Spectrum Slicing for Sparse Hermitian Definite Matrices Based on Zolotarev's Functions

Yingzhou Li♯, Haizhao Yang†,

♯ ICME, Stanford University
† Department of Mathematics, Duke University

March 1, 2017

Abstract

This paper proposes an efficient method for computing selected generalized eigenpairs of a sparse Hermitian definite matrix pencil (A, B). Based on Zolotarev's best rational function approximations of the signum function and conformal maps, we construct the best rational function approximation of a rectangular function supported on an arbitrary interval. This new best rational function approximation is applied to construct spectrum filters of (A, B). Combining fast direct solvers and the shift-invariant GMRES, a hybrid fast algorithm is proposed to apply spectral filters efficiently. Assuming that the sparse Hermitian matrices A and B are of size $N \times N$ with $O(N)$ nonzero entries, the computational cost for computing $O(1)$ interior eigenpairs is bounded by that of solving a shifted linear system $(A - \sigma B)x = b$. Utilizing the spectrum slicing idea, the proposed method computes the full eigenvalue decomposition of a sparse Hermitian definite matrix pencil via solving $O(N)$ linear systems. The efficiency and stability of the proposed method are demonstrated by numerical examples of a wide range of sparse matrices. Compared with existing spectrum slicing algorithms based on contour integrals, the proposed method is faster and more reliable.

Keywords. Generalized eigenvalue problem, spectrum slicing, rational function approximation, sparse Hermitian matrix, Zolotarev's function, shift-invariant GMRES.

AMS subject classifications: 44A55, 65R10 and 65T50.

1 Introduction

Given a sparse Hermitian definite matrix pencil (A, B) (i.e., A and B are Hermitian and B is positive-definite) in $\mathbb{F}^{N\times N}$, where $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$, the generalized eigenvalue problem computes the following matrix decomposition

$$AX = BX\Lambda, \tag{1}$$

where $X \in \mathbb{F}^{N\times N}$ consists of N eigenvectors $\{x_j\}_{1\le j\le N}$, and Λ is a diagonal matrix with diagonal entries $\{\lambda_j\}_{1\le j\le N}$ being the real eigenvalues. Throughout this paper, we assume that eigenvalues are ordered, i.e., $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_N$. The spectrum slicing method [3, 30, 11, 23, 25, 26, 28, 32, 35] slices the spectrum of $B^{-1}A$ into m intervals $\{(a_j, b_j)\}_{1\le j\le m}$ covering the entire spectrum, i.e., $[\lambda_1, \lambda_N] \subset \cup_{j=1}^{m}(a_j, b_j)$, and solves the interior generalized eigenvalue problem on each interval (a_j, b_j) independently. Naturally, the original generalized eigenvalue problem can be divided into m jobs and solved in parallel. If more than m processes are available, they can be organized as m disjoint process groups and each process group solves an interior generalized eigenvalue problem further in parallel.

Therefore, this paper focuses on solving the interior generalized eigenvalue problem as follows. Given a matrix pencil (A, B), an interval (a, b), and the number of eigenvalues of (A, B) in (a, b) (denoted as n_λ[1]), identify all interior eigenvalues $\{\lambda_j\}_{1\le j\le n_\lambda}$, and their corresponding eigenvectors $\{x_j\}_{1\le j\le n_\lambda}$ such that,

$$Ax_j = \lambda_j B x_j \quad \text{and} \quad a < \lambda_j < b, \quad j = 1, 2, \ldots, n_\lambda. \tag{2}$$

The interior generalized eigenvalue problem not only can be applied to solve the full generalized eigenvalue problem via the spectrum slicing idea, but also is a stand-alone problem encountered in many fields in science and engineering (such as computational chemistry, control theory, material science, etc.), where a partial spectrum is of interest. For a full implementation of incorporating an interior eigensolver to the spectrum slicing problem, readers are referred to [3, 5, 28, 35] for detailed descriptions.

1.1 Related Work

A powerful tool for solving the interior generalized eigenvalue problem is the subspace iteration method accelerated by spectrum filters. Let $P_{ab}(A, B)$ be an approximate spectrum projector onto the eigen-subspace of the matrix pencil (A, B) corresponding to eigenvalues in (a, b), i.e.,

$$\text{span}\left(P_{ab}(A, B)\right) \approx \text{span}\left(x_1, x_2, \ldots, x_{n_\lambda}\right), \tag{3}$$

where $x_1, x_2, \ldots, x_{n_\lambda}$ are the interior generalized eigenvectors of (A, B) as defined in (2). Here, we assume that the interval endpoints a and b are not eigenvalues of the matrix pencil (A, B). A typical subspace iteration method contains the following main steps after initializing the guess of eigenvectors:

1. Apply the spectrum projector $P_{ab}(A, B)$ to the basis of the current guess of the eigenvectors;
2. Orthonormalize the filtered results to obtain an approximate eigen-subspace;
3. Apply the Rayleigh-Ritz procedure to obtain the estimation of the desired eigenpairs;
4. Use the eigenvectors provided by the Rayleigh-Ritz procedure as a new guess of the eigenvectors.

All these steps are executed repeatedly until the eigenpairs converge. The number of iterations depends on the approximation accuracy of the spectrum projector. Ideally, if one could construct $P_{ab}(A, B) = S_{ab}(B^{-1}A)$, where $S_{ab}(x)$ is a rectangular function with a support on (a, b), the subspace iteration method converges in one iteration. However, $S_{ab}(B^{-1}A)$ is not available unless we know the full generalized eigendecomposition of the matrix pencil (A, B). A feasible way is to design a filter function $R_{ab}(x)$ as a good approximation of $S_{ab}(x)$ such that we can evaluate $P_{ab}(A, B) = R_{ab}(B^{-1}A)$ without the full eigendecomposition of (A, B). If such a filter function $R_{ab}(x)$ is available, the subspace iteration method converges quickly.

In the spectrum slicing, there are mainly two kinds of spectrum filters: polynomial filters [11, 28] and rational filters [3, 30, 31, 23, 25, 26, 32, 35], approximating the spectrum projector using a polynomial or a rational function $R_{ab}(x)$. The difficulty in designing an appropriate filter comes from the dilemma that: an accurate approximation to the spectrum projector requires a polynomial

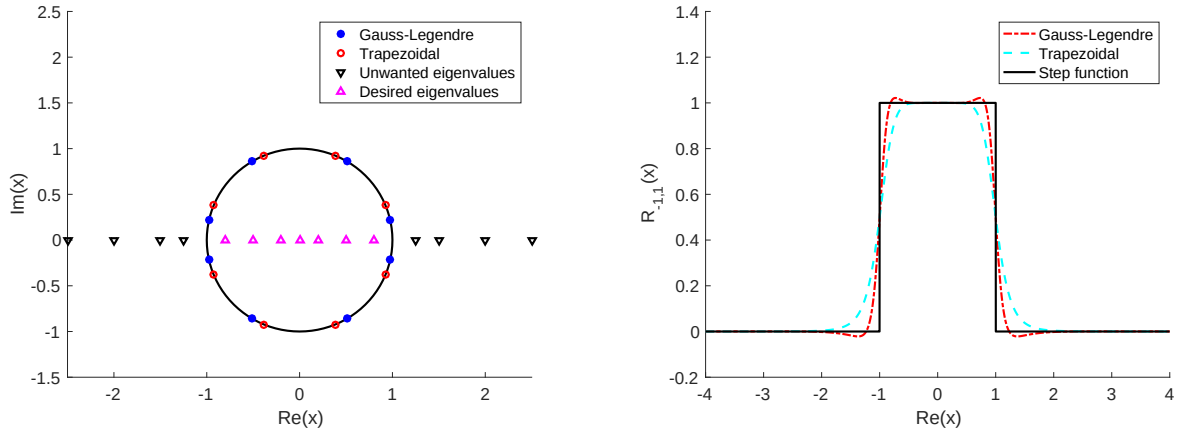
[1] Through out the paper, we assume that the exact number of eigenvalues, n_λ, is known a priori, which would simplify the presentation of the method. While, in practice, an estimated number of n_λ is enough for the algorithm.

of high degree or a rational function with many poles; however this in turn results in expensive computational cost in applying the spectrum projector $R_{ab}(B^{-1}A)$.

In general, a rational filter can be written as follows

$$R_{ab}(x) = \alpha_0 + \sum_{j=1}^{p} \frac{\alpha_j}{x - \sigma_j}, \tag{4}$$

where $\{\alpha_j\}_{0\le j\le p}$ are weights, $\{\sigma_j\}_{1\le j\le p}$ are poles, and p is the number of poles. Hence, applying the spectrum projector $R_{ab}(B^{-1}A)$ to a vector v requires solving p linear systems $\{(A - \sigma_j B)^{-1}Bv\}_{1\le j\le p}$. Therefore, a large number p makes it expensive to apply the approximate spectrum projector $R_{ab}(B^{-1}A)$. A natural idea is to solve the linear systems $\{(A - \sigma_j B)^{-1}Bv\}_{1\le j\le p}$ in parallel. However, for the purposes of energy efficiency and numerical stability, an optimal p is always preferred. Extensive effort has been made to developing rational functions with p as small as possible while keeping the accuracy of the approximation.



(a) (b)

Figure 1: (a) An example of a unit circle contour Γ centered at the origin, i.e., the desired spectrum range is $(-1, 1)$, together with eight Gauss-Legendre quadrature points (solid blue circle) and eight Trapezoid quadrature points (red circle). The desired eigenvalues (pink up triangular) are inside the contour whereas the unwanted eigenvalues (black down triangular) are outside. (b) A rectangular function in solid black line together with rational functions corresponding to the quadratures from (a).

Many rational filters in the literature were constructed by discretizing the contour integral on the complex plane,

$$\pi(x) = \frac{1}{2\pi\imath} \oint_{\Gamma} \frac{1}{x - z}\, \mathrm{d}z, \quad x \notin \Gamma \tag{5}$$

with an appropriate quadrature rule (e.g., the Gauss-Legendre quadrature rule [23], the trapezoidal quadrature rule [29, 35], and the Zolotarev quadrature rule [8]). Here Γ is a closed contour on the complex plane intersecting the real axis at $z = a$ and $z = b$ with all desired eigenvalues inside (a, b) and other eigenvalues outside (See Figure 1 (left) for an example). Suppose $\{\sigma_j\}_{1\le j\le p}$ and $\{w_j\}_{1\le j\le p}$ are the quadrature points and weights in the discretization of the contour Γ, respectively, the contour integral (5) is discretized as a rational function

$$R(x) = \sum_{j=1}^{p} \frac{w_j}{2\pi\imath(x - \sigma_j)} = \alpha_0 + \sum_{j=1}^{p} \frac{\alpha_j}{x - \sigma_j}, \tag{6}$$

where $\alpha_0 = 0$, and $\alpha_j = \frac{w_j}{2\pi\imath}$ for $j = 1, 2, \ldots, p$. Some other methods advanced with conformal maps [9, 13] and optimization [30, 32] can also provide good rational filters.

