Abstract

The calculation of a segment of eigenvalues and their corresponding eigenvectors of a Hermitian matrix or matrix pencil has many applications. A new density-matrix-based algorithm has been proposed recently and a software package FEAST has been developed. The density-matrix approach allows FEAST's implementation to exploit a key strength of modern computer architectures, namely, multiple levels of parallelism. Consequently, the software package has been well received, especially in the electronic structure community. Nevertheless, theoretical analysis of FEAST has been lagging. For instance, the FEAST algorithm has not been proven to converge. This paper offers a detailed numerical analysis of FEAST. In particular, we show that the FEAST algorithm can be understood as the standard subspace iteration algorithm in conjunction with the Rayleigh-Ritz procedure. The novelty of FEAST is that it does not iterate directly with the original matrices, but instead iterates with an approximation to the spectral projector onto the eigenspace in question. Analysis of the numerical nature of this approximate spectral projector and the resulting subspaces generated in the FEAST algorithm establishes the algorithm's convergence. This paper shows that FEAST is resilient against rounding errors and establishes properties that can be leveraged to enhance the algorithm's robustness. We also outline a possible extension of FEAST to handle non-Hermitian problems. Further investigations into numerical quadrature rules suitable for approximating spectral projectors are also worthwhile.

Keywords: Generalized eigenvalue problem, subspace iteration, spectral projection
AMS Classifications: 15A18, 65F15

FEAST as Subspace Iteration Accelerated by Approximate Spectral Projection

Ping Tak Peter Tang*[1] and Eric Polizzi†[2]

[1]Intel Corporation, 2200 Mission College Blvd, Santa Clara, CA 95054
[2]Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA 01003

September 13, 2013

1 Introduction

Solving matrix eigenvalue problems is crucial in many scientific and engineering applications. For problems of moderate size, robust solvers are well developed and widely available [2] and are sometimes referred to as direct solvers [8]. These solvers typically calculate the entire spectrum of the matrix or matrix pencil in question. In many applications, especially for those where the underlying linear systems are large and sparse, often only selected segments of the spectrum are of interest. Polizzi recently proposed a density-matrix-based algorithm [21] named FEAST for these kinds of eigenvalue problems that are Hermitian. Unlike well-known Krylov subspace methods (see for example [5, 7, 17, 20]) which maintain subspaces of increasing dimensions, the FEAST algorithm maintains a basis for a fixed-dimension subspace but updates it per iteration. In this view, it is similar to the non-expanding subspace version of an eigensolver based on trace minimization [25, 26] but with a different subspace update strategy. From an implementation point of view, this new approach is similar to spectral divide-and-conquer [3, 4] in that the calculation is expressed in terms of high-level building blocks that can much better exploit the advantages of modern computing architectures. In this case, the high-level building block is a numerical-quadrature based technique to approximate an exact spectral projector. This building block consists of solving independent linear systems, each for multiple right hand sides. A software package FEAST [22] based on this approach has been made available since 2009. Nevertheless, theoretical analysis of FEAST has been lagging its software development. In particular, there is no theoretical study available on the conditions under which FEAST converges, and if so, at what rate.

This paper shows that the FEAST algorithm can be understood as a standard subspace iteration in conjunction with the Rayleigh-Ritz procedure, an approach which is explained in standard references such as [8], page 157, or [24], page 115. The important variation is that the subspace iteration in FEAST is carried out on an approximate spectral projector obtained by numerical quadrature. Our analysis shows that the quadrature approximation perturbs the projector's eigenvalues but not the eigenvectors. Consequently, the convergence of subspace iteration and hence of FEAST can be established by the same approaches shown in [24]. By exploring the structure of the generated subspaces, the numerical characteristics of FEAST, some of which subtle, can be much better understood. Furthermore, FEAST's robustness can be enhanced by new techniques made possible by these structure analyses – techniques such as estimating the number of eigenvalues present or judging if the dimension chosen for the subspace in question is appropriate. This paper puts the FEAST algorithm on a more solid foundation. Furthermore, we outline at the end of this paper how FEAST can be extended to handle non-Hermitian problems.

*peter.tang@intel.com
†polizzi@ecs.umass.edu

2 Overview

Throughout this paper, we consider the generalized Hermitian eigenvalue problem (GHEP) specified by two n-by-n Hermitian matrices A and B, with B being positive definite:

$$A = A^*, \quad \text{and} \quad B = C^*C, \quad C \text{ invertible},$$

where Z^* stands for complex-conjugate transposition of an arbitrary matrix Z. GHEP is that of determining eigenvalues and eigenvectors λ and $\boldsymbol{x}$, respectively, where

$$A\boldsymbol{x} = \lambda B\boldsymbol{x}.$$

We state some well-known properties of GHEP germane to us. All n eigenvalues (counting multiplicities) are real, and that we can choose a set of eigenvectors that are B-orthonormal. In matrix notation, let Λ be a diagonal matrix whose diagonal is composed of the n eigenvalues arranged in some order. Then there is a matrix $X \in \mathbb{C}^{n\times n}$ such that

$$A\,X = B\,X\,\Lambda, \quad \text{and} \quad X^*\,B\,X = I \iff X^{-1} = X^*B \quad \text{and} \quad A = B\,X\,\Lambda\,X^*\,B.$$

Each column, $\boldsymbol{x}$, of X is an eigenvector. Because $AX = BX\Lambda \iff B^{-1}AX = X\Lambda$,

$$M \stackrel{\text{def}}{=} B^{-1}A = X\,\Lambda\,X^{-1} = X\,\Lambda\,X^*B. \tag{1}$$

This paper focuses on the following problem. Given an interval $\mathcal{I} = [\lambda_-, \lambda_+]$ on the real line, compute all the eigenpairs $(\lambda, \boldsymbol{x})$, $M\boldsymbol{x} = \lambda\boldsymbol{x}$, where $\lambda \in \mathcal{I}$.

Let e be the number of eigenvalues (counting multiplicities) in $\mathcal{I}$ and $\Lambda_\mathcal{I}$ be an $e \times e$ diagonal matrix whose diagonal consists of the eigenvalues of interest. Let $X_\mathcal{I}$ be a set of corresponding eigenvectors that are B-orthonormal: $X_\mathcal{I}^*BX_\mathcal{I} = I_e$ (the $e \times e$ identity matrix). Our strategy is motivated by the spectral projector onto the invariant subspace $\text{span}(X_\mathcal{I})$, which is the matrix $X_\mathcal{I}X_\mathcal{I}^*B$. Suppose we could compute $(X_\mathcal{I}X_\mathcal{I}^*B)\boldsymbol{q}$ for any n-vector $\boldsymbol{q}$, then one can apply $X_\mathcal{I}X_\mathcal{I}^*B$ to a set of random vectors $Q = [\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_p]$. Clearly, $\text{span}((X_\mathcal{I}X_\mathcal{I}^*B)Q) \subseteq \text{span}(X_\mathcal{I})$. If it happens that $\text{rank}((X_\mathcal{I}X_\mathcal{I}^*B)Q) = \text{rank}(X_\mathcal{I})$, then $\text{span}((X_\mathcal{I}X_\mathcal{I}^*B)Q) = \text{span}(X_\mathcal{I})$. One can then obtain a basis Y for $\text{span}(X_\mathcal{I})$, for instance by performing a rank-revealing factorization of $(X_\mathcal{I}X_\mathcal{I}^*B)Q$. Now that $Y = X_\mathcal{I}W^{-1}$ for some invertible $W \in \mathbb{C}^{e\times e}$, we can define two (reduced-size) matrices $\widehat{A}, \widehat{B} \in \mathbb{C}^{e\times e}$:

$$\widehat{A} \stackrel{\text{def}}{=} Y^*\,A\,Y = (W^{-1})^*\,\Lambda_\mathcal{I}\,W^{-1}, \qquad \widehat{B} \stackrel{\text{def}}{=} Y^*\,B\,Y = (W^{-1})^*\,W^{-1},$$

thus $\widehat{A}W = \widehat{B}W\Lambda_\mathcal{I}$. Therefore, solving the reduced-size generalized eigenvalue problem $(\widehat{A}, \widehat{B})$ for $\Lambda_\mathcal{I}$ and W yields the desired eigenvalues $\Lambda_\mathcal{I}$ and eigenvectors $X_\mathcal{I} = YW$.

While the (exact) spectral projector is not readily available, we will construct an approximate projector with the following properties.

1. There is a rational function $\rho(\mu) : \mathbb{C} \to \mathbb{C}$ with no poles on the real line. Furthermore, $\rho(\mu) \in \mathbb{R}$ whenever $\mu \in \mathbb{R}$.

2. Apply the rational function $\rho(\mu)$ to Λ and M in the standard way (see page 1 of [12] for example) yields $\rho(M) = X\rho(\Lambda)X^{-1} = X\rho(\Lambda)X^*B$. In particular, for every eigenpair $(\lambda, \boldsymbol{x})$ of M, $(\rho(\lambda), \boldsymbol{x})$ is an eigenpair of $\rho(M)$.

3. $\rho(\mu)$ for $\mu \in \mathbb{R}$ is a good approximation to the indicator function of $\mathcal{I}$, which is the function that maps $\mathcal{I}$ to 1 and the rest to 0. Consequently $\rho(M)$ is a good approximation to $X_\mathcal{I}X_\mathcal{I}^*B$.

If $\rho(M)$ approximates $X_\mathcal{I}X_\mathcal{I}^*B$ well, then the subspace iteration process described in Algorithm A is a natural adaptation of the previous discussions on the exact spectral projector.

Had $\rho(M)$ be simply M, Algorithm A corresponds to straightforward subspace iteration with Rayleigh-Ritz procedure. Had $\rho(M)$ be the exact spectral projection and p happens to be e, then Algorithm A converges in one step provided $Y_{(0)}$ has full rank. When $\rho(M)$ is based on a rational function $\rho(\mu)$ constructed via a Gaussian-Legendre quadrature, Algorithm A is the FEAST algorithm as stated in [21] whose convergence behavior will be analyzed in the rest of this paper as follows.

Algorithm A (Accelerated Subspace Iteration with Rayleigh-Ritz)

1: Pick p random n-vectors $Q_{(0)} = [\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_p]$. Set $k \leftarrow 1$.
2: **repeat**
3: Approximate subspace projection: $Y_{(k)} \leftarrow \rho(M) \cdot Q_{(k-1)}$.
4: Form reduced system: $\widehat{A}_{(k)} \leftarrow Y_{(k)}^* A Y_{(k)}$, $\widehat{B}_{(k)} \leftarrow Y_{(k)}^* B Y_{(k)}$.
5: Solve p-dimension eigenproblem: $\widehat{A}_{(k)}\widehat{X}_{(k)} = \widehat{B}_{(k)}\widehat{X}_{(k)}\widehat{\Lambda}_{(k)}$ for $\widehat{\Lambda}_{(k)}$, $\widehat{X}_{(k)}$.
6: Set $Q_{(k)} \leftarrow Y_{(k)}\widehat{X}_{(k)}$, in particular $Q_{(k)}^* B\, Q_{(k)} = I_p$.
7: $k \leftarrow k + 1$.
8: **until** Appropriate stopping criteria

- Section 3 constructs a function $\rho : \mathbb{C} \to \mathbb{C}$ for a specified $\mathcal{I} = [\lambda_-, \lambda_+]$. The properties of the function $\rho(\mu)$ for μ restricted on the real line are studied, and the implications for the matrix function, our approximate subspace projector $\rho(M)$, are explained.
- Section 4 establishes that the distances from an eigenvector of interest to $\mathrm{span}(Q_{(k)})$ converge to zero, where $Q_{(k)}$ is generated according to Algorithm A. The first theorem there is a straightforward generalization of Theorem 5.2 from [24], taking into account (1) the special properties of $\rho(M)$, and (2) that we are dealing with a generalized eigenvalue problem. The second theorem examines the impact on convergence when the application of $\rho(M)$ to vectors, and in particular to the $Q_{(k)}$s, contains error. This study is relevant because application of $\rho(M)$ involves solutions of linear systems (see Section 3 for details).
- The Rayleigh-Ritz procedure is needed to derive the actual desired eigenpairs from merely a basis $Q_{(k)}$ of the subspace $\mathcal{Q}_{(k)} = \mathrm{span}(Q_{(k)})$ that is close to the desired eigenvectors. Section 5 analyzes the convergence properties of eigenpairs, taking into account the idiosyncrasies of $\mathcal{Q}_{(k)}$ due to the use of $\rho(M)$. We also show that eigenvalues of $\widehat{B}_{(k)}$ offer accurate estimates of e, the number of eigenvalues inside $\mathcal{I}$. These properties can be exploited in an enhanced version of FEAST.
- Section 6 presents a number of computational examples to illustrate key aspects of the preceding analysis as well as numerical subtleties.

3 Approximate Spectral Projector $\rho(M)$

Given an interval $\mathcal{I} = [\lambda_-, \lambda_+]$ on the real line, $\lambda_- < \lambda_+$, we will construct a rational function $\rho : \mathbb{C} \to \mathbb{C}$ such that $\rho(\mu) \in \mathbb{R}$ for $\mu \in \mathbb{R}$, and that the function $\rho(\mu)$ restricted on the real line is a good approximation to the indicator function of $\mathcal{I}$. To accomplish this, we use a Cauchy integral representation of the indicator function and construct $\rho(\mu)$ based on a numerical quadrature rule.

3.1 Construction of $\rho(\mu)$

Let $\mathcal{C}$ be the circle centered at $c = (\lambda_+ + \lambda_-)/2$ with radius $r = (\lambda_+ - \lambda_-)/2$. Define the function $\pi(\lambda)$ by the contour integral (in the counter clockwise direction)

$$\pi(\mu) = \frac{1}{2\pi\iota} \oint_{\mathcal{C}} \frac{1}{z - \mu}\, dz, \qquad \mu \notin \mathcal{C}. \tag{2}$$

The Cauchy integral theorem shows that $\pi(\mu) = 1$ for $|\mu - c| < r$ and $\pi(\mu) = 0$ for $|\mu - c| > r$. We use a numerical quadrature to approximate the integral in Equation 2. To this end, we define the parametrization $\phi(t)$, $-1 \le t \le 3$:

$$\phi(t) = c + r\, e^{\iota \frac{\pi}{2}(1+t)}, \quad \text{and} \quad \phi'(t) = \iota \frac{\pi}{2} r\, e^{\iota \frac{\pi}{2}(1+t)}. \tag{3}$$

Thus,

$$\begin{aligned}\pi(\mu) &= \frac{1}{2\pi\iota}\int_{-1}^{3}\frac{\phi'(t)}{\phi(t)-\mu}\,dt,\\ &= \frac{1}{2\pi\iota}\left[\int_{-1}^{1}\frac{\phi'(t)}{\phi(t)-\mu}\,dt+\int_{-1}^{1}\frac{\phi'(2-t)}{\phi(2-t)-\mu}\,dt\right],\\ &= \frac{1}{2\pi\iota}\int_{-1}^{1}\left[\frac{\phi'(t)}{\phi(t)-\mu}\;-\;\frac{\overline{\phi'(t)}}{\overline{\phi(t)}-\mu}\right]dt. \end{aligned} \tag{4}$$

We restrict ourselves to Gauss-Legendre quadratures on $[-1,1]$ (see for example [29]). A q-point Gauss-Legendre quadrature rule is defined by a set of weight-node pairs (w_k,t_k), $k=1,2,\ldots,q$, where $w_k>0$ and $-1<t_k<1$. The set is symmetric in that both (w_k,t_k) and $(w_k,-t_k)$ are present. The choice of the weight-node pairs are meant to make $\sum_{k=1}^{q}w_k f(t_k)$ approximate $\int_{-1}^{1}f(t)\,dt$ well for continuous function $f:[-1,1]\to\mathbb{C}$. Moreover, for any polynomial f of degree at most $2q-1$, the q-term summation produces the exact integral. In particular, $\sum_{k=1}^{q}w_k=2$ (by taking $f\equiv 1$).

In a usual setting, a quadrature aims at producing a single value that approximates a specific definite integral of an integrand. Here, it corresponds to approximating $\pi(\mu)$ for a specific fixed μ. But if we use the same quadrature rule for all possible μ, we have in fact defined a function of μ. This is how we define our $\rho(\mu)$. Let (w_k,t_k), $k=1,2,\ldots,q$, be the q-point Gauss-Legendre rule of choice. We define the function $\rho(\mu)$ by the quadrature formula applied to the integral of Equation 4:

$$\rho(\mu)\stackrel{\text{def}}{=}\frac{1}{2\pi\iota}\sum_{k=1}^{q}\left(w_k\phi'(t_k)\frac{1}{\phi(t_k)-\mu}-w_k\overline{\phi'(t_k)}\frac{1}{\overline{\phi(t_k)}-\mu}\right)=\sum_{k=1}^{q}\left(\sigma_k\frac{1}{\phi_k-\mu}+\overline{\sigma_k}\frac{1}{\overline{\phi_k}-\mu}\right), \tag{5}$$

$\phi_k=\phi(t_k)$ and $\sigma_k=w_k\phi'(t_k)/(2\pi\iota)$. Note that $\rho:\mathbb{C}\to\mathbb{C}$ is a rational function in partial fraction form. The $2q$ poles of $\rho(\mu)$ are ϕ_k and $\overline{\phi_k}$ for $k=1,2,\ldots,q$. Because $-1<t_k<1$, these poles are all complex valued. Consequently, $\rho(\mu)$ is defined for all $\mu\in\mathbb{R}$. From Equation 5, $\rho(\mu)=\overline{\rho(\mu)}$ for $\mu\in\mathbb{R}$. Thus $\rho(\mu)\in\mathbb{R}$ for $\mu\in\mathbb{R}$.

3.2 Computing $\rho(M)Q$

Consider our matrix $M=B^{-1}A$ and a function $f(x)=\alpha/(\beta-x)$, α,β constant where $\beta I-M$ invertible. It is common to define the function f of M, $f(M)$, as the matrix $\alpha\,(\beta I-M)^{-1}$ (see page 1 of [12]). Since M is diagonalizable, $M=X\,\Lambda\,X^{-1}$,

$$\begin{aligned} f(M) &\stackrel{\text{def}}{=} \alpha\,(\beta I-M)^{-1}, &(6)\\ &= \alpha\,(\beta XX^{-1}-X\,\Lambda\,X^{-1})^{-1},\\ &= \alpha\,X\,(\beta I-\Lambda)^{-1}\,X^{-1},\\ &= X\,f(\Lambda)\,X^{-1}, &(7)\end{aligned}$$

where $f(\Lambda)$ is the standard definition of a function of a diagonal matrix: namely replacing each diagonal entry λ of Λ with $f(\lambda)$. Clearly, for each eigenpair $(\lambda,\boldsymbol{x})$ of M, $(f(\lambda),\boldsymbol{x})$ is an eigenpair of $f(M)$.

As none of the ϕ_ks are on the real line while M's eigenvalues are all real, $\phi_k I-M$, $\overline{\phi_k}I-M$, $k=1,2,\ldots,q$, are all invertible. Following Equation 6, we have

$$\rho(M)=\sum_{k=1}^{q}\sigma_k(\phi_k I-M)^{-1}+\sum_{k=1}^{q}\overline{\sigma_k}(\overline{\phi_k}I-M)^{-1}=\sum_{k=1}^{q}\sigma_k(\phi_k B-A)^{-1}B+\sum_{k=1}^{q}\overline{\sigma_k}(\overline{\phi_k}B-A)^{-1}B.$$

Therefore, for any $Q\in\mathbb{C}^{n\times p}$,

$$\begin{aligned}\rho(M)Q &= \textstyle\sum_{k=1}^{q}\sigma_k(\phi_k B-A)^{-1}BQ+\sum_{k=1}^{q}\overline{\sigma_k}(\overline{\phi_k}B-A)^{-1}BQ, \quad \text{in general,}\\ &= 2\textstyle\sum_{k=1}^{q}\operatorname{Re}\left(\sigma_k(\phi_k B-A)^{-1}BQ\right), \quad \text{if } A,\ B,\text{ and } Q \text{ are real valued.}\end{aligned} \tag{8}$$

Application of $\rho(M)$ to Q involves, in general, solutions of $2q$ linear systems of equations with p right-hand-sides each, but q linear systems only if A, B, and Q are all real matrices.

Substituting ρ for f in Equation 7 gives

$$\rho(M) = X\rho(\Lambda)X^{-1} = X\rho(\Lambda)X^*B \tag{9}$$

because $M = X\Lambda X^{-1} = X\Lambda X^*B$. This implies that $(\rho(\lambda), \boldsymbol{x})$ is an eigenpair of $\rho(M)$ for any eigenpair $(\lambda, \boldsymbol{x})$ of M. Suppose $\rho(\lambda) = 1$ for all the e eigenvalues λ of M that lie inside $\mathcal{I} = [\lambda_-, \lambda_+]$ and $\rho(\lambda) = 0$ for all those $n - e$ that lie outside, then $\rho(M)$ is in fact the exact spectral projector $X_{\mathcal{I}}X_{\mathcal{I}}^*B$. In general, for any n-vector $\boldsymbol{q}$,

$$\boldsymbol{q} = \sum_{\lambda\in\mathrm{eig}(M)} \alpha_\lambda \boldsymbol{x}_\lambda \implies \rho(M)\boldsymbol{q} = \sum_{\lambda\in\mathrm{eig}(M)} \alpha_\lambda \rho(\lambda)\boldsymbol{x}_\lambda. \tag{10}$$

Suppose the scalar function $\rho(\mu)$ approximates the indicator function well in the sense that $\rho(\lambda) \approx 1$ for eigenvalues λ inside $\mathcal{I}$ and $|\rho(\lambda)| \ll 1$ for those eigenvalues λ outside of $\mathcal{I}$. Then $\rho(M)$ approximates the behavior of the exact projector $X_{\mathcal{I}}X_{\mathcal{I}}^*B$: $\rho(M)\boldsymbol{q}$ leaves almost invariant the component of $\boldsymbol{q}$ in $\mathrm{span}(X_{\mathcal{I}})$ while almost annihilating the component of $\boldsymbol{q}$ in the complementary eigenspace. We will now study more closely how well $\rho(\mu)$ approximates the indicator function.

