|_H \left[ \sqrt{1 - \|R\|_H} + (1 - \|R\|_H) \right]^{-1},$$

which implies

$$\|T\|_H \leq \|R\|_H \left[ \sqrt{1 - \delta^2} + (1 - \delta^2) \right]^{-1}. \tag{54}$$

It is standard (see e.g., [22, p. 33]) to use $R$ to define an isometry $W = (I - R)^{-1/2} [QP + (I - Q)(I - P)]$ on $\mathcal{H}$, which maps $E$ one-to-one onto $E_h$, and whose inverse is

$$W^{-1} = W^* = [QP + (I - P)(I - Q)] (I - R)^{-1/2}. \tag{55}$$

Note that the spectrum of $A_E$ and the unitarily equivalent $WA_E W^*|_{E_h}$ are identical.

Let $D = WA_E W^*|_{E_h} - A_{E_h}$, a selfadjoint operator on $E_h$. By [22, Theorem V.4.10],

$$\text{dist}(\Lambda, \Lambda_h) \leq \|D\|_{E_h} \|H\| = \sup_{0 \neq f \in E_h} \frac{|(Df, f)_H|}{(f, f)_H}. \tag{56}$$

For $f \in E_h$, we have

$$(Df, f)_H = a(W^*f, W^*f) - a(f, f) = a(PW^*f, PW^*f) - a(Qf, Qf) = \text{Re } a(PW^*f + Qf, PW^*f - Qf).$$

Observing that (55) implies $PW^* = PQ(I - R)^{-1/2}$, we split

$$(Df, f)_H = \text{Re } a((PW^* + Q)f, PQ [(I - R)^{-1/2} - I] f) + \text{Re } a((PW^* + Q)f, (P - I)Qf).$$
Labelling the two terms on the right as $t_1$ and $t_2$, we proceed to estimate them. The first is estimated using (54):

$$|t_1| = |\text{Re} a((PW^* + Q)f, PQTf)| \leq c_a \|PW^* + Q\|_{\mathcal{H}} \frac{\|PQ\|_{\mathcal{H}} \|R\|_{\mathcal{H}} \|f\|^2_{\mathcal{H}}}{\sqrt{1-\delta^2} + (1-\delta^2)}$$

$$\leq \frac{2c_a}{\sqrt{1-\delta^2} + (1-\delta^2)} \text{gap}_{\mathcal{H}}(E, E_h)^2 \|f\|^2_{\mathcal{H}},$$

where $c_a$ is any constant satisfying $|a(f, g)| \leq c_a \|f\|_{\mathcal{H}} \|g\|_{\mathcal{H}}$ for all $f, g \in E + E_h$. Before estimating $t_2$, recall that $P$ commutes with $A$, so it commutes with $|A|$ and $U$ [22, p.335ff], and moreover, it commutes with $|A|^{1/2}$ [22, Theorem V.3.35]. Hence for all $x, y \in \text{dom}(a)$, we have $a((P - I)x, y) = a(x, (P - I)y)$. Applying this to $t_2$,

$$|t_2| = |\text{Re} a((P - I)(PW^* + Q)f, (P - I)Qf)|$$

$$= |\text{Re} a((P - I)Qf, (P - I)Qf)| \leq c_a \|P - I\|_{H^1}^2 \|f\|^2_{H^1} \leq c_a \text{gap}_{\mathcal{H}}(E, E_h)^2 \|f\|^2_{\mathcal{H}}.$$

Adding the estimates for $|t_1|$ and $|t_2|$ and using it in (56), the proof is finished. \qed

5. A MODEL OPERATOR AND ITS DISCRETIZATION

The purpose of this section is to provide an example for application and illustration of the theory in the previous sections. The specific partial differential operator $A$ considered is the Dirichlet operator. The specific discretization considered is the DPG discretization. A practical reason for considering this discretization is that it allows one to approximate $R(z)$ by solving a sparse Hermitian positive definite system (even when $z - A$ is indefinite) using efficient iterative solvers. Another practical reason is that it offers a built-in (a posteriori) error estimator in the present approximation, thus immediately suggesting a straightforward technique for eigenpair error control. More pertinent to the theory in this paper is the additional reason that the DPG discretization offers an interesting example that generates a non-selfadjoint $S_N^h$ (and yet the Ritz values in $\Lambda_h$ converge at double the rate of convergence of the eigenspace $E_h$ of $S_N^h$).

5.1. The Dirichlet Operator. We focus on a simple model problem with the Laplace operator

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

where $\Omega \subset \mathbb{R}^d$ ($d = 2, 3$) is a bounded polyhedral domain with Lipschitz boundary. Note that here $\mathcal{V}$ is normed by $\|A^{1/2}u\|_{\mathcal{H}} = \|\text{grad} u\|_{L^2(\Omega)} = |u|_{H^1(\Omega)}$, 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

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

Hence this fits into Example 3.4, so Assumption 1 holds.

To calculate the application of the resolvent $u = R(z)v$, we need to solve the operator equation $(z - A)u = v$. In weak form, this equation may be stated as the problem of finding $u \in H_0^1(\Omega)$ satisfying

$$b(u, w) = (v, w)_{\mathcal{H}} \quad \text{for all } w \in H_0^1(\Omega),$$


$57$
where \( b(u, v) = z(u, v) = a(u, v) \). Let \( d(z) \) denote the distance from \( z \) to \( \Sigma(A) \) in \( \mathbb{C} \). A standard resolvent estimate for selfadjoint operators (see, e.g. [22, p. 272]) is

\[
\| R(z) v \|_{L^2(\Omega)} \leq d(z)^{-1} \| v \|_{L^2(\Omega)}.
\]

(58)

As a first step in the analysis, we obtain such an estimate in the \( \mathcal{V} \)-norm.

**Lemma 5.1.** For all \( v \in L^2(\Omega) \),

\[
\| R(z) v \|_{H^1(\Omega)}^2 \leq \left( \frac{|z|}{d(z)^2} + \frac{1}{d(z)} \right) \| v \|_{L^2(\Omega)}^2.
\]

(59)

**Proof.** Since the first estimate (58) is standard (see e.g. [22, §V.5]) we focus on proving the second. Letting \( u = R(z)v \) and putting \( w = u \) in (57),

\[
\| u \|_{H^1(\Omega)}^2 = z \| u \|_{L^2(\Omega)}^2 - \int_{\Omega} v \bar{u} \leq |z| \| u \|_{L^2(\Omega)}^2 + \| v \|_{L^2(\Omega)} \| u \|_{L^2(\Omega)}
\]

\[
\leq \frac{|z|}{d(z)^2} \| v \|_{L^2(\Omega)}^2 + \frac{1}{d(z)} \| v \|_{L^2(\Omega)}^2
\]

where we have used (58) in the last step.