1.2 Contribution

Based on Zolotarev's best rational function approximations of the signum function and conformal maps, we construct the best rational function $R_{ab}(x)$ approximating a rectangular function supported on an arbitrary interval (a, b). Combining fast direct solvers and the shift-invariant GMRES, a hybrid fast algorithm is proposed to apply the spectrum filter $R_{ab}(B^{-1}A)$ to given vectors.

Suppose $a \in (a_-, a_+)$ and $b \in (b_-, b_+)$ respectively, and no eigenvalue lies in (a_-, a_+) and (b_-, b_+). The proposed rational filter $R_{ab}(x)$ is constructed via the composition of Zolotarev's functions as follows

$$R_{ab}(x) = \frac{Z_{2r}(\widehat{Z}_{2r}(T(x); \ell_1); \ell_2) + 1}{2}, \tag{7}$$

where $Z_{2r}(x; \ell)$ is the Zolotarev's function of type $(2r - 1, 2r)$, $\widehat{Z}_{2r}(x; \ell)$ is the scaled Zolotarev's function

$$\widehat{Z}_{2r}(x; \ell) = \frac{Z_{2r}(x; \ell)}{\max_{x \in [\ell, 1]} Z_{2r}(x; \ell)}, \tag{8}$$

and $T(x)$ is a Möbius transformation of the form

$$T(x) = \gamma \frac{x - \alpha}{x - \beta} \tag{9}$$

with $\alpha \in (a_-, a_+)$ and $\beta \in (b_-, b_+)$ such that

$$T(a_-) = -1, \quad T(a_+) = 1, \quad T(b_-) = \ell_1, \quad \text{and } T(b_+) = -\ell_1. \tag{10}$$

In the above construction, the variables α, β, γ, ℓ_1, and ℓ_2 are all determined by a_-, a_+, b_-, and b_+.

The novelty of the proposed rational filter in (7) is to construct a high-order rational function via the composition of two Zolotarev's functions and a Möbius transformation. This new construction can significantly improve the approximation accuracy for a rectangular function approximation even if r is small, as compared to other methods via a single Zolotarev's functions in [8]. Similar composition ideas have been applied to the signum function approximation (e.g., polar decomposition of matrices [17], full diagonalization of matrices [18], and the density matrix purification [16, 19, 21]), and the square root function approximation for accelerating Herons iteration [4, 20].

An immediate challenge arises from applying the composition of functions $R_{ab}(B^{-1}A)$ in (7) to given vectors when A and B are sparse matrices. Directly computing $R_{ab}(B^{-1}A)$ will destroy the sparsity of A and B since $R_{ab}(B^{-1}A)$ is dense. Fortunately, the function composition structure in (7) admits a hybrid fast algorithm for the matrix-vector multiplication (matvec) $R_{ab}(B^{-1}A)V$, where A and B are sparse Hermitian matrices of size N with $O(N)$ nonzero entries, B is positive definite, and V is a tall skinny matrix of size N by $O(1)$. We apply the multifrontal method [6, 14] to solve the sparse linear systems involved in $\widehat{Z}_{2r}(T(B^{-1}A); \ell_1)V$. The multifrontal method consists of two parts: the factorization of sparse matrices and the application of the factorization. Once sparse factors have been constructed, evaluating $\widehat{Z}_{2r}(T(B^{-1}A); \ell_1)V$ is efficient; in this sense, the multifrontal method converts the dense matrix $\widehat{Z}_{2r}(T(B^{-1}A); \ell_1)$ into an operator with a fast application. Since the Zolotarev's function well approximates the signum function, the matrix $\widehat{Z}_{2r}(T(B^{-1}A); \ell_1)$ has a condition number close to 1. Therefore, the computation $Z_{2r}(\widehat{Z}_{2r}(T(B^{-1}A); \ell_1); \ell_2)V$ can be

carried out efficiently by the GMRES iterative method. As we shall see later, by the shift-invariant property of Krylov subspace, the computational time can be further reduced in the GMRES.

When we incorporate the above hybrid fast algorithm into the subspace iteration method, the factorization time of the multifrontal method can be treated as precomputation, since all the multi-shift linear systems in every iteration remain unchanged. Since $Z_{2r}(\widehat{Z}_{2r}(T(B^{-1}A);\ell_1);\ell_2)$ is able to approximate the desired spectrum projector of (A,B) accurately, the subspace iteration method usually needs only one or two iterations to identify desired eigenpairs up to an 10^{-10} relative error. Hence, the dominant computing time in the proposed interior eigensolver is the factorization time in the multifrontal method.

1.3 Organization

In what follows, we introduce the subspace iteration, the best rational filter, and the hybrid fast algorithm in Section 2. In Section 3, extensive numerical examples of a wide range of sparse matrices are presented to demonstrate the efficiency of the proposed algorithms. Finally, we conclude this paper with a short discussion in Section 4.

2 Spectrum slicing via rational filters

First, we recall a standard subspace iteration accelerated by a rational filter for interior generalized eigenvalue problems in Section 2.1. Second, we introduce the best rational filter $R_{ab}(x)$ in (7) and show its efficiency of approximating the rectangular function $S_{ab}(x)$ on the interval

$$(-\infty, a_-] \cup [a_+, b_-] \cup [b_+, \infty), \tag{11}$$

where (a_-, a_+) and (b_-, b_+) are eigengaps around a and b, i.e., there is no eigenvalue inside these two intervals. Third, the hybrid fast algorithm for evaluating the matvec $R_{ab}(B^{-1}A)V$ is introduced in Section 2.3.

Notation	Description
N	Size of the matrix
$\mathbb{F}$	Either $\mathbb{R}$ or $\mathbb{C}$
A, B	Sparse Hermitian definite matrix of size $N \times N$
(A, B)	Matrix pencil
(a, b)	Interval of interest on the spectrum of (A, B)
$(a_-, a_+), (b_-, b_+)$	Eigengaps around a and b respectively
n_λ	Number of eigenvalues in the interval
k	Oversampling constant

Table 1: Commonly used notations.

Throughout this paper, we adopt MATLAB notations for submatrices and indices. For example, given a matrix A and two index sets, s_1 and s_2, $A(s_1, s_2)$ represents the submatrix of A with the row indices in s_1 and column indices in s_2. Another important notation from MATLAB is ":". With the same setting, $A(s_1, :)$ represents the submatrix of A with row indices in s_1 and all columns. Besides usual MATLAB notations, we summarize a few notations that would be used in the rest of the paper without further explanation in Table 1.

2.1 Subspace iteration with rational filters

Various subspace iteration methods have been proposed and analyzed in the literature. For the completeness of the presentation, we introduce a standard one in conjunction of a rational filter. Algorithm 1 below is a combination of the CIRR algorithm [26] and the FEAST algorithm [23] for the interior generalized eigenvalue problem for a matrix pencil (A, B) on a spectrum interval $[a, b]$.

Algorithm 1: A standard subspace iteration method

input : Sparse Hermitian matrix pencil (A, B), a spectrum range (a, b), the number of eigenpairs n_λ, and a rational filter $R_{ab}(x)$

output: A diagonal matrix Λ with diagonal entries being the eigenvalues of (A, B) on (a, b), V are the corresponding eigenvectors

1 Generate orthonormal random vectors $Q \in \mathbb{F}^{N\times(n_\lambda+k)}$.
2 **while** *not convergent* **do**
3 $\quad Y = R_{ab}(B^{-1}A)Q$
4 $\quad$ Orthonormalize Y
5 $\quad$ Compute $\tilde{A} = Y^*AY$ and $\tilde{B} = Y^*BY$
6 $\quad$ Solve $\widetilde{A}\widetilde{Q} = \widetilde{\Lambda}\widetilde{B}\widetilde{Q}$ for $\widetilde{\Lambda}$ and $\widetilde{Q}$
7 $\quad$ Update $Q = Y\widetilde{Q}$
8 **end**
9 $\mathcal{I} = \{i \mid a < \widetilde{\Lambda}(i, i) < b\}$
10 $\Lambda = \widetilde{\Lambda}(\mathcal{I}, \mathcal{I})$
11 $V = Q(:, \mathcal{I})$

In the original CIRR algorithm, the rational filter and the Rayleigh-Ritz procedure are only applied once without iterations. This requires a highly accurate rational filter to guarantee the good accuracy of the eigensolver. Repeating the CIRR procedure for a few times can relax the accuracy requirement and allows an approximate rational filter. As for the FEAST eigensolver [23], there is no orthonormalization step in Line 4 in Algorithm 1. This step helps to improve the stability of the FEAST algorithm. Similar convergence analysis for FEAST algorithm [29] can be applied here as well.

The main cost in Algorithm 1 is to compute $Y = R_{ab}(B^{-1}A)Q$, since any other steps scale at most linearly in N or even independent of N. If the rational function $R_{ab}(x)$ is not a good approximation to the rectangular function $S_{ab}(x)$, it might take many iterations for Algorithm 1 to converge. Our goal is to get an accurate rational function approximation $R_{ab}(x)$ so that only a small number of iterations is sufficient to estimate the eigenpairs of (A, B) with machine accuracy. The method to achieve the goal will be discussed in the next two subsections.

2.2 Best rational filter by Zolotarev's functions

Matrix function evaluation is an important topic in numerical linear algebra. Let $f(x)$ be a scalar function and $A \in \mathbb{F}^{N\times N}$ be a Hermitian matrix, where $A = X\Lambda X^*$ with X being the eigenvectors of A and Λ being a diagonal matrix consisting eigenvalues $\{\lambda_j\}_{1\le j\le N}$ of A. The matrix function $f(A)$ is defined as

$$f(A) = Xf(\Lambda)X^* = X \begin{pmatrix} f(\lambda_1) & & & \\ & f(\lambda_2) & & \\ & & \ddots & \\ & & & f(\lambda_N) \end{pmatrix} X^*. \tag{12}$$

If $f(x) = S_{ab}(x)$, then $f(\lambda_j) = 1$ for $\lambda_j \in (a, b)$ and $f(\lambda_j) = 0$ for $\lambda_j \notin (a, b)$. In such a case, $f(A) = Xf(\Lambda)X^* = VV^*$, where V is the orthonormal vectors being the eigenvectors of A with their corresponding eigenvalues on (a, b). Hence, $f(A)$ is a projector onto the eigen-subspace corresponding to the eigenvalues on (a, b). However, without the full diagonalization of the matrix A, such a matrix function $f(x) = S_{ab}(x)$ cannot be evaluated. This is also true for most other functions. Two well-known exceptions are the polynomial and the rational function. We focus on rational functions in this paper.

