Zolotarev quadrature rules and load balancing for the FEAST eigensolver

Güttel, Stefan and Polizzi, Eric and Tang, Peter and Viaud, Gautier

2014

MIMS EPrint: **2014.39**
ZOLOTAREV QUADRATURE RULES AND LOAD BALANCING
FOR THE FEAST EIGENSOLVER

STEFAN GÜTTEL∗, ERIC POLIZZI†, PING TAK PETER TANG‡, AND GAUTIER VIALAUD§

Abstract. The FEAST method for solving large sparse eigenproblems is equivalent to subspace
iteration with an approximate spectral projector and implicit orthogonalization. This relation allows
to characterize the convergence of this method in terms of the error of a certain rational approx-
imation to an indicator function. We propose improved rational approximants leading to FEAST
variants with faster convergence, in particular, when using rational approximants based on the work
of Zolotarev. Numerical experiments demonstrate the possible computational savings especially for
pencils whose eigenvalues are not well separated and when the dimension of the search space is only
slightly larger than the number of wanted eigenvalues. The new approach improves both convergence
robustness and load balancing when FEAST runs on multiple search intervals in parallel.

Key words. generalized eigenproblem, FEAST, quadrature, Zolotarev, filter design, load bal-
ancing

AMS subject classifications. 65F15, 41A20, 65Y05

1. Introduction. The FEAST method [29] is an algorithm for computing a few
eigenpairs (λ, x) of a large sparse generalized eigenproblem

\begin{equation}
Ax = \lambda Bx,
\end{equation}

where \( A \in \mathbb{C}^{N \times N} \) is Hermitian and \( B \in \mathbb{C}^{N \times N} \) is Hermitian positive definite. This
method belongs to the class of contour-based eigensolvers which have attracted much
attention over the past decade. Contour-based methods utilize integrals of the form

\begin{equation}
C_j := \frac{1}{2\pi i} \int_{\Gamma} \gamma^j (\gamma B - A)^{-1} B \, d\gamma = \frac{1}{2\pi i} \int_{\Gamma} \gamma^j (\gamma I - M)^{-1} d\gamma,
\end{equation}

where \( M = B^{-1} A \) and \( \Gamma \) is a contour in the complex plane enclosing the wanted
eigenvalues of \((A, B)\). Typically a quadrature rule is then applied to evaluate this contour integral numerically.

Probably the first practical method which combined contour integrals and quadra-
ture was presented by Delves and Lyness [9], although this was for the (related) pur-
pose of finding roots of scalar analytic functions (see also [3] for an overview of various
methods for this purpose). The method presented by Sakurai and Sugiura in [35] (see
also [19, 34]) makes use of the moments \( \mu_j = u^* C_j v \) for solving (1.1). This is done
by constructing a matrix pencil of small size whose eigenvalues correspond to the tar-
geted ones of the original system. The procedure terminates after the reduced system
is constructed and its eigenvalues are obtained. In this sense the method in [35], some-
times referred to as SS method, is non-iterative in nature (although the possibility of incorporating a FEAST-like outer iteration is mentioned in [33]). The SS method

∗School of Mathematics, The University of Manchester, Alan Turing Building, Oxford Road, M13
9PL Manchester, United Kingdom, stefan.guettel@manchester.ac.uk
†Department of Electrical and Computer Engineering, University of Massachusetts, Amherst,
USA, polizzi@ecs.umass.edu. This author was partially supported by the National Science Foun-
dation under Grant #ECCS-0846457, and also acknowledges travel support from the EPSRC Network
Grant EP/103112X/1.
‡Intel Corporation, USA, peter.tang@intel.com
§Ecole Centrale Paris, France, gautier.viaud@ecp.fr
based on explicit moments may become numerically unstable and the so-called CIRR method [36] tries to address this problem by using explicit Rayleigh–Ritz projections for Hermitian eigenproblems. A block-version of CIRR applicable to non-Hermitian eigenproblems was presented in [18].

Expressed in terms of moments, FEAST uses only the zeroth moment matrix $C_0$, which corresponds to the spectral projector onto the invariant subspace associated with the eigenvalues enclosed by $\Gamma$. Since this projector can be computed only approximately, FEAST must be an iterative algorithm: it applies an approximate spectral projector repeatedly, progressively steering the search space into the direction of an invariant subspace containing the wanted eigenvectors. The original paper [29] demonstrated the effectiveness of the approach without analysis of convergence, which was then completed only very recently in [37].