\( \square \)

The following “inf-sup” condition plays a basic role in the study of any finite element discretization of problem (57).

**Theorem 5.2.** There exists a \( C > 0 \) depending only on \( \Omega \) such that for all \( v \in H^1_0(\Omega) \),

\[
c_0(z) \| v \|_{H^1(\Omega)} \leq \sup_{w \in H^1_0(\Omega)} \frac{|b(v, w)|}{\| w \|_{H^1(\Omega)}}
\]

where \( c_0(z) \) is defined by \( c_0(z)^{-1} = C \left[ 1 + |z|^2 d(z)^{-1} \left( 1 + (1 + |z|)d(z)^{-1} \right) \right] \).

**Proof.** Given any \( v \) in \( H^1_0(\Omega) \), consider the unique \( w \in H^1_0(\Omega) \) satisfying \( w = R(\bar{z}) \bar{w} \), or equivalently,

\[
\int_{\Omega} (z s \bar{w} - \nabla s \cdot \nabla \bar{w}) \, dx = \int_{\Omega} z s \bar{w}
\]

for all \( s \) in \( H^1_0(\Omega) \). Then by Lemma 5.1,

\[
\| w \|_{L^2(\Omega)} \leq \frac{1}{d(z)} \| \bar{w} \|_{L^2(\Omega)}, \quad \| w \|_{H^1(\Omega)} \leq \left( \frac{|z|}{d(z)^2} + \frac{1}{d(z)} \right)^{1/2} \| \bar{w} \|_{L^2(\Omega)}.
\]

Hence, with \( C_1(z)^2 = |z|^2 d(z)^{-1} \left( 1 + (1 + |z|)d(z)^{-1} \right) \), we have

\[
\| v - w \|_{H^1(\Omega)}^2 \leq 2 \| v \|_{H^1(\Omega)}^2 + 2 \| w \|_{L^2(\Omega)}^2 + \| w \|_{H^1(\Omega)}^2 \leq 2 \| v \|_{H^1(\Omega)}^2 + 2 C_1(z)^2 \| v \|_{L^2(\Omega)}^2
\]

\[
\leq 2 c_0(z)^2 \| v \|_{H^1(\Omega)}^2
\]

Moreover, (60) implies that \( b(v, v - w) = b(v, v) - b(v, w) = b(v, v) - z \| v \|_{L^2(\Omega)}^2 = - \| v \|_{H^1(\Omega)}^2 \).

Hence

\[
\sup_{w \in H^1_0(\Omega)} \frac{|b(v, w)|}{\| w \|_{H^1(\Omega)}} \geq \frac{|b(v, v - w)|}{\| v - w \|_{H^1(\Omega)}} \geq \frac{|v|^2_{H^1(\Omega)}}{\sqrt{2} c_0(z) \| v \|_{H^1(\Omega)}}
\]

so the proof is finished using the Poincaré inequality. \( \square \)
In view of the inf-sup condition of Theorem 5.2, one possible approach for discretizing the resolvent \( R(z) \) is by applying the standard Lagrange finite element method to (57). Its certainly possible to analyze this approximation. However we prefer to solve Hermitian positive definite systems instead of indefinite operators (like \( z - A \)). This motivates us to study a least-squares type finite element method for approximating the resolvent in the next subsection.

5.2. The DPG resolvent discretization. We now assume that \( \Omega \) is partitioned by a conforming simplicial finite element mesh \( \Omega_h \). As is usual in finite element theory, while the mesh need not be regular, the shape regularity of the mesh is reflected in the estimates.

To describe the DPG discretization of \( z - A \), we begin by introducing the nonstandard variational formulation on which it is based. We will be brief as the method is described in detail in previous works [11, 12]. Define

\[
H^1(\Omega_h) = \prod_{K \in \Omega_h} H^1(K), \quad Q = H(\text{div}, \Omega)/ \prod_{K \in \Omega_h} H_0(\text{div}, K).
\]

On every mesh element \( K \) in \( \Omega_h \), the trace \( q \cdot n|_{\partial K} \) is in \( H^{-1/2}(\partial K) \) for any \( q \) in \( H(\text{div}, K) \).

We denote by \( \langle q \cdot n, v \rangle_{\partial K} \) the action of this functional on the trace \( v|_{\partial K} \) for any \( v \) in \( H^1(K) \).

Next, for any \( u \in H^1_0(\Omega), q \in Q \) and \( v \in H^1(\Omega_h) \), set

\[
b_h((u, q), v) = \sum_{K \in \Omega_h} \int_K (zu\bar{v} - \text{grad } u \cdot \text{grad } \bar{v}) \, dx + \langle q \cdot n, \bar{v} \rangle_{\partial K}.
\]

This sesquilinear form gives rise to a well-posed Petrov-Galerkin formulation as will be clear from the discussion below.

For the DPG discretization, we use the following finite element subspaces. Let \( L_h \) denote the Lagrange finite element subspace of \( H^1_0(\Omega) \) consisting of continuous functions, which when restricted to any \( K \) in \( \Omega_h \) is in \( P_p(K) \) for some \( p > 1 \). Here and throughout \( P_\ell(K) \) denote the set of polynomials of total degree at most \( \ell \) restricted to \( K \). Let \( RT_h \subset H(\text{div}, \Omega) \) denote the well-known Raviart-Thomas finite element subspace consisting of functions whose restriction to any \( K \in \Omega_h \) is a polynomial in \( RT_p(K) = P_{p-1}(K)^n + xP_{p-1}(K) \), where \( x \) is the coordinate vector. Then we set \( Q_h = \{q_h \in Q : q_h|_K \in RT_p(K) + H_0(\text{div}, K)\} \). Finally, let \( Y_h \subset H^1(\Omega_h) \) consist of functions which when restricted to any \( K \in \Omega_h \) lie in \( P_{p+n+1}(K) \).