In what follows, we introduce basic definitions and theorems for rational function approximations. Let $\mathcal{P}_r$ denote the set of all polynomials of degree r. A rational function $R(x)$ is said to be of type (r_1, r_2) if $R(x) = \frac{P(x)}{Q(x)}$ with $P(x) \in \mathcal{P}_{r_1}$ and $Q(x) \in \mathcal{P}_{r_2}$. We denote the set of all rational functions of type (r_1, r_2) as $\mathcal{R}_{r_1,r_2}$. For a given function $f(x)$ and a rational function $R(x)$, the approximation error in a given domain Ω is quantified by the infinity norm

$$\|f - R\|_{L^\infty(\Omega)} = \sup_{x\in\Omega} |f(x) - R(x)| . \tag{13}$$

A common problem in the rational function approximation is the minimax problem that identifies $R(x) \in \mathcal{R}_{r_1,r_2}$ satisfying

$$R = \arg \min_{g\in\mathcal{R}_{r_1,r_2}} \|f - g\|_{L^\infty(\Omega)} . \tag{14}$$

More specifically, the minimax problem of interest in this paper is

$$R = \arg \min_{g\in\mathcal{R}_{2r-1,2r}} \|\mathrm{sign}\,(x) - g(x)\|_{L^\infty([-1,-\ell]\cup[\ell,1])} , \tag{15}$$

where r is a given integer and $\ell \in (0, 1)$ is given parameter. The problem in (15), has a unique solution and the explicit expression of the solution is given by Zolotarev [36]. We denote this best rational approximation to the signum function by $Z_{2r}(x; \ell)$. To be more precise, the following theorem summarizes one of Zolotarev's conclusions which is rephrased by Akhiezer in Chapter 9 in [2], and by Petrushev and Popov in Chapter 4.3 in [22].

Theorem 2.1 (Zolotarev's function)**.** *The best uniform rational approximant of type* $(2r, 2r)$ *for the signum function* $sign\,(x)$ *on the set* $[-1, -\ell] \cup [\ell, 1]$, $0 < \ell < 1$, *is given by*

$$Z_{2r}(x; \ell) := Mx \frac{\prod_{j=1}^{r-1}(x^2 + c_{2j})}{\prod_{j=1}^{r}(x^2 + c_{2j-1})} \in \mathcal{R}_{2r-1,2r}, \tag{16}$$

where $M > 0$ *is a unique constant such that*

$$\min_{x\in[-1,-\ell]} Z_{2r}(x; \ell) + 1 = \min_{x\in[\ell,1]} Z_{2r}(x; \ell) - 1. \tag{17}$$

The coefficients $c_1, c_2, \ldots, c_{2r-1}$ *are given by*

$$c_j = \ell^2 \frac{sn^2\left(\frac{jK'}{2r}; \ell'\right)}{cn^2\left(\frac{jK'}{2r}; \ell'\right)}, \quad j = 1, 2, \ldots, 2r - 1, \tag{18}$$

where $sn(x; \ell')$ *and* $cn(x; \ell')$ *are the Jacobi elliptic functions (see [1, 2]),* $\ell' = \sqrt{1 - \ell^2}$, *and* $K' = \int_0^{\pi/2} \frac{d\theta}{\sqrt{1-(\ell')^2 \sin^2\theta}}$.

By Add. E in [1], the maximum approximation error $\delta(2r,\ell) := \|\mathrm{sign}\,(x) - Z_{2r}(x;\ell)\|_{L^\infty}$ is attained at $2r+1$ points $x_1 := \ell < x_2 < \cdots < x_{2r} < x_{2r+1} := 1$ on the interval $[\ell, 1]$ and also $2r+1$ points $x_{-j} := -x_j$, $j = 1, 2, \ldots, 2r+1$, on the interval $[-1, -\ell]$. The function $\mathrm{sign}\,(x) - Z_{2r}(x;\ell)$ equioscillates between the x_j's; in particular,

$$1 - Z_{2r}(x_j;\ell) = (-1)^{j+1}\delta(2r,\ell), \quad j = 1, 2, \ldots, 2r+1. \tag{19}$$

The approximation error of Zolotarev's functions as an approximant to $\mathrm{sign}\,(x)$ decreases exponentially with degree $2r$ ([22] Section 4.3), i.e.

$$\delta(2r,\ell) \approx C\rho^{-2r} \tag{20}$$

for some positive C and $\rho > 1$ that depends on ℓ. In more particular, Gončar [7] gave the following quantitative estimation on the approximation error, $\delta(2r,\ell)$:

$$\frac{2}{\rho^{2r}+1} \le \delta(2r,\ell) \le \frac{2}{\rho^{2r}-1}, \tag{21}$$

where

$$\rho = \exp\left(\frac{\pi K(\mu')}{4K(\mu)}\right), \tag{22}$$

$\mu = \frac{1-\sqrt{\ell}}{1+\sqrt{\ell}}$, $\mu' = \sqrt{1-\mu^2}$, and $K(\mu) = \int_0^{\pi/2} \frac{\mathrm{d}\theta}{\sqrt{1-(\mu)^2\sin^2\theta}}$ is the complete elliptic integral of the first kind for the modulus μ.

Even though the approximation error $\delta(2r,\ell)$ decreases exponentially in r, the decay rate of $\delta(2r,\ell)$ in r might still be slow if ρ is small. In fact, ρ could be small in many applications when eigengaps are small. As we shall see later, if the eigenvalues cluster together, ℓ should be very small and hence ρ is small by (33). As we have discussed earlier in the introduction of this paper, it is not practical to use a large r due to the computational expense and numerical instability. This motivates the study of the composition of Zolotarev's functions in $\mathcal{R}_{2r-1,2r}$, which constructs a high order Zolotarev's function in $\mathcal{R}_{(2r)^2-1,(2r)^2}$. Such a composition has a much smaller approximation error

$$\delta(4r^2,\ell) \approx C\rho^{-4r^2}. \tag{23}$$

For simplicity, let us use the rescaled Zolotarev's function defined by

$$\widehat{Z}_{2r}(x;\ell) = \frac{Z_{2r}(x;\ell)}{\max_{x\in[\ell,1]} Z_{2r}(x;\ell)}. \tag{24}$$

Note that $\max_{x\in[\ell,1]} \widehat{Z}_{2r}(x;\ell) = 1$, and $\widehat{Z}_{2r}(x;\ell)$ maps the set $[-1,-\ell]\cup[\ell,1]$ onto $[-1, -\widehat{Z}_{2r}(\ell;\ell)]\cup[\widehat{Z}_{2r}(\ell;\ell), 1]$. Hence, if one defines a composition via

$$S(x;\ell_1) = Z_{2r}(\widehat{Z}_{2r}(x;\ell_1);\ell_2), \tag{25}$$

where $\ell_2 = \widehat{Z}_{2r}(\ell_1;\ell_1)$, then $S(x;\ell_1) \in \mathcal{R}_{(2r)^2-1,(2r)^2}$ is the best uniform rational approximant of type $((2r)^2,(2r)^2)$ for the signum function $\mathrm{sign}\,(x)$ on the set $[-1,-\ell_1]\cup[\ell_1,1]$. This optimal approximation is an immediate result of a more general theorem as follows.

Theorem 2.2. *Let $\widehat{Z}_{2r_1}(x;\ell_1) \in \mathcal{R}_{2r_1-1,2r_1}$ be the rescaled Zolotarev's function corresponding to $\ell_1 \in (0,1)$, and $Z_{2r_2}(x;\ell_2) \in \mathcal{R}_{2r_2-1,2r_2}$ be the Zolotarev's function corresponding to $\ell_2 := \widehat{Z}_{2r_1}(\ell_1;\ell_1)$. Then*

$$Z_{2r_2}(\widehat{Z}_{2r_1}(x;\ell_1);\ell_2) = Z_{(2r_1)(2r_2)}(x;\ell_1). \tag{26}$$

The proof of Theorem 2.2 is similar to Theorem 3 in [18]. Hence, we leave it to readers.

Finally, given a desired interval (a, b) and the corresponding eigengaps, (a_-, a_+) and (b_-, b_+), we construct a uniform rational approximant $R_{ab}(x) \in \mathcal{R}_{(2r)^2,(2r)^2}$ via the Möbius transformation $T(x)$ as follows

$$R_{ab}(x) = \frac{S(T(x); \ell_1) + 1}{2} = \frac{Z_{2r}(\widehat{Z}_{2r}(T(x); \ell_1); \ell_2) + 1}{2}, \tag{27}$$

where $\ell_2 = \widehat{Z}_{2r}(1; \ell_1)$ and

$$T(x) = \gamma \frac{x - \alpha}{x - \beta} \tag{28}$$

with $\alpha \in (a_-, a_+)$ and $\beta \in (b_-, b_+)$ such that

$$T(a_-) = -1, \quad T(a_+) = 1, \quad T(b_-) = \ell_1, \quad \text{and } T(b_+) = -\ell_1. \tag{29}$$

We would like to emphasize that the variables α, β, γ, ℓ_1, and ℓ_2 are determined by a_-, a_+, b_-, and b_+ via solving the equations in (29) in the above construction. In practice, a_-, a_+, b_-, and b_+ can be easily calculated from a and b. We fixed the buffer region $(a_-, a_+) \cup (b_-, b_+)$ first according to the eigengaps of target matrices and construct a Möbius transformation adaptive to this region. This adaptive idea is natural but does not seem to have been considered before in the literature. The following theorem shows that $R_{ab}(x)$ in (27) is indeed the best uniform rational approximant of type $((2r)^2, (2r)^2)$.

Theorem 2.3. *The rational function $R_{ab}(x)$ given in* (27) *satisfies the following properties:*

1) $R_{ab}(x)$ is the best uniform rational approximant of type $((2r)^2, (2r)^2)$ of the rectangular function $S_{ab}(x)$ on

$$\Omega = (-\infty, a_-] \cup [a_+, b_-] \cup [b_+, \infty), \tag{30}$$

where (a_-, a_+) and (b_-, b_+) are eigengaps.