3.3 Properties of $\rho(\mu)$ and $\rho(M)$.

As the spectrum of M is real and $\rho(M) = X\rho(\Lambda)X^{-1}$, it suffices to study $\rho(\mu)$ for $\mu \in \mathbb{R}$. As noted previously, $\rho(\mu) \in \mathbb{R}$ for $\mu \in \mathbb{R}$. Moreover, it suffices to study the reference function $\rho_{\mathrm{ref}}(\mu)$ that corresponds to the interval $[-1, 1]$. This is because a general $\rho(\mu)$ that corresponds to $\mathcal{I} = [\lambda_-, \lambda_+]$ with center c and radius r is given by the simple relationship $\rho(\mu) = \rho_{\mathrm{ref}}((\mu - c)/r)$ due to our choice of parametrization (Equation 3). Equation 5 shows that for $\mu \in \mathbb{R}$,

$$\begin{aligned} \rho_{\mathrm{ref}}(\mu) &= \frac{1}{2}\sum_{k=1}^{q} w_k \,\mathrm{Re}\left(\frac{\phi_k}{\phi_k - \mu}\right), \\ &= \frac{1}{2}\sum_{k=1}^{q} w_k \,\frac{1 + \mu s_k}{1 + 2\mu s_k + \mu^2}\,, \qquad s_k = \sin(\pi t_k/2). \end{aligned} \tag{11}$$

As noted previously, for each weight-node pair (w_k, t_k) where $t_k > 0$, there is a pair $(w_{k'}, t_{k'})$ where $w_{k'} = w_k$ and $t_{k'} = -t_k$. Note also that $s_k = \sin(\pi t_k/2)$, and thus summing the pair

$$w_k \frac{1 + \mu s_k}{1 + 2\mu s_k + \mu^2} + w_{k'} \frac{1 + \mu s_{k'}}{1 + 2\mu s_{k'} + \mu^2} = w_k \left(\frac{1 + \mu s_k}{1 + 2\mu s_k + \mu^2} + \frac{1 - \mu s_k}{1 - 2\mu s_k + \mu^2}\right)$$

yields an even function. For $t_k = s_k = 0$,

$$w_k \frac{1 + \mu s_k}{1 + 2\mu s_k + \mu^2} = \frac{w_k}{1 + \mu^2}$$

is also an even function. As a result, $\rho_{\mathrm{ref}}(\mu)$ is an even function. It suffices to study $\rho_{\mathrm{ref}}(\mu)$ for $\mu \geq 0$.

Before we present proofs on several properties of $\rho_{\mathrm{ref}}(\mu)$, let us examine several illustrative figures. Figure 1 suggests that for a reference interval $\mathcal{I} = [-1, 1]$ and the quadrature rule choice of $q = 8$, eigen-components that correspond to eigenvalues $|\lambda| \geq 1.6$ will be attenuated by roughly 4 orders of magnitudes or more. Figures 1 and 2 suggest that $\rho_{\mathrm{ref}}(\mu) \geq 1/2$ for $\mu \in \mathcal{I}$ while $|\rho_{\mathrm{ref}}(\mu)| \leq 1/2$ for $|\mu| \geq 1$. Figure 3 illustrates the attenuation properties of six specific Gauss-Legendre quadrature rules.

Theorem 1. *For a Gauss-Legendre quadrature of any choice $q \geq 1$, with parametrization $\phi(t) = e^{-\iota\frac{\pi}{2}(1+t)}$, the followings hold.*

1. $\rho_{\mathrm{ref}}(0) = 1$ *and* $\rho_{\mathrm{ref}}(1) = 1/2$.
2. $\rho'_{\mathrm{ref}}(1) < 0$.
3. $\rho_{\mathrm{ref}}(\mu) \geq 1/2$ *for* $\mu \in [-1, 1]$.

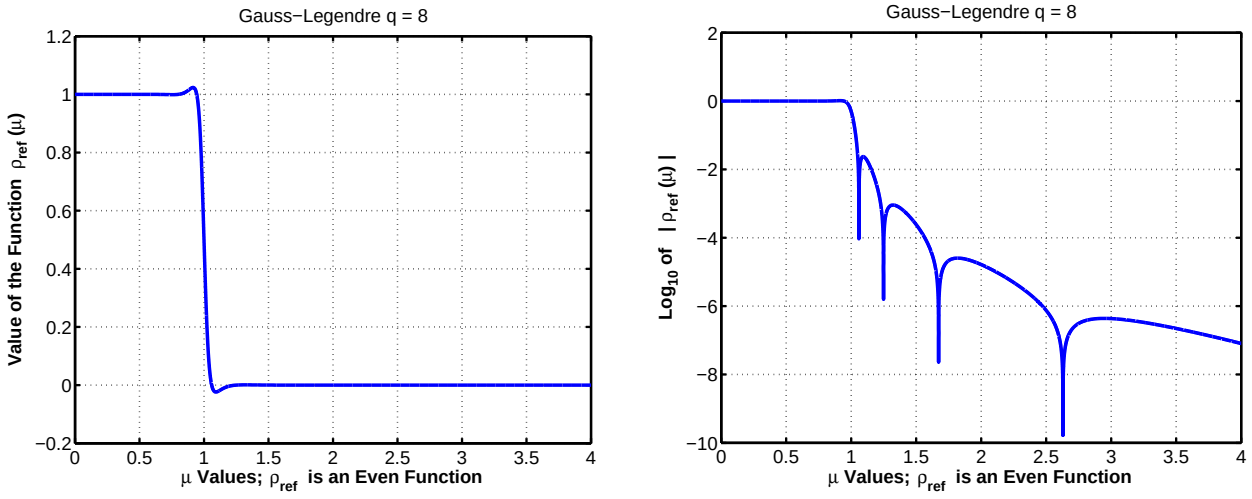

Figure 1: Gauss-Legendre quadrature with $q = 8$. The left is $\rho_{\rm ref}(\lambda)$ in normal scale, and the right is $|\rho_{\rm ref}(\lambda)|$ in logarithmic scale. Because $\rho(-\mu) = \rho(\mu)$, it suffices to show $\rho(\mu)$ for $\mu \geq 0$. The picture on the left conveys the general idea that $\rho_{\rm ref}(\lambda)$ approximates 1 inside $[\lambda_-, \lambda_+] = [-1, 1]$, and 0, outside. The picture on the right illustrates how $|\rho_{\rm ref}(\lambda)|$ gets very small very soon beyond the reference $[\lambda_-, \lambda_+]$ of $[-1, 1]$.

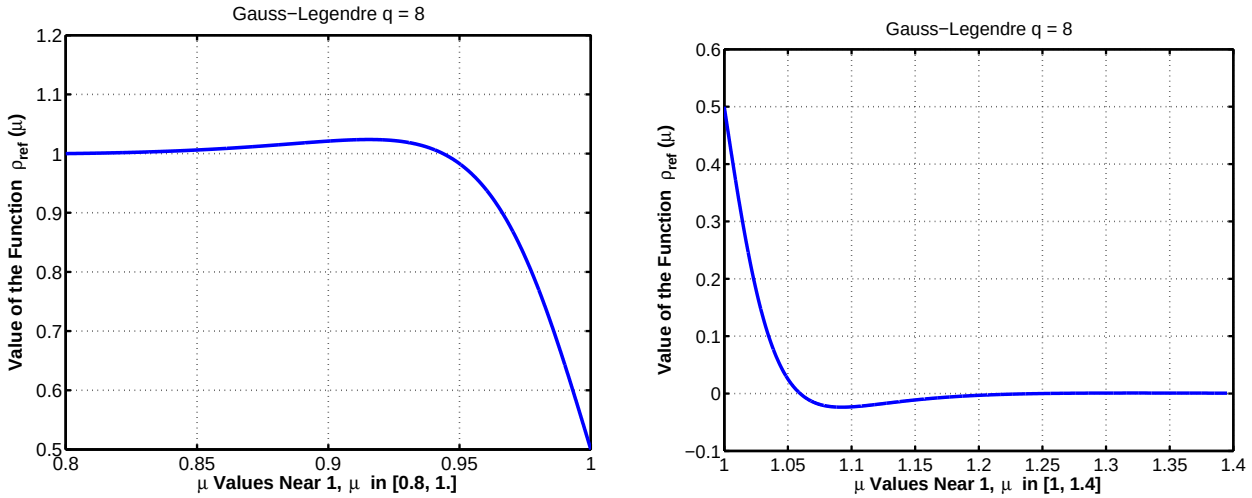

Figure 2: This figure shows more closely the Gauss-Legendre quadrature with $q = 8$ on the left and right of the boundary at 1. The left graph shows $\rho(\mu)$ stays near 1 inside the interval $\mathcal{I} = [-1, 1]$ and decreases to $1/2$ at $\mu = 1$. The right shows $\rho(\mu)$ decreases from $1/2$ to near zero soon after μ leaves the interval $\mathcal{I}$.

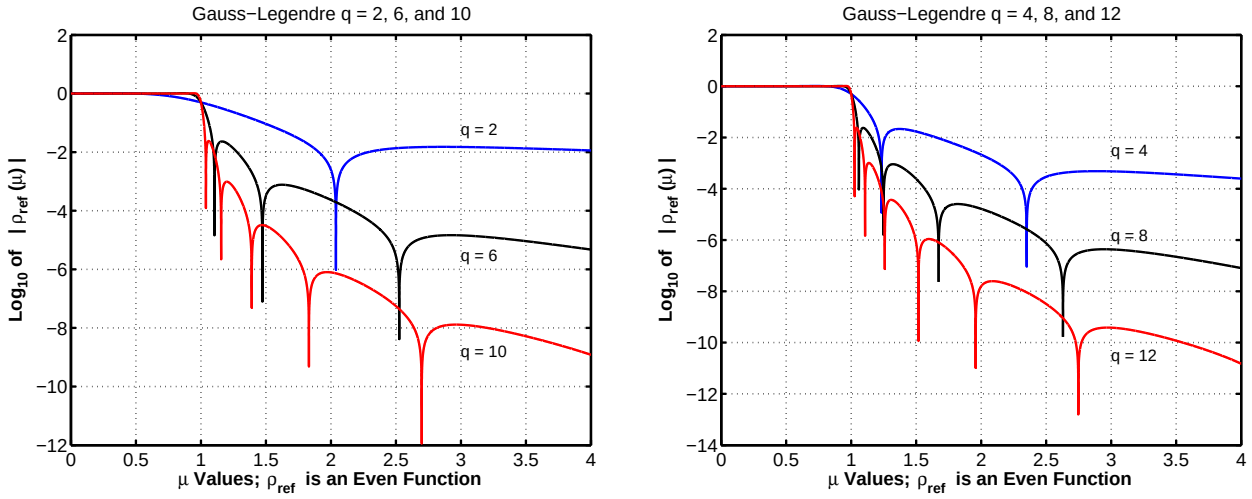

Figure 3: This figure illustrates the decay properties of $|\rho_{\text{ref}}(\mu)|$ of different degree q of Gauss-Legendre quadratures. The Y-axes are in logarithmic scale.

Proof. *Let (w_k, t_k), $k = 1, 2, \ldots, q$, be the weight-node pairs. Recall Equation 11*

$$\rho_{\text{ref}}(\mu) = \frac{1}{2}\sum_{k=1}^{q} w_k \, \frac{1+\mu s_k}{1+2\mu s_k+\mu^2}\,, \qquad s_k = \sin(\pi t_k/2).$$

As noted before, $\sum_{k=1}^{q} w_k = 2$. Thus

$$\rho_{\text{ref}}(0) = \frac{1}{2}\sum_{k=1}^{q} w_k = 1, \quad \text{and} \quad \rho_{\text{ref}}(1) = \frac{1}{2}\sum_{k=1}^{q} w_k \frac{1}{2} = \frac{1}{2}.$$

This establishes (1).

Next, note that $-1 < t_k < 1$ implies $-1 < s_k < 1$ and $1 - s_k^2 > 0$. Thus,

$$\begin{aligned}\rho_{\text{ref}}(\mu) &= \frac{1}{2}\sum_{k=1}^{q} w_k \, \frac{1+\mu s_k}{(\mu+s_k)^2+c_k^2}, \quad c_k^2 = 1-s_k^2,\\ &= \frac{1}{2}\sum_{k=1}^{k} w_k g_k(\mu), \quad g_k(\mu) \stackrel{\text{def}}{=} \frac{1+\mu s_k}{(\mu+s_k)^2+c_k^2}.\end{aligned}$$

Each of the $g_k(\mu)$s is a rational function with a strictly positive denominator. Simple differentiation shows

$$g_k'(\mu) = -\frac{s_k + 2\mu + s_k\mu^2}{[(\mu+s_k)^2+c_k^2]^2}.$$

Hence

$$g_k'(1) = -\frac{1}{2(1+s_k)} < 0 \quad \textit{for all } k \implies \rho_{\text{ref}}'(1) = \frac{1}{2}\sum_{k=1}^{q} w_k g_k'(1) < 0.$$

This establishes (2).

Finally, the formula for $g_k'(\mu)$ shows that if $s_k \geq 0$, then $g_k'(\mu) \leq 0$ for all $\mu \in [0,1]$. Therefore, for those ks where $1 > s_k \geq 0$, $g_k(\mu)$ are decreasing functions on $[0,1]$ and $g_k(\mu) \geq g_k(1) = 1/2$ for $\mu \in [0,1]$. If $s_k < 0$,

$$g_k'(\mu) = \frac{|s_k|\mu^2 - 2\mu + |s_k|}{[(\mu+s_k)^2+c_k^2]^2}.$$

The numerator is a convex quadratic that attains its minimum at $\mu = |s_k|^{-1} > 1$. Thus $g_k'(\mu)$ changes sign at most once in $[0,1]$. But $g_k'(0) = |s_k| > 0$ and $g_k'(1) < 0$ together imply that $g_k'(\mu)$ changes sign exactly once

in $[0,1]$. *This means that* $g_k(\mu)$ *attains a maximum in the interior* $(0,1)$ *and attains its minimum in* $[0,1]$ *at either* $\mu = 0$ *or* $\mu = 1$. *However,* $g_k(0) = 1$ *while* $g_k(1) = 1/2$. *Thus we must have* $g_k(\mu) \geq 1/2$ *for those* k*s that* $s_k < 0$. *In conclusion, because* $w_k > 0$ *and* $\min_{\mu\in[0,1]} g_k(\mu) = g_k(1)$ *for all* k, $\rho_{\rm ref}(\mu) \geq \rho_{\rm ref}(1) = 1/2$. *This establishes (3) and completes the proof.*

Theorem 1 shows that the matrix $\rho(M)$, $\rho(\mu) = \rho_{\rm ref}((\mu - c)/r)$, when applied to a vector preserves rather well those eigencomponents corresponding to eigenvalue in $\mathcal{I}$: those components are never attenuated by a factor smaller than $1/2$. The fact that $\rho'_{\rm ref}(1) < 0$ shows that $\rho_{\rm ref}(\mu)$ decreases from $1/2$ at $\mu = 1$ as μ gets larger than 1, as illustrated by Figure 2. As $\rho_{\rm ref}(\mu)$ is a rational function without poles on the real line, a dense numerical sampling on a finite interval $[0, L]$ gives a good description of $\rho_{\rm ref}(\mu)$ for $|\mu| \leq L$. The characteristics of action of $\rho(M)$ vectors for eigencomponents corresponding to eigenvalues in $c+\alpha L$, $|\alpha| \leq r$ can be well understood. However, we must ensure that $|\rho_{\rm ref}(\mu)|$ be small as $|\mu| \to \infty$ so that $\rho(M)$ does indeed attenuate all eigencomponents that are not of interest. One cannot study the behavior of $\rho_{\rm ref}(\mu)$ on the entire real line by numerical sampling alone. The next theorem shows that there exists a decreasing function $\xi(\mu)$ on $\mu \in (1,\infty)$ such that $|\rho_{\rm ref}(\mu)| \leq \xi(\mu)$ for all $\mu > 1$. Furthermore, $\xi(\mu)$ can be evaluated numerically. Thus we can evaluate $\xi(\mu)$ at any point $\mu = L$ to yield a bound on $|\rho_{\rm ref}(\mu)|$ for all $\mu \geq L$.

Theorem 2. *For a* q*-point Gauss-Legendre quadrature rule, there exists a function* $\xi_q(\mu)$, $\mu \in (1,\infty)$ *such that* $\xi_q(\mu)$ *is decreasing in* μ *and for any* $\mu_0 > 1$, $|\rho_{\rm ref}(\mu)| \leq \xi_q(\mu_0)$ *for all* $\mu \geq \mu_0$.

Proof. *Consider a* q*-point Gauss-Legendre rule.* $\rho_{\rm ref}(\mu)$ *is the value obtained by applying this quadrature rule to the definite integral*

$$\begin{aligned}\pi(\mu) &= \frac{1}{2}\int_{-1}^{1} \frac{1+\mu\sin(\pi t/2)}{1+2\mu\sin(\pi t/2)+\mu^2}\,dt,\\ &= \int_{-1}^{1} f_\mu(t)\,dt, \quad f_\mu(t) \stackrel{\rm def}{=} \frac{1}{2}\frac{1+\mu\sin(\pi t/2)}{1+2\mu\sin(\pi t/2)+\mu^2}. \end{aligned} \tag{12}$$

The subscript μ *in* f_μ *emphasizes that* μ *is considered as a parameter and* t *is the independent variable. For any fixed* $\mu > 1$, $1 + 2\mu\sin(\pi t/2) + \mu^2 > 0$ *for all* t *and that* $f_\mu(t)$ *is infinitely differentiable (with respect to* t*). The error in Gauss-Legendre rule is well known (see for example [29]):*

$$\int_{-1}^{1} f_\mu(t)\,dt - \rho_{\rm ref}(\mu) = \frac{f_\mu^{(2q)}(t_0)}{(2q)!}\int_{-1}^{1} P_q^2(t)\,dt$$

for some $t_0 \in (-1,1)$ *and* $P_q(t)$ *is the monic Legendre polynomial of degree* q. *It is known that [1] the* L^2*-norm of the* q*-degree Legendre polynomial with leading coefficient* $(\Pi_{j=1}^{q}(2j-1))/q!$ *is* $(q+1/2)^{-1/2}$. *Therefore*

$$\int_{-1}^{1} P_q^2(t)\,dt = \frac{2^{2q+1}(q!)^4}{(2q+1)((2q)!)^2}.$$

Because $\int_{-1}^{1} f_\mu(t)\,dt = 0$ *for* $\mu > 1$, *we have*

$$|\rho_{\rm ref}(\mu)| \leq K_q \max_{t\in[-1,1]} \left|f_\mu^{(2q)}(t)\right|, \qquad K_q \stackrel{\rm def}{=} \frac{2^{2q+1}(q!)^4}{(2q+1)((2q)!)^3}. \tag{13}$$

We will complete the proof by constructing a function

$$\xi_q(\mu) \geq K_q \max_{t\in[-1,1]} \left|f_q^{(2q)}(t)\right|$$

where $\xi_q(\mu)$ *is decreasing in* μ *for* $\mu > 1$. *This construction is via a simple application of the chain rule. Define*

$$F_\mu(s) \stackrel{\rm def}{=} \frac{1}{2}\frac{1+\mu s}{1+2\mu s+\mu^2}, \qquad \text{and} \qquad s(t) \stackrel{\rm def}{=} \sin(\pi t/2).$$

Thus $f_\mu(t) = F_\mu(s(t))$. *Note that* $|s| \leq 1$ *for* $|t| \leq 1$. *Let us first examine the derivative of* F *(with respect to* s*). Simple differentiation shows*

$$F_\mu^{(k)}(s) = (-1)^{k+1}\,k!\,2^{k-2}\,\frac{\mu^2-1}{(1+2s\mu+\mu^2)^{k+1}}\,.$$

Since (1) $1+2s\mu+\mu^2$ is positive and increasing in $\mu \in (1,\infty)$ for any fixed value of $s \in [-1,1]$, and (2) for any fixed $\mu > 1$, $1+2s\mu+\mu^2$ for the range $|s| \le 1$ attains its minimum value of $1-2\mu+\mu^2 = (\mu-1)^2$ at $s=-1$, we have

$$\max_{s\in[-1,1]} |F_\mu^{(k)}(s)| = k!\, 2^{k-2}\, \frac{\mu^2-1}{(\mu-1)^{2(k+1)}} \stackrel{\text{def}}{=} A_k(\mu).$$

$A_k(\mu)$ as defined above is a decreasing function in $\mu \in (1,\infty)$ as its derivative is easily seen to be always negative. Hence, for any $\mu_0 > 1$

$$\max_{s\in[-1,1]} |F_\mu^{(k)}(s)| \le A_k(\mu_0) \qquad \textit{for all } \mu \ge \mu_0. \tag{14}$$