We now define the approximation of the resolvent action \( u = R(z)f \) by the DPG method, denoted by \( u_h = R_h(z)f \), for any \( f \in L^2(\Omega) \). The function \( u_h \) is in \( L_h \). Together with \( \varepsilon_h \in Y_h \) and \( q_h \in Q_h \), it satisfies

\begin{align}
\int_{\Omega} \varepsilon_h \bar{\eta}_h \, dx + b_h((u_h, q_h), \eta_h) &= \int_{\Omega} f \, \bar{\eta}_h \, dx, & \text{for all } \eta_h \in Y_h, \label{eqn:61a} \\
b_h((w_h, r_h), \varepsilon_h) &= 0, & \text{for all } w_h \in L_h, r_h \in Q_h. \label{eqn:61b}
\end{align}

The distance between \( u \) and \( u_h \) is bounded in the next theorem. In the theorem, note that unlike standard finite element methods there is no assumption that \( h \) is sufficiently small, i.e., the method is stable in the pre-asymptotic regime.

**Theorem 5.3.** Suppose \( z \) lies in a bounded set of diameter \( D \) in the complex plane. There is a \( C > 0 \) depending only on \( D \) and the shape regularity of the mesh \( \Omega_h \) such that for all \( f \in L^2(\Omega) \),

\[
\|R(z)f - R_h(z)f\|_V \leq C \frac{d(z)}{d(z)} \left[ \inf_{w_h \in L_h} \|u - w_h\|_{H^1(\Omega)} + \inf_{q_h \in RT_h} \|q - q_h\|_{H(\text{div}, \Omega)} \right]
\]
where \( u = R(z)f \) and \( q = \text{grad } u \).

Proof. The proof proceeds by verifying the sufficient conditions for convergence of DPG methods known in the existing literature. The result of [13, Theorem 2.1] immediately gives the stated result, provided we verify its three conditions, reproduced below. The first is the uniqueness condition

\[
\{(w, r) \in H^1_0(\Omega) \times Q : b_h((w, r), \eta) = 0, \text{ for all } \eta \in H^1(\Omega) \} = \{0\}.
\]

The second condition is that there are \( C_1, C_2 > 0 \) such that

\[
C_1 \|\eta\|_{H^1(\Omega)} \leq \sup_{(w, r) \in H^1_0(\Omega) \times Q} \frac{|b_h((w, r), \eta)|}{\|w\|_{H^1_0(\Omega)} \|r\|_{H^1_0(\Omega)} \|\eta\|_{H^1(\Omega)}} \leq C_2 \|\eta\|_{H^1(\Omega)}
\]

for all \( \eta \in H^1(\Omega) \). The third condition is the existence of a bounded linear operator \( \Pi_h : H^1(\Omega) \rightarrow Y_h \) such that

\[
b_h((w_h, r_h), \eta - \Pi_h \eta) = 0.
\]

Once these conditions are verified, the estimate of the theorem follows with \( C/d(z) = C_2/\|\Pi\|/C_1 \).

It is possible to verify conditions (62a) and (62b) on \( b_h(\cdot, \cdot) \) using the properties of \( b(\cdot, \cdot) \). Specifically, using [8, eq. (10b)] and [8, Theorem 3.3], we conclude that the inf-sup condition for \( b \) that we proved in Theorem 5.2 implies an inf-sup condition for \( b_h \), namely the lower inequality of (62b) holds with

\[
\frac{1}{C_1^2} = \frac{1}{c_0(z)^2} + \left( \frac{1 + |z|}{c_0(z)} + 1 \right)^2.
\]

Combining with the continuity estimate of \( b_h \) with \( C_2 = 1 + |z| \), we obtain that \( C_2/C_1 \) is \( O(d(z)^{-1}) \). Finally, Condition (62c) follows from the operator constructed in [13, Lemma 3.2] whose norm is a constant bounded independently of \( z \).

5.3. FEAST iterations with the DPG discretization. To approximate \( E \), we enclose the eigenvalues of interest, \( \lambda_1, \ldots, \lambda_m \), within a circular or ellipsoidal contour \( \Gamma \), choose any of the corresponding filters in (26c), and apply the filtered subspace iteration. By bounding the approximation estimates on the right hand side of the estimate of Theorem 5.3, we can prove convergence. To this end, we need the next regularity assumption:

Assumption 4. Suppose there are constants \( C_{\text{reg}}, s > 0 \) such that the solution \( u^f \in V \) of the Dirichlet problem \(-\Delta u^f = f\) admits the regularity estimate

\[
\|u^f\|_{H^{1+s}(\Omega)} \leq C_{\text{reg}} \|f\|_V \quad \text{for any } f \in V.
\]

Suppose also that there is a number \( s_E \geq s \) such that

\[
\|u^f\|_{H^{1+s_E}(\Omega)} \leq C_{\text{reg}} \|f\|_V \quad \text{for any } f \in E.
\]

Standard regularity results for elliptic operators (see, e.g. [14, 15]) yield that \( \text{dom}(A) \supset H^{1+s}(\Omega) \) for some \( s > 0 \) depending on the geometry of \( \Omega \). For example, if \( \Omega \) is convex, we may take \( s = 1 \) in (63); and if \( \Omega \subset \mathbb{R}^2 \) is non-convex, with its largest interior angle at a corner being \( \pi/\alpha \) 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 (64). 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 $\Omega_L = (0, 2) \times (0, 2) \setminus [1, 2] \times [1, 2]$, also yield smooth eigenfunctions of $\Omega_L$. But not all eigenfunctions of $\Omega_L$ are so regular.

Set $S_N^h$ by (39) with $R_h(z)$ equal to the DPG discretization of the resolvent. We start the iterations with a subspace $E_h^{(0)}$ of the Lagrange finite element subspace $V_h = L_h$ satisfying $\dim(P_hE_h^{(0)}) = m$. Then compute

$$E_h^{(\ell)} = S_N^hE_h^{(\ell-1)}, \quad \ell = 1, 2, \ldots.$$  \hspace{1cm} (65)

**Theorem 5.4.** Suppose Assumption 4 holds. Then, there is an $h > 0$ such that for all $h < h_0$, the subspace iterates $E_h^{(\ell)}$ converge (in gap$_V$) to a space $E_h$ satisfying

$$\text{gap}_V(E, E_h) \leq C h^\min(p,s_E)$$ \hspace{1cm} (66)

$$\text{dist}(\Lambda, \Lambda_h) \leq C h^2\min(p,s_E)$$ \hspace{1cm} (67)

where $C$ is independent of $h$ (but depends on $W$, $C_N$, $p$, $d(z_k)$, $\lambda_i$, $C_{\text{reg}}$, and the shape regularity of the mesh).