2) The error curve $e(x) := S_{ab}(x) - R_{ab}(x)$ equioscillates on Ω with the maximal error

$$\delta_0 := \max_{x \in \Omega} |e(x)| = \min_{g \in \mathcal{R}_{(2r)^2,(2r)^2}} \|S_{ab}(x) - g(x)\|_{L^\infty(\Omega)} \tag{31}$$

and

$$\frac{2}{\rho^{(2r)^2} + 1} \le \delta_0 \le \frac{2}{\rho^{(2r)^2} - 1}, \quad \rho = \rho(\ell_1) > 1, \tag{32}$$

where

$$\rho(\ell_1) = \exp\left(\frac{\pi K(\mu')}{4K(\mu)}\right), \tag{33}$$

$\mu = \frac{1-\sqrt{\ell_1}}{1+\sqrt{\ell_1}}$, $\mu' = \sqrt{1-\mu^2}$, and $K(\mu) = \int_0^{\pi/2} \frac{d\theta}{\sqrt{1-\mu^2 \sin^2 \theta}}$ is the complete elliptic integral of the first kind for the modulus μ.

Proof. Note that inserting a rational transformation of type $(1, 1)$ into a rational function of type $((2r)^2 - 1, (2r)^2)$ results in a rational function of type $((2r)^2, (2r)^2)$. Since $S(x; \ell_1) \in \mathcal{R}_{(2r)^2-1,(2r)^2}$ and Möbius transform $T(x) \in \mathcal{R}_{1,1}$, we know $R_{ab}(x) \in \mathcal{R}_{(2r)^2,(2r)^2}$. In the following proof, we will first show that $R_{ab}(x)$ is the best uniform rational approximant of type $((2r)^2, (2r)^2)$ to the rectangular function on Ω and then derive the error estimator.

Suppose $R_{ab}(x)$ is not the best uniform rational approximant of type $((2r)^2, (2r)^2)$ of the rectangular function $S_{ab}(x)$ on

$$\Omega = (-\infty, a_-] \cup [a_+, b_-] \cup [b_+, \infty), \tag{34}$$

then there exists another rational function $\tilde{R}(x)$ in $\mathcal{R}_{(2r)^2,(2r)^2}$ such that

$$\left\| S_{ab}(x) - \tilde{R}(x) \right\|_{L^\infty(\Omega)} < \| S_{ab}(x) - R_{ab}(x) \|_{L^\infty(\Omega)} .$$

Let $T^{-1}(x)$ denote the inverse transform of the Möbius transformation $T(x)$ in Equation (28), and we have $T^{-1} \in \mathcal{R}_{1,1}$. Note that inserting a rational transformation of type $(1,1)$ into a rational function of type $((2r)^2,(2r)^2)$ results in a rational function of type $((2r)^2,(2r)^2)$. Hence, $2\tilde{R}(T^{-1}(x)) - 1$ is a rational function approximant in $\mathcal{R}_{(2r)^2,(2r)^2}$ of the signum function $\operatorname{sign}(x)$ on the set $[-1,-\ell_1] \cup [\ell_1,1]$. Note that the Möbius transformations $T(x)$ and $T^{-1}(x)$ are bijective maps that do not change the approximation errors, we have

$$\begin{aligned} & \left\| \operatorname{sign}(x) - (2\tilde{R}(T^{-1}(x)) - 1) \right\|_{L^\infty([-1,-\ell_1]\cup[\ell_1,1])} = 2 \left\| S_{ab}(x) - \tilde{R}(x) \right\|_{L^\infty(\Omega)} \\ < \;& 2 \| S_{ab}(x) - R_{ab}(x) \|_{L^\infty(\Omega)} = \| \operatorname{sign}(x) - S(x;\ell_1) \|_{L^\infty([-1,-\ell_1]\cup[\ell_1,1])} . \end{aligned}$$

The inequality

$$\left\| \operatorname{sign}(x) - (2\tilde{R}(T^{-1}(x)) - 1) \right\|_{L^\infty([-1,-\ell_1]\cup[\ell_1,1])} < \| \operatorname{sign}(x) - S(x;\ell_1) \|_{L^\infty([-1,-\ell_1]\cup[\ell_1,1])}$$

conflicts with the fact that $S(x;\ell_1) \in \mathcal{R}_{(2r)^2-1,(2r)^2}$ is the best rational approximant (among all rational functions of type $((2r)^2,(2r)^2)$) of the signum function on $[-1,-\ell_1]\cup[\ell_1,1]$ by Equation (25) and (26). Hence, our previous assumption that $R_{ab}(x)$ is not the best uniform rational approximant of type $((2r)^2,(2r)^2)$ of the rectangular function $S_{ab}(x)$ on Ω is false. This proves the first statement of Theorem 2.3.

The error inequalities in Property 2) follow from Gončar's quantitative estimation on the approximation error of Zolotarev's functions in [7] and the bijective transformation in (28).

□

An immediate result of Theorem 2.3 is

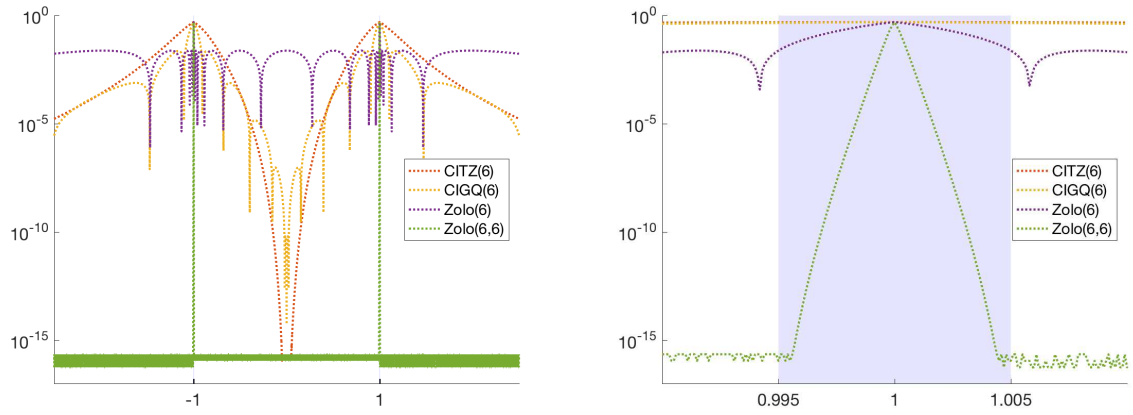
$$\delta_0 = \| S_{ab}(x) - R(x) \|_{L^\infty(\Omega)} = C_{4r^2} \rho^{-4r^2}, \tag{35}$$

with $1 \le \frac{2}{1+\rho^{-(2r)^2}} \le C_{4r^2} \le \frac{2}{1-\rho^{-(2r)^2}}$.

To illustrate this improvement, we compare the performance of the proposed rational filter in (27) with other existing rational filters that are constructed by discretizing the complex value contour integral

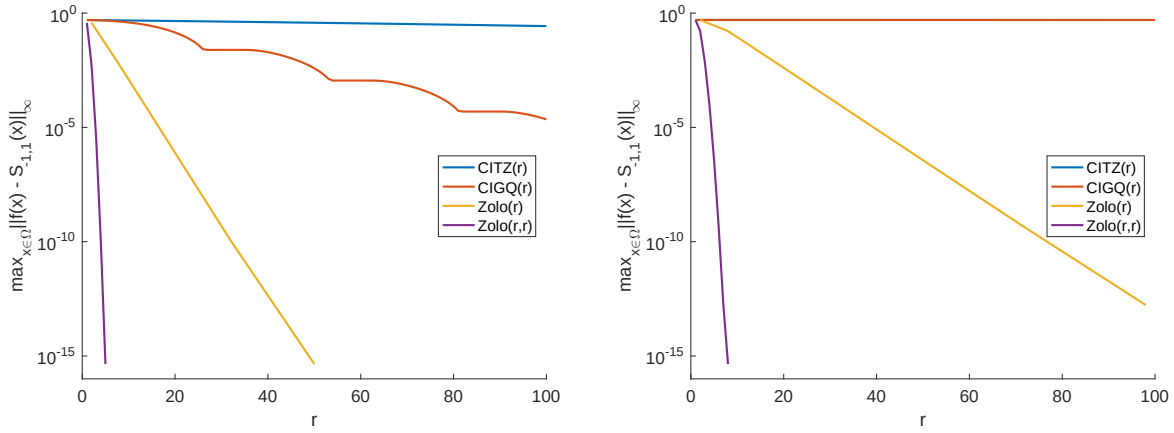
$$\pi(x) = \frac{1}{2\pi\imath} \oint_\Gamma \frac{1}{x-z} \, \mathrm{d}z, \quad x \notin \Gamma \tag{36}$$

with an appropriate quadrature rule (e.g., the Gauss-Legendre quadrature rule [23] and the trapezoidal quadrature rule [29, 35]). Since the dominant cost of applying all these filters is the sparse matrix factorization, we fix the number of matrices to be factorized and compare the approximation error of various filters. The results in Figure 2 verifies the advantage of the proposed rational filter over existing rational filters and shows that 6 matrix factorizations are enough to construct the composition of Zolotarev's rational function approximating a rectangular function within a machine accuracy. Here the eigengaps are 10^{-2}. Figure 3 further explores the decay for the errors in L^∞ norm for different methods. Figure 3a is the decay property for problem with eigengaps 10^{-2} whereas Figure 3b shows the decay property for problem with eigengaps 10^{-6}.



(a) (b)

Figure 2: This figure shows the approximation error of various rational filters as an approximation of the rectangular function supported on $(-1, 1)$. The eigengaps around 1 and -1 is set to be 10^{-2}. These functions include: the trapezoidal filter [29, 35] (denoted as CITZ(r), where r is the number of poles), the Gauss-Legendre filter [23] (denoted as CIGQ(r), where r is the number of poles), the Zolotarev approximation (denoted as Zolo(r), where r is the degree), and the proposed Zolotarev filter via compositions (denoted as Zolo(r, r), where r is the degree). (a) shows the approximation on $[-2.5, 2.5]$ and (b) zooms in on $[0.99, 1.01]$. Light purple areas are the buffer areas in which it is not necessary to condiser the approximation accuracy because of the eigengaps.



(a) (b)

Figure 3: This figure shows the approximation error against degree r for various rational functions as an approximation of the rectangular function supported on $(-1, 1)$. Legends are referred to Figure 2. The eigengaps around -1 and 1 are set to be 10^{-2} in (a) and 10^{-6} in (b). The approximation errors of Zolo(r, r) decay significantly faster than other methods. In (b), the line for CITZ is overwritten by that of CIGQ.

2.3 A hybrid algorithm for applying the best rational filter

In this section, we introduce a hybrid algorithm for applying the best rational filter $R_{ab}(x)$ constructed in Section 2.2, i.e., computing the matvec $R_{ab}(B^{-1}A)V$ when A and B are sparse Hermitian matrices in $\mathbb{F}^{N\times N}$ and V is a vector in $\mathbb{F}^N$. Recall that the rational filter $R_{ab}(x)$ is constructed by

$$R_{ab}(x) = \frac{Z_{2r}(\widehat{Z}_{2r}(T(x);\ell_1);\ell_2)+1}{2}. \tag{37}$$

Hence, it is sufficient to show how to compute $Z_{2r}(\widehat{Z}_{2r}(T(B^{-1}A);\ell_1);\ell_2)V$ efficiently.