Differentiation of $s(t) = \sin(\pi t/2)$ is simple:

$$|s^{(k)}(t)| = \begin{cases} \left(\frac{\pi}{2}\right)^k |\sin(\pi t/2)| & k \textit{ even}, \\ \left(\frac{\pi}{2}\right)^k |\cos(\pi t/2)| & k \textit{ odd}. \end{cases}$$

Note also that, using a half-angle formula,

$$\max_{t\in[-1,1]} |\sin(\pi t/2)\cos(\pi t/2)| = \max_{t\in[-1,1]} |\sin(\pi t)/2| = 1/2.$$

We can now construct a bound on the derivative of $f_\mu(t)$. Apply the chain rule formula for higher derivative (often attributed to Faà di Bruno [14]) to the composite function $f_\mu(t) = F_\mu(s(t))$:

$$\frac{d^n}{dt^n} F_\mu(s(t)) = \sum_{(k_1,k_2,\ldots,k_n)\in\mathcal{S}_n} \frac{n!}{k_1!\,k_2!\cdots k_n!}\, F_\mu^{(k)}(s)\, \Pi_{j=1}^n \left(\frac{s^{(j)}(t)}{j!}\right)^{k_j}, \tag{15}$$

where

$$\mathcal{S}_n = \{\ (k_1,k_2,\ldots,k_n) \mid k_j \ge 0,\ k_1+2k_2+3k_3+\cdots+nk_n = n\ \}$$

is the set of all nonnegative solutions to the Diophantine equation $k_1+2k_2+\cdots nk_n = n$, and k is defined as $k = \sum_{j=1}^n k_j$. For each solution $(k_1,\ldots,k_n) \in \mathcal{S}_n$ we define

$$\begin{aligned} k_{\text{even}} &\stackrel{\text{def}}{=} k_2 + k_4 + \cdots + k_{2\lfloor n/2\rfloor}, \\ k_{\text{odd}} &\stackrel{\text{def}}{=} (\sum_{j=1}^n k_j) - k_{\text{odd}} = k - k_{\text{even}}, \\ k_{\text{sc}} &\stackrel{\text{def}}{=} \min\{k_{\text{even}}, k_{\text{odd}}\}. \end{aligned}$$

Therefore,

$$\begin{aligned} \left|\Pi_{j=1}^n \left(s^{(j)}(t)\right)^{k_j}\right| &= \left(\frac{\pi}{2}\right)^{k_1+2k_2+\cdots+nk_n} |\sin(\pi t/2)|^{k_{\text{even}}}\ |\cos(\pi t/2)|^{k_{\text{odd}}}, \\ &= \left(\frac{\pi}{2}\right)^n |\sin(\pi t/2)|^{k_{\text{even}}}\ |\cos(\pi t/2)|^{k_{\text{odd}}}, \\ &\le \left(\frac{\pi}{2}\right)^n \left(\frac{1}{2}\right)^{k_{\text{sc}}}, \qquad \textit{for all } t \in [-1,1]. \end{aligned} \tag{16}$$

Finally, we define

$$\alpha(k_1,k_2,\ldots,k_n) \stackrel{\text{def}}{=} \frac{n!}{k_1!\,k_2!\cdots k_n!}\, \Pi_{j=1}^n \left(\frac{1}{j!}\right)^{k_j},$$

and arrive at

$$\begin{aligned} K_q \max_{t\in[-1,1]} \left|f_\mu^{(2q)}(t)\right| &\le K_q \left(\frac{\pi}{2}\right)^{2q} \sum_{(k_1,k_2,\ldots,k_{2q})\in\mathcal{S}_{2q}} \alpha(k_1,k_2,\ldots,k_{2q})\, A_k(\mu)\, (1/2)^{k_{\text{sc}}}, \\ &\stackrel{\text{def}}{=} \xi_q(\mu). \end{aligned} \tag{17}$$

This bound is a synthesis of Equations 13 through 16. The function $\xi_q(\mu)$ is decreasing in μ as each of the $A_k(\mu)$ is a decreasing function. Finally, for any $\mu_0 > 1$, for all $\mu \ge \mu_0$, we have

$$|\rho_{\text{ref}}(\mu)| \le \xi_q(\mu) \le \xi_q(\mu_0).$$

That completes the proof.

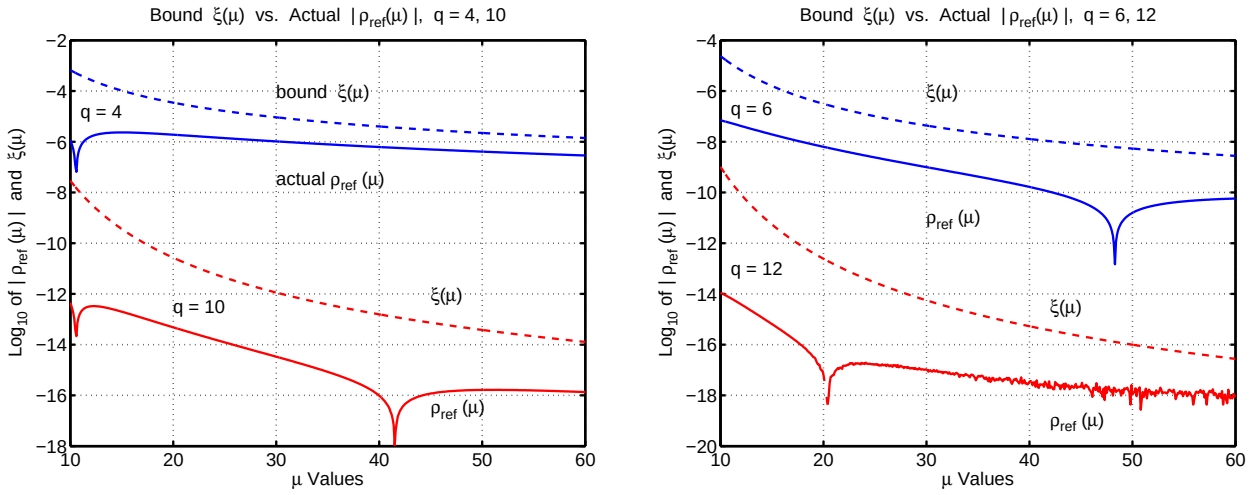

Figure 4: Upper bound functions that are decreasing are useful in bounding tail ends of $|\rho_{\text{ref}}(\mu)|$. For example, the graph of the left shows $\xi(30) \approx 10^{-12}$ for the $q = 10$ quadrature rule. This implies $|\rho_{\text{ref}}(\mu)| \leq 10^{-12}$ for all $\mu \geq 30$.

Figure 4 illustrates the upper bound function $\xi_q(\mu)$ for several values of q. Note that $\xi_q(\mu)$ given in Equation 17 is easy to compute numerically. Theorem 2 allows us to assess the maximum of the infinite tail end $|\rho_{\text{ref}}(\mu)|$ on, for example, $[y, \infty)$ for some specific y by numerical sampling on just a finite interval because

$$\max_{\mu\in[y,\infty)} |\rho_{\text{ref}}(\mu)| \leq \max\left\{ \max_{\mu\in[y,y+L]} |\rho_{\text{ref}}(\mu)|,\ \xi_q(y+L) \right\}$$

for any $L > 0$. Table 1 tabulates the decay of the tail ends of $\rho_{\text{ref}}(\mu)$ for several specific q values. It also tabulates the maximum value of $\rho_{\text{ref}}(\mu)$ on the interval $[-1, 1]$.

q	$\max \rho_{\text{ref}}(\mu)$ for $\mu \in [-1,1]$	y where $\max_{\mu\in[y,\infty)} \lvert\rho_{\text{ref}}(\mu)\rvert \leq \frac{1}{2}10^{-j}, j = 1,2,\ldots,7$						
		10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}
4	1.022	1.17	1.84	2.24	6.08	8.96	44.90	145.80
6	1.023	1.08	1.35	1.84	2.32	4.00	5.50	11.39
8	1.024	1.05	1.20	1.45	1.64	2.29	2.59	4.28
10	1.024	1.03	1.13	1.29	1.39	1.74	2.20	2.58
12	1.024	1.03	1.10	1.21	1.28	1.50	1.79	1.96

Table 1: For the tabulated quadratures, $1/2 \leq \rho_{\text{ref}}(\mu) \leq 1.024$ for $\mu \in [-1, 1]$ (cf. Theorem 1). The rightmost columns show the "relative attenuation" as comparison of $\max_{\mu\in[y,\infty)} |\rho_{\text{ref}}(\mu)|$ with $\min_{\mu\in[-1,1]} |\rho_{\text{ref}}(\mu)| = 1/2$.

As stated previously, $M = X\Lambda X^{-1}$ and $\rho(M) = X\rho(\Lambda)X^{-1}$. If $\rho(\lambda) = 1$ for $\lambda \in \text{eig}(M) \cap [\lambda_-, \lambda_+]$ and $\rho(\lambda) = 0$ for $\lambda \in \text{eig}(M)\backslash[\lambda_-, \lambda_+]$, then $\rho(M)$ is the exact subspace projector $X_{\mathcal{I}}X_{\mathcal{I}}^*B$. Table 1 illustrates quantitatively that our construction $\rho(M)$ based on Gaussian-Legendre quadrature is an approximate subspace projector. For example, for a quadrature rule of $q = 8$ and a general $\mathcal{I} = [\lambda_-, \lambda_+]$ with $c = (\lambda_+ + \lambda_-)/2$, $r = (\lambda_+ - \lambda_-)/2$, we have $1 \approx \rho(\lambda) \in [1/2, 1.024]$ for $\lambda \in \mathcal{I}$ and $|\rho(\lambda)| \leq \frac{1}{2}\,10^{-3}$ for $|\lambda - c| \geq 1.45r$.

4 Robust Subspace Convergence of FEAST

The previous analysis suggests that a Gauss-Legendre quadrature accelerator $\rho(M)$ should be effective in Algorithm A as the eigenvalues of interest of M are now the (highly) dominant eigenvalues of $\rho(M)$. We restate Algorithm A as Algorithm FEAST with $\rho(M)$ being derived specifically from Gauss-Legendre quadratures using a circular path parametrized as in Equation 3. A varieties of functions ρ can be constructed by different

quadrature rules (e.g. Trapezoidal Rule) on different contours (e.g. ellipses). It is also possible to construct ρ from a purely function approximation approach, for example via Zolotarev approximation [30].

Algorithm FEAST (as given in [21])

1: Specify $\mathcal{I} = [\lambda_-, \lambda_+]$ and a Gauss-Legendre quadrature choice of q.
2: Pick p and p random n-vectors $Q_{(0)} = [\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_p]$. Set $k \leftarrow 1$.
3: **repeat**
4: Approximate subspace projection (see Equation 8): $Y_{(k)} \leftarrow \rho(M) \cdot Q_{(k-1)}$.
5: Form reduced system: $\widehat{A}_{(k)} \leftarrow Y_{(k)}^* A Y_{(k)}$, $\widehat{B}_{(k)} \leftarrow Y_{(k)}^* B Y_{(k)}$.
6: Solve p-dimension eigenproblem: $\widehat{A}_{(k)}\widehat{X}_{(k)} = \widehat{B}_{(k)}\widehat{X}_{(k)}\widehat{\Lambda}_{(k)}$ for $\widehat{\Lambda}_{(k)}$, $\widehat{X}_{(k)}$.
7: Set $Q_{(k)} \leftarrow Y_{(k)}\widehat{X}_{(k)}$, in particular $Q_{(k)}^* B\, Q_{(k)} = I_p$.
8: $k \leftarrow k + 1$.
9: **until** Appropriate stopping criteria

Basic convergence properties of subspace iteration for eigenvalue problem are well known (see for example [8, 24]). Nonetheless, utilization of an approximate subspace projector leads to special properties. $\rho(M)$ approximates a low-rank operator, and thus we customize our analysis to focus on subspaces within the generated subspace $\mathcal{Q}_{(k)} = \mathrm{span}(Q_{(k)})$. The first theorem in this section, Theorem 3, generalizes Theorem 5.2 of [24] to cover generalized eigenvalue problems. Our second theorem, Theorem 4, takes into account that the application of $\rho(M)$ is inexact as it involves solutions of linear systems. We show that subspace convergence is not affected in any fundamental way, and hence this section's title of robust convergence.

As the decay of the function $\rho(\mu)$ is closely tied to the property of the approximate subspace projector $\rho(M)$, we adopt the following convention. Whenever there is an underlying fixed choice of $\rho(M)$, we number the eigenpairs $(\lambda_j, \boldsymbol{x}_j)$ so that $|\rho(\lambda_1)| \geq |\rho(\lambda_2)| \geq \cdots \geq |\rho(\lambda_n)|$. In particular, if there are e eigenvalues in $\mathcal{I}$, then

$$\rho(\lambda_1) \geq \rho(\lambda_2) \geq \cdots \geq \rho(\lambda_e) \geq 1/2 > |\rho(\lambda_{e+1})| \geq \cdots \geq |\rho(\lambda_n)|.$$

We denote the eigenvalues of $\rho(M)$ by $\gamma_j \stackrel{\mathrm{def}}{=} \rho(\lambda_j)$, $j = 1, 2, \ldots, n$. Our analysis involves examining sections of the eigenvectors $X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]$ and the corresponding eigenvalues. We set up some simplifying notations: For any integer ℓ, $1 \leq \ell < n$,

$$\begin{array}{lcllcl} X_\ell &=& [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_\ell], & X_{\ell'} &=& [\boldsymbol{x}_{\ell+1}, \boldsymbol{x}_{\ell+2}, \ldots, \boldsymbol{x}_n], \\ \Lambda_\ell &=& \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_\ell), & \Lambda_{\ell'} &=& \mathrm{diag}(\lambda_{\ell+1}, \lambda_{\ell+2}, \ldots, \lambda_n), \\ \Gamma_\ell &=& \mathrm{diag}(\gamma_1, \gamma_2, \ldots, \gamma_\ell), & \Gamma_{\ell'} &=& \mathrm{diag}(\gamma_{\ell+1}, \gamma_{\ell+2}, \ldots, \gamma_n). \end{array}$$

With these notations, $X_\ell X_\ell^* B$ is the B-orthogonal projector onto $\mathrm{span}(\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_\ell\})$, $X_{\ell'} X_{\ell'}^* B$ is the B-orthogonal projector onto $\mathrm{span}(\{\boldsymbol{x}_{\ell+1}, \boldsymbol{x}_{\ell+2}, \ldots, \boldsymbol{x}_n\})$, $I = X_\ell X_\ell^* B + X_{\ell'} X_{\ell'}^* B$, and

$$\rho(M) = X_\ell\, \Gamma_\ell\, X_\ell^* B + X_{\ell'}\, \Gamma_{\ell'}\, X_{\ell'}^* B. \tag{18}$$

All norms $\|\cdot\|$, unless explicitly stated using subscripts, are 2-norms.

Theorem 3. *Consider Algorithm FEAST. Suppose $|\gamma_p| > 0$ and that the $p \times p$ matrix $X_p^* B Q_{(0)}$ is invertible. Let m be any integer $m \leq p$. Then there is a constant α such that the followings hold for $k = 1, 2, \ldots$.*

- *The subspace $\mathcal{Q}_{(k)} \stackrel{\mathrm{def}}{=} \mathrm{span}(Q_{(k)})$ is of dimension p.*
- *$Q_{(k)}$ is of the form $Q_{(k)} = [\ (X_m + X_{p'} E_{(k)}) L_{(k)}^{-1} \qquad V_{(k)}\ ]\, U_{(k)}$ where $U_{(k)}$ is a $p \times p$ unitary matrix. $E_{(k)}, L_{(k)}, V_{(k)}$ are of conforming dimensions of $(n-p) \times m, m \times m, n \times (p-m)$, respectively. ($V_{(k)}$ is non-existent for $m = p$.) Moreover, $E_{(k)} = \Gamma_{p'}\, E_{(k-1)}\, \Gamma_m^{-1}$.*
- *For each $j = 1, 2, \ldots, m$, there is a vector $\boldsymbol{s}_j \in \mathcal{Q}_{(k)}$ such that $\|\boldsymbol{x}_j - \boldsymbol{s}_j\|_B \leq \alpha |\gamma_{p+1}/\gamma_j|^k$. The B-norm of an n-vector $\boldsymbol{y}$ is defined in the standard way: $\|\boldsymbol{y}\|_B \stackrel{\mathrm{def}}{=} (\boldsymbol{y}^*\, B\, \boldsymbol{y})^{1/2}$. In particular, $\|(I - P_{(k)})\boldsymbol{x}_j\|_B \leq \alpha |\gamma_{p+1}/\gamma_j|^k$ where $P_{(k)}$ is the B-orthogonal projection onto the space $\mathcal{Q}_{(k)}$.*

Proof. *$|\gamma_p| > 0$ implies that Γ_i is invertible for any $i = 1, 2, \ldots, p$. Suppose $X_p^* B\, Q_{(k-1)}$ is invertible for some $k = 1, 2, \ldots$, then*

$$\begin{aligned} Y_{(k)} &= \rho(M) Q_{(k-1)}, \\ &= X_p \Gamma_p (X_p^* B\, Q_{(k-1)}) + X_{p'} \Gamma_{p'} (X_{p'}^* B\, Q_{(k-1)}), \\ &= \left(X_p + X_{p'} (\Gamma_{p'} X_{p'}^* B\, Q_{(k-1)}) Z^{-1} \right) Z, \end{aligned}$$

where $Z = \Gamma_p (X_p^* B\, Q_{(k-1)})$ is invertible. Thus $Y_{(k)}$ has full column rank, which implies $\widehat{B}_{(k)} = Y_{(k)}^* B Y_{(k)}$ is Hermitian positive definite. The dimension-p generalized eigenvalue problem specified by $\widehat{A}_{(k)}$ and $\widehat{B}_{(k)}$ has linearly independent eigenvectors, that is, the matrix $\widehat{X}_{(k)}$ is invertible. Consequently, $Q_{(k)} = Y_{(k)} \widehat{X}_{(k)}$ has full column rank, $\mathcal{Q}_{(k)}$ is of dimension p, and the matrix $X_p^* B Q_{(k)}$ is invertible. Since by assumption $X_p^* B Q_{(0)}$ is invertible, we conclude by induction that the subspaces $\mathcal{Q}_{(k)}$ are all of dimension p for $k = 1, 2, \ldots$.

Define $U_{(0)} \stackrel{\text{def}}{=} X_p^* B Q_{(0)}$. $U_{(0)}$ is invertible by assumption (but not necessarily unitary).

$$\begin{aligned} Q_{(0)} &= (X_p X_p^* B + X_{p'} X_{p'}^* B) Q_{(0)}, \\ &= (X_p + X_{p'} X_{p'}^* B Q_{(0)} U_{(0)}^{-1}) U_{(0)}. \end{aligned}$$

Given an integer $m \leq p$, partition the p columns of $Q_{(0)} U_{(0)}^{-1}$ into m and $p - m$ columns:

$$Q_{(0)} = [\ (X_m + X_{p'} E_{(0)}) \quad V_{(0)}\ ]\, U_{(0)},$$

where $E_{(0)}$ is simply the first m columns of $X_{p'}^* B Q_{(0)} U_{(0)}^{-1}$. Consequently,

$$Y_{(1)} = \rho(M) Q_{(0)} = [\ (X_m + X_{p'} E_{(1)}) \Gamma_m \quad \rho(M)\, V_{(0)}\ ]\, U_{(0)},$$

$E_{(1)} = \Gamma_{p'}\, E_{(0)}\, \Gamma_m^{-1}$. $X_m + X_{p'} E_{(1)}$ is of full rank m as X_m is linearly independent with $X_{p'}$. Since B is Hermitian positive definite, there is a matrix $L_{(1)}$ such that

$$L_{(1)}^* L_{(1)} = (X_m + X_{p'} E_{(1)})^*\, B\, (X_m + X_{p'} E_{(1)}),$$

which implies $(X_m + X_{p'} E_{(1)}) L_{(1)}^{-1}$ is B-orthonormal. Any $L_{(1)}$ that satisfies the above equation is acceptable. Note that $L_{(1)}$ is unique up to a unitary equivalence. For any specifically chosen $L_{(1)}$, complete a B-orthonormal basis for $\operatorname{span}(Y_{(1)})$ to yield $[\ (X_m + X_{p'} E_{(1)}) L_{(1)}^{-1} \quad V_{(1)}\ ]$. Since $Q_{(1)}$ is B-orthonormal as well, there must be a $p \times p$ unitary matrix $U_{(1)}$ such that

$$Q_{(1)} = [\ (X_m + X_{p'} E_{(1)}) L_{(1)}^{-1} \qquad V_{(1)}\ ]\, U_{(1)}.$$

We have shown that at $k = 1$, $Q_{(k)}$ is of the form

$$Q_{(k)} = [\ (X_m + X_{p'} E_{(k)}) L_{(k)}^{-1} \qquad V_{(k)}\ ]\, U_{(k)}, \tag{19}$$

where $U_{(k)}$ is unitary and $Q_{(k)}$ B-orthonormal. Suppose now Equation 19 holds for some $k \geq 1$.

$$Y_{(k+1)} = \rho(M) Q_{(k)} = [\ (X_m + X_{p'} E_{(k+1)}) \Gamma_m L_{(k)}^{-1} \quad \rho(M)\, V_{(k)}\ ]\, U_{(k)},$$

where

$$E_{(k+1)} = \Gamma_{p'}\, E_{(k)}\, \Gamma_m^{-1}.$$

Let $L_{(k+1)}$ be such that

$$L_{(k+1)}^* L_{(k+1)} = (X_m + X_{p'} E_{(k+1)})^*\, B\, (X_m + X_{p'} E_{(k+1)}).$$

This makes $(X_m + X_{p'} E_{(k+1)}) L_{(k+1)}^{-1}$ B-orthonormal. Complete a B-orthonormal basis with $V_{(k+1)}$ for $\operatorname{span}(Y_{(k+1)}) = \operatorname{span}(Q_{(k+1)})$. Hence, there must be a $p \times p$ unitary matrix $U_{(k+1)}$ such that

$$Q_{(k+1)} = [\ (X_m + X_{p'} E_{(k+1)}) L_{(k+1)}^{-1} \qquad V_{(k+1)}\ ]\, U_{(k+1)}.$$

By induction, we have established the second item of this theorem.