**Proof.** This proof proceeds by verifying Assumption 2 and applying the prior general results. For any $z \in \Gamma$, we apply the estimate of Theorem 5.3 to $u = R(z)v$ and $u_h = R_h(z)v$. Then we use the standard finite element approximation estimates for the Lagrange and Raviart-Thomas spaces to get

$$\|u - u_h\|_{H^1(\Omega)} \leq \frac{C}{d(z)} \left[ h^r \|u\|_{H^{1+r}(\Omega)} + h^r \|q\|_{H^r(\Omega)} + h^r \|\text{div} q\|_{H^r(\Omega)} \right].$$ \hspace{1cm} (68)

for $r \leq p$. Throughout this proof, we use $C$ to denote a generic constant with the properties stated in the theorem, but whose value in each occurrence may differ.

Note that function $u = R(z)v$ satisfies (57). Hence by Theorem 5.2, $c_0(z)\|u\|_{H^1(\Omega)} \leq \|v\|_{L^2(\Omega)}$ where $c_0(z) = O(d(z))$. Moreover, since $\Delta u = v - zu$, the regularity assumption (63) implies that

$$\|u\|_{H^{1+r}(\Omega)} \leq \frac{C}{d(z)} \|v\|_{H^1(\Omega)}$$ \hspace{1cm} (69)

for any $r \leq s$. Since $q = \text{grad} u$ and $\text{div} q = v - zu$, the estimate (68) now yields

$$\|u - u_h\|_{H^1(\Omega)} \leq \frac{C}{d(z)} \left[ h^r \|u\|_{H^{1+r}(\Omega)} + h^r \left( \|v\|_{H^r(\Omega)} + \|z\|_{H^r(\Omega)} \right) \right]$$ \hspace{1cm} (70)

for $r \leq \min(p,s)$. Restricting $r = \min(p,s,1)$ so that $\|v\|_{H^r(\Omega)} \leq C\|v\|_{H^1(\Omega)}$, the regularity estimate (69) implies that $\|u - u_h\|_{H^1(\Omega)} \leq Cd(z)^{-2}h^r\|v\|_{H^1(\Omega)}$, i.e.,

$$\|R(z)v - R_h(z)v\|_{H^1(\Omega)} \leq \frac{C}{d(z)^2} h^\min(p,s,1) \|v\|_{H^1(\Omega)}$$ \hspace{1cm} (71)

for all $v \in V$.

This verifies (40) of Assumption 2. Hence by Theorem 4.4, the iterates $E_h^{(\ell)}$ of (65) converges to $E_h$.

Next, we apply Theorem 4.2, from which (44) follows. To bound the right hand side of (44), we repeat the argument leading to (71) for $v \in E$, using the enhanced regularity assumption (64) for the eigenfunctions. Thus we obtain (70) with $r = \min(p,s_E)$. Now, writing $v$ as a linear combination of a fixed basis of $E$, viewing it as the solution of a
Dirichlet problem with an \( f \in E \), and using the eigenspace regularity estimate (64) for each eigenfunction, we obtain \( \|v\|_{H^r(\Omega)} \leqslant \|v\|_{H^{1+r}(\Omega)} \leqslant C \|v\|_{H^1(\Omega)} \). Thus,

\[
(72) \quad \| R(z)v - R_h(z)v \|_{H^1(\Omega)} \leqslant \frac{C}{d(z)2} h_{\min(p,s_E)} \|v\|_{H^1(\Omega)} \quad \text{for all } v \in E
\]

and all \( z \in \Gamma \). Using this bound for the right hand side of (44), the proof of (66) is complete.

Finally, note that Assumption 3 also holds (see Example 4.6). Hence Theorem 4.10, combined with (66), yields (67).

\[ \square \]

**Remark 5.5.** Consider a variant of (65) where the mesh changes at each iteration, for instance, in an adaptive algorithm. Then, the mesh size \( h_\ell \) varies at each iteration \( \ell \), and we get approximate eigenspace iterates \( E_h^{(\ell)} = S_N h_\ell E_h^{(\ell-1)} \) for \( \ell = 1, 2, \ldots \). We may apply Theorem 3.8 to analyze (65), viewing it as a perturbed subspace iteration, with \( \Delta_\ell = S_N - S_N h_\ell \). As an example, consider using Filter 1, for which \( \|S_N\|_V > 1/2 \) (by (14) and Lemma 3.7). Then (41), (50), and (71) imply that the estimate of Theorem 3.8 holds with

\[
\epsilon \leqslant \frac{2CW}{\min_{k=1,\ldots,N} d(z_k)^2} h_{\min(p,s_E)}.
\]

Such estimates can be extended to incorporate errors in iterative solvers or other linear algebra errors in computing the action of \( R_h(z) \).

### 6. Numerical Experiments

We implemented the FEAST iteration with the DPG approximation of the spectral projector considered in Section 5.2. All numerical experiments are conducted using [1], which builds a hierarchy of Python classes representing approximations of spectral projectors. The DPG discretization is implemented using a python interface into an existing well-known C++ finite element library called NGSolve [31]. We do not write out the details of the FEAST algorithm because they can be found in our public code [1] or in many previous papers (see e.g., [28, Algorithm 1.1] and [18]). 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 our experiments reported below, we do so. 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 per iteration for Filters 1 and 3. Similarly, only \( N/2 + 2 \) solves are needed for Filters 2 and 4.