For the sake of numerical stability and parallel computing, a rational function is usually evaluated via a partial fraction representation in terms of a sum of fractions involving polynomials of low degree. For the Zolotarev's function $Z_{2r}(x;\ell)$ introduced in (16), we have the following partial fraction representation.

Proposition 2.4. *The function $Z_{2r}(x;\ell)$ as in* (16) *can be reformulated as*

$$Z_{2r}(x;\ell) = Mx\frac{\prod_{j=1}^{r-1}(x^2+c_{2j})}{\prod_{j=1}^{r}(x^2+c_{2j-1})} = Mx\left(\sum_{j=1}^{r}\frac{a_j}{x^2+c_{2j-1}}\right), \tag{38}$$

where

$$a_j = \frac{b_j}{c_{2r-1}-c_{2j-1}} \tag{39}$$

for $j=1,\dots,r-1$, and

$$a_r = 1-\sum_{j=1}^{r-1}\frac{b_j}{c_{2r-1}-c_{2j-1}}. \tag{40}$$

Here

$$b_j = (c_{2j}-c_{2j-1})\prod_{k=1,k\neq j}^{r-1}\frac{c_{2k}-c_{2j-1}}{c_{2k-1}-c_{2j-1}} \tag{41}$$

for $j=1,\dots,r-1$, $\{c_j\}$ and M are given in (18).

Proof. First we prove that we have the following partial fraction representation

$$Mx\prod_{j=1}^{r-1}\frac{x^2+c_{2j}}{x^2+c_{2j-1}} = Mx\left(1+\sum_{j=1}^{r-1}\frac{b_j}{x^2+c_{2j-1}}\right), \tag{42}$$

where

$$b_j = (c_{2j}-c_{2j-1})\prod_{k=1,k\neq j}^{r-1}\frac{c_{2k}-c_{2j-1}}{c_{2k-1}-c_{2j-1}} \tag{43}$$

for $j=1,\dots,r-1$. Since any rational function has a partial fraction form and the coefficients $\{c_{2j-1}\}$ are distinct, Equation (42) holds. One can verify (42) by multiplying x^2+c_{2j-1} to both sides and set $x=\imath\sqrt{c_{2j-1}}$.

By (42), we have

$$Z_{2r}(x;\ell) = Mx\frac{\prod_{j=1}^{r-1}(x^2+c_{2j})}{\prod_{j=1}^{r}(x^2+c_{2j-1})} = Mx\left(1+\sum_{j=1}^{r-1}\frac{b_j}{x^2+c_{2j-1}}\right)\frac{1}{x^2+c_{2r-1}}. \tag{44}$$

Hence, simple partial fraction representations of

$$\frac{b_j}{(x^2+c_{2j-1})(x^2+c_{2r-1})}=\frac{b_j}{c_{2r-1}-c_{2j-1}}\left(\frac{1}{x^2+c_{2j-1}}-\frac{1}{x^2+c_{2r-1}}\right) \tag{45}$$

for $j=1,\ldots,r-1$ complete the proof of the proposition. □

If complex coefficients are allowed, the following corollary can be derived from Proposition 2.4 directly.

Corollary 2.5. *The function $Z_{2r}(x;\ell)$ as in* (16) *can be reformulated as*

$$Z_{2r}(x;\ell)=\frac{M}{2}\sum_{j=1}^{r}\left(\frac{a_j}{x+\imath\sqrt{c_{2j-1}}}+\frac{a_j}{x-\imath\sqrt{c_{2j-1}}}\right), \tag{46}$$

where a_j and c_{2j-1} are as defined in Proposition 2.4.

By Proposition 2.4, we obtain the partial fraction representation of $Z_{2r}(T(x);\ell)$ as follows, where $T(x)$ is a Möbius transformation $T(x)=\gamma\frac{x-\alpha}{x-\beta}$.

Proposition 2.6. *The function $Z_{2r}(T(x);\ell)$ can be reformulated as*

$$Z_{2r}(T(x);\ell)=M\sum_{j=1}^{r}\frac{a_j\gamma}{\gamma^2+c_{2j-1}}+M\sum_{j=1}^{r}\left(\frac{w_j}{x-\sigma_j}+\frac{\bar{w}_j}{x-\bar{\sigma}_j}\right). \tag{47}$$

where

$$\sigma_j=\frac{\gamma\alpha+\imath\sqrt{c_{2j-1}}\beta}{\gamma+\imath\sqrt{c_{2j-1}}},\quad w_j=\frac{a_j(\sigma_j-\beta)}{2(\gamma+\imath\sqrt{c_{2j-1}})}. \tag{48}$$

Proof. We further decompose (38) as complex rational functions,

$$Z_{2r}(x;\ell)=M\sum_{j=1}^{r}\frac{a_j}{2}\left(\frac{1}{x+\imath\sqrt{c_{2j-1}}}+\frac{1}{x-\imath\sqrt{c_{2j-1}}}\right). \tag{49}$$

Substitute the Möbius transformation into (49),

$$\begin{aligned}Z_{2r}(T(x);\ell)=&M\sum_{j=1}^{r}\frac{a_j}{2}\left(\frac{x-\beta}{\gamma(x-\alpha)+\imath\sqrt{c_{2j-1}}(x-\beta)}+\frac{x-\beta}{\gamma(x-\alpha)-\imath\sqrt{c_{2j-1}}(x-\beta)}\right)\\ =&M\sum_{j=1}^{r}\frac{a_j}{2}\left(\frac{\frac{x-\beta}{\gamma+\imath\sqrt{c_{2j-1}}}}{x-\frac{\gamma\alpha+\imath\sqrt{c_{2j-1}}\beta}{\gamma+\imath\sqrt{c_{2j-1}}}}+\frac{\frac{x-\beta}{\gamma-\imath\sqrt{c_{2j-1}}}}{x-\frac{\gamma\alpha-\imath\sqrt{c_{2j-1}}\beta}{\gamma-\imath\sqrt{c_{2j-1}}}}\right).\end{aligned} \tag{50}$$

We denote

$$\sigma_j:=\frac{\gamma\alpha+\imath\sqrt{c_{2j-1}}\beta}{\gamma+\imath\sqrt{c_{2j-1}}}=\frac{(\gamma^2\alpha+c_{2j-1}\beta)+\imath\sqrt{c_{2j-1}}(\beta-\alpha)\gamma}{\gamma^2+c_{2j-1}}. \tag{51}$$

Readers can verify that

$$\bar{\sigma}_j=\frac{\gamma\alpha-\imath\sqrt{c_{2j-1}}\beta}{\gamma-\imath\sqrt{c_{2j-1}}}, \tag{52}$$

where $\bar{\sigma}_j$ is the complex conjugate of σ_j. Equation (50) can be rewritten as,

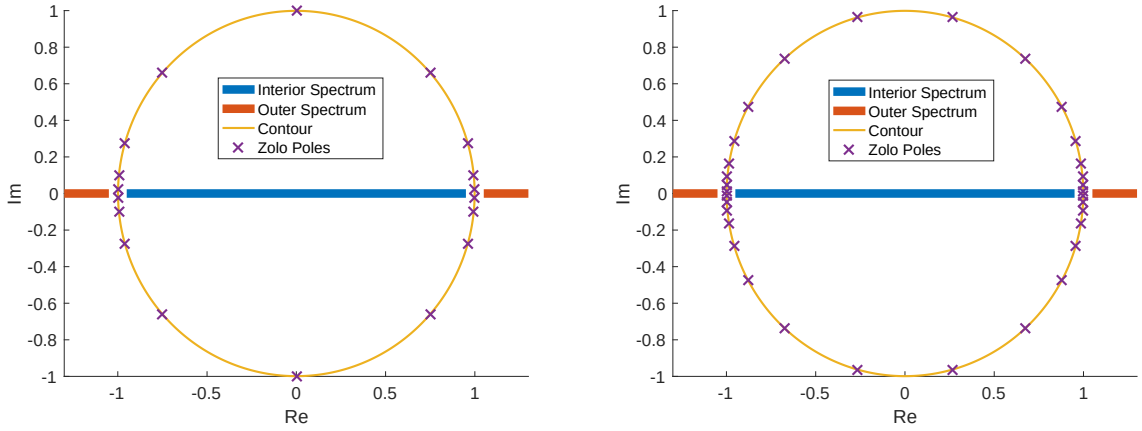
$$Z_{2r}(T(x);\ell) = M\sum_{j=1}^{r}\frac{a_j\gamma}{\gamma^2+c_{2j-1}} + M\sum_{j=1}^{r}\left(\frac{w_j}{x-\sigma_j}+\frac{\bar{w}_j}{x-\bar{\sigma}_j}\right), \tag{53}$$

where

$$w_j = \frac{a_j(\sigma_j-\beta)}{2(\gamma+\imath\sqrt{c_{2j-1}})}. \tag{54}$$

□

Remark 2.7. *In the rest of this paper, we denote the constants associated with $Z_{2r}(x;\ell_2)$ as a_j, c_{2j-1}, σ_j, and w_j for $j=1,\ldots,r$; and the constants associated with $\widehat{Z}_{2r}(x;\ell_1)$ as $\widehat{a}_j, \widehat{c}_{2j-1}, \widehat{\sigma}_j$, and $\widehat{w}_j$ for $j=1,\ldots,r$.*



(a) (b)

Figure 4: The corresponding contour integral discretization of Zolotarev's function composed with Möbius transformation. The eigengaps here are $(-1.1,-0.9)$ and $(0.9,1.1)$ and the contour is a circle centered at origin with radius 0.998749. The discretization points are calculated as Proposition 2.6. (a) $r=9$ provides 18 poles; (b) $r=16$ provides 32 poles.

Proposition 2.6 can be viewed as a discretization of a contour at poles, σ_j and $\widehat{\sigma}_j$ with weights w_j and $\widehat{w}_j$ for $j=1,\ldots,r$. The contour is a circle centered on the real axis cutting through the eigengaps. Figure 4 demonstrate an example with eigengaps $(-1.1,-0.9)$ and $(0.9,1.1)$. The calculated contour is a circle centered at origin with radius 0.998749. Meanwhile, the pole locations and the corresponding weights are provided by Proposition 2.6. Figure 4a adopts $r=9$ which is also the composition of two Zolotarev's functions with degree 3 whereas Figure 4b adopts $r=16$ which is also the composition of two Zolotarev's functions with degree 4.