Finally, each of the m column vectors of $X_m + X_{p'} E_{(k)}$ is in the subspace $\mathcal{Q}_{(k)}$, and $E_{(k)} = \Gamma_{p'}^k \, E_{(0)} \, \Gamma_m^{-k}$ for $k = 1, 2, \ldots$. Let $\alpha \stackrel{\text{def}}{=} \|E_{(0)}\|$. For each $k = 1, 2, \ldots$, and $j = 1, 2, \ldots, m$, let $\boldsymbol{e}_j$ be the j-th column of $E_{(k)}$, and define $\boldsymbol{s}_j \stackrel{\text{def}}{=} \boldsymbol{x}_j + X_{p'} \boldsymbol{e}_j$. Clearly $\boldsymbol{s}_j \in \mathcal{Q}_{(k)}$ and

$$\|\boldsymbol{x}_j - \boldsymbol{s}_j\|_B = \|X_{p'} \boldsymbol{e}_j\|_B = \|\boldsymbol{e}_j\| \le \alpha \left| \frac{\gamma_{p+1}}{\gamma_j} \right|^k .$$

In particular,

$$\|(I - P_{(k)}) \boldsymbol{x}_j\|_B = \min_{\boldsymbol{s} \in \mathcal{Q}_{(k)}} \|\boldsymbol{x}_j - \boldsymbol{s}\|_B \le \alpha \left| \frac{\gamma_{p+1}}{\gamma_j} \right|^k .$$

This proves the final item of the theorem.

Theorem 3 together with Table 1 illustrate the effectiveness of FEAST. For example, let $\mathcal{I} = [\lambda_-, \lambda_+]$ be the specified interval, with center $c = (\lambda_+ + \lambda_-)/2$ and radius $r = (\lambda_+ - \lambda_-)/2$. If M's spectrum is distributed somewhat uniformly, then the interval $[c-1.45r, c+1.45r]$ would have roughly 1.45 times as many eigenvalues as there are in $\mathcal{I}$. For a quadrature rule of $q = 8$ and that p happens to be $p \ge 1.45e$, then for $j = 1, 2, \ldots, e$, we would expect $|\gamma_{p+1}/\gamma_j|$ be be as small as 10^{-3}. So that for each j, there exist elements $\boldsymbol{s}_{(k)} \in \mathcal{Q}_{(k)}$, $k = 1, 2, \ldots$, such that $\|\boldsymbol{x}_j - \boldsymbol{s}_{(k)}\|_B \to 0$ at a rate of 10^{-3k}.

Application of $\rho(M)$ in general incurs more error compared to standard subspace iterations ($\rho(\mu) \equiv 1$) or a low-degree polynomial accelerator ($\rho(\mu)$ is a low-degree polynomial). We now examine Algorithm FEAST when Step 4 is replaced by

$$Y_{(k)} \leftarrow \rho(M) Q_{(k-1)} + \Delta_{(k-1)}, \qquad \Delta_{(k-1)} \text{ is a } n \times p \text{ matrix.}$$

Our analysis below shows that convergence is not affected in any fundamental way. Without the error term, Theorem 3 shows that the generated subspaces $\mathcal{Q}_{(k)}$ contains elements of the form $\boldsymbol{x}_j + X_{p'} \boldsymbol{e}$, $1 \le j \le m$. The component $X_{p'} \boldsymbol{e}$ will be attenuated in the iterative process. The presence of the error term $\Delta_{(k)}$ in effect introduces a small component of the form $X_{m'} \boldsymbol{f}$ at each iteration. Once introduced, however, this component will also be attenuated as the iterative process proceeds. Roughly speaking, the one dominant error term is the one most recently introduced, and convergence of subspace iteration remains robust. The next lemma examines the form of the generated subspaces and Theorem 4 quantifies them.

Lemma 1. *Suppose for any $k = 1, 2, \ldots$, $Q_{(k-1)}$ is of the form*

$$[\, (X_m + X_{p'} E_{(k-1)} + X_{m'} F_{(k-1)}) L_{(k-1)}^{-1} \qquad V_{(k-1)} \,] \, U_{(k-1)}$$

for some $m \le p$, where $U_{(k-1)}$ and $L_{(k-1)}$ are invertible matrices of dimensions $p \times p$ and $m \times m$, respectively. The matrices $E_{(k-1)}$, $F_{(k-1)}$, and $V_{(k-1)}$ are of conforming dimensions. Define the terms $\xi_{(k-1)}$, $\zeta_{(k-1)}$, $\psi_{(k-1)}$ via the following partitioning involving the error term $\Delta_{(k-1)}$:

$$X^{-1} \, \Delta_{(k-1)} \, U_{(k-1)}^{-1} \left[\begin{array}{c|c} L_{(k-1)} & \\ \hline & I \end{array} \right] = \left[\begin{array}{cc|c} & & \\ & \xi_{(k-1)} & \psi_{(k-1)} \\ & \zeta_{(k-1)} & \end{array} \right], \qquad \xi_{(k-1)} \text{ is } m \times m.$$

If $\Gamma_m + \xi_{(k-1)}$ is invertible and $Y_{(k)} = \rho(M) Q_{(k-1)} + \Delta_{(k-1)}$ remains full column rank, then $Q_{(k)}$ is of the form

$$Q_{(k)} = [\, (X_m + X_{p'} E_{(k)} + X_{m'} F_{(k)}) L_{(k)}^{-1} \qquad V_{(k)} \,] \, U_{(k)},$$

where $U_{(k)}$ is unitary, and

$$L_{(k)}^* \, L_{(k)} = (X_m + X_{p'} \, E_{(k)} + X_{m'} \, F_{(k)})^* \, B \, (X_m + X_{p'} \, E_{(k)} + X_{m'} \, F_{(k)}).$$

In particular, $\|L_{(k)}\| \le \left(1 + (\|E_{(k)}\| + \|F_{(k)}\|)^2\right)^{1/2}$. Furthermore,

$$\begin{aligned} E_{(k)} &= \Gamma_{p'} \, E_{(k-1)} \, (\Gamma_m + \xi_{(k-1)})^{-1}, \qquad \text{and} \\ F_{(k)} &= \Gamma_{m'} \, F_{(k-1)} \, (\Gamma_m + \xi_{(k-1)})^{-1} + \zeta_{(k-1)} \, (\Gamma_m + \xi_{(k-1)})^{-1}. \end{aligned}$$

Proof. *By assumption,*

$$\begin{aligned}\rho(M)\,Q_{(k-1)} &= [\,(X_m\,\Gamma_m + X_{p'}\,\Gamma_{p'}\,E_{(k-1)} + X_{m'}\,\Gamma_{m'}\,F_{(k-1)})L_{(k-1)}^{-1} \quad \rho(M)\,V_{(k-1)}\,]\,U_{(k-1)},\\ \Delta_{(k-1)} &= [\,(X_m\,\xi_{(k-1)} + X_{m'}\,\zeta_{(k-1)})\,L_{(k-1)}^{-1} \quad X\,\psi_{(k-1)}\,]\,U_{(k-1)}.\end{aligned}$$

Note the following m column vectors are in $\mathrm{span}(Y_{(k)})$*,* $Y_{(k)} = \rho(M)\,Q_{(k-1)} + \Delta_{(k-1)}$*:*

$$\begin{aligned}&X_m\,(\Gamma_m + \xi_{(k-1)}) + X_{p'}\,\Gamma_{p'}\,E_{(k-1)} + X_{m'}\,(\Gamma_{m'}\,F_{(k-1)} + \zeta_{(k-1)})\\ &\quad = \quad (X_m + X_{p'}\,E_{(k)} + X_{m'}\,F_{(k)})\,(\Gamma_m + \xi_{(k-1)}),\end{aligned}$$

where

$$\begin{aligned}E_{(k)} &= \Gamma_{p'}\,E_{(k-1)}\,(\Gamma_m + \xi_{(k-1)})^{-1}, \qquad \text{and}\\ F_{(k)} &= \Gamma_{m'}\,F_{(k-1)}\,(\Gamma_m + \xi_{(k-1)})^{-1} + \zeta_{(k-1)}\,(\Gamma_m + \xi_{(k-1)})^{-1}.\end{aligned}$$

We can B-orthonormalize these m columns via

$$(X_m + X_{p'}\,E_{(k)} + X_{m'}\,F_{(k)})\,L_{(k)}^{-1}$$

where

$$L_{(k)}^{*}L_{(k)} = (X_m + X_{p'}\,E_{(k)} + X_{m'}\,F_{(k)})^{*}\,B\,(X_m + X_{p'}\,E_{(k)} + X_{m'}\,F_{(k)}).$$

Recall that $B = C^{*}C$ and CX is unitary. Therefore

$$\begin{aligned}\|L_{(k)}\|^2 &\le \|(X_m + X_{p'}\,E_{(k)} + X_{m'}\,F_{(k)})^{*}\,B\,(X_m + X_{p'}\,E_{(k)} + X_{m'}\,F_{(k)}).\|,\\ &\le 1 + 2\|E_{(k)}\|\|F_{(k)}\| + \|E_{(k)}\|^2 + \|F_{(k)}\|^2,\\ &= 1 + (\|E_{(k)}\| + \|F_{(k)}\|)^2.\end{aligned}$$

Complete a B-orthonormal basis for $\mathrm{span}(Y_{(k)})$ *by adding an appropriate $p-m$ columns $V_{(k)}$. Since $Q_{(k)}$ is B-orthonormal for $k = 1, 2, \ldots$, there is a $p \times p$ unitary matrix $U_{(k)}$ such that*

$$Q_{(k)} = [\,(X_m + X_{p'}E_{(k)} + X_{m'}F_{(k)})L_{(k)}^{-1} \qquad V_{(k)}\,]\,U_{(k)}.$$

This completes the proof.

The next Lemma is a collection of several technical details needed in subsequent discussions.

Lemma 2. *Assume $|\gamma_p| > 0$ and consider any integer $m \le p$. Let $\alpha, \beta, \beta', \delta$ be positive numbers where*

$$\delta < 1, \quad \beta(1+\delta) \le 1/2, \quad \tau \stackrel{\text{def}}{=} 1 - \beta'(1+\delta) > 0.$$

Define the sequence $e_{\delta,0} \stackrel{\text{def}}{=} \alpha$, $f_{\delta,0} \stackrel{\text{def}}{=} 0$, and for $k = 1, 2, \ldots$,

$$e_{\delta,k} \stackrel{\text{def}}{=} \beta(1+\delta)e_{\delta,k-1}, \qquad f_{\delta,k} \stackrel{\text{def}}{=} \beta'(1+\delta)f_{\delta,k-1} + \delta(1+\delta)/2.$$

The followings hold.

1. *$e_{\delta,k} = \alpha(\beta(1+\delta))^k$ decreases as k increases. As long as $e_{\delta,k-1} \ge 2\delta$, $e_{\delta,k} + f_{\delta,k} \le e_{\delta,k-1} + f_{\delta,k-1}$.*
2. *$f_{\delta,k} \le \delta(1+\delta)/(2\tau)$ for all $k = 1, 2, \ldots$.*
3. *Consider any $m \times m$ matrix ξ. If $\|\xi\| \le \delta|\gamma_m|/2$, then $\|(I + \Gamma_m^{-1}\xi)^{-1}\| < 1+\delta$, $\|(I + \Gamma_m^{-1}\xi)^{-1} - I\| < \delta$, and the matrix $\Gamma_m + \xi$ is invertible with $(\Gamma_m + \xi)^{-1} = (I + \Gamma_m^{-1}\xi)^{-1}\,\Gamma_m^{-1}$.*
4. *Consider a Hermitian matrix of an arbitrary dimension ℓ of the form $I + \xi$, $\xi^{*} = \xi$ and $\|\xi\| \le 1/2$. Then $(I+\xi)^{1/2}$ and $(I+\xi)^{-1/2}$ are well defined and*

$$\|(I+\xi)^{\pm 1/2} - I\| \le \|\xi\|.$$

Proof. *That $e_{\delta,k} = \alpha(\beta(1+\delta))^k$ is clear, and $e_{\delta,k}$ is obviously a decreasing function in k because $\beta(1+\delta) \leq 1/2$ by assumption. For $k \geq 1$ and $e_{\delta,k-1} \geq 2\delta$, we have*

$$\begin{aligned} e_{\delta,k} + f_{\delta,k} &= \beta(1+\delta)e_{\delta,k-1} + \beta'(1+\delta)f_{\delta,k-1} + \delta(1+\delta)/2, \\ &\leq e_{\delta,k-1}/2 + f_{\delta,k-1} + \delta(1+\delta)/2, \quad \textit{because } \beta(1+\delta) \leq 1/2 \textit{ and } \beta'(1+\delta) < 1, \\ &\leq e_{\delta,k-1}/2 + f_{\delta,k-1} + \delta, \quad \textit{because } \delta < 1, \\ &\leq e_{\delta,k-1}/2 + f_{\delta,k-1} + e_{\delta,k-1}/2, \quad \textit{because } e_{\delta,k-1} \geq 2\delta, \\ &= e_{\delta,k-1} + f_{\delta,k-1}. \end{aligned}$$

This proves (1).

Next,

$$f_{\delta,k} < \frac{\delta}{2}(1+\delta)\sum_{\ell=0}^{\infty}(\beta'(1+\delta))^\ell \leq \frac{\delta(1+\delta)}{2(1-\beta'(1+\delta))} = \frac{\delta(1+\delta)}{2\tau},$$

which proves (2).

To prove (3), note that $\|\xi\| \leq \delta|\gamma_m|/2$ implies $\|\Gamma_m^{-1}\xi\| \leq \frac{\delta}{2} < 1/2$. Consequently, $(I+\Gamma_m^{-1}\xi)^{-1} = \sum_{\ell=0}^{\infty}(-1)^\ell(\Gamma_m^{-1}\xi)^\ell$ and

$$\|(I+\Gamma_m^{-1}\xi)^{-1}\| = \|\sum_{\ell=0}^{\infty}(-1)^\ell(\Gamma_m^{-1}\xi)^\ell\| \leq 1 + \frac{\delta}{2}\sum_{\ell=0}^{\infty}\left(\frac{\delta}{2}\right)^\ell < 1+\delta.$$

Similarly,

$$\|(I+\Gamma_m^{-1}\xi)^{-1} - I\| = \|\sum_{\ell=1}^{\infty}(-1)^\ell(\Gamma_m^{-1}\xi)^\ell\| \leq \frac{\delta}{2}\sum_{\ell=0}^{\infty}\left(\frac{\delta}{2}\right)^\ell < \delta.$$

Moreover, $\Gamma_m + \xi = \Gamma_m(I+\Gamma_m^{-1}\xi)$ must be invertible as $(\Gamma_m+\xi)^{-1} = (I+\Gamma_m^{-1}\xi)^{-1}\Gamma_m^{-1}$.

Finally, $\|\xi\| \leq 1/2$ implies $\|I+\xi\| \in [1-\|\xi\|, 1+\|\xi\|] \subseteq [1/2, 3/2]$. Thus $I+\xi$ is Hermitian positive definite, with an eigendecomposition ZDZ^*, $D = \mathrm{diag}(d_1, d_2, \ldots, d_\ell)$, $d_j \in [1/2, 3/2]$ for all $j = 1, 2, \ldots, \ell$. $(I+\xi)^{\pm 1/2} = Z\,D^{\pm 1/2}\,Z^*$.

$$\|(I+\xi)^{1/2} - I\| = \|D^{1/2} - I\| \leq \max_{|x|\leq\|\xi\|}\left|(1+x)^{1/2} - 1\right| \leq \|\xi\|,$$

where the last inequality holds because of $\|\xi\| \leq 1/2$. Similarly,

$$\|(I+\xi)^{-1/2} - I\| = \|D^{-1/2} - I\| \leq \max_{|x|\leq\|\xi\|}\left|(1+x)^{-1/2} - 1\right| \leq \|\xi\|,$$

where the last inequality holds because of $\|\xi\| \leq 1/2$. This completes the proof of the lemma.

Theorem 4. *Consider Algorithm FEAST where application of $\rho(M)$ to Q results in $\rho(M)Q+\Delta$. Specifically, Step 4 of FEAST becomes $Y_{(k)} \leftarrow \rho(M)Q_{(k-1)} + \Delta_{(k-1)}$. Suppose $|\gamma_p| > 0$, $U_{(0)} \stackrel{\text{def}}{=} X_p^* B Q_{(0)}$ is invertible, and that $\mathcal{Q}_{(k)}$ for all k have dimension p even in the presence of errors $\Delta_{(k)}$s. Let m be any integer $m \leq p$ and $E_{(0)}$ be the $(n-p)\times m$ matrix, the first m columns of $(X_{p'}^* B Q_{(0)})U_{(0)}^{-1}$. Define $\alpha \stackrel{\text{def}}{=} \|E_{(0)}\|$. Suppose there is a constant δ, $0 < \delta < 1$ such that the computational errors $\Delta_{(k)}$ always satisfy*

$$\|\Delta_{(k)}\| < \min\left\{\frac{\delta|\gamma_m|}{2\|C\|\|U_{(0)}^{-1}\|}, \frac{\delta|\gamma_m|}{2\|C\|\sqrt{1+\alpha^2}}\right\},$$

and that $|\gamma_{p+1}/\gamma_m|(1+\delta) \leq 1/2$ and $\tau \stackrel{\text{def}}{=} 1 - |\gamma_{m+1}/\gamma_m|(1+\delta) > 0$. Define the sequence $e_{\delta,k}, f_{\delta,k}$, $k = 0, 1, 2, \ldots$, as in Lemma 2 using the α and δ here, and $|\gamma_{p+1}/\gamma_m|$ as β, $|\gamma_{m+1}/\gamma_m|$ as β'. The followings hold.

1. *For $k = 1$, as well as for all subsequent $k = 2, 3, \ldots$, as long as $e_{\delta,k-1} \geq 2\delta$:*

$$Q_{(k)} = [\ (X_m + X_{p'}E_{(k)} + X_{m'}F_{(k)})L_{(k)}^{-1} \qquad V_{(k)}\ ]\,U_{(k)},$$

where $U_{(k)}$ is unitary of dimension $p \times p$,

$$\begin{aligned} E_{(k)} &= \Gamma_{p'}\, E_{(k-1)} (\Gamma_m + \xi_{(k-1)})^{-1}, \quad \text{and} \\ F_{(k)} &= \Gamma_{m'}\, F_{(k-1)} (\Gamma_m + \xi_{(k-1)})^{-1} + \zeta_{(k-1)} (\Gamma_m + \xi_{(k-1)})^{-1}. \end{aligned}$$

Furthermore, $\|E_{(k)}\| \le e_{\delta,k}$ and $\|F_{(k)}\| \le f_{\delta,k}$.

2. *For $k = 1$, as well as for all subsequent $k = 2, 3, \ldots$, as long as $e_{\delta,k-1} \ge 2\delta$, there are m vectors $\boldsymbol{s}_j \in \mathcal{Q}_{(k)}$, $j = 1, 2, \ldots, m$, such that*

$$\|\boldsymbol{x}_j - \boldsymbol{s}_j\|_B \le \alpha \left(\left| \frac{\gamma_{p+1}}{\gamma_m} \right| (1+\delta) \right)^k + \frac{\delta(1+\delta)}{2\tau},$$

and also, alternatively,

$$\|\boldsymbol{x}_j - \boldsymbol{s}_j\|_B \le \alpha \left| \frac{\gamma_{p+1}}{\gamma_j} \right|^k + \alpha\delta \left| \frac{\gamma_{p+1}}{\gamma_m} \right|^k \sum_{\ell=0}^{k-1} (1+\delta)^\ell + \frac{\delta(1+\delta)}{2\tau}.$$

Both bounds hold with $\|\boldsymbol{x}_j - \boldsymbol{s}_j\|_B$ replaced by $\|(I - P_{(k)})\boldsymbol{x}_j\|_B$.