In Subsection 6.1 we study iteration errors for a simple matrix eigenproblem, considering the implications of Theorem 3.8 concerning eigenvalue errors. Here, the subspace iteration is started with a random subspace having the dimension of the true eigenspace, and run until the errors in the eigenvalue approximations are on the order of machine precision. All four filters are considered for this example. In Subsections 6.2-6.4 we consider convergence rates with respect to the discretization parameter \( h \) for the DPG discretization, illustrating the estimates in Theorem 5.4. For these experiments, we use Filter 1, start the subspace iterations with a random subspace of overestimated dimension, and let the algorithm truncate the basis vectors that generate Ritz values outside of the prescribed contour \( \Gamma \). We stop the iterations when successive Ritz values differ by less than \( 10^{-9} \).
6.1. Iterative convergence for a tridiagonal matrix. We apply the FEAST iteration on the matrix eigenvalue problem \( Ax = \lambda x \) where \( A \) is the \( n \times n \) tridiagonal matrix having stencil \((-1,2,-1)\). Up to scaling, this matrix arises from a uniform-grid finite-difference discretization of the 1D Dirichlet Laplace eigenproblem, but in this subsection we shall ignore the discretization error because our focus is on iterative error. The eigenvalues \( \lambda_j \) of \( A \), and corresponding eigenvectors \( e_j \), are known explicitly:

\[
\lambda_j = 2 - 2 \cos(\pi j/(n+1)) \quad , \quad [e_j]_i = \sqrt{2/(n+1)} \sin(\pi i j/(n+1)) .
\]

The eigenvectors are scaled so that they have unit length in \( \mathbb{R}^n \), \( \|e_j\|^2 = e_j^T e_j = 1 \). We also introduce the associated energy norm, \( \|x\|^2 = x^T A x \).

Setting \( n = 100 \) we consider the eigenvalues in the interval \( (y-\gamma, y+\gamma) \), where \( y = 1/3 \) and \( \gamma = 1/18 \). These eigenvalues, per our above enumeration, are \( \Lambda = \{\lambda_j : j \in J\} \) where \( J = \{18,19,20\} \). We study the performance of the FEAST algorithm when it iteratively finds \( \Lambda \). At the \( \ell \)th step, the algorithm computes a basis for the approximate eigenspace consisting of three unit vectors \( e_j^{(\ell)} \) for \( j \in J \). Corresponding eigenvalue approximations are obtained by \( \lambda_j^{(\ell)} = \|e_j^{(\ell)}\|^2 \). Note that

\[
(73) \quad \lambda_j^{(\ell)} - \lambda_j = \|e_j - e_j^{(\ell)}\|^2 - \lambda_j \|e_j - e_j^{(\ell)}\|^2 .
\]

Theorem 3.8 is applicable here with \( \Delta_{\ell} = 0 \) and with \( \| \cdot \|_Y \) set to either \( \| \cdot \| \) or \( \| \cdot \|_V \). It guarantees that there are \( q_j^{(\ell)} \) in the span of \( \{e_j^{(\ell)} : j \in J\} \) satisfying \( \|e_j - q_j^{(\ell)}\|_Y \leq \kappa_j^{(\ell)} \|e_j - q_j^{(0)}\|_V \). One might expect \( \|e_j - q_j^{(\ell)}\|_Y \) to converge similarly, and therefore by (73), \( \lambda_j^{(\ell)} \) to converge to \( \lambda_j \) at the squared rate \( \kappa_j^2 \). This is precisely the convergence behavior we observe for this experiment, as described below.

Using \( N = 8 \) and \( \rho = 3/2 \) (as in Figures 1-2), we compute eigenvalue iterative errors \( \text{ERR}(\ell) = |\lambda_j - \lambda_j^{(\ell)}| \) and error ratios \( \text{RAT}(\ell) = \text{RAT}(\ell) = \text{ERR}(\ell)/\text{ERR}(\ell-1) \) for each eigenvalue in \( \Lambda \) and each of the four filters. A subset of this data is given in Table 1. The ratio \( \text{RAT}(\ell) \) represents the numerically observed contraction rate for the eigenvalue error at step \( \ell \). Ratios that agree with the exact contraction constants \( \kappa_j^2 \) to at least four significant digits are given in bold in the table. This experiment suggests that the iterative error bound from Theorem 3.8 in the case of zero perturbations may be asymptotically sharp.

As mentioned early on, \( \kappa_i \leq \hat{\kappa} \). For this reason, \( \hat{\kappa} \) is often referred to as the worst-case convergence factor (see [18, Equation 5.1]). For this example, we have \( \delta = \min_{\lambda \in \Sigma(4) \setminus \Lambda} |y - \lambda|/\gamma - 1 = |y - \lambda_{17}|/\gamma - 1 \approx 0.0832 \), which yields the squares of the worst-case convergence factors \( \hat{\kappa}^2 \) for Filters 1-4 reported in the second column of the table. They clearly overestimate the individual squared contraction factors \( \kappa_j^2 \) (in bold). Hence we conclude that \( \hat{\kappa} \) provides a pessimistic assessment of the practical convergence behavior.

6.2. Discretization errors on the unit square. Set \( \Omega = (0,1) \times (0,1) \) and consider the circular contour with radius \( \gamma = 40 \) and center \( y = 20 \). The exact eigenvalues of the Dirichlet operator of Section 5 within this contour are known to be \( \lambda_1 = 2\pi^2 \) (of multiplicity 1) and \( \lambda_2 = \lambda_3 = 5\pi^2 \) (of multiplicity 2), i.e., \( \Lambda = \{2\pi^2,5\pi^2\} \). The corresponding exact eigenfunctions \( e_i \) for \( i = 1,2, \) and 3, span \( E \). The well-known expressions for \( e_i \) show that they are in \( H^s(\Omega) \) for any \( s > 0 \), so the estimates of Theorem 5.4 hold with any \( s_E > 0 \).

We begin the numerical studies using a coarse triangulation \( \mathcal{Q}_h \) of \( \Omega \) whose element diameters are not larger than \( 1/4 \). By connecting the midpoints of edges, each mesh triangle can be split into four congruent sub-triangles, thus refining the mesh to obtain another mesh of
Table 1. Eigenvalue iterative errors (ERR) and observed contraction rates (RAT) for each of the eigenvalues $\Lambda_1, \Lambda_2, \Lambda_3$ and each of the four filters for the tridiagonal matrix example.