With these propositions ready, we now introduce the hybrid algorithm for computing the matvec $R_{ab}(B^{-1}A)V = Z_{2r}(\widehat{Z}_{2r}(T(B^{-1}A);\ell_1);\ell_2)V$. This hybrid algorithm consists of two parts of linear system solvers: an inner part and an outer part. The inner part implicitly computes the matvec $\widehat{Z}_{2r}(T(B^{-1}A);\ell_1)V := GV$ via fast direct solvers. Once GV has been implicitly computed, the matrix G can be viewed as an operator with fast application algorithm, where each application costs nearly $O(N)$ operations in many applications. The outer part computes $Z_{2r}(G;\ell_2)V$ using a GMRES method when the fast matvec GV is available. Since the matrix G has singular values

greater than $\ell_2 = \widehat{Z}_{2r}(\ell_1;\ell_1)$, which is a number close to 1, a few steps of iterations in GMRES method are enough to solve the linear systems in the matvec $Z_{2r}(G;\ell_2)V$ accurately. In practice, the iteration number varies from 6 to 25.

In particular, by Proposition 2.6,

$$\begin{aligned} GV =& \widehat{Z}_{2r}(T(B^{-1}A);\ell_1)V \\ =& \widehat{M}\sum_{j=1}^{r}\frac{\widehat{a}_j\gamma}{\gamma^2+\widehat{c}_{2j-1}}V + \widehat{M}\sum_{j=1}^{r}\left(\left(\widehat{w}_j\left(A-\widehat{\sigma}_jB\right)^{-1}BV\right)+\left(\overline{\widehat{w}}_j\left(A-\overline{\widehat{\sigma}}_jB\right)^{-1}BV\right)\right) \\ =& \widehat{M}\sum_{j=1}^{r}\frac{\widehat{a}_j\gamma}{\gamma^2+\widehat{c}_{2j-1}}V + \widehat{M}\sum_{j=1}^{r}\left(\left(\widehat{w}_j\left(A-\widehat{\sigma}_jB\right)^{-1}BV\right)+\left(\overline{\widehat{w}}_j\left(A-\widehat{\sigma}_jB\right)^{-*}BV\right)\right). \end{aligned} \tag{55}$$

The third equality holds since A and B are Hermitian matrices. Hence, evaluating $\widehat{Z}_{2r}(T(B^{-1}A);\ell_1)V$ boils down to solving r linear systems of the form

$$(A-\widehat{\sigma}_jB)x = y \tag{56}$$

for $j = 1,\ldots,r$. This is a set of r sparse linear systems. Since the operator G is involved in an outer function, where it is repeatedly applied, a fast and efficient algorithm for applying G is necessary. This can also be rephrased as "a fast and efficient algorithm for solving (56) is necessary". There are two groups of efficient algorithms for solving (56): direct solvers and iterative solvers with efficient preconditioners.

Fast direct solvers for sparse linear system as $A-\widehat{\sigma}_jB$ usually contains two phases. The first phase (termed as the pre-factorization phase) factorizes the sparse matrix into a product of a sequence of lower and upper triangular sparse matrices. The second phase (termed as the solving phase) solves the sequence of triangular sparse matrices efficiently against vectors. The computational complexities for both the pre-factorization and the solving phase vary from method to method, also heavily rely on the sparsity pattern of the matrix. For simplicity, we denoted the computational complexity for the pre-factorization and the solving phase as F_N and S_N respectively for matrices of size $N\times N$. Usually, F_N is of higher order in N than S_N. Particularly, we adopt the multifrontal method (MF) [6, 14] as the general direct sparse solver for all numerical examples in this paper. For sparse matrices of size $N\times N$ from two-dimensional PDEs, the computational complexities for MF are $F_N = O(N^{3/2})$ and $S_N = O(N\log N)$. While, for three-dimensional problems, MF requires $F_N = O(N^2)$ and $S_N = O(N^{4/3})$ operations.

Iterative solvers with efficient preconditioners is another efficient way to solve sparse linear systems. The construction of preconditioners is the pre-computation phase whereas the iteration together with applying the preconditioner is the solving phase. Similarly to the direct solver, the choices of iterative solvers and preconditioners highly depend on sparse matrices. For elliptic PDEs, GMRES could be used as the iterative solver for $A-\widehat{\sigma}_jB$, and MF with reduced frontals [34, 27, 10, 12] could provide good preconditioners.

Once the fast application of G is available, we apply the classical generalized minimal residual method (GMRES) together with the shift-invariant property of the Krylov subspace (See [24] Section 7.3) to evaluate $R_{ab}(B^{-1}A)V = Z_{2r}(\widehat{Z}_{2r}(T(B^{-1}A);\ell_1);\ell_2)V = Z_{2r}(G;\ell_2)V$. In more particular, by Corollary 2.5, we have

$$Z_{2r}(G;\ell_2)V = M\sum_{j=1}^{r}\frac{a_j}{2}\left(\left(G+\imath\sqrt{c_{2j-1}}I\right)^{-1}V+\left(G-\imath\sqrt{c_{2j-1}}I\right)^{-1}V\right), \tag{57}$$

where I is the identify matrix. Hence, to evaluate $Z_{2r}(G;\ell_2)V$, we need to solve multi-shift linear systems of the form

$$(G \pm \imath\sqrt{c_{2j-1}}I)x = y \tag{58}$$

with $2r$ shifts $\pm\imath\sqrt{c_{2j-1}}$ for $j = 1,\ldots,r$. These systems are solved by the multi-shift GMRES method efficiently. In each iteration, only a single evaluation of GV is needed for all shifts. Meanwhile, since G has a condition number close to 1, only a few iterations are sufficient to solve the multi-shift systems to a high accuracy. Let the number of columns in V be $O(n_\lambda)$ and the number of iterations be m. The complexity for evaluating the rational filter $R_{ab}(B^{-1}A)V$ is $O(mn_\lambda S_N)$.

Algorithm 2: A hybrid algorithm for the rational filter $R_{ab}(B^{-1}A)$

input : A sparse Hermitian definite matrix pencil (A,B), a spectrum range (a,b), vectors V, tolerance ϵ

output: $R_{ab}(B^{-1}A)V$ as defined in (37)

1 Estimate eigengaps (a_-,a_+) and (b_-,b_+) for a and b respectively.

2 Given ϵ, estimate the order of Zolotarev's functions, r

3 Solve (29) for ℓ_1 and Möbius transformation parameter γ,α,β.

4 Calculate function coefficients, $\widehat{M},\widehat{a}_j,\widehat{w}_j,\widehat{\sigma}_j,\widehat{c}_{2j-1}$ and ℓ_2,M,a_j,c_{2j-1} for $j=1,\ldots,r$

5 **for** $j = 1,2,\ldots,r$ **do**

6 Pre-factorize $A-\widehat{\sigma}_jB$ as K_j

7 **end**

8 Generate algorithm for operator

$$GV = \widehat{M}\sum_{j=1}^{r}\frac{\widehat{a}_j\gamma}{\gamma^2+\widehat{c}_{2j-1}}V + \widehat{M}\sum_{j=1}^{r}\left(\widehat{w}_jK_j^{-1}BV + \overline{\widehat{w}}_jK_j^{-*}BV\right)$$

9 Apply the multi-shift GMRES method for solving linear systems $(G\pm\imath\sqrt{c_{2j-1}}I)^{-1}V$ with $j=1,\ldots,r$

10 $R_{ab}(B^{-1}A)V = \frac{M}{2}\sum_{j=1}^{r}\frac{a_j}{2}\left(\left(G+\imath\sqrt{c_{2j-1}}I\right)^{-1}V + \left(G-\imath\sqrt{c_{2j-1}}I\right)^{-1}V\right) + \frac{1}{2}V$

Algorithm 2 summarizes the hybrid algorithm introduced above for applying the rational filter $R_{ab}(B^{-1}A)$ in (37) to given vectors V. By taking Line 1-8 in Algorithm 2 as precomputation and inserting Line 9-10 in Algorithm 2 into Line 3 in Algorithm 1, we obtain a complete algorithm for solving the interior generalized eigenvalue problem on a given interval (a,b). When the matrix pencil (A,B) consists of sparse complex Hermitian definite matrices, the dominant cost of the algorithm is the pre-factorization of r matrices in (56) or Line 6 in Algorithm 2.

3 Numerical examples

In this section, we will illustrate three examples based on different collections of sparse matrices. The proposed method in this paper is denoted as ZoloEig for short. All numerical examples are performed on a desktop with Intel Core i7-3770K 3.5 GHz, 32 GB of memory together with MATLAB R2016a.

The estimation error for the eigenvalues is measured in two ways: the relative error of eigenvalues and the relative error of eigenvalue decomposition. The relative error of eigenvalues is defined as follows,

$$e_\lambda = \frac{\left\|\widetilde{\lambda}-\lambda\right\|_F}{\|\lambda\|_F}, \tag{59}$$

where $\|\cdot\|_F$ is the Frobenius norm, $\widetilde{\lambda}$ is a vector of computed eigenvalues, and λ is a vector of true eigenvalues, which is calculated via the dense eigenvalue decomposition for the entire matrix. In the following sections, e_λ^{zolo} and e_λ^{feast} will be used to denote the relative error of eigenvalues from the proposed ZoloEig algorithm and the FEAST algorithm [23], respectively.

In many problems, the true eigenvalues are not available. A relative error of the eigenvalue decomposition is reported instead, which is defined as follows,

$$e_{\lambda,X} = \frac{\|AX - BX\Lambda\|_F}{\|AX\|_F}, \tag{60}$$

where (A, B) is the matrix pencil of size N by N, $\Lambda \in \mathbb{R}^{k\times k}$ is a diagonal matrix with diagonal entries being the eigenvalues in the given interval, and $X \in \mathbb{F}^{N\times k}$ is the corresponding eigenvectors. This relative error of the eigenvalue decomposition is also used in ZoloEig stopping criteria.

3.1 Spectrum of Hamiltonian Operators

The first example is a Hamiltonian operator,

$$H = -\frac{1}{2}\Delta + V, \tag{61}$$

with a Dirichlet boundary condition, where V is the potential field containing Gaussian wells. The main goal of this example is to illustrate the efficiency and complexity of the proposed new algorithm. The operator is discretized with a local scheme that results in a sparse matrix. The multifrontal method is naturally designed for inverting such sparse matrices. Therefore, the claimed complexity of the ZoloEig algorithm can be rigorously verified for the operator in (61). Detailed settings and numerical results are given for the 2D problem only. The operator (61) is discretized by a 5-point finite different scheme in $[0, 1)^2$. 4×4 Gaussian wells with depth 16 and radius 0.2 are regularly placed as the potential field. Figure 5 shows the potential field V in 2D and the spectrum of the matrix with size $N = 65,536$.