Proof. *From definitions of $E_{(0)}$ and $U_{(0)}$,*

$$\begin{aligned} Q_{(0)} &= [\; X_m + X_{p'} E_{(0)} + X_{m'} F_{(0)} \quad V_{(0)} \;]\, U_{(0)}, \\ \Delta_{(0)} &= [\; X_m \xi_{(0)} + X_{m'} \zeta_{(0)} \qquad\quad X \psi_{(0)} \;]\, U_{(0)}, \end{aligned}$$

where, $F_{(0)}$ is the zero matrix of dimension $(n-m) \times m$,

$$X^{-1}\, \Delta_{(0)}\, U_{(0)}^{-1} = \left[\begin{array}{c|c} \xi_{(0)} & \psi_{(0)} \\ \zeta_{(0)} & \end{array}\right].$$

Because $\|\xi_{(0)}\|, \|\zeta_{(0)}\| \le \|C\| \|\Delta_{(0)}\| \|U_{(0)}^{-1}\| < \delta |\gamma_m|/2$, Lemma 2 shows that $\Gamma_m + \xi_{(0)}$ is invertible. Lemma 1 shows that (1) holds for $k = 1$. We use an induction argument. Suppose the followings hold for $k = 1, 2, \ldots, K-1$ for some $K \ge 2$:

$$\begin{aligned} Q_{(k)} &= [\; (X_m + X_{p'} E_{(k)} + X_{m'} F_{(k)}) L_{(k)}^{-1} \qquad V_{(k)} \;]\, U_{(k)}, \\ E_{(k)} &= \Gamma_{p'}\, E_{(k-1)}\, (\Gamma_m + \xi_{(k-1)})^{-1}, \\ F_{(k)} &= \Gamma_{m'}\, F_{(k-1)}\, (\Gamma_m + \xi_{(k-1)})^{-1} + \zeta_{(k-1)}\, (\Gamma_m + \xi_{(k-1)})^{-1}, \end{aligned} \tag{20}$$

where $\|L_{(k)}\| \le \sqrt{1 + (e_{\delta,k} + f_{\delta,k})^2}$, $\|E_{(k)}\| \le e_{\delta,k}$, $\|F_{(k)}\| \le f_{\delta,k}$. Suppose $e_{\delta,K-1} \ge 2\delta$, then $e_{\delta,K-2} \ge 2\delta$. Lemma 2 shows that $e_{\delta,K-1} + f_{\delta,K-1} \le e_{\delta,K-2} + f_{\delta,K-2} \le \cdots \le e_{\delta,0} + f_{\delta,0} = \alpha$. Consequently, $\|L_{(K-1)}\| \le \sqrt{1+\alpha^2}$. Examining the partition

$$X^{-1}\, \Delta_{(K-1)}\, U_{(K-1)}^{-1} \left[\begin{array}{c|c} L_{(K-1)} & \\ \hline & I \end{array}\right] = \left[\begin{array}{c|c} \xi_{(K-1)} & \psi_{(K-1)} \\ \zeta_{(K-1)} & \end{array}\right],$$

and noting that $U_{(k)}$ are all unitary for $k \ge 1$, we conclude that

$$\begin{aligned} \|\xi_{(K-1)}\|, \|\zeta_{(K-1)}\| &\le \|C\|\, \|\Delta_{(K-1)}\|\, \|L_{(K-1)}\|, \\ &\le \|C\|\, \|\Delta_{(K-1)}\|\, \sqrt{1+\alpha^2}, \\ &< \delta\, |\gamma_m|/2. \end{aligned}$$

Lemma 2 shows that $\Gamma_m + \xi_{(K-1)}$ is invertible. And hence by Lemma 1, Equation 20 holds for $k = K$ as well. Furthermore,

$$\begin{aligned} \|E_{(K)}\| &\le |\gamma_{p+1}|\, \|E_{(K-1)}\|\, \|(I + \Gamma_m^{-1} \xi_{(K-1)})^{-1}\|\, \|\Gamma_m^{-1}\|, \\ &\le |\gamma_{p+1}/\gamma_m| (1+\delta) e_{\delta,K-1}, \\ &= e_{\delta,K}, \\ \|F_{(K)}\| &\le |\gamma_{m+1}/\gamma_m|\, (1+\delta)\, \|F_{(K-1)}\| + \delta(1+\delta)/2, \\ &\le |\gamma_{m+1}/\gamma_m|\, (1+\delta)\, f_{\delta,K-1} + \delta\, (1+\delta)/2, \\ &= f_{\delta,K}. \end{aligned}$$

This establishes the first point of the theorem.

For $k = 1, 2, \ldots$, and as long as $e_{\delta,k-2} \geq 2\delta$, let $\boldsymbol{s}_j$ be the j-th column of $X_m + X_{p'} E_{(k)} + X_{m'} F_{(k)}$. The first bound is easy to obtain:

$$\begin{aligned}
\|\boldsymbol{x}_j - \boldsymbol{s}_j\|_B &\leq \|X_{p'} E_{(k)} + X_{m'} F_{(k)}\|_B, \\
&\leq \|X_{p'} E_{(k)}\|_B + \|X_{m'} F_{(k)}\|_B, \\
&= \|E_{(k)}\| + \|F_{(k)}\|, \\
&\leq e_{\delta,k} + f_{\delta,k}, \\
&\leq \alpha \left(\left| \frac{\gamma_{p+1}}{\gamma_m} \right| (1+\delta) \right)^k + \frac{\delta(1+\delta)}{2\tau}.
\end{aligned}$$

This bound is independent of the specific value of j, $1 \leq j \leq m$, but is given in terms of m. We can refine this by examining the j-th column of $E_{(k)}$ more closely. Denote the columns of $E_{(k)}$ by

$$E_{(k)} = [\ \boldsymbol{e}_1^{(k)} \boldsymbol{e}_2^{(k)}\ \cdots\ \boldsymbol{e}_m^{(k)}\ ].$$

Noting that

$$E_{(k)} = \Gamma_{p'}\, E_{(k-1)}\, (\Gamma_m + \xi_{(k-1)})^{-1} = \Gamma_{p'}\, E_{(k-1)}\, \Gamma_m^{-1} + \Gamma_{p'}\, E_{(k-1)}\, \left((I + \Gamma_m^{-1} \xi_{(k-1)})^{-1} - I \right) \Gamma_m^{-1},$$

$$\begin{aligned}
\|\boldsymbol{e}_j^{(k)}\| &\leq \left| \frac{\gamma_{p+1}}{\gamma_j} \right| \|\boldsymbol{e}_j^{(k-1)}\| + \left| \frac{\gamma_{p+1}}{\gamma_m} \right| \delta\, e_{\delta,k-1}, \\
&\leq \left| \frac{\gamma_{p+1}}{\gamma_j} \right| \left(\left| \frac{\gamma_{p+1}}{\gamma_j} \right| \|\boldsymbol{e}_j^{(k-2)}\| + \left| \frac{\gamma_{p+1}}{\gamma_m} \right| \delta\, e_{\delta,k-2} \right) + \left| \frac{\gamma_{p+1}}{\gamma_m} \right| \delta\, e_{\delta,k-1}, \\
&\leq \left| \frac{\gamma_{p+1}}{\gamma_j} \right|^2 \|\boldsymbol{e}_j^{(k-2)}\| + \delta \left| \frac{\gamma_{p+1}}{\gamma_m} \right|^2 e_{\delta,k-2} + \delta \left| \frac{\gamma_{p+1}}{\gamma_m} \right| e_{\delta,k-1}, \qquad \textit{(because } \left| \tfrac{\gamma_{p+1}}{\gamma_j} \right| \leq \left| \tfrac{\gamma_{p+1}}{\gamma_m} \right| \textit{)}, \\
&\leq \cdots\cdots\cdots, \\
&\leq \left| \frac{\gamma_{p+1}}{\gamma_j} \right|^k \|\boldsymbol{e}_j^{(0)}\| + \delta \sum_{\ell=0}^{k-1} \left| \frac{\gamma_{p+1}}{\gamma_m} \right|^{k-\ell} e_{\delta,\ell}, \\
&\leq \left| \frac{\gamma_{p+1}}{\gamma_j} \right|^k \|\boldsymbol{e}_j^{(0)}\| + \alpha\, \delta \sum_{\ell=0}^{k-1} \left| \frac{\gamma_{p+1}}{\gamma_m} \right|^{k-\ell} \left| \frac{\gamma_{p+1}}{\gamma_m} \right|^{\ell} (1+\delta)^{\ell}, \\
&\leq \alpha \left| \frac{\gamma_{p+1}}{\gamma_j} \right|^k + \alpha\, \delta \left| \frac{\gamma_{p+1}}{\gamma_m} \right|^k \sum_{\ell=0}^{k-1} (1+\delta)^{\ell}.
\end{aligned}$$

Therefore, an alternative bound on $\|\boldsymbol{x}_j - \boldsymbol{s}_j\|_B$ is

$$\|\boldsymbol{x}_j - \boldsymbol{s}_j\|_B \leq \alpha \left| \frac{\gamma_{p+1}}{\gamma_j} \right|^k + \alpha\, \delta \left| \frac{\gamma_{p+1}}{\gamma_m} \right|^k \sum_{\ell=0}^{k-1} (1+\delta)^{\ell} + \frac{\delta(1+\delta)}{2\tau}.$$

Both bounds apply to $\|(I - P_{(k)})\boldsymbol{x}_j\|_B$ as by definition it is $\min_{\boldsymbol{s} \in \mathcal{Q}_{(k)}} \|\boldsymbol{x}_j - \boldsymbol{x}\|_B$. This completes the proof of Theorem 4.

Theorems 3 shows that for each $\boldsymbol{x}_j$ of the m eigenvectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$, its distance to $\mathcal{Q}_{(k)}$ decreases to zero at rate of $|\gamma_{p+1}/\gamma_j|^k$. Theorem 4 shows that the error in applying $\rho(M)$ does not affect the convergence fundamentally. The convergence rate is degraded (very slightly) to $|\gamma_{p+1}/\gamma_m|^k (1+\delta)^k$, and the distance may only decrease down to a certain nonzero threshold, of the order δ that is commensurate with the accuracy of the linear solvers used to compute $\rho(M)Q$. In particular, iterative solvers are suitable for implementing $\rho(M)$.

5 Convergence of Eigenvalues and Residuals

The previous section shows that if the subspace dimension p in Algorithm FEAST is chosen large enough so that $|\gamma_{p+1}/\gamma_e| \ll 1$, then the generated subspaces $\mathcal{Q}_{(k)} = \mathrm{span}(Y_{(k)}) = \mathrm{span}(Q_{(k)})$ will capture rapidly the eigenvectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_e$. In fact, if $|\gamma_{p+1}/\gamma_m| \ll 1$ for some m, $e < m \le p$, the subspace will also capture the additional eigenvectors $\boldsymbol{x}_{e+1}, \boldsymbol{x}_{e+2}, \ldots, \boldsymbol{x}_m$ very well. This scenario is typical when there are eigenvalues outside of $[\lambda_-, \lambda_+]$ but close to the boundaries. This means that $\gamma_e \approx \gamma_{e+1} \approx \cdots \approx \gamma_m$ for some $m > e$. Thus $|\gamma_{p+1}/\gamma_e| \ll 1$ implies $|\gamma_{p+1}/\gamma_m| \ll 1$ as well.

Now to complete the story, we must show how to make use of these subspaces that have presumably captured the wanted eigenvectors, to actually obtaining the target eigenpairs $(\lambda_j, \boldsymbol{x}_j)$, $j = 1, 2, \ldots, e$. Specifically, we show that $m \ge e$ of the p eigenvalues of $\widehat{\Lambda}_{(k)}$ converge to $\lambda_1, \lambda_2, \ldots, \lambda_m$, so as the corresponding vectors in $Q_{(k)} = Y\widehat{X}_{(k)}$. Consider Steps 5 to 7 of FEAST at a particular iteration, omitting the subscript k, we have

$$\widehat{X}^* Y^* A Y \widehat{X} = (\widehat{X}^* Y^* B Y \widehat{X})\,\widehat{\Lambda} \;\Rightarrow\; Q^* A Q = \widehat{\Lambda}. \tag{21}$$

Theorems 3 and 4 show that Q is of the form

$$Q = [\;\; (X_m + X_{m'}G)L^{-1} \quad V \;\;]\,U,$$

where G encapsulates the E and F terms: $X_{m'}\,G = X_{p'}\,E + X_{m'}\,F$. Q is B-orthonormal, and U is unitary of dimension $p \times p$. The next lemma analyzes the structure of $\mathrm{span}(Q)$, which is key to convergence of eigenpairs (Theorem 5) and to useful properties of the $\widehat{B}$ and Y matrices (Theorem 6).

Lemma 3. *Consider a $n \times p$ B-orthonormal matrix Q of the form*

$$Q = [\;(X_m + X_{m'}G)L^{-1} \quad V\;]\,U$$

for some $m \le p$ where $\|G\| = \epsilon \le 1/2$, V is of dimension $n \times (p-m)$, U is unitary of dimension $p \times p$, and $L^{-1} = \left[\;(X_m + X_{m'}G)^*\,B\,(X_m + X_{m'}G)\;\right]^{-1/2}$. Then

1. *$\|L^{-1} - I_m\| \le \epsilon^2$. V can be represented as $V = X_m\,S + X_{m'}\,H\,R$ where $\|S\| \le \epsilon$, H is $(n-m)\times(p-m)$ where $H^*H = I_{p-m}$, and $\|R - I_{p-m}\| \le \epsilon^2$.*

2. $$\begin{aligned} X^*\,B\,Q &= \left(\left[\begin{array}{c|c} I_m & \\ \hline & H \end{array}\right] + \Theta\right) U, \\ \Theta &= \left[\begin{array}{c|c} \Theta_{11} & \Theta_{12} \\ \hline \Theta_{21} & \Theta_{22} \end{array}\right], \quad \Theta_{11} \text{ is } m \times m, \end{aligned}$$

   $\|\Theta_{11}\|, \|\Theta_{22}\| \le \epsilon^2$, $\|\Theta_{12}\|, \|\Theta_{21}\| \le (1+\epsilon^2)\epsilon$.

3. *Given any $n \times n$ real diagonal matrix $D = \mathrm{diag}(d_1, d_2, \ldots, d_n) = \mathrm{diag}(D_m, D_{m'})$,*

$$\begin{aligned} (Q^*\,B\,X)\,D\,(X^*\,B\,Q) &= U^* \left(\left[\begin{array}{c|c} D_m & \\ \hline & H^*\,D_{m'}\,H \end{array}\right] + \Delta\right) U, \\ \Delta &= \left[\begin{array}{c|c} \Delta_{11} & \Delta_{12} \\ \hline \Delta_{12}^* & \Delta_{22} \end{array}\right], \quad \Delta_{11} \text{ is } m \times m, \end{aligned}$$

   $\|\Delta_{11}\|, \|\Delta_{22}\| \le 4\|D\|\epsilon^2$, $\|\Delta_{12}\| \le 4\|D\|\epsilon$.

Proof. *Since $L^{-1} = (I_m + G^*G)^{-1/2}$, $\|G^*G\| \le \epsilon^2 \le 1/4 \le 1/2$, Lemma 2 shows that $\|L^{-1} - I_m\| \le \epsilon^2$. Q is B-orthonormal, and so is QU^* as U is a unitary matrix by assumption. Represent V in the basis vectors X_m and $X_{m'}$:*

$$V = (X_m X_m^* B + X_{m'} X_{m'}^* B)V = X_m S + X_{m'} T.$$

Note that CV, CX_m, and $CX_{m'}$ have orthonormal columns, and CX_m is orthogonal with $CX_{m'}$. Thus $\|S\|, \|T\| \le 1$. V being B-orthogonal with the first m columns of QU^* implies $S + G^*T = 0$. Therefore, $\|S\| = \|-G^*T\| \le \|G^*\| = \epsilon$. $V^*\,B\,V = I_{p-m}$ then implies

$$T^*T = I_{p-m} + (-S^*S), \qquad \|-S^*S\| \le \epsilon^2 \le 1/2.$$

Lemma 2 shows that $R \stackrel{\text{def}}{=} (T^*T)^{1/2}$ *satisfies* $\|R - I_{p-m}\| \le \epsilon^2$. *Clearly,* $H \stackrel{\text{def}}{=} TR^{-1}$ *leads to* $H^*H = I_{p-m}$. *Summarizing,* $V = X_m\, S + X_{m'}\, H\, R$, $\|S\| \le \epsilon$, $H^*H = I_{p-m}$, *and* $\|R - I_{p-m}\| \le \epsilon^2$. *This establishes the first point of the lemma.*

Since $Q = [\ (X_m + X_{m'}G)L^{-1} \quad X_m\, S + X_{m'}\, HR\ ]\, U$,

$$\begin{aligned} X^*\, B\, Q &= \begin{bmatrix} L^{-1} & S \\ GL^{-1} & HR \end{bmatrix} U, \\ &= \left(\left[\begin{array}{c|c} I_m & \\ \hline & H \end{array}\right] + \begin{bmatrix} L^{-1} - I_m & S \\ GL^{-1} & H(R - I_{p-m}) \end{bmatrix} \right) U, \\ &= \left(\left[\begin{array}{c|c} I_m & \\ \hline & H \end{array}\right] + \Theta \right) U, \end{aligned}$$

$\|\Theta_{11}\|, \|\Theta_{22}\| \le \epsilon^2$, $\|\Theta_{21}\| \le (1+\epsilon^2)\epsilon$, $\|\Theta_{12}\| \le \epsilon \le (1+\epsilon^2)\epsilon$. *This establishes the second point of the lemma.*

Let $D = \mathrm{diag}(D_m, D_{m'})$ *be any* $n \times n$ *real diagonal matrix. Using the structure of* X^*BQ *just established, we have*

$$\begin{aligned} (Q^*\, B\, X)\, D\, (X^*\, B\, Q) &= U^* \left(\left[\begin{array}{c|c} D_m & \\ \hline & H^*\, D_{m'}\, H \end{array}\right] + \Theta^* D + D\Theta + \Theta^* D \Theta \right) U, \\ &= U^* \left(\left[\begin{array}{c|c} D_m & \\ \hline & H^*\, D_{m'}\, H \end{array}\right] + \Delta \right) U. \end{aligned}$$

Bounding Δ *in a straightforward manner yields*

$$\begin{array}{rcccl} \|\Delta_{11}\| & \le & \|D\| \left(2\epsilon^2 + \epsilon^4 + (1+\epsilon^2)^2\epsilon^2\right) & \le & 4\|D\|\epsilon^2, \\ \|\Delta_{22}\| & \le & \|D\| \left(2\epsilon^2 + \epsilon^4 + \epsilon^2\right) & \le & 4\|D\|\epsilon^2, \\ \|\Delta_{12}\| & \le & \|D\| \left((1+\epsilon^2)^2\epsilon + (1+\epsilon^2)\epsilon\right) & \le & 4\|D\|\epsilon, \end{array}$$

all making use of the assumption $\epsilon \le 1/2$.

Theorem 5. *Consider Algorithm FEAST that exhibits subspace convergence as in Theorem 4. For any iteration* k, $Q_{(k)}$ *is represented as in Lemma 3*

$$Q_{(k)} = [\ (X_m + X_{m'}G_{(k)}){L_{(k)}}^{-1} \qquad X_m\, S_{(k)} + X_{m'}\, H_{(k)}R_{(k)}\ ]\, U_{(k)}.$$

Denote $\|G_{(k)}\|$ *by* ϵ_k *and define the spectral gap* η_k *as*

$$\eta_k \stackrel{\text{def}}{=} \begin{cases} \min_{\lambda \in \mathrm{eig}(\Lambda_m), \mu \in \mathrm{eig}(H^*_{(k)} \Lambda_{m'} H_{(k)})} |\lambda - \mu| / \|\Lambda\| & \text{if } m < p, \\ \infty & \text{if } m = p. \end{cases}$$

As long as $\epsilon_k \le 1/2$, *there are* m *eigenpairs* $(\widehat{\lambda}_j, \widehat{\boldsymbol{x}}_j)$ *among the* p *eigenpairs in* $(\widehat{\Lambda}_{(k)}, \widehat{X}_{(k)})$ *such that*

$$|\lambda_j - \widehat{\lambda}_j| \le 4\|\Lambda\| \left(\epsilon_k^2 + \min\{\epsilon_k, 4\epsilon_k^2/\eta_k\}\right),$$

and

$$\|A\boldsymbol{q}_j - \widehat{\lambda}_j B \boldsymbol{q}_j\| \le 12\|C\|\, \|\Lambda\| \epsilon_k\, (1 + 1/\eta_k).$$

Proof. *As displayed in Equation 21,* $Q^*_{(k)} A Q_{(k)} = \widehat{\Lambda}_{(k)}$ *so that the eigenvalues of* $\widehat{\Lambda}_{(k)}$ *are those of* $Q^*_{(k)} A Q_{(k)}$. *Because* $A = BX\, \Lambda\, X^*B$, *Lemma 3 shows that*

$$\begin{aligned} Q^*_{(k)} A Q_{(k)} &= (Q^*_{(k)} BX)\, \Lambda\, (XBQ_{(k)}), \\ &= U^*_{(k)}\, (W + \Delta_{\text{off}} + \Delta_{\text{diag}})\, U_{(k)}, \\ \mathrm{eig}(Q^*_{(k)} A Q_{(k)}) &= \mathrm{eig}\, (W + \Delta_{\text{off}} + \Delta_{\text{diag}}), \end{aligned}$$

where

$$W = \left[\begin{array}{c|c} \Lambda_m & \\ \hline & H^*_{(k)} \Lambda_{m'} H_{(k)} \end{array}\right], \quad \Delta_{\text{off}} = \left[\begin{array}{c|c} & \Delta_{12} \\ \hline \Delta_{12}^* & \end{array}\right], \quad \Delta_{\text{diag}} = \left[\begin{array}{c|c} \Delta_{11} & \\ \hline & \Delta_{22} \end{array}\right],$$

where Δ_{off} and Δ_{diag} are small perturbations: $\|\Delta_{\text{diag}}\| \leq 4\|\Lambda\|\epsilon_k^2$ and $\|\Delta_{\text{off}}\| \leq 4\|\Lambda\|\epsilon_k$. Of the p eigenvalues of W, m of them are $\lambda_1, \lambda_2, \ldots, \lambda_m$. We analyze the eigenvalues of $W + \Delta_{\text{off}} + \Delta_{\text{diag}}$ by standard Hermitian perturbation theory (see for example [9, 19, 28, 18]). First, there are m eigenvalues $\lambda'_1, \lambda'_2, \ldots, \lambda'_m$ of $W + \Delta_{\text{off}}$ such that

$$|\lambda_j - \lambda'_j| \leq \min\{\|\Delta_{\text{off}}\|, \|\Delta_{\text{off}}\|^2/(\eta_k\|\Lambda\|)\} \leq 4\|\Lambda\| \min\{\epsilon_k, 4\epsilon_k^2/\eta_k\}. \tag{22}$$