<table>
<thead>
<tr>
<th>Filter</th>
<th>$\hat{\kappa}^2$</th>
<th>ERR $\Lambda_1$</th>
<th>RAT $\Lambda_1$</th>
<th>ERR $\Lambda_2$</th>
<th>RAT $\Lambda_2$</th>
<th>ERR $\Lambda_3$</th>
<th>RAT $\Lambda_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filter 3</td>
<td>1.563e-01</td>
<td>3.706e-01</td>
<td>3.706e-06</td>
<td>3.706e-06</td>
<td>3.706e-06</td>
<td>3.706e-06</td>
<td>3.706e-06</td>
</tr>
</tbody>
</table>

Figure 4. Results for the unit square
maximal element diameter \( h/2 \). Repeating this process, we obtain a sequence of finer meshes. When the FEAST algorithm using the DPG resolvent approximation of degree \( p \) (see Section 5.2) exits, it returns approximate eigenpairs \((\lambda_{i,h}, e_{i,h})\) on each such mesh. The values \( \lambda_{i,h} \) form the set \( \Lambda_h \). The span of the functions \( e_{i,h} \) is taken to be the computed approximate eigenspace \( E_h \). Let

\[
\delta_i = \min_{0 \neq e \in E} |e_{i,h} - e|_{H^1(\Omega)}.
\]

Since \( \delta_i/|e_{i,h}|_{H^1(\Omega)} \leq \text{gap}_p(E, E_h) \), we expect \( \delta_i \) to decrease with \( h \) per the estimate (66) of Theorem 5.4. Since \( E \) does not consist of finite element functions, instead of computing \( \delta_i \) exactly, we compute and report its approximation \( \delta_{i}^{h} \) obtained by replacing \( e \) in (74) by its interpolant into the finite element space. Our numerical observations indicate that each \( \delta_{i}^{h} \) goes to 0 at the rate \( O(h^p) \) for \( i = 1, 2, 3 \). For brevity, we only include one plot showing how the sum \( \delta_1^{h} + \delta_2^{h} + \delta_3^{h} \) behaves for decreasing \( h \) in Figure 4a. These results are obtained using Filter 1 (set by (10) and (11a)) with \( N = 8 \).

We also studied how \( |\lambda_i - \lambda_{i,h}| \) converges. We observed that as \( h \to 0 \), the eigenvalue error \( |\lambda_i - \lambda_{i,h}| \to 0 \) at the rate \( O(h^{2p}) \). Each of these errors are bounded by the Hausdorff distance \( \text{dist}(\Lambda, \Lambda_h) \), which is displayed in Figure 4b. Clearly, these convergence rates are in accordance with (67) of Theorem 5.4.

6.3. Convergence rates on an L-shaped domain. The error estimates of Theorem 5.4 are expressed in terms of the worst-case regularity of functions in the eigenspace \( E \) to be approximated. The following example illustrates that, when the eigenspace \( E \) associated with a given filter contains eigenfunctions of different regularities, the convergence rates of individual eigenfunction and eigenvalue approximations are not necessarily dictated by the worst-case regularity in \( E \), but may converge according to the regularities of the associated eigenfunctions.

Let \( \Omega = (0, 2) \times (0, 2) \setminus [1, 2] \times [1, 2] \) be the L-shaped domain. For this domain, the eigenvalues and eigenfunctions are not explicitly known in general, though \( \lambda = (m^2 + n^2)\pi^2 \) is an eigenvalue with eigenvector \( e = \sin(m\pi x)\sin(n\pi y) \) for each \( m, n \geq 1 \), as noted in Section 5.3. Although we are guaranteed that all eigenfunctions \( e \) satisfy \( e \in H^s(\Omega) \) for any \( s < 1 + 2/3 \), the regularity of any particular eigenfunction for this domain cannot generally be determined a priori.

In [33] the first several eigenvalues for the L-shaped domain are computed to high accuracy. The first four of their computed values up to the significant digits reported there are 9.6397238, 15.197252, 19.739209 (\( 2\pi^2 \)), and 29.521481. We use the first three as the “true eigenvalues” for our convergence study. We note that \( \lambda_3 = 2\pi^2 \) and \( e_3 = \sin(\pi x)\sin(\pi y) \), so the third eigenfunction is smooth. Taking \( p = 2 \), we consider the individual eigenvalue errors.

<table>
<thead>
<tr>
<th>( h )</th>
<th>( \lambda_1 ) ERR NOC</th>
<th>( \lambda_2 ) ERR NOC</th>
<th>( \lambda_3 ) ERR NOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>2(^{-2})</td>
<td>4.85e-02</td>
<td>1.34e-02</td>
<td>2.36e-02</td>
</tr>
<tr>
<td>2(^{-3})</td>
<td>2.01e-02 1.27</td>
<td>2.18e-03 2.61</td>
<td>3.76e-03 2.65</td>
</tr>
<tr>
<td>2(^{-4})</td>
<td>7.74e-03 1.37</td>
<td>1.97e-04 3.47</td>
<td>2.36e-04 3.99</td>
</tr>
<tr>
<td>2(^{-5})</td>
<td>3.05e-03 1.34</td>
<td>2.18e-05 3.17</td>
<td>1.48e-05 3.99</td>
</tr>
<tr>
<td>2(^{-6})</td>
<td>1.21e-03 1.34</td>
<td>2.81e-06 2.96</td>
<td>9.27e-07 4.00</td>
</tr>
</tbody>
</table>

Table 2. Eigenvalue errors (ERR) and numerical order of convergence (NOC) for the smallest three eigenvalues on the L-shaped domain.
\[ |\lambda_i - \lambda_{i,h}|, \ i = 1, 2, 3, \] for a sequence of uniformly refined meshes, as we did for the square domain, starting with an initial quasi-uniform mesh having characteristic element diameter \( h = 1/4 \). We use Filter 1, with \( y = 15, \gamma = 6, \) and \( N = 4 \) for the FEAST iteration. For each refinement level \( h \), we record the eigenvalue errors \( \text{ERR} = \text{ERR}(h) \), and numerical order of convergence \( \text{NOC} = \text{NOC}(h) = \log(\text{ERR}(2h)/\text{ERR}(h))/\log(2) \), in Table 2. It is clear that the three eigenvalue approximations are converging at different rates, with the apparent convergence order for \( \lambda_1 \) being approximately \( 4/3 \), which is consistent with \( \epsilon_1 \) having almost \( H^{1+2/3}(\Omega) \)-regularity. Since \( p = 2 \), the highest convergence order that can be expected is 4, and the apparent convergence order for \( \lambda_3 \) is consistent with this.