Consistent with [37], we use the fact that the FEAST method is equivalent to subspace iteration with implicit orthogonalization applied with a rational matrix function $r_m(M)$. In the original FEAST derivation [29], the rational function $r_m(z)$ was obtained via quadrature approximation of an indicator function $f(z)$ represented as

$$f(z) = \int_{\Gamma} \frac{d\gamma}{\gamma - z},$$

where $\Gamma$ is a contour enclosing the wanted eigenvalues of $M$. We will show that the convergence of FEAST is governed by the separation of the wanted and unwanted eigenvalues of $r_m(M)$, and that this separation is determined by the accuracy of the quadrature approximation $r_m$ for $f$. We then use this argument to motivate our new choice of $r_m$, which is not based on contour integration but on a rational approximant constructed by Zolotarev.

Zolotarev's rational functions are ubiquitous in the design of electronic filters (see, e.g., [6, 40]) and in this context often referred to as elliptic filters or Cauer filters [7]. Examples from numerical analysis where these functions have proven useful are the choice of optimal parameters in the ADI method [22, 42], for the solution of Lyapunov matrix equations [32, 5], the construction of optimal finite-difference grids [20], in parameter selection problems with rational Krylov methods for matrix functions [16], or for the optimization of time steps in the Crank–Nicolson method [25], see also [38]. The use of elliptic filters in contour integration eigensolvers has been considered in recent papers by Murikami [26, 27]. The use of Zolotarev rational functions (or equivalently, elliptic filters) in the context of FEAST is also very natural, and initial tests prior to writing this paper were performed in the master thesis [41].

The outline of this paper is as follows. In Section 2 we briefly review the FEAST method and its connection with subspace iteration. In Section 3 we compare two different quadrature approaches that are commonly used, namely the approach based on mapped Gauss quadrature as proposed by Polizzi [29], and another one based on the trapezoid rule which is close in spirit to that of Sakurai and coauthors (see, e.g., [35, 18]). We also derive a relation between the rational functions obtained from the trapezoid rule on ellipsoidal contours and so-called type-1 Chebyshev filters. While the trapezoid rule seems most natural, Gauss quadrature turns out to be advantageous if the wanted and unwanted eigenvalues of $M$ are not well separated. In Section 4 we derive an improved quadrature rule based on the optimal Zolotarev approximation to the sign function, and compare it in Section 5 with the previous quadrature rules. In Section 6 we discuss the implications of the Zolotarev approach on the load balancing problem which arises when FEAST runs on multiple search intervals.
The induced $B$-norm is defined as $\|x\|_B := \langle x, x \rangle^{1/2}$.

Here is a step-by-step listing of the FEAST method:

1. Choose $n < N$ random column vectors to form $Y_0 := [y_1, \ldots, y_n] \in \mathbb{C}^{N \times n}$.
2. Set $k := 1$.
3. Compute $Z_k := r_m(M)Y_{k-1} \in \mathbb{C}^{N \times n}$.
4. Compute $\hat{A}_k := Z_k^*AZ_k$ and $\hat{B}_k := Z_k^*BZ_k$.
5. Compute a $\hat{B}_k$-orthonormal matrix $W_k \in \mathbb{C}^{n \times n}$ and the diagonal matrix $D_k = \text{diag}(\vartheta_1, \ldots, \vartheta_n)$ such that $\hat{A}_kW_k = \hat{B}_kW_kD_k$.
6. Set $Y_k := Z_kW_k$.
7. If $Y_k$ has not converged, set $k := k + 1$ and goto Step 3.

For the rational matrix function $r_m(M)$ in Step 3 to be well-defined we assume here and in the following that none of the poles of $r_m$ coincides with an eigenvalue of $M$. When $r_m$ has a partial fraction expansion

$$r_m(z) = \sum_{j=1}^{2m} \frac{w_j}{z_j - z},$$

then Step 3 amounts to the solution of $2m$ decoupled linear systems which can be solved in parallel (with an appropriate choice of $r_m$ only $m$ linear systems need to be solved in some cases, see Section 3):

$$Z_k = r_m(M)Y_{k-1} = \sum_{j=1}^{2m} w_j (z_j B - A)^{-1} (BY_{k-1}).$$

In the original formulation of FEAST in [29] the $B$-factor in $(BY_{k-1})$ is not applied in Step 3 but at the end of each loop. This makes a difference only in the first iteration.