In the case $m = p$, $\Delta_{\text{off}} = 0$ and the definition of $\eta_k = \infty$ correctly reflects that $|\lambda_j - \lambda'_j| = 0$. Next, apply the standard Weyl perturbation theorem on $(W + \Delta_{\text{off}}) + \Delta_{\text{diag}}$ where Δ_{diag} is the perturbation term. There are m eigenvalues $\widehat{\lambda}_1, \widehat{\lambda}_2, \ldots, \widehat{\lambda}_m$ of $W + \Delta_{\text{off}} + \Delta_{\text{diag}}$ such that

$$|\lambda'_j - \widehat{\lambda}_j| \leq \|\Delta_{\text{diag}}\| \leq 4\|\Lambda\|\epsilon_k^2. \tag{23}$$

Combining Equations 22 and 23 gives

$$|\lambda_j - \widehat{\lambda}_j| \leq 4\|\Lambda\| \left(\epsilon_k^2 + \min\{\epsilon_k, 4\epsilon_k^2/\eta_k\}\right).$$

Moving on to examine the residual of the approximate eigenvector $\boldsymbol{q}_j$, note that

$$\boldsymbol{q}_j = [\ (X_m + X_{m'}G_{(k)})L_{(k)}{}^{-1} \quad X_m\, S_{(k)} + X_{m'}\, H_{(k)}R_{(k)}\ ]\, \boldsymbol{u}_j,$$

$\boldsymbol{u}_j$ being the j-th column of $U_{(k)}$.

$$\begin{aligned} Q_{(k)}^* A Q_{(k)} - \widehat{\Lambda} = 0 \quad &\Longrightarrow \quad U_{(k)}^*\ (\ W + \Delta_{\text{off}} + \Delta_{\text{diag}}\ )\, U_{(k)} - \widehat{\Lambda} = 0, \\ &\Longrightarrow \quad (\ W + \Delta_{\text{off}} + \Delta_{\text{diag}}\ )\, \boldsymbol{u}_j - \widehat{\lambda}_j\, \boldsymbol{u}_j = 0, \\ &\Longrightarrow \quad \|W\, \boldsymbol{u}_j - \widehat{\lambda}_j\, \boldsymbol{u}_j\| \leq 4\|\Lambda\|\epsilon_k(1+\epsilon_k). \end{aligned}$$

Partition $\boldsymbol{u}_j$ into its top m and bottom $p-m$ elements: $\boldsymbol{u}_j = \begin{bmatrix} \boldsymbol{u}_j^{(t)} \\ \boldsymbol{u}_j^{(b)} \end{bmatrix}$.

$$\eta_k\|\Lambda\|\|\boldsymbol{u}_j^{(b)}\| \leq \|H_{(k)}^*\Lambda_{m'}H_{(k)}\,\boldsymbol{u}_j^{(b)} - \widehat{\lambda}_j\,\boldsymbol{u}_j^{(b)}\| \leq 4\|\Lambda\|\epsilon_k(1+\epsilon_k).$$

Thus $\|\boldsymbol{u}_j^{(b)}\| \leq 4\epsilon_k(1+\epsilon_k)/\eta_k$. Furthermore,

$$\|(\Lambda_m - \widehat{\lambda}_j I)\boldsymbol{u}_j^{(t)}\| \leq 4\|\Lambda\|\epsilon_k(1+\epsilon_k).$$

Estimating the residual:

$$\begin{aligned} A\boldsymbol{q}_j &= A\,[\ (X_m + X_{m'}G_{(k)})L_{(k)}^{-1} \quad V_{(k)}\ ]\, \boldsymbol{u}_j, \\ &= (BX\,\Lambda\,X^*B)\,[\ (X_m + X_{m'}G_{(k)})L_{(k)}^{-1} \quad V_{(k)}\ ]\, \boldsymbol{u}_j, \\ &= B\left((X_m\Lambda_m + X_{m'}\Lambda_{m'}G_{(k)})L_{(k)}^{-1}\boldsymbol{u}_j^{(t)} + X\Lambda(X^*BV_{(k)})\boldsymbol{u}_j^{(b)}\right), \\ \widehat{\lambda}_j B\boldsymbol{q}_j &= B\left(\widehat{\lambda}_j\,(X_m + X_{m'}G_{(k)})L_{(k)}^{-1}\boldsymbol{u}_j^{(t)} + \widehat{\lambda}_j V_{(k)}\,\boldsymbol{u}_j^{(b)}\right), \\ A\boldsymbol{q}_j - \widehat{\lambda}_j\boldsymbol{q}_j &= B\Big(X_m(\Lambda_m - \widehat{\lambda}_j I)L_{(k)}^{-1}\boldsymbol{u}_j^{(t)} + X_{m'}(\Lambda_{m'} - \widehat{\lambda}_j I)G_{(k)}L_{(k)}^{-1}\boldsymbol{u}_j^{(t)} + \\ &\qquad (X\Lambda(X^*BV_{(k)}) - \widehat{\lambda}_j V_{(k)})\boldsymbol{u}_j^{(b)}\Big), \\ &= C^*\Big(CX_m(\Lambda_m - \widehat{\lambda}_j I)L_{(k)}^{-1}\boldsymbol{u}_j^{(t)} + CX_{m'}(\Lambda_{m'} - \widehat{\lambda}_j I)G_{(k)}L_{(k)}^{-1}\boldsymbol{u}_j^{(t)} + \\ &\qquad (CX\Lambda(X^*BV_{(k)}) - \widehat{\lambda}_j CV_{(k)})\boldsymbol{u}_j^{(b)}\Big). \end{aligned}$$

Note that $\|CX_m\|$, $\|CX_{m'}\|$, $\|CX\|$, and $\|X^*BV\|$ are all of unity as CX is unitary and V is B-orthonormal. Estimating $\|\Lambda_{m'} - \widehat{\lambda}_j I\| \leq 2\|\Lambda\|$ and using bounds of $\|L_{(k)}^{-1}\|$ and $\boldsymbol{u}_j^{(b)}$, we have

$$\begin{aligned} \|A\boldsymbol{q}_j - \widehat{\lambda}_j B\boldsymbol{q}_j\| &\leq \|C\|\|\Lambda\|\left(4\epsilon_k(1+\epsilon_k)(1+\epsilon_k^2) + 2\epsilon_k(1+\epsilon_k^2) + 8\epsilon_k(1+\epsilon_k)/\eta_k\right), \\ &\leq 12\|C\|\|\Lambda\|\epsilon_k\,(1 + 1/\eta_k), \end{aligned}$$

using $\epsilon_k \leq 1/2$. This completes the proof.

Theorem 5 shows that if the subspace dimension p is large enough, we would expect some m, $m \geq e$, eigenvalues among the p values of $\widehat{\Lambda}_{(k)}$ converge to the actual eigenvalues of $AX = BX\Lambda$. Furthermore, e of these eigenvalues are inside $\mathcal{I} = [\lambda_-, \lambda_+]$. If the spectral gaps η_k are never too small, the convergence rate of eigenvalues are essentially $|\gamma_{p+1}/\gamma_m|^{2k}$ while the residual vectors norms $\|A\boldsymbol{q}_j - \widehat{\lambda}_j B\boldsymbol{q}_j\|$ decrease at the rate $|\gamma_{p+1}/\gamma_m|^k$.

However, all we can conclude about the remaining $p-m$ eigenvalues (when $m < p$) is that they are close to the eigenvalues of $H_{(k)}^* \Lambda_{m'} H_{(k)}$, which can change at each iteration. As $H_{(k)}$ has orthonormal columns, each of these $p-m$ eigenvalues, $\mu \in \mathrm{eig}(H_{(k)}^* \Lambda_{m'} H_{(k)})$, satisfies $\min_{j>m} \lambda_j \leq \mu \leq \max_{j>m} \lambda_j$. In particular, some or all of them can fall inside $\mathcal{I}$. Hence there may be more than e eigenvalues of $\widehat{\Lambda}_{(k)}$ that fall inside $\mathcal{I}$. Our general experience is that, a posteriori, exactly e of the eigenvalues of $\widehat{\Lambda}_{(k)}$ fall inside $\mathcal{I}$. Nevertheless, knowing the value of e, a priori, can be exploited to help monitor convergence. It turns out that the value e can be accurately estimated as a by-product of Algorithm FEAST. Theorem 6 shows that the distribution of $\widehat{B}_{(k)}$'s eigenvalues offer a good estimate of e. More important, this distribution gives us an indication if the choice of p is too small. Due to the nature of $\rho(M)$, $p < e$ would in general lead to nonconvergence of FEAST. For example, consider $Q_{(0)} = X_e W$, where W is $e \times p$, $p < e$ and $W^*W = I_p$. Suppose $\rho(M)$ is the exact spectral projector, $\rho(M) = X_e X_e^*$. Let W^*AW have the spectral decomposition VDV^*. Algorithm FEAST will simply get stuck after the first iteration at $Q_{(k)} = X_e WV$ and $\widehat{\Lambda}_{(k)} = D$. The key reason is that by design $\rho(M)$ maps all the possibly distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_e$ to almost identically 1. We note that nonconvergence due to $p < e$ was observed in Experiment 3.1 of [15].

Theorem 6. *Consider Algorithm FEAST that exhibits subspace convergence as in Theorem 4. For any iteration k, $Q_{(k)}$ is represented as in Lemma 3*

$$Q_{(k)} = [\ (X_m + X_{m'} G_{(k)}) {L_{(k)}}^{-1} \quad X_m\, S_{(k)} + X_{m'}\, H_{(k)} R_{(k)}\ ]\, U_{(k)},$$

where $\|G_{(k)}\| = \epsilon_k$ is small. The eigenvalues of $\widehat{B}_{(k+1)}$ are close to those of $\mathrm{diag}(\Gamma_m^2, H_{(k)}^* \Gamma_{m'}^2 H_{(k)})$. The eigenvalues of $Z \stackrel{\mathrm{def}}{=} Q_{(k)}^* B Y_{(k+1)}$ are close to those of $\mathrm{diag}(\Gamma_m, H_{(k)}^* \Gamma_{m'} H_{(k)})$. In particular, the number e of target eigenvalues in $[\lambda_-, \lambda_+]$ can be estimated by the number of $\widehat{B}_{(k+1)}$'s eigenvalues that are no less than $1/4$, or the number of Z's eigenvalues no less than $1/2$.

Proof. *Given*

$$Q_{(k)} = [\ (X_m + X_{m'} G_{(k)}) {L_{(k)}}^{-1} \quad X_m\, S_{(k)} + X_{m'}\, H_{(k)} R_{(k)}\ ]\, U_{(k)}.$$

$$\begin{aligned}
\widehat{B}_{(k+1)} &= Y_{(k+1)}^*\, B\, Y_{(k+1)}, \\
&= Q_{(k)}^*\, \rho^*(M)\, B\, \rho(M)\, Q_{(k)}, \\
&= (Q_{(k)}^* BX)\, \Gamma X^* BX \Gamma\, (X^* B Q_{(k)}), \qquad \textit{because } \rho(M) = X\Gamma X^* B, \\
&= (Q_{(k)}^* BX)\, \Gamma^2\, (X^* B Q_{(k)}), \\
&= U_{(k)}^* \left(\mathrm{diag}(\Gamma_m^2,\, H_{(k)}^* \Gamma_{m'}^2 H_{(k)}) + \Delta\right) U_{(k)}, \qquad \textit{by Lemma 3.}
\end{aligned}$$

Clearly, the eigenvalues of $\widehat{B}_{(k+1)}$ are close to those of $\mathrm{diag}(\Gamma_m^2,\, H_{(k)}^* \Gamma_{m'}^2 H_{(k)})$. Similarly,

$$\begin{aligned}
Z &= Q_{(k)}^*\, B\, Y_{(k+1)}, \\
&= (Q_{(k)}^* BX)\, \Gamma\, (X^* B Q_{(k)}), \\
&= U_{(k)}^* \left(\mathrm{diag}(\Gamma_m,\, H_{(k)}^* \Gamma_{m'} H_{(k)}) + \Delta\right) U_{(k)}, \qquad \textit{by Lemma 3.}
\end{aligned}$$

Note however that $\gamma_1 \geq \cdots \gamma_e \geq 1/2 > |\gamma_{e+1}| \geq \cdots \geq |\gamma_n|$ and

$$\min_{j>m'} \gamma_j^2 \ \leq\ \mathrm{eig}(H_{(k)}^* \Gamma_{m'}^2 H_{(k)}) \ \leq\ \max_{j>m'} \gamma_j^2$$

because $H_{(k)}$ has orthonormal columns. For small $\|G_{(k)}\|$, the number of $\widehat{B}_{(k+1)}$'s eigenvalues no smaller than $1/4$ is an accurate estimate of e. Similar arguments shows that the number e can be estimated by counting the eigenvalues of Z that are no less than $1/2$.

Let us elaborate on a number of details related to FEAST as all the main theoretical properties have been presented.

1. The algorithm requires a choice of the subspace dimension, p. If the user has an educated guess of e, the actual number of eigenvalues in the search interval, p can be set to be about 1.5 times of that. Otherwise, a somewhat arbitrary choice is set. The following discussions are germane.

   - Whenever $p \geq e$, convergence is possible and the rate is generally determined by $|\gamma_{p+1}/\gamma_e|$. Tables 3 through 6 of Section 6 are illustrations. The examples there exhibit rates consistent with $|\gamma_{p+1}/\gamma_e|$. Theoretically, $p = e$ does not lead to nonconvergence. Nevertheless, in practice $|\gamma_{p+1}/\gamma_e| = |\gamma_{e+1}/\gamma_e|$ will be close to 1, rendering convergence slow. This slow convergence was observed in Experiment 3.1 of [15].
   - In general, a $p \geq e$ where $|\gamma_{p+1}/\gamma_e| \ll 1$ is desirable as it leads to fast convergence. As shown in Theorem 3, as long as $|\gamma_p| > 0$ and $X^*BQ_{(0)}$ has full column rank, $\rho(M)Q_{(k)}$ is of full rank p for all iterations k. Note that $\rho(M)$ is not the exact spectral projector and is theoretically almost always invertible. This is because $\rho(\mu)$ is a rational function and only has a small numbers of zeros on the real line (see Section 3.3). So in all likelihood $\rho(\Lambda)$ is invertible. However, $\rho(\mu)$ decays rapidly, so from a numerical point of view, $\rho(M)$ could be numerically rank deficient if p is chosen too large[1], for example, $p \geq 4e$. Consequently, $Y_{(1)} = \rho(M)Q_{(0)}$ will be rank deficient, leading to a semi-definite $\widehat{B}_{(1)}$. The conservative precautionary approach is to perform a SVD or a rank-revealing QR factorization [10, 13] on $Y_{(1)}$ to possibly reduce the value of p before proceeding further. However, we found that the greedy approach of letting LAPACK's Cholesky factorization on $\widehat{B}_{(1)}$ proceed naturally works very well in practice. If the factorization fails at the K-th column, we reset $p \leftarrow K - 1$ and use these first p columns of $Y_{(1)}$. That this strategy is effective has to do with the randomness of $Q_{(0)}$. While $\rho(M)$ is numerically rank deficient (of low rank), each $\rho(M)\boldsymbol{q}_j$ is a random mixture of all the columns of $\rho(M)$. QR without column pivoting on such a $\rho(M)Q_{(0)}$ is an effective "greedy" rank-revealing algorithm. The review article [11] and references thereof contain much information about recent works on randomized algorithms.
   - If $p < e$, then as discussed previously, FEAST in general will fail to converge. If $p \geq e$ but $|\gamma_{p+1}/\gamma_e|$ close to unity, convergence will be slow. We exploit Theorem 6 to protect against both scenarios. In practice, we compute the eigenvalues of $\widehat{B}_{(2)}$. If the minimum eigenvalue is bigger than threshold/4 for some "threshold" less than 1, for example, 1/10, we warn against p being set too small. For p not considered too small, our experience shows that the count of $\widehat{B}_{(2)}$'s eigenvalues not smaller than 1/4 to match e, the number of eigenvalues in the search interval.

2. The search interval $\mathcal{I} = [\lambda_-, \lambda_+]$.

   - This is an obvious feature for parallelism. One would be able to locate the eigenvalues and eigenvectors within different search intervals independently and simultaneously. The convergence theory established here shows that as long as $|\gamma_{p+1}/\gamma_e|$ is suitably small, eigenvalues within each search interval can be obtained, each with an eigenvector that results in small residual. (See Theorem 5 for detailed conditions.) Computed eigenvectors within one search interval are mutually B-orthogonal (assuming an accurate eigensolver is used for the reduced problem).
   - A natural application of the previous point is to partition one search interval $\mathcal{I} = [\lambda_-, \lambda_+]$ into a sequence of connecting intervals $\mathcal{I}^{(k)} = [\lambda^{(k-1)}, \lambda^{(k)}]$, $k = 1, 2, \ldots, K$, where $\lambda_- = \lambda^{(0)} < \lambda^{(1)} < \cdots < \lambda^{(K)} = \lambda_+$. Each $\mathcal{I}^{(k)}$ is tackled independently. Convergence theory applies on each subinterval. However, in the case when there is a cluster of eigenvalues around a break point, say $\lambda^{(k)}$, there will be a natural loss of B-orthogonality between the computed eigenvectors associated to the clusters on the left interval $\mathcal{I}^{(k)} = [\lambda^{(k-1)}, \lambda^{(k)}]$ and to those on the right $\mathcal{I}^{(k+1)} = [\lambda^{(k)}, \lambda^{(k+1)}]$. This phenomenon is due to sensitivity of eigenvectors associated with a cluster of eigenvalues (see [8] Section 5.2 for example). But Algorithm FEAST offers a natural strategy to handle this situation. If there is indeed a cluster of eigenvalues around one of the points of a search interval and if FEAST is indeed converging at a reasonable rate, then p must have been chosen large enough to include the number of eigenvalues in the cluster (including those outside of the search

[1] What makes p too large is obviously dependent of the actual distribution of the eigenvalues of the problem in question.

interval). Thus in the end, the computations for $\mathcal{I}^{(k)}$ and $\mathcal{I}^{(k+1)}$ will each have obtained all the clustered eigenvalues and a complete set of B-orthonormal eigenvectors. It suffices to adopt one of these two set of eigenvectors. Section 6.4 illustrates this idea.

3. Convergence criteria.

   - The original implementation FEAST Version 1.0 only monitors convergence of eigenvalues and does so through the surrogate of "trace," namely the sum of all the computed eigenvalues $\widehat{\Lambda}_{(k)}$ that fall inside $\mathcal{I}$. Our analysis exposes two issues of this strategy. First, eigenvalues in general converge faster than the residual norm. Thus the algorithm may terminate before the latter is driven down as small as level as is achievable. Second, it is possible that some of $\widehat{\Lambda}_{(k)}$'s eigenvalues are "spurious." These spurious eigenvalues generally do not converge, and monitoring them will only defeat FEAST's convergence test, resulting in a false negative.
   - FEAST Version 2.1 [23] corrected both problems. First, it offers the user an option to set convergence thresholds for either eigenvalues or residuals. Second, with the estimator of e, the actual number of eigenvalues inside the search interval $\mathcal{I}$, the absence of spurious eigenvalues is easily recognized. When the presence of spurious eigenvalues is detected, they are identified with the help of residual norms. The trace consists of the sum of computed eigenvalues without the spurious ones. Convergence of eigenvalues is monitored by the surrogate $|\text{trace}_{(k-1)} - \text{trace}_{(k)}|/max\{|\lambda_-|, |\lambda_+|\}$.

FEAST Version 2.1 [23] incorporated these improvements and is outlined below.

Algorithm FEAST with Estimate

Specify $\mathcal{I} = [\lambda_-, \lambda_+]$ and a Gauss-Legendre quadrature choice of q.
Pick p and p random n-vectors $Q_{(0)} = [\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_p]$. Set $k \leftarrow 1$.
repeat
    Approximate subspace projection (see Equation 8): $Y_{(k)} \leftarrow \rho(M) \cdot Q_{(k-1)}$.
    Form reduced system: $\widehat{A}_{(k)} \leftarrow Y^*_{(k)} A Y_{(k)}$, $\widehat{B}_{(k)} \leftarrow Y^*_{(k)} B Y_{(k)}$.
    if $k = 2$ **then**
        Compute $\widehat{B}_{(k)}$'s p eigenvalues.
        If minimum eigenvalue $\geq$ thres/4, report that p is probably too small.
        (Note: thres ≤ 1. FEAST Version 2.1 uses thres $= 1$.)
        Otherwise, set $\widehat{e}$ to be number of eigenvalues $\geq 1/4$. ($\widehat{e}$ estimates e.)
    end if
    Solve p-dimension eigenproblem: $\widehat{A}_{(k)}\widehat{X}_{(k)} = \widehat{B}_{(k)}\widehat{X}_{(k)}\widehat{\Lambda}_{(k)}$ for $\widehat{\Lambda}_{(k)}$, $\widehat{X}_{(k)}$.
    if $k = 1$ and the above fails due to non-definite $\widehat{B}_{(1)}$ **then**
        Reduce p to the last column of $\widehat{B}_{(1)}$ before Cholesky fails.
    end if
    Set $Q_{(k)} \leftarrow Y_{(k)}\widehat{X}_{(k)}$, in particular $Q^*_{(k)}\, B\, Q_{(k)} = I_p$.
    $k \leftarrow k + 1$.
until Stopping criteria based on the $\widehat{e}$ smallest residuals or trace of $\widehat{e}$ computed eigenvalues

6 Numerical Experiments

We present a number of numerical experiments to illustrate various aspects of our previous analyses. To this end, we utilize primarily synthetic, controlled, examples. Given the problem dimension n and a search interval $[\lambda_-, \lambda_+]$, we generate Λ, the diagonal matrix containing the eigenvalues, somewhat randomly, except for special placements of some of the eigenvalues near the boundaries of $[\lambda_-, \lambda_+]$. Random unitary matrices are the basic ingredient of our test matrices. With a specified condition number κ, random matrix C is generated as $U\Sigma V^*$ where U and V are random unitary matrices and Σ are random singular values so as to make the condition number of C equal κ. The matrix B is constructed as $B = C^*C$. The eigenvectors X are constructed by solving $CX = W$ where W is a random unitary matrix. Finally, the matrix A is constructed as $A = (BX)\,\Lambda\,(BX)^*$. This way,

$$AX = BX\Lambda,$$

and (X, Λ) is the eigenpair of the generalized eigenproblem defined by A and B.