### 6.4. Schrödinger operator.

In this experiment, we consider the Schrödinger operator with a potential \( V \), set by

\[
\mathcal{H} = L^2(\mathbb{R}^2), \quad A = -\Delta + V, \quad V = -50e^{-(x^2+y^2)}
\]

In mathematical physics, the eigenmodes of such operators are of interest, and we are guided by the known theoretical results there. It follows from [27, Theorem XIII.15] that the essential spectrum of \( A \) is \([0, \infty)\), and it is clear that \( \Sigma(A) \) is bounded below by \(-50\). Away from the essential spectrum, we search for eigenvalues of finite multiplicity (or localized “bound states”) along the negative real axis. Because the potential \( V \) is negative, there is at least one negative eigenvalue, and an estimate of Chadan et al. [9, Equation (44)] guarantees that the joint multiplicity of all negative eigenvalues is finite.

Since it is infeasible to compute on an unbounded domain with finite elements, a standard approximation technique is to consider \( A \) with zero Dirichlet boundary conditions on the bounded domain \([-D, D]^2\) for sufficiently large \( D \). Because the essential spectrum of \( A \) is \([0, \infty)\) and there are finitely many negative eigenvalues, we conclude that the eigenvalues of the operator restricted to \([-D, D]^2\) will accumulate at zero only from above as \( D \to \infty \).

We used Filter 1 with \( N = 8 \) on the circular contour of center \( y = -40 \) and radius \( \gamma = 30 \). Mesh element diameters were constrained to be never more than \( h = 0.4 \). We performed several experiments varying \( D \) and \( p \), starting each one with a subspace spanned by 10 random vectors. In all cases, the iterations converged to a six-dimensional eigenspace. The basis functions for this eigenspace, computed in the case \( D = 5 \) and \( p = 5 \), denoted by \( e_1, \ldots, e_6 \), are displayed in Figure 5 in both surface and patch plots. The surface plots of the eigenfunctions are overlaid with that of the potential well \( V \) to illustrate the localized behaviour of the eigenfunctions. Table 3 reports the corresponding eigenvalues in this case as well as for higher \( p \). The first and the fourth eigenvalues seem to be simple, while the remaining eigenvalues seem to be approximating two distinct eigenvalues each of multiplicity two. Doubling \( D \) from 5 to 10 did not change these eigenvalues up to seven significant digits, as shown in the last row of the table.

<table>
<thead>
<tr>
<th>( p )</th>
<th>( D )</th>
<th>( \lambda_1 )</th>
<th>( \lambda_2 )</th>
<th>( \lambda_3 )</th>
<th>( \lambda_4 )</th>
<th>( \lambda_5 )</th>
<th>( \lambda_6 )</th>
</tr>
</thead>
</table>

Table 3. Eigenvalues corresponding to the bound Schrödinger states.
Figure 5. Different views of the eigenmodes of the two-dimensional Schrödinger example. (The first row shows surface plots of the real part of the eigenmodes at their corresponding eigenvalue heights, overlaid with the potential well. The second row shows patch plots of the same functions.)

Appendix A. Remarks on the Zolotarev Filter

In the recent work of [18], the authors build on results of Zolotarev [34] to construct a rational approximation \( r_m^Z \) of type \((2m, 2m)\) of the indicator function \( \text{ind}_{[-G,G]} \) for the interval \([-G,G]\), where \( 0 < G < 1 \). The number \( G \) here is not to be confused with the operator \( G \) in Section 3. More specifically, it is proved in [18, Corollary 4.2] that \( r_m^Z \) is the best uniform rational approximation of type \((2m, 2m)\) for \( \text{ind}_{[-G,G]} \) on \( \omega = (-\infty, -G^{-1}] \cup [-G, G] \cup [G^{-1}, \infty) \). Furthermore,

\[
E'_m = \max_{x \in \omega} |\text{ind}_{[-G,G]}(x) - r_m^Z(x)| = \max_{|x| \leq G} |1 - r_m^Z(x)| = \max_{|x| \geq G^{-1}} |r_m^Z(x)| .
\]

The value \( E'_m \) is attained at \( \pm G, \pm G^{-1}, \pm \infty \), and at \( 4(m - 1) \) other points. The poles of \( r_m^Z \) all lie on the unit circle, and come in conjugate pairs.

It is clear from the above discussion that \( r_m^Z \) does not decay at infinity, in contrast to filters based on the trapezoid rule, for example. The authors of [18] identify this lack of decay at infinity of \( r_m^Z \) as a reason that the trapezoid approach (or a related approach based on Gaussian quadrature) may yield more rapid FEAST convergence than the Zolotarev approach in some cases. In a footnote, the authors mention a vertically-shifted version of \( r_m^Z \) that vanishes at infinity, which we denote by \( r_m^{SZ} \),

\[
r_m^{SZ} = r_m^Z + (-1)^m E'_m ,
\]

and we consider this variant as well in the discussion below.

The relative gap \((G^{-1} - G)/G = G^{-2} - 1\) plays the same role as our relative spectral gap \( \delta \). By shifting and scaling our interval of interest, we can make the identifications \( \gamma = G \) and \( \delta = G^{-2} - 1 \). Having done so, it becomes clear that what is called the worst-case convergence factor in [18, Equation 5.1] is precisely the contraction factor \( \tilde{\kappa} \) in (8). In Table 4, we provide
Table 4. Worst-case convergence factors, $\hat{\kappa}^E$, $\hat{\kappa}^Z$, $\hat{\kappa}^{SZ}$, for (17b) and the regular and shifted Zolotarev filters, for various values of $G$ (with $\delta = G^{-2} - 1$ given in parentheses). We have taken $\rho = 1.01$ for $\hat{\kappa}^E$, and $N = 2m$.