Note that the columns of $Y_k$ for $k \geq 1$ are $B$-orthogonal because the eigenvector matrix $W_k$ of the reduced pencil $(\hat{A}_k, \hat{B}_k)$ computed in Step 5 is $\hat{B}_k$-orthonormal and

$$Y_k^*BY_k = W_k^*Z_k^*BZ_kW_k = W_k^*\hat{B}_kW_k = I_n.$$

The orthogonalization procedure is therefore implicitly built into the Rayleigh–Ritz extraction procedure. FEAST can hence be viewed and analyzed as a subspace iteration with implicit orthogonalization run with the matrix $r_m(M)$; see, e.g., [31, §5.2].

The following results are adopted from [41] and [37]. We include them for completeness and to motivate our derivations in the following sections. Let the eigenpairs $(\lambda_j, x_j)$ of $M$ be ordered such that

$$(2.1) \quad |r_m(\lambda_1)| \geq |r_m(\lambda_2)| \geq \cdots \geq |r_m(\lambda_N)|.$$
We introduce the following notations. For any integer \( n, 1 \leq n < N \):
\[
X_n = [x_1, x_2, \ldots, x_n] \quad \in \mathbb{C}^{N \times n},
\]
\[
X_n' = [x_{n+1}, x_{n+2}, \ldots, x_N] \quad \in \mathbb{C}^{N \times (N-n)},
\]
\[
\Lambda_n = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \quad \in \mathbb{R}^{n \times n},
\]
\[
\Lambda_n' = \text{diag}(\lambda_{n+1}, \lambda_{n+2}, \ldots, \lambda_N) \quad \in \mathbb{R}^{(N-n) \times (N-n)}.
\]

With these notations, the matrix \( X_nX_n^*B \) corresponds to the \( B \)-orthogonal projector onto \( \text{span}(\{x_1, x_2, \ldots, x_n\}) \), and likewise \( X_n'X_n'^*B \) is the \( B \)-orthogonal projector onto \( \text{span}(\{x_{n+1}, x_{n+2}, \ldots, x_N\}) \), for any \( 1 \leq n < N \). In particular, since \( X^*BX = I \) implies \( X^{-1} = X^*B \), the eigendecomposition of \( M \) and \( r_m(M) \) can be written as
\[
M = X_n\Lambda_nX_n^*B + X_n'\Lambda_n'X_n'^*B,
\]
and
\[
r_m(M) = X_n r_m(\Lambda_n) X_n^*B + X_n' r_m(\Lambda_n') X_n'^*B,
\]
for any \( 1 \leq n < N \). Note also that \( X_nX_n^*B + X_n'X_n'^*B = I \). The following lemma provides a characterization of \( \text{span}(Z_k) \).

**Lemma 2.1.** Consider the FEAST method as described in Steps 1–7 previously. Suppose \( |r_m(\lambda_n)| > 0 \), and that \( Y_0 \) in Step 1 is such that the \( n \times n \) matrix \( X_n^*BY_0 \) is invertible. Then the matrices \( Z_k \) of Step 3 always maintain full column rank \( n \) and
\[
\text{span}(Z_k) = \text{span}(r_m^k(M)Y_0)
\]
for all iterations \( k \geq 1 \).

**Proof.** We will first use an induction argument to show that the matrices \( Z_k \) have full column rank and that the matrices \( W_k \) are invertible. Suppose \( X_n^*BY_{k-1} \) is invertible for some \( k \geq 1 \). Then the \( n \times n \) matrix
\[
X_n^*BZ_k = X_n^*B[X_n r_m(\Lambda_n) X_n^*B + X_n' r_m(\Lambda_n') X_n'^*B] Y_{k-1}
= r_m(\Lambda_n) X_n^*BY_{k-1}
\]
is invertible because \( r_m(\Lambda_n) = \text{diag}(r_m(\lambda_1), \ldots, r_m(\lambda_n)) \) is invertible by the assumption \( |r_m(\lambda_n)| > 0 \). In particular \( Z_k \) has full column rank. This means that \( B_k = Z_k^*BZ_k \) is positive definite, resulting in an invertible matrix \( W_k \). Hence \( X_n^*BY_{k} = (X_n^*BZ_k)W_k \) is also invertible. By assumption, \( X_n^*BY_0 \) is invertible and hence by induction \( Z_k \) has full column rank and \( W_k \) is invertible for \( k \geq 1 \).