In this numerical example, we compare the performance of the ZoloEig algorithm with the MATLAB built-in function Eigs. The Eigs function with a flavor 'sm' in MATLAB applies the shift-invert implicitly restarted Arnoldi method to compute the smallest eigenvalues (in terms of magnitudes). When the input matrix is a sparse matrix, Eigs uses a fast direct solver to apply the shifted inverse matrix in the Arnoldi method. Hence, the running time of Eigs depends both on the factorization time in the fast direct solver and the iteration time in the Arnoldi method. In this shift-invert algorithm, it is sufficient to factorize one matrix, while in the ZoloEig algorithm there are r sparse matrix factorization. Hence, we expect that the factorization time of Eigs is smaller than that of the ZoloEig, but the iteration time of Eigs should be larger than that of the ZoloEig, especially in the case of computing many selected eigenpairs.

Figure 6 shows the running time and the relative error of eigenvalues, e_λ. The 2D problem size varies from 32^2 to 512^2 and the corresponding matrix size varies from $1,024$ to $262,144$. The order r is 4 in the ZoloEig and the subspace iteration is turned off, i.e., we only apply the rational filter once in Algorithm 1. For each matrix, we provide the smallest (in terms of magnitudes) 96 eigenvalues and the ZoloEig is executed 10 times with different initial random vectors. In Figure 6a and Figure 6b, these results are presented in a bar plot manner: the vertical bars indicate the largest and the smallest values, whereas the trend line goes through the mean values. As we can read from Figure 6a, the running time for ZoloEig has a scaling close to linear while the MATLAB default function Eigs scales quadratically. When the problem size is greater than 10^4, ZoloEig is faster than MATLAB Eigs. The ZoloEig algorithm is supposed to scale as $F_N = N^{3/2}$ for 2D

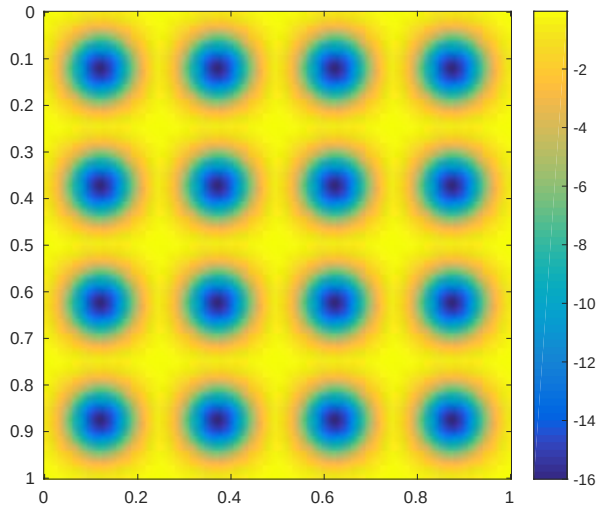


Figure 5: 2D potential field with 4×4 Gaussian wells, each of which is with depth 16 and radius 0.2.

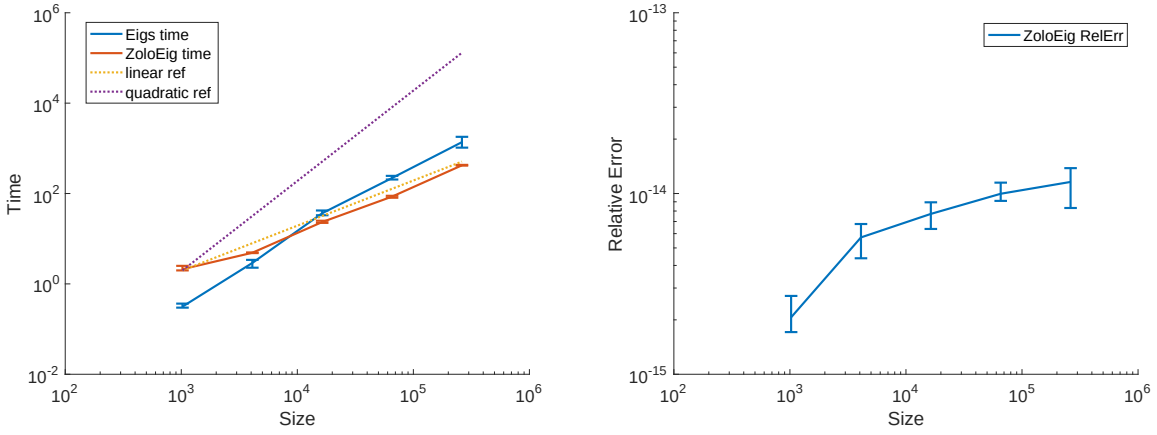
problems. However, in this example, the discretization scheme is very localized and the running time for 4 multifrontal factorizations is relatively small comparing to that of the application to many vectors. Therefore, the linear application time dominate the running time when the matrix size is small. Figure 6b shows the relative error of the eigenvalues, which increases mildly as the problem size increases. In this example, machine accuracy is achieved without subspace iteration.

3.2 Florida Sparse Matrix Collection

In the second example, the proposed algorithm is applied to general sparse Hermitian matrices from the Florida sparse matrix collection. In order to show the broad applicability of the algorithm, all Hermitian matrices with size between 200 and $6,000$ in the collection are tested. The full list of these matrices can be found in the test file "test_eigs_Florida.m" in the MATLAB toolbox. For each of these matrices, we randomly choose an interval (a, b) containing 96 eigenvalues.

In these examples, we compare the performance of the ZoloEig algorithm with the FEAST algorithm based on the contour integral method with trapezoidal rule. The subspace refinement is turned off again, aiming at testing the approximation accuracies of the Zolotarev's rational function and the discretized contour integral. The order r in the Zolotarev's rational function is 4 and the contour integral method has 16 poles. Hence, both the ZoloEig and FEAST algorithms use the same order of rational functions in the approximation.

Figure 7 visualizes the results of both the ZoloEig and the FEAST algorithms. Figure 7a includes the running time of the MATLAB default dense eigensolver Eig, FEAST and ZoloEig. The running time of Eig aligns with the cubic scaling reference, whereas the running times of both FEAST and ZoloEig align with the linear scaling reference. As explained in previous examples, for a problem of small size, the linear part in both FEAST and ZoloEig dominates the running time. The outliers of each line in these figures are caused by different sparsity densities and patterns of sparse matrices. According to Figure 7a, the running time of FEAST is constantly larger than ZoloEig. In Figure 7b, the relative error of FEAST is larger than ZoloEig for most matrices. Based



(a) (b)

Figure 6: The running time and the relative error for 2D Hamiltonian operator with Gaussian wells solved via ZoloEig. The relative error here is the relative error of eigenvalues defined in (59).

on the right part of Figure 7b, FEAST fails for some sparse matrices, where the relative error is close to 1. Meanwhile, the relative error of ZoloEig is smaller than 1e-4 in all cases and the overall accuracy is about 1e-10. This observation supports that the composition of Zolotarev's rational functions is a better way to approximate rectangular functions.

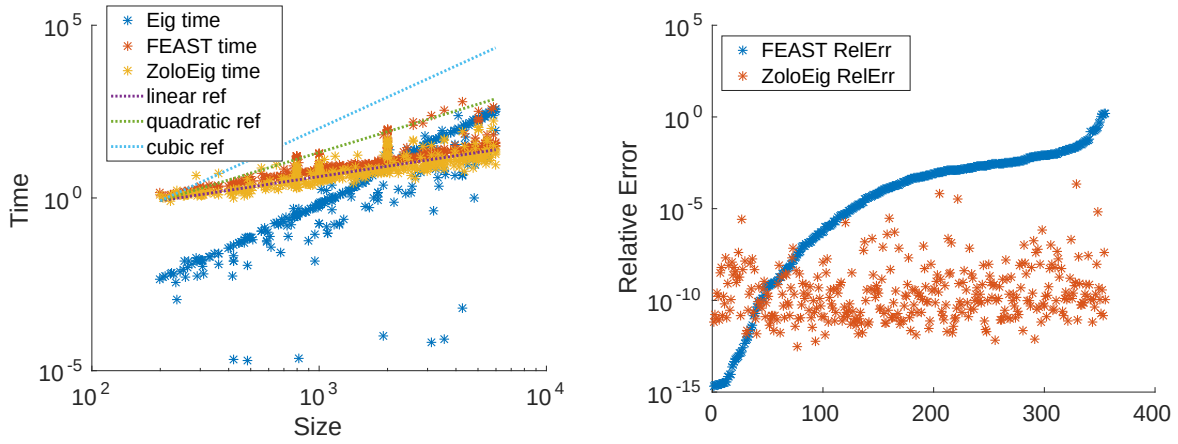
3.3 Hamiltonian of Silicon Bulk

The last example is a sparse Hermitian definite matrix pencil, (A, B), generated by SIESTA (a quantum chemistry software). For a silicon bulk in 3D with y^3 supercell of cubic Si, a DZP basis set with radius 4 Å is adopted to discretize the system, where $y = 2, 3, 4, 5$. In the spectrum slicing problem, the interval is chosen to contain the smallest 93 eigenvalues. The ZoloEig algorithm with $r = 3$ and contour integral method with 16 poles are compared side-by-side. In both algorithms, the subspace refinement is turned on. The iteration accuracy is $1e-8$.

			ZoloEig				FEAST			
y	N	δ_λ	r	$e_{\lambda,X}$	N_{iter}	Time(sec)	N_{pole}	$e_{\lambda,X}$	N_{iter}	Time(sec)
2	832	6.20e-02	(3,3)	5.08e-10	2	5.04e+00	16	3.96e-09	7	3.26e+01
3	2808	6.74e-02	(3,3)	1.60e-09	1	1.98e+01	16	3.22e-09	6	1.87e+02
4	6656	3.98e-03	(3,3)	3.93e-09	1	1.20e+02	16	3.08e-03	14	1.37e+03
5	13000	7.32e-03	(3,3)	1.52e-10	2	3.09e+02	16	6.77e-09	45	9.33e+03

Table 2: Numerical results for generalized eigenvalue problem. y is the number of unit cell on each dimension, N is the size of the sparse matrix, δ_λ is the eigengap between the 93th and 94th eigenvalues, r is the order used in ZoloEig, $e_{\lambda,X}$ is the relative error of eigenvalue decomposition defined in (60), N_{iter} is the number of iterations in the subspace refinement and Time is evaluated in second.

Table 2 includes the detail information of the numerical results. Given the similar eigengaps, ZoloEig outperforms FEAST. In ZoloEig, the tolerance 1e-8 is achieved within two iterations of subspace refinement, while FEAST requires much more iterations. And the running time of ZoloEig



(a) (b)

Figure 7: Running time and relative error for matrices in Florida matrix collection solved via ZoloEig and FEAST. The relative error here is the relative error of eigenvalues defined in (59).