<table>
<thead>
<tr>
<th>$N$</th>
<th>$G (\delta)$</th>
<th>$\hat{\kappa}^E$</th>
<th>$\hat{\kappa}^Z$</th>
<th>$\hat{\kappa}^{SZ}$</th>
<th>$G (\delta)$</th>
<th>$\hat{\kappa}^E$</th>
<th>$\hat{\kappa}^Z$</th>
<th>$\hat{\kappa}^{SZ}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>1.292e-5</td>
<td>5.126e-4</td>
<td>1.026e-3</td>
<td>2.515e-3</td>
<td>5.838e-4</td>
<td>7.377e-3</td>
<td>1.486e-2</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>6.191e-11</td>
<td>1.312e-7</td>
<td>2.625e-7</td>
<td>2.491e-6</td>
<td>2.682e-8</td>
<td>5.363e-5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>0.50</td>
<td>1.178e-21</td>
<td>8.613e-15</td>
<td>0.75</td>
<td>1.976e-12</td>
<td>3.595e-10</td>
<td>7.191e-10</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>(3.0000)</td>
<td>5.052e-27</td>
<td>2.206e-18</td>
<td>(7.778e-1)</td>
<td>1.724e-15</td>
<td>1.316e-12</td>
<td>2.633e-12</td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>9.269e-72</td>
<td>2.599e-48</td>
<td>5.198e-48</td>
<td>5.381e-41</td>
<td>6.557e-33</td>
<td>1.311e-32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>4.247e-4</td>
<td>3.700e-4</td>
<td>7.397e-4</td>
<td>1.013e-3</td>
<td>5.838e-4</td>
<td>1.167e-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>5.187e-6</td>
<td>5.030e-6</td>
<td>1.006e-5</td>
<td>1.934e-5</td>
<td>9.966e-6</td>
<td>1.993e-5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>0.88</td>
<td>6.143e-8</td>
<td>6.340e-8</td>
<td>0.90</td>
<td>3.572e-7</td>
<td>1.702e-7</td>
<td>3.404e-7</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>(2.913e-1)</td>
<td>7.190e-10</td>
<td>9.302e-10</td>
<td>(2.346e-1)</td>
<td>6.516e-9</td>
<td>2.907e-9</td>
<td>5.814e-9</td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>5.460e-26</td>
<td>2.597e-25</td>
<td>5.194e-25</td>
<td>1.978e-23</td>
<td>5.423e-24</td>
<td>1.085e-23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>4.001e-1</td>
<td>1.362e-1</td>
<td>3.155e-1</td>
<td>8.916e-1</td>
<td>3.578e-1</td>
<td>1.114</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>0.98</td>
<td>3.346e-3</td>
<td>2.741e-5</td>
<td>0.998</td>
<td>2.828e-1</td>
<td>8.268e-4</td>
<td>1.652e-3</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>(4.123e-2)</td>
<td>6.316e-4</td>
<td>1.667e-6</td>
<td>(4.012e-3)</td>
<td>1.757e-1</td>
<td>1.178e-4</td>
<td>2.356e-4</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>1.370e-7</td>
<td>1.390e-12</td>
<td>2.780e-12</td>
<td>1.481e-2</td>
<td>6.936e-9</td>
<td>1.387e-8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>4.875e-10</td>
<td>1.231e-16</td>
<td>2.463e-16</td>
<td>2.769e-3</td>
<td>1.050e-11</td>
<td>2.100e-11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>9.552e-1</td>
<td>1.106e-1</td>
<td>1.992e-1</td>
<td>9.954e-1</td>
<td>2.069e-1</td>
<td>3.429e-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>0.9998</td>
<td>8.400e-1</td>
<td>5.086e-3</td>
<td>0.99998</td>
<td>9.818e-1</td>
<td>1.588e-2</td>
<td>3.126e-2</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>(4.001e-4)</td>
<td>7.680e-1</td>
<td>1.136e-3</td>
<td>(4.000e-5)</td>
<td>9.719e-1</td>
<td>4.670e-3</td>
<td>9.384e-3</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>4.222e-1</td>
<td>6.439e-7</td>
<td>1.288e-7</td>
<td>8.970e-1</td>
<td>1.080e-5</td>
<td>2.161e-5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>2.663e-1</td>
<td>4.413e-9</td>
<td>8.827e-9</td>
<td>8.312e-1</td>
<td>1.896e-7</td>
<td>3.971e-7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A comparison of the contraction factors related to the ellipse filter (17b), with those related to $r^Z_m$ and $r^{SZ}_m$. More specifically, we compare

$$\hat{\kappa}^E = \frac{\rho^N + \rho^{-N} + 2}{2TN\left(\frac{a^2 + a^{-1}}{2N^2}\right) - (\rho^N + \rho^{-N})}, \quad \hat{\kappa}^Z = \frac{E'_m}{1 - E'_m}, \quad \hat{\kappa}^{SZ} = \frac{2E'_m}{1 - (1 - (-1)^m)E'_m}.$$

The comparisons in Table 4 are done for various values of $G$ and $N = 2m$, and include the values of $G$ and $m$ that appear in [18, Table 1]. For $\hat{\kappa}^E$ there is no optimal value of $\rho$, so we choose $\rho = 1.01$, similar to the choice $S = 1.01$ in [18, Table 1] for many of the ellipse-based filters they consider. We see that the Zolotarev contraction factors tend be smaller than $\hat{\kappa}^E$ for $G \geq 0.9$, but the reverse holds for more modest $G$. We also see that it is possible for $\hat{\kappa}^{SZ}$ to exceed 1 when $G$ is sufficiently close to 1 and $m$ is odd.

**Remark A.1.** In [18, Corollary 4.2], upper and lower bounds are provided for $E'_m$ based on similar results from [24] for errors in the rational approximation of the sign function. Although these bounds hold for nearly all cases considered in Table 4, they are not valid in general. More specifically, the bounds given in [18, Corollary 4.2] imply that, for any fixed $m$, $1 \leq \lim_{G \to 1} E'_m \leq 2$. But it is clear that $E'_m \leq 1/2$ for all $m \in \mathbb{N}$ and $0 < G < 1$, because
the best rational approximation of type \((2m, 2m)\) is certainly not worse than the constant approximation \(1/2\). In [5], the authors also note the mistake in these bounds, and related bounds in a few other papers, and identify and correct the source of these mistakes.

The incorrect bounds on \(E'_m\) are used to establish an upper-bound on \(\hat{\kappa}^Z\), and it appears that this upper bound is used in [18, Table 1] to provide the given values of \(\hat{\kappa}^Z\). This explains the (typically slight) discrepancy between some of their reported values of \(\hat{\kappa}^Z\) and the ones we report.

Remark A.2. Although the focus of our analysis is on the circle and ellipse filters considered in Lemmas 2.1 and 2.3, Lemma 3.7 and Theorem 3.8 apply to more general to rational filters of the form

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

(77)

where \(z_k \notin \Sigma(A)\) and \(w_N \in \mathbb{R}\), provided the strict separation condition (7) holds. This more general form includes Zolotarev filters, for example.

References


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

University of Zagreb, Bijenička 30, 10000 Zagreb, Croatia
E-mail address: luka.grubisic@math.hr

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