6.1 Approximate spectral projector via quadrature

A crucial property of the quadrature-based approximate spectral projector $\rho(M)$ is that it preserves M's eigenvectors, $M = B^{-1}A$, and changes only its eigenvalues from Λ to $\Gamma = \rho(\Lambda)$ (see Equations 9 and 10):

$$\rho(M) \stackrel{\text{def}}{=} \sum_{k=1}^{K} \sigma_k (\gamma_k B - A)^{-1} \cdot B = X\, \rho(\Lambda) X^* B = X\, \rho(\Lambda)\, X^{-1}.$$

We generate matrices A, B, Λ, X (as outlined previously) of dimension $n = 300$ with the elements of Λ to be uniformly distributed in $[-30, 30]$. The C matrices used in generating $B = C^*C$ have condition number 100. We used Gauss-Legendre quadrature rule with 6, 8 and 10 quadrature points on $[-1, 1]$. For each test system and quadrature rule, we compute

$$\epsilon \stackrel{\text{def}}{=} \max_{1 \le j \le n} \frac{\|\boldsymbol{e}_j\|}{\|M\|}, \quad \boldsymbol{e}_j = \rho(M)\boldsymbol{x}_j - \rho(\lambda_j)\boldsymbol{x}_j.$$

For each quadrature rule, 200 test cases are generated and Table 2 tabulates the maximum, mean, and standard deviation of these 200 ϵs.

Statistics of $\{\|\rho(M)\boldsymbol{x}_j - \rho(\lambda_j)\boldsymbol{x}_j\|/\|M\|\}$	Quadrature Points of Gauss-Legendre 6	8	10
Maximum	5.5×10^{-15}	9.0×10^{-15}	1.0×10^{-14}
Mean	2.1×10^{-16}	2.4×10^{-16}	2.9×10^{-16}
Standard Deviation	5.5×10^{-16}	8.0×10^{-16}	9.8×10^{-16}

Table 2: *Key Property of Quadrature-Based Approximate Spectral Projector.* For each eigenpair $(\lambda_j, \boldsymbol{x}_j)$ of M, we check if indeed $\rho(M)\boldsymbol{x}_j \approx \rho(\lambda_j)\boldsymbol{x}_j$.

6.2 Convergence of subspace iteration using approximate spectral projector

To illustrate Theorem 3, we generate a complex generalized problem of dimension $n = 500$. We use Gauss-Legendre quadrature with 8 points on $[-1, 1]$. $[\lambda_-, \lambda_+]$ is set to $[15, 17]$. The n eigenvalues are generated as follows. We pick four eigenvalues in $[15, 17]$ by picking three randomly with uniform distribution in the region $[15.2, 16.8]$. The fourth is set to be 17. This guarantees that $|\rho(\lambda_j)| \approx 1$ for $j = 1, 2, 3$, and $|\rho(\lambda_4)| = 1/2$. These 4 eigenvalues are the only ones in $[15, 17]$ and hence $e = 4$. Next, five eigenvalues are set to be in the interval $(17, 18]$ such that the values of $|\rho(\lambda_j)|$ are $2^{-\ell}$ for $\ell = 3, 5, 7, 9, 11$. The remaining 491 eigenvalues are chosen randomly with uniform distribution on the set $[-40, 14] \cup [18, 60]$. The iteration of *Algorithm Subspace Iteration* is carried out with $p = 8$. With this choice of p, $|\gamma_{p+1}/\gamma_j|$ is 2^{-11} for $j = 1, 2, 3$, and 2^{-10}, 2^{-8}, up to 2^{-2} for the next 5 eigenvalues. Since the problem is generated, the eigenvectors $\boldsymbol{x}_j$ are known, and the projectors $P_{(k)} = Q_{(k)}Q_{(k)}^*B$ are easy to compute. We examine the quantities $\|(I - P_{(k)})\boldsymbol{x}_j\|_B$ for each of $j = 1, 2, \ldots, 8$ for 5 iterations $k = 1, 2, \ldots, 5$. Indeed these norms decrease in a way consistent with what the theorem predicts, except when the ultimate threshold of machine precision is reached. Table 3 tabulates the result.

To illustrate Theorem 4, we repeat the same experimental setting except that we added artificial errors to the linear solvers. To every solution z of equation of the form of Equation 8:

$$\sigma_k(\phi_k B - A)\boldsymbol{z} = B\boldsymbol{q},$$

we modify $\boldsymbol{z}$ by a random error of $2^{15}u$, u being the machine precision,

$$\boldsymbol{z} \leftarrow \boldsymbol{z} + 2^{15} u\, \|\boldsymbol{z}\|\, \Delta, \quad \text{each element of the } n\text{-vector } \Delta \text{ is uniformly random in } [-1/2, 1/2].$$

According to the bound of Theorem 4, which we restate here (see that section for details)

$$\|\boldsymbol{s}_j - \boldsymbol{x}_j\|_B \le \alpha \left|\frac{\gamma_{p+1}}{\gamma_j}\right|^k + \alpha\, \delta \left|\frac{\gamma_{p+1}}{\gamma_m}\right|^k \sum_{\ell=0}^{k-1} (1+\delta)^\ell + \frac{\delta(1+\delta)}{2\tau},$$

j	$\log_2\left\|\frac{\gamma_9}{\gamma_j}\right\|$	$\log_2\|(I-P_{(k)})x_j\|_B$ at Iteration k = 1	2	3	4	5
1	-11	-12.9	-23.8	-34.7	-43.3	-43.2
2	-11	-12.0	-22.8	-33.8	-43.3	-43.3
3	-11	-11.6	-22.4	-33.4	-43.4	-43.4
4	-10	-12.7	-22.6	-32.5	-41.2	-41.1
5	-8	-7.3	-15.1	-23.1	-31.1	-38.5
6	-6	-5.0	-10.8	-16.8	-22.8	-28.8
7	-4	-3.2	-7.0	-11.0	-15.0	-19.8
8	-2	-1.1	-3.0	-5.0	-6.9	-8.9

Table 3: *Subspace Convergence, Complex GHEP.* The convergence rate is consistent with the factor $|\gamma_{p+1}/\gamma_j|$. This test problem is designed with $p = 8$ and $\gamma_{p+1} = 2^{-11}$. There are 4 eigenvalues in $[\lambda_-, \lambda_+] = [15, 17]$ and the nature of $\rho(M)$ is that $1^+ \geq \gamma_j \geq 1/2$ for all $\gamma_j \in [\lambda_-, \lambda_+]$. Convergence rate for the targets are hence quite uniform, unlike standard subspace iteration applied with the original matrices.

where $\delta \approx 2^{15}u$ in our case here. We expect the overall convergence rate not to be affected. The ultimate accuracy limit is degraded commensurate with the artificial errors injected here. The parameter m is flexible. Thus for each eigenvalues λ_j, we can apply the bound with $m = j$. The bound suggests that the actual convergence limit is affected by the last term with the factor $1/|\gamma_m|$. Table 4 is consistent with these predictions. We note that the data in Experiment 3.3 of [15] is consistent with Theorems 4 and 5.

j	$\log_2\left\|\frac{\gamma_9}{\gamma_j}\right\|$	$\log_2\|(I-P_{(k)})x_j\|_B$ at Iteration k = 2	3	4	5	6	7	8	9
1	-11	-21.35	-32.34	-34.39	-34.50	-34.40	-34.46	-34.36	-34.34
2	-11	-22.97	-33.66	-34.34	-34.38	-34.33	-34.37	-34.33	-34.38
3	-11	-23.04	-33.68	-34.37	-34.35	-34.38	-34.42	-34.44	-34.29
4	-10	-19.51	-29.51	-33.40	-33.35	-33.39	-33.44	-33.39	-33.40
5	-8	-16.22	-24.22	-31.17	-31.38	-31.31	-31.39	-31.36	-31.32
6	-6	-12.20	-18.21	-24.21	-29.16	-29.31	-29.43	-29.45	-29.34
7	-4	-7.48	-11.48	-15.49	-19.49	-23.49	-26.96	-27.36	-27.36
8	-2	-5.02	-7.02	-9.01	-11.01	-13.01	-15.00	-17.00	-19.00

Table 4: *Subspace Convergence with Error in Linear System Solutions.* Note that the overall convergence rate is not affected by errors injected into the solutions of linear systems. The ultimate accuracy achieved is consistent with the error bound of Theorem 4. By Iteration 8, the generated subspaces have captured the best they ever can the eigenvectors 1 to 7. The ultimate achievable accuracy degrades by a factor of 2 from eigenvectors 3 to 7, consistent with the factor of $1/|\gamma_j|$, $j = 3, \ldots, 7$.

6.3 Eigenvalue and residual norm convergence

We illustrate important aspects of Algorithm FEAST as stated in Theorem 5. The first example is complex GHEP, dimension 500, with $[\lambda_-, \lambda_+] = [15, 17]$. We generate $e = 5$ eigenvalues well inside this interval. Eigenvalues outside of $[15, 17]$ are generated randomly except for a few specially placed so that $\gamma = 2^{-3,-5,-7,-9}$. Had p be set to $5 = e$, the convergence rate would be somewhat slow. With p set to 8, convergence rate for the target eigenpairs will be linear with a factor of 2^{-9}. The implication is that the three "collaterals" pair will also converge, except at a slower rate. This example reflects a typical scenario according to our experience with actual applications. There are often eigenvalues outside but quite close to the boundaries of $[\lambda_-, \lambda_+]$. As a result, the successful p will be strictly bigger than e and that the iterations will also

Convergence of Eigenvalues and Residual

j	$\log_2\left\|\frac{\gamma_9}{\gamma_j}\right\|$	$\log_2 \frac{\|\lambda_j-\tilde{\lambda}_j\|}{\|M\|}$ at Iteration k = 1	2	3	$\log_2 \frac{\|A\tilde{x}_j-\tilde{\lambda}_j B\tilde{x}_j\|}{\|M\|}$ at Iteration k = 2	3	4	5	6
1	-9	-33.74	-51.15	-64.04	-28.57	-37.56	-46.57	-52.93	-52.95
2	-9	-32.07	-49.50	-62.32	-27.83	-36.82	-45.83	-52.95	-52.86
3	-9	-34.40	-51.88	-61.90	-29.23	-38.22	-47.22	-53.15	-53.10
4	-9	-36.59	-54.23	-62.11	-30.46	-39.46	-48.46	-53.14	-53.14
5	-9	-34.79	-52.18	-61.23	-29.58	-38.57	-47.58	-52.88	-52.91
above are e target eigenvalues; below are "collaterals"									
6	-6	-30.73	-40.64	-52.61	-24.92	-30.91	-36.91	-42.91	-48.54
7	-4	-24.27	-30.73	-38.74	-20.03	-24.04	-28.05	-32.06	-36.07
8	-2	-24.55	-28.95	-32.99	-19.96	-22.08	-24.08	-26.09	-28.10

Table 5: *Convergence of Eigenvalues and Residual Vectors.* This table represents a typical scenario. Subspace dimension p is bigger than e but the "extra" dimensions also capture additional invariant subspaces, albeit slower. Note the eigenvalues converge linearly at the rate of $(\gamma_9/\gamma_j)^2$, while residuals do so at that of $|\gamma_9/\gamma_j|$.

obtain extra eigenpairs that can be called "collaterals." Table 5 shows the numerical details. The ratios are $|\gamma_{p+1}/\gamma_j| = 2^{-9}$ for the target eigenpairs. Note that eigenvalues accuracies improve by 2^{-18} per iteration as suggested by Theorem 5. This is typical, especially when the collaterals converge. In this event, unless the gap between the target and collateral eigenvalues are small, Theorem 5 predicts linear convergence of eigenvalues with the factor $(\gamma_{p+1}/\gamma_j)^2$.

The next example underlines the fact that $p \geq e$ suffices for convergence, and in particular for the case $p = e$. We generated two GHEP each of dimension 500 and $[\lambda_-, \lambda_+] = [15, 17]$. We place 5 eigenvalues in the interior, and 490 eigenvalues well separated from $[\lambda_-, \lambda_+]$. In the first test case, we place a cluster of 5 eigenvalues around the point μ where $\rho(\mu) = 2^{-3}$, and in the second case, around μ such that $\rho(\mu) = 2^{-7}$. We set $p = 5$ for both problems. Table 6 shows convergence for both cases at rates that correspond to the two different gaps. Along the same line, a setting of $p = e$ will result in slow convergence in practice as $|\gamma_{p+1}/\gamma_e| = |\gamma_{e+1}/\gamma_e|$ and will likely be close to unity. This observation is consistent with the slow convergence observed for $p = e$ in Figure 2 of [15].

The second example is similar to the first: complex GHEP, dimension 500. We generate $e = 5$ eigenvalues well inside this interval. Eigenvalues outside of $[15, 17]$ are generated randomly except for five specially-placed ones. One is placed so that $\gamma = 2^{-9}$, and four others are placed so that γ is strictly bigger than, but extremely close to, 2^{-9}. The other 491 eigenvalues are random but at least 0.5 away from $[15, 17]$. By setting $p = 9$, the convergence rate of the targets eigenvalues are expected to be linear with a factor $(2^{-9})^2 = 2^{-18}$ or smaller. But the collaterals do not converge. Table 7 exhibits this phenomenon.

The relationship between the subspace dimension p and the actual number of targets e can be subtle. In a typical scenario, $p > e$ and that the collaterals will also converge, except at a slower speed. But in the case when the collaterals do not converge, one might think that there is no fundamental harm in carrying them along except for a moderate increase of computational cost. Theorem 5 suggests some potential problems. Consider the previous example where the 9-dimensional subspaces capture the e target eigenvectors well, but not much of anything else. The reduced systems carry with them two subsystems. One is approximately Λ_e, and the other of the form $H^*\Lambda_{e'}H$ (in the notations of our theorems). If one is unlucky to have the eigenvalues of the second subsystem closely approximating some of the targets, convergence speed of target eigenvalues may be reduced to improvement of $|\gamma_{p+1}/\gamma_e|$ per step, as opposed to $|\gamma_{p+1}/\gamma_e|^2$. More important, some of the eigenvectors may actually be wrong! The residual may not converge to zero. We illustrate this phenomenon in the next example. For simplicity, we use a real-valued simple eigenvalue problem of dimension 500. We place just one eigenvalue $\lambda = 16$ in the middle of $[\lambda_-, \lambda_+] = [15, 17]$ but place two eigenvalues at $15 - \zeta$ and $17 + \zeta$ so that $\rho(15 - \zeta) = \rho(17 + \zeta) = 2^{-9}$. The remaining 497 eigenvalues are randomly generated except at least at a distance 3 away from $[15, 17]$. We set $p = 2$ and thus the target eigenvalue should converge at least

		Convergence of Eigenvalues and Residual								
		$\log_2 \frac{\lvert\lambda_j-\tilde{\lambda}_j\rvert}{\lVert M\rVert}$ at Iteration k =				$\log_2 \frac{\lVert A\tilde{x}_j-\tilde{\lambda}_j B\tilde{x}_j\rVert}{\lVert M\rVert}$ at Iteration k =				
j	$\log_2 \left\lvert\frac{\gamma_6}{\gamma_j}\right\rvert$	1	2	3	4	3	4	5	6	7
1	-3	-14.18	-20.03	-26.04	-32.05	-12.79	-15.80	-18.80	-21.80	-24.81
2	-3	-11.99	-17.74	-23.75	-29.76	-11.71	-14.71	-17.72	-20.72	-23.72
3	-3	-13.68	-20.77	-26.79	-32.80	-13.23	-16.25	-19.25	-22.26	-25.26
4	-3	-14.29	-20.42	-26.43	-32.44	-13.13	-16.14	-19.14	-22.15	-25.15
5	-3	-14.61	-20.67	-26.68	-32.68	-13.45	-16.46	-19.46	-22.47	-25.47
Above and below are two problems. The gaps are different.										
1	-7	-15.65	-29.64	-43.63	-51.58	-24.43	-31.43	-38.42	-45.41	-49.10
2	-7	-17.52	-31.50	-45.49	-50.99	-25.53	-32.53	-39.53	-46.50	-49.18
3	-7	-14.95	-28.92	-42.91	-50.58	-24.34	-31.34	-38.34	-45.33	-49.14
4	-7	-16.03	-30.02	-44.02	-51.32	-25.04	-32.04	-39.03	-46.01	-49.18
5	-7	-15.19	-29.18	-43.17	-50.58	-24.79	-31.78	-38.78	-45.77	-49.09

Table 6: *Convergence of Eigenvalues and Residual Vectors.* This table demonstrates convergence when $p = e$. Rate is fundamentally determined by the gap $|\gamma_{p+1}/\gamma_e|$. The two test problems here illustrate different convergence rates due to different gaps. In practice, however, $p = e$ will likely results in slow convergence unless all eigenvalues outside of the search interval $\mathcal{I} = [\lambda_-, \lambda_+]$ are far from it.

by 2^{-9} per iteration, but usually at 2^{-18} per step. We contrivedly start the iterations with two vectors, one close to the target eigenvector, and the other about the middle of the two eigenvectors associated with $15-\zeta$ and $17+\zeta$. That is, the Raleigh quotient with this vector is exactly 16. Table 8 illustrates the problem with a small gap between Λ_m and $H^*\Lambda_{m'}H$. As exhibited there, one of the two eigenvalues of the reduced system converge to 16, albeit only improving by 2^{-9} per step. Neither residual vector converges in any practical sense.

6.4 Multiple search intervals and splitting of clusters

Given several search intervals, FEAST can compute eigenpairs within a search interval totally independently. A natural use for this property is to split one large interval into several smaller ones, offering parallelism. As in Experiment 4.1 in [15], we apply this approach to the generalized Hermitian eigenvalue problem specified by the matrix pair `bcsstk11` and `bcsstm11` from Matrix Market[2]. We set $\mathcal{I}$ to $[0, 3.85 \times 10^7]$ and partition it into K equal-length (sub)intervals, $K = 1, 2, 3, 4, 5, 10$. Table 9 summarizes the result.

In this next example, FEAST computes eigenpairs of two attaching intervals $[1, 2]$ and $[2, 3]$ of a complex Hermitian eigenvalue problem ($B = I$) of dimension 500. A cluster of eigenvalues $2 \pm \ell \times 10^{-10}$, $\ell = 1, 2, \ldots, 5$, is placed around 2. In addition, there are 5 eigenvalues randomly placed in each of the interiors: $[1.2, 1.8]$ and $[2.2, 2.8]$. The remaining 480 eigenvalues are placed randomly outside of $[1, 3]$ separated by a distance of at least 0.5. Although there are 10 eigenvalues in each of the two search intervals, any $p \leq 15$ is detected as small by Algorithm FEAST with Estimate as all of $\widehat{B}_{(k)}$'s eigenvalues are large, due to a large $|\gamma_{p+1}|$. Both search intervals are handled with $p = 16$. In each search interval, all the associated spectrum together with the entire cluster, 15 eigenpairs in total, are obtained accurately in the sense of residuals at the level of machine roundoff by the fourth iteration.

We now number the 20 eigenvalues inside $[1, 3]$ from small to large. Denote the computed eigenpairs on the "left" and "right" intervals $[1, 2]$ and $[2, 3]$ by $(\widehat{\mu}_i, \boldsymbol{u}_i)$, $(\widehat{\nu}_j, \boldsymbol{v}_j)$, $1 \leq i \leq 15$, and $6 \leq j \leq 20$. Indices from 6 to 15 correspond to those of the eigenvalue cluster. Each of the two sets of 15 eigenvectors are mutually orthonormal. Table 10 shows the orthogonality properties across intervals. The natural strategy in handling two intervals sharing a cluster is to adopt the complete set of eigenpairs for the cluster from just one of the two intervals: $\{(\widehat{\mu}_i, \boldsymbol{u}_i) | 1 \leq i \leq 15\} \cup \{(\widehat{\nu}_j, \boldsymbol{v}_j) | 16 \leq j \leq 20\}$ or $\{(\widehat{\mu}_i, \boldsymbol{u}_i) | 1 \leq i \leq 5\} \cup \{(\widehat{\nu}_j, \boldsymbol{v}_j) | 6 \leq j \leq 20\}$.