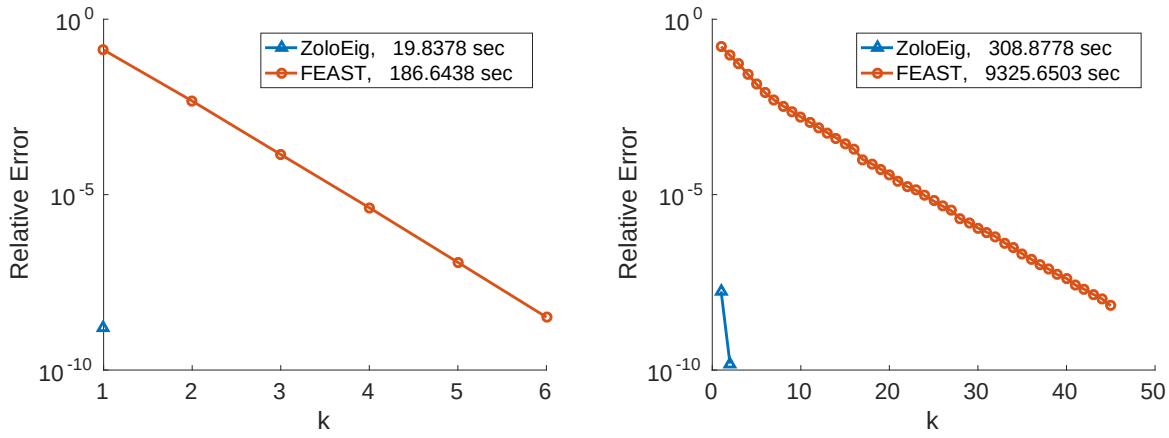
is also on average 10 times faster than FEAST. For the case $y = 4$, FEAST converges slowly and we stop it at the 14th iteration. Figure 8a and Figure 8b show the iteration accuracy for $y = 3$ and $y = 5$, respectively. According to the figure, both algorithms decay exponentially with respect to the number of iterations. However, ZoloEig is much faster than FEAST since the approximation for each iteration is more accurate even if the orders of the rational functions used in the function composition here is $r = 3$ (i.e., the total order of the composed rational function is 9) for ZoloEig, and the order of the rational function in FEAST is 16. This example demonstrates the advantage of the optimal rational function approximation via function compositions again.

4 Conclusion

This paper proposed an efficient method for computing selected eigenpairs of a sparse Hermitian definite matrix pencil (A, B) in the generalized eigenvalue problem. First, based on the best rational function approximations of signum functions by Zolotarev, the best high-order rational filter in a form of function compositions is proposed. Second, taking advantage of the shift-invariant property of Krylov subspaces in iterative methods and the matrix sparsity in sparse direct solvers, a hybrid fast algorithm is proposed to apply the best rational filter in the form of function compositions. Assuming that the sparse Hermitian matrices A and B are of size $N \times N$ and contains $O(N)$ nonzero entries, the computational cost for computing $O(1)$ eigenpairs is $O(F_N)$, where F_N is the operation count for solving the shifted linear system $(A - \sigma B)x = b$ using sparse direct solvers.

It is worth pointing out that the proposed rational filter can also be applied efficiently if an efficient dense direct solver or an effective iterative solver for solving the multi-shift linear systems in (56) is available. The proposed rational function approximation can also be applied as a pre-conditioner for indefinite sparse linear system solvers [33] and the orbital minimization method in electronic structure calculation [15]. These will be left as future works.

Acknowledgments. Y. Li was partially supported by the National Science Foundation under award DMS-1328230 and the U.S. Department of Energy's Advanced Scientific Computing Research program under award DE-FC02-13ER26134/DE-SC0009409. H. Y. is partially supported by the National Science Foundation under grants ACI-1450280 and thank the support of the AMS-Simons



(a) (b)

Figure 8: Subspace refinement iteration accuracy for FEAST and ZoloEig. The iterations for $y = 3$ and $y = 5$ are reported here.

travel award. The authors would like to thank Lexing Ying and Jianfeng Lu for fruitful discussions, and thank Fabiano Corsetti for seting up Silicon Bulk examples.

References

[1] N. I. Akhiezer. *Theory of approximation.* F. Unger Pub. Co., New York, 1956.

[2] N. I. Akhiezer. *Elements of the theory of elliptic functions.* American Mathematical Soc., 1990.

[3] H. M. Aktulga, L. Lin, C. Haine, E. G. Ng, and C. Yang. Parallel eigenvalue calculation based on multiple shiftinvert Lanczos and contour integral based spectral projection method. *Parallel Comput.*, 40(7):195–212, 2014.

[4] D. Braess. On rational approximation of the exponential and the square root function. In *Ration. Approx. Interpolat.*, pages 89–99. Springer Berlin Heidelberg, 1984.

[5] E. Di Napoli, E. Polizzi, and Y. Saad. Efficient estimation of eigenvalue counts in an interval. *Numer. Linear Algebr. with Appl.*, 23(4):674–692, aug 2016.

[6] I. S. Duff and J. K. Reid. The multifrontal solution of indefinite sparse symmetric linear equations. *ACM Trans. Math. Softw.*, 9(3):302–325, sep 1983.

[7] A. A. Gončar. Zolotarev Problems Connected with Rational Functions. *Math. USSR-Sbornik*, 7(4):623–635, apr 1969.

[8] S. Güttel, E. Polizzi, P. T. P. Tang, and G. Viaud. Zolotarev Quadrature Rules and Load Balancing for the FEAST Eigensolver. *SIAM J. Sci. Comput.*, 37(4):A2100–A2122, jan 2015.

[9] N. Hale, N. J. Higham, and L. N. Trefethen. Computing $A^\alpha, \log(A)$, and Related Matrix Functions by Contour Integrals. *SIAM J. Numer. Anal.*, 46(5):2505–2523, jan 2008.

[10] K. L. Ho and L. Ying. Hierarchical interpolative factorization for elliptic operators: differential equations. *Commun. Pure Appl. Math.*, 2015.

[11] R. Li, Y. Xi, E. Vecharynski, C. Yang, and Y. Saad. A Thick-Restart Lanczos Algorithm with Polynomial Filtering for Hermitian Eigenvalue Problems. *SIAM J. Sci. Comput.*, 38(4):A2512–A2534, jan 2016.

[12] Y. Li and L. Ying. Distributed-memory Hierarchical Interpolative Factorization. *Preprint*, 2016.

[13] L. Lin, J. Lu, L. Ying, and W. E. Pole-based approximation of the Fermi-Dirac function. *Chinese Ann. Math. Ser. B*, 30:729–742, nov 2009.

[14] J. W. H. Liu. The multifrontal method for sparse matrix solution: theory and practice. *SIAM Rev.*, 34(1):82–109, mar 1992.

[15] J. Lu and H. Yang. Preconditioning orbital minimization method for planewave discretization. *SIAM Multiscale Modeling and Simulation*, to appear.

[16] D. A. Mazziotti. Towards idempotent reduced density matrices via particle-hole duality: McWeeny's purification and beyond. *Phys. Rev. E*, 68(6):066701, dec 2003.

[17] Y. Nakatsukasa, Z. Bai, and F. Gygi. Optimizing Halley's Iteration for Computing the Matrix Polar Decomposition. *SIAM J. Matrix Anal. Appl.*, 31(5):2700–2720, jan 2010.

[18] Y. Nakatsukasa and R. W. Freund. Computing Fundamental Matrix Decompositions Accurately via the Matrix Sign Function in Two Iterations: The Power of Zolotarev's Functions. *SIAM Rev.*, 58(3):461–493, jan 2016.

[19] A. M. N. Niklasson. Expansion algorithm for the density matrix. *Phys. Rev. B*, 66(15):155115, oct 2002.

[20] I. Ninomiya. Best rational starting approximations and improved Newton iteration for the square root. *Math. Comput.*, 24:391–404, 1970.

[21] A. H. R. Palser and D. E. Manolopoulos. Canonical purification of the density matrix in electronic-structure theory. *Phys. Rev. B*, 58(19):12704–12711, nov 1998.

[22] P. P. Petrushev and V. A. Popov. *Rational approximation of real functions*. Cambridge University Press, Cambridge, 1987.

[23] E. Polizzi. Density-matrix-based algorithm for solving eigenvalue problems. *Phys. Rev. B*, 79(11):115112, mar 2009.

[24] Y. Saad. *Iterative methods for sparse linear systems*, volume 8 of *Stud. Comput. Math.* Society for Industrial and Applied Mathematics, second edition, 2003.

[25] T. Sakurai and H. Sugiura. A projection method for generalized eigenvalue problems using numerical integration. *J. Comput. Appl. Math.*, 159(1):119–128, 2003.

[26] T. Sakurai and H. Tadano. CIRR: a Rayleigh-Ritz type method with contour integral for generalized eigenvalue problems. *Hokkaido Math. J.*, 36(4):745–757, nov 2007.

[27] P. G. Schmitz and L. Ying. A fast nested dissection solver for Cartesian 3D elliptic problems using hierarchical matrices. *J. Comput. Phys.*, 258:227–245, 2014.

[28] G. Schofield, J. R. Chelikowsky, and Y. Saad. A spectrum slicing method for the Kohn-Sham problem. *Comput. Phys. Commun.*, 183(3):497–505, 2012.

[29] P. T. P. Tang, J. Kestyn, and E. Polizzi. A new highly parallel non-Hermitian eigensolver. In *Proc. High Perform. Comput. Symp.*, pages 1–9. Society for Computer Simulation International, 2014.

[30] M. Van Barel. Designing rational filter functions for solving eigenvalue problems by contour integration. *Linear Algebra Appl.*, 502:346–365, 2016.

[31] M. Van Barel and P. Kravanja. Nonlinear eigenvalue problems and contour integrals. *J. Comput. Appl. Math.*, 292:526–540, 2016.

[32] Y. Xi and Y. Saad. Computing Partial Spectra with Least-Squares Rational Filters. *SIAM J. Sci. Comput.*, 38(5):A3020–A3045, jan 2016.

[33] Y. Xi and Y. Saad. A rational function preconditioner for indefinite sparse linear systems. *SIAM Journal on Scientific Computing*, to appear.

[34] J. Xia. Efficient structured multifrontal factorization for general large sparse matrices. *SIAM J. Sci. Comput.*, 35(2):A832–A860, 2013.

[35] X. Ye, J. Xia, R. H. Chan, S. Cauley, and V. Balakrishnan. A fast contour-integral eigensolver for non-Hermitian matrices. Technical report, 2016.

[36] E. Zolotarev. Application of elliptic functions to questions of functions deviating least and most from zero. *Zap. Imp. Akad. Nauk. St. Petersbg.*, 30(5):1–59, 1877.