[2] `http://math.nist.gov/MatrixMarket`

		Convergence of Eigenvalues $\log_2 \frac{\lvert\lambda_j - \tilde{\lambda}_j\rvert}{\lVert M\rVert}$ at Iteration					
j	$\log_2 \left\lvert\frac{\gamma_9}{\gamma_j}\right\rvert$	1	2	3	4	5	6
1	-9	-38.77	-56.77	-62.87	-63.00	-64.45	-66.45
2	-9	-35.38	-53.39	-61.17	-62.87	-62.65	-62.55
3	-9	-37.50	-55.51	-62.41	-62.37	-63.45	-61.81
4	-9	-36.77	-54.78	-65.55	-65.45	-63.13	-63.00
5	-9	-43.19	-61.21	-63.17	-62.21	-62.13	-64.87
above are e target eigenvalues; below are "collaterals"							
6	-0.0023	-33.53	-35.55	-35.56	-35.56	-35.56	-35.56
7	-0.0017	-32.36	-37.61	-37.62	-37.62	-37.63	-37.63
8	-0.0012	-31.26	-36.75	-36.77	-36.77	-36.77	-36.77
9	-0.0006	-30.29	-35.07	-35.07	-35.07	-35.07	-35.07

Table 7: *Non-Convergence of Collaterals.* There are 5 targets, and subspace dimension p is set to 9. The ratios $|\gamma_{p+1}/\gamma_j| \approx 1$ for $j = 6, 7, 8, 9$ and thus the collateral eigenvalues do not converge. These iterations would have been successful even if p was set to be just 5.

$p = 2, \left\lvert\frac{\gamma_3}{\gamma_1}\right\rvert = 2^{-9}$	Convergence Hampered by Spurious Eigenvalues Examine $\log_2(\delta_k/\lVert A\rVert)$ at Iterations $k =$							
δ_k is	1	2	3	4	5	6	7	8
$\min_{j=1,2} \lvert\tilde{\lambda}_j - 16\rvert$	-18.45	-27.46	-37.60	-47.42	-46.00	-46.42	-46.00	-46.19
$\min_{j=1,2} \lVert A\tilde{x}_j - \tilde{\lambda}_j\tilde{x}_j\rVert$	-6.13	-6.13	-6.20	-12.66	-15.84	-16.13	-15.92	-16.46

Table 8: $O(\epsilon^k)$ *Convergence of Eigenvalues and Non-Convergence of Residual.* This artificial example is set up so that there is only one eigenvalue, $\lambda = 16$, in the target interval. With $p = 2$ the ratio $\gamma_{p+1}/\gamma_1 = 2^{-9}$. In fact, $\gamma_2/\gamma_1 = 2^{-9}$ as well. The collateral space is affecting the overall convergence. Convergence of eigenvalue falls back to 2^{-9} per iteration, not at the often enjoyed speed of 2^{-18} per iteration. More importantly, the residual vector is not really converging. The $1/\eta$ factor in Theorem 5 is realistic. For this example, convergence will be restored to the perfect situation had p be set to 1.

6.5 Estimation of eigenvalue count

The number of eigenvalues in $[\lambda_-, \lambda_+]$ is valuable information; but guessing what that number is by counting the number of computed eigenvalues of reduced systems that fall inside $[\lambda_-, \lambda_+]$ is an unsound practice. Theorem 6 suggests that we can instead count the number of $\widehat{B}$'s eigenvalues $\geq 1/4$. In the following complex GHEP example of dimension 48, we generated 8 eigenvalues inside $[\lambda_-, \lambda_+] = [15, 17]$. We place on each side of $[\lambda_-, \lambda_+]$ 20 random eigenvalues of similar distribution to increase the chance of "spurious" eigenvalues. We set p to 12. Table 11 shows that the distribution of $\widehat{B}$'s eigenvalues is a much more robust indication of e than that of computed eigenvalues of reduced systems.

Along the same line, the next example in Table 12 shows that we can get an early indication that p is set too small by the eigenvalues of $\widehat{B}_{(k)}$. The example's setting is similar to the previous one, except p is set to 6, which is 2 less than the number of eigenvalues inside $[\lambda_-, \lambda_+] = [15, 17]$. The actual computed eigenvalues do not converge, which was to be expected.

	Number of equal-length partition of $[0, 3.85 \times 10^7]$					
	1	2	3	4	5	10
$\text{ortho}_{\text{all}}$	3.45×10^{-15}	2.73×10^{-14}	2.53×10^{-14}	2.64×10^{-14}	2.76×10^{-13}	3.55×10^{-13}
$\max_k \text{ortho}_k$	3.45×10^{-15}	5.25×10^{-15}	4.41×10^{-15}	5.42×10^{-15}	8.75×10^{-15}	5.49×10^{-15}
$\min_k \text{ortho}_k$	3.45×10^{-15}	2.84×10^{-15}	2.51×10^{-15}	1.58×10^{-15}	1.87×10^{-15}	9.80×10^{-16}

Table 9: *Matrix Market test problem using multiple search intervals.* Each of the (sub)interval is computed with $q = 16$. At most 3 applications of $\rho(M)$ were required for convergence for both eigenvalues and residual vectors to machine precision. We report the multual B-orthogonality within one subinterval and across all subintervals: $\text{ortho}_k \stackrel{\text{def}}{=} \max_{i,j} |\boldsymbol{x}_i^* B \boldsymbol{x}_j|$ for all distinct computed eigenvectors from the k-th subinterval, $k = 1, 2, \ldots, K$. The measure $\text{ortho}_{\text{all}}$ is defined similarly, except computed eigenvectors are drawn from all subintervals.

		$\max \lvert \boldsymbol{u}_i{}^* \boldsymbol{v}_j \rvert$
$6 \le i \le 10$	$11 \le j \le 15$	1.60×10^{-7}
$1 \le i \le 5$	$6 \le j \le 20$	1.25×10^{-14}
$1 \le i \le 15$	$16 \le j \le 15$	1.30×10^{-14}

Table 10: *Splitting a cluster into two search intervals.* FEAST is applied on two intervals $[1, 2]$ and $[2, 3]$, each having 10 eigenvalues, but the middle 10 of these 20 eigenvalues are clustered around 2, five on the left and five to the right. With subspace dimension set to $p = 16$, computation on each search interval produced 15 accurate eigenpairs: the 10 eigenpairs belong to its assigned interval and the 5 clustering ones in its neighbor. We number the computed eigenpairs on $[1, 2]$ as $(\widehat{\mu}_i, \boldsymbol{u}_i)$, $i = 1, 2, \ldots, 15$, and those on $[2, 3]$ as $(\widehat{\nu}_j, \boldsymbol{v}_j)$, $j = 6, 7, \ldots, 20$. $\widehat{\mu}_i = \widehat{\nu}_i$ up to machine roundoff for $i = 6, 7, \ldots, 15$. This table examines the orthogonality properties of the computed eigenvectors. The first row illustrates the fundamental nature of sensitivity of eigenvectors of clustering eigenvalues. Rows 2 and 3 show that one can adopt the entire cluster computed from either search interval to obtain a complete spectrum for $[1, 2] \cup [2, 3] = [1, 3]$.

7 Conclusions

We have shown that quadrature-based approximate spectral projectors are superb tools to be used with the standard subspace iteration method. This combination is the essence of the recently proposed FEAST algorithm and software ([21, 22]). Our detailed analysis establishes FEAST's convergence properties and shows how its robustness can be further enhanced as methods for counting target eigenvalues and detecting inappropriate subspace dimension are identified. Eigenproblems of large-and-sparse systems fit FEAST naturally as it can tolerate less-accurate solutions of linear systems, allowing the use of iterative linear solvers (see Example 3 in [21]).

Extension of the present work to non-Hermitian problems is a natural next step. Consider for now a simple non-Hermitian eigenvalue problem for a diagonalizable matrix A with an eigendecomposition $A = X \Lambda Y^*$, $XY^* = I$, where Λ is a diagonal matrix and X is a set of right eigenvectors. (Y is a set of left eigenvectors.) Hermitian FEAST is shown here to be subspace iteration with a special accelerator. Subspace iteration, however, is applicable for non-Hermitian problems, either focusing on the right (or left) eigenspace as in [27] or on both eigenspaces ([6] or [31] page 609). Furthermore, our approximate spectral projector accelerator is applicable to non-Hermitian matrices as well: Suppose $\mathcal{C}$ is a simple region (e.g. an ellipse) containing a spectrum of interest. Let $\rho(\mu)$ be of the form $\sum_{k=1}^{q} \alpha_k/(\beta_k - \mu)$ where none of the β_k's are in A's spectrum. Then $\rho(A) = X \rho(\Lambda) Y^*$. Provided $|\rho(\lambda)| \approx 1$ for $\lambda \in \text{eig}(A) \cap \mathcal{C}$ and $|\rho(\mu)| \ll 1$ for $\lambda \in \text{eig}(A) \setminus \mathcal{C}$, $\rho(A)$ approximates the (right) spectral projector $X_{\mathcal{C}} Y_{\mathcal{C}}^*$. ($\rho^*(A)$ approximates the left spectral projector $Y_{\mathcal{C}} X_{\mathcal{C}}^*$.) The function $\rho(\mu)$ can be constructed by quadrature rules applied to the Cauchy integral corresponding to $\mathcal{C}$. For example, define the parametrization for ellipses (similar to Equation 3):

$$\phi_a(t) = \cos\left(\frac{\pi}{2}(1+t)\right) + \iota a \sin\left(\frac{\pi}{2}(1+t)\right), \quad -1 \le t \le 3,$$

	$p = 12$ eigenvalues, μ_j, of $\widehat{B}_{(k)}$ at Iteration $k =$			
j	2	3	4	5
1	1.031865	1.042407	1.042588	1.042589
2	1.021825	1.037967	1.041980	1.042530
3	0.999644	1.000591	1.000626	1.000662
4	0.998992	0.999999	1.000000	1.000000
5	0.998206	0.999997	1.000000	1.000000
6	0.996776	0.999936	0.999999	1.000000
7	0.929957	0.999593	0.999857	0.999909
8	0.872044	0.989169	0.998930	0.999845
9	0.201241	0.210605	0.211077	0.211304
10	0.137805	0.146882	0.150190	0.153307
11	0.086650	0.095591	0.098854	0.104075
12	0.050975	0.078311	0.084910	0.088754
$\#\mu_j \geq 1/4$	8	8	8	8
$\#\tilde{\lambda}_j \in [\lambda_-, \lambda_+]$	10	9	9	9

Table 11: *Eigenvalue Count of $\widehat{B}_{(k)}$ to Estimate e.* In this example, there are exactly $e = 8$ eigenvalues in $[\lambda_-, \lambda_+] = [15, 17]$, $p = 12$. The computed eigenvalues of reduced problem may have more than 8 falling inside $[\lambda_-, \lambda_+]$. But the eigenvalue count of $\widehat{B}_{(k)}$ estimates e correctly from Iteration 2 onwards.

	Computed eigenvalues $\tilde{\lambda}_j$ of reduced system at Iteration $k =$				Eigenvalues μ_j of $\widehat{B}_{(k)}$, as a monitor, at Iteration $k =$			
j	2	3	4	5	2	3	4	5
1	15.30888	15.30960	15.30709	15.30358	1.03473	1.03600	1.03768	1.03997
2	15.53633	15.54323	15.54300	15.54067	1.01624	1.01788	1.02042	1.02356
3	16.18678	16.21928	16.22838	16.23129	1.00012	1.00016	1.00016	1.00016
4	16.54569	16.58401	16.58713	16.58817	0.99987	1.00002	1.00002	1.00002
5	16.59844	16.63769	16.66165	16.66953	0.99871	0.99984	0.99984	0.99985
6	16.81077	16.83859	16.85640	16.86408	0.74160	0.89480	0.96709	0.99060

Table 12: *Eigenvalue Count of $\widehat{B}_{(k)}$ to Judge p.* In this example, there are $e = 8$ eigenvalues in $[\lambda_-, \lambda_+] = [15, 17]$. But p is set too small at $p = 6$. Computed eigenvalues will not converge in general. This table illustrates that a too-small-p can be detected by examining $\widehat{B}_{(k)}$'s eigenvalues as early as at the second iteration. The symptom is that none of $\widehat{B}_{(k)}$'s eigenvalues are less than 1/4.

for a parameter a, $0 < a < \infty$. A parameter $a < 1$ corresponds to a flat ellipse, and $a > 1$, tall. One can construct a rational function $\rho(\mu)$ by applying a quadrature rule to $\pi(\mu)$ in Equation 4, with $\phi_a(t)$ in place of $\phi(t)$. Because A's spectrum can be complex, we need to study $\rho(\mu)$'s behavior for complex μ. Figure 5 illustrates that indeed the quadrature approach is effective. While more rigorous analysis is needed, the above discussions, supported by positive early experimental results in [16], make the idea of a non-Hermitian FEAST credible.

On a different note, we have used Gauss-Legendre quadrature as our numerical integrator of choice for $\rho(\lambda)$'s accurate approximation to the characteristic function $\pi(\lambda)$ on $[\lambda_-, \lambda_+]$ (see Equation 2). Nevertheless, accurate approximation of $\pi(\lambda)$ is by no means the only relevant property of an integrator suitable for FEAST. Investigation of other quadrature rules are worthwhile. One observation is that $|\rho(\lambda)|$ needs not approximate 1 very well on a large portion of $[\lambda_-, \lambda_+]$ or decay to zero outside of $[\lambda_-, \lambda_+]$ remarkably, both phenomena of which Gauss-Legendre possesses. It suffices to have, for example, $\rho(\lambda)$ fluctuates as long as $1^+ \geq \rho(\lambda) \geq \eta \gg 0$ on $[\lambda_-, \lambda_+]$ while keeping $|\rho(\lambda)|$ uniformly small outside $[\lambda_- - \delta, \lambda_+ + \delta]$ for some small $\delta > 0$. Another observation is that it is valuable to have a quadrature rule that provides increasing accuracy by progressively adding more nodes (while maintaining the existing ones). This would require us to find an alternative to Gauss-Legendre. In short, opportunities for further work are ample.

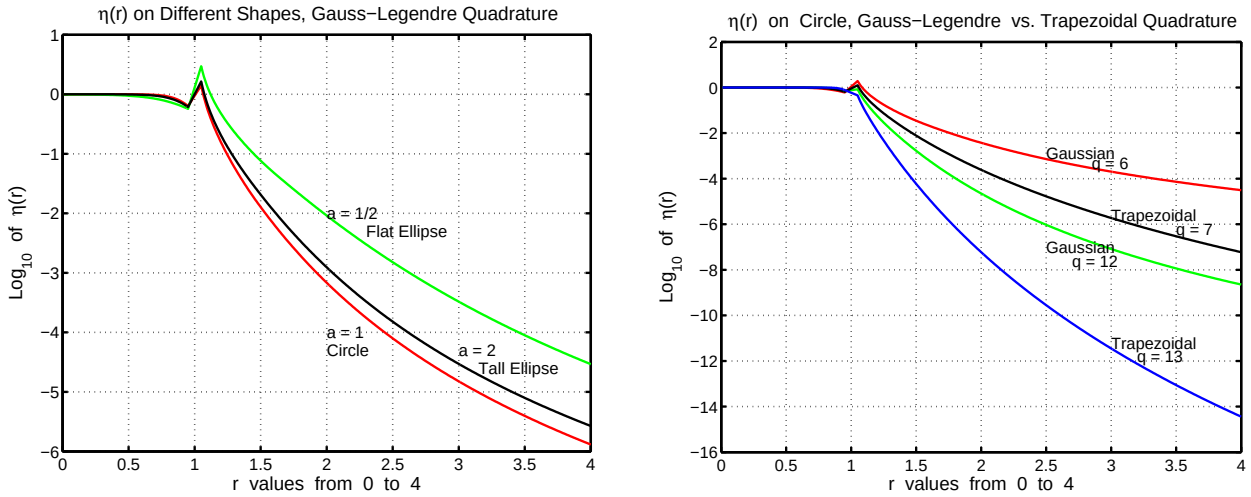

Figure 5: Let $\mu(r,t) = r[\cos(\frac{\pi}{2}(1+t))+\iota a\sin(\frac{\pi}{2}(1+t))]$ for some fixed $a > 0$. Define $\eta(r)$ to be $\min_t |\rho(\mu(r,t))|$ for $0 \leq r \leq 1-0.01$, and $\max_t |\rho(\mu(r,t))|$ for $r \geq 1+0.01$. The function $\eta(r)$ serves as an indicator of ρ's behavior as an approximate spectral projector. The plot on the left shows $\rho(\mu)$'s behavior on the complex plane via $\eta(r)$ for Gauss-Legendre $q = 8$ on different elliptical shapes. The plot on the right shows $\rho(\mu)$'s behavior on a circular search region for Gauss-Legendre and Trapezoidal quadratures of several different degrees.

8 Acknowledgments

We acknowledge the many fruitful discussions with Prof. Ahmed Sameh and Dr. Faisal Saied of Purdue University as well as Dr. Victor Kostin and Dr. Sergey Kuznetsov of Intel Corporation. In addition, Sergey Kuznetsov's rigorous testing of multiple versions of the FEAST software is invaluable.

References

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions*. Dover, Mineola, 1965.

[2] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorenson. *LAPACK Users Guide*. SIAM, Philadelphia, 3rd edition, 1999.

[3] A. Bai, J. Demmel, J. Dongarra, A. Petitet, H. Robinson, and K. Stanley. The spectral decomposition of nonsymmetric matrices on distributed memory parallel computers. *SIAM Journal on Scientific Computing*, 18(5):1446–1461, September 1997.

[4] Z. Bai and J. Demmel. Using the matrix sign function to compute invariant subspaces. *SIAM Journal on Matrix Analysis and Applications*, 10(1):205–225, January 1998.

[5] Z. Bai, J. Demmel, A. Ruhe, and H. van der Vorst. *Templates for the Solution of Algebraic Eigenvalue Problems*. SIAM, Philadelphia, 2000.

[6] F. L. Bauer. On modern matrix iteration processes of Bernoulli and Graeffe types. *Journal of the Association of Computing Machinery*, 5:246–257, 1958.

[7] J. Cullum and R. A. Willoughby. *Lanczos Algorithms for Large Symmetric Eigenvalue Computations*. Birkhäuser, Boston, 1985.

[8] J. Demmel. *Applied Numerical Linear Algebra*. SIAM, Philadelphia, 1997.

[9] G. H. Golub and C. F. van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 2nd edition, 1989.

[10] M. Gu and S. Eisenstat. Efficient algorithm for computing a strong rank-revealing QR factorization. *SIAM Journal on Scientific Computing*, 17(4):848–869, 1996.

[11] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

[12] N. J. Higham. *Functions of Matrices*. SIAM, Philadelphia, 2008.

[13] Y. P. Hong and C.-T. Pan. Rank-revealing QR factorizations and the singular value decomposition. *Mathematics of Computation*, 58(197):213–232, 1992.

[14] W. P. Johnson. The curious history of Faà di Bruno's formula. *American Mathematical Monthly*, 109:217–234, March 2002.

[15] L. Krämer, E. Di Napoli, M. Galgon, B. Lang, and P. Bientinesi. Dissecting the FEAST algorithm for generalized eigenvalue problems. *Journal of Computational and Applied Mathematics*, 244:1–9, May 2013.

[16] S. E. Laux. Solving complex band structure problems with the FEAST eigenvalue algorithm. *Physical Review B*, 86(075103), 2012.

[17] R. Lehoucq and D. Sorensen. Deflation techniques for an implicitly restarted Arnoldi iteration. *SIAM Journal on Matrix Analysis and Applications*, 17:789–821, 1996.

[18] C. Li and R. Li. A note on eigenvalues of perturbed Hermitian matrices. *Linear Algebra and Its Applications*, 295:221–229, 2005.

[19] R. Mathias. Quadratic residual bounds for the Hermitian eigenvalue problem. *SIAM Journal on Matrix Analysis and Applications*, 19:541–550, 1998.

[20] B. Parlett. *The Symmetric Eigenvalue Problem*. SIAM, Philadelphia, 1998.

[21] E. Polizzi. Density-matrix-based algorithm for solving eigenvalue problems. *Physical Review B*, 79(115112), 2009.

[22] E. Polizzi. The FEAST solver. `http://www.ecs.umass.edu/~polizzi/feast/`, 2009.

[23] E. Polizzi. Latest version of the free version of FEAST. `http://www.feast-solver.org`, 2013.

[24] Y. Saad. *Numerical Methods for Large Eigenvalue Problems*. SIAM, Philadelphia, 2011.

[25] A. Sameh and Z. Tong. The trace minimization method for the symmetric generalized eigenvalue problem. *Journal on Computational and Applied Mathematics*, 123:155–175, 2000.

[26] A. H. Sameh and J. A. Wisniewski. A trace minimization algorithm for the generalized eigenvalue problem. *SIAM Journal on Numerical Analysis*, 19(6):1243–1259, 1982.

[27] G. W. Stewart. Simultaneous iteration for computing invariant subspaces of non-Hermitian matrices. *Numerische Mathematik*, 25:123–136, 1976.

[28] G. W. Stewart and J.-G. Sun. *Matrix Perturbation Theory*. Academic Press, Boston, 1990.

[29] J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis*. Springer-Verlag, New York, 3rd edition, 2010.

[30] G. Viaud. The FEAST algorithm for generalised eigenvalue problems. Master's thesis, University of Oxford, Oxford, England, 2012.

[31] J. H. Wilkinson. *The Algebraic Eigenvalue Problem*. Clarendon Press, Oxford, 1965.