Finally, by definition we have \( Z_1 = r_m(M)Y_0 \), and it is easy to see that
\[
Z_k = r_m^k(M)Y_0 W_1 W_2 \cdots W_{k-1} \quad \text{for} \quad k \geq 2.
\]
Consequently, \( \text{span}(Z_k) = \text{span}(r_m^k(M)Y_0) \) for \( k \geq 1 \) as claimed. \( \square \)

The following theorem is a straightforward adaptation of [31, Thm. 5.2] (see [4] for the original result).

**Theorem 2.2.** Let the eigenpairs \((\lambda_j, x_j)\) of \( M \) be ordered such that (2.1) is satisfied and consider the FEAST method as described in Steps 1–7 previously. Suppose that \( |r_m(\lambda_n)| > 0 \) and \( Y_0 \) in Step 1 is such that the \( n \times n \) matrix \( X_n^*BY_0 \) is invertible. Let \( P_k \) be the \( B \)-orthogonal projector onto the subspace \( \text{span}(Z_k) \). Then for each \( j = 1, 2, \ldots, n \) there is a constant \( \alpha_j \) such that
\[
\|(I - P_k)x_j\|_B \leq \alpha_j \left| \frac{r_m(\lambda_{n+1})}{r_m(\lambda_j)} \right|^k \quad \text{for iterations} \quad k \geq 1.
\]
In particular, \( \|I - P_k\|_B \to 0 \) as long as \( |r_m(\lambda_1)| > |r_m(\lambda_{n+1})| \).

Proof. As observed previously, \( I = X_nX_n^*B + X_n'X_n'B \). Therefore
\[
Y_0 = (X_nX_n^*B + X_n'X_n'B)Y_0
= (X_n + X_n'(X_n^*BY_0)(X_n^*BY_0)^{-1})X_n^*BY_0.
\]
Hence \( \text{span}(Y_0) = \text{span}(X_n + X_n'W) \), where \( W \) is the \((N - n) \times n\) matrix
\[
W = (X_n^*BY_0)(X_n^*BY_0)^{-1}.
\]
Writing \( W \) as \([w_1, w_2, \ldots, w_{n}]\), the vector \( x_j + X_n'w_j \) is an element of \( \text{span}(Y_0) \). Define the constant \( \alpha_j \) as \( \|w_j\|_2 \). By Lemma 2.1, \( \text{span}(Z_k) = \text{span}(r_k^k(M)Y_0) \) and thus \( r_k^k(M)(x_j + X_n'w_j) \in \text{span}(Z_k) \). But \( r_k^k(M) = X r_m^k(A) X^{-1} \) and thus
\[
r_k^k(M)(x_j + X_n'w_j) = r_k^k(\lambda_j) x_j + X_n' r_k^k(\lambda_{n'}) w_j.
\]
Therefore, the vector \( x_j + X_n'\tilde{w}_j \) is an element of \( \text{span}(Z_k) \), where
\[
\tilde{w}_j = \text{diag}\left( \frac{r_k^k(\lambda_{n+1})}{r_k^k(\lambda_j)}, \frac{r_k^k(\lambda_{n+2})}{r_k^k(\lambda_j)}, \ldots, \frac{r_k^k(\lambda_N)}{r_k^k(\lambda_j)} \right) w_j.
\]
Hence \( \|\tilde{w}_j\|_2 \leq \alpha_j \left|\frac{r_m(\lambda_{n+1})}{r_m(\lambda_j)}\right|^k \). Therefore, inside \( \text{span}(Z_k) \) lies a vector \( x_j + e_j \) with \( \|e_j\|_B = \|\tilde{w}_j\|_2 \leq \alpha_j \left|\frac{r_m(\lambda_{n+1})}{r_m(\lambda_j)}\right|^k \). Finally,
\[
\|(I - P_k)x_j\|_B = \min_{z \in \text{span}(Z_k)} \|x_j - z\|_B
\leq \|e_j\|_B
\leq \alpha_j \left|\frac{r_m(\lambda_{n+1})}{r_m(\lambda_j)}\right|^k,
\]
which completes the proof. \( \square \)

We learn from Theorem 2.2 that fast convergence can be achieved for a wanted eigenpair \((\lambda_j, x_j)\) if the ratio \( |r_m(\lambda_{n+1})/r_m(\lambda_j)| \) is small \((j \leq n)\). This is an approximation problem which we will investigate closer in the following sections.

3. Two simple quadrature rules. We assume without loss of generality that the pencil \((A, B)\) has been transformed linearly to \((\alpha A - \beta B, B)\) such that the wanted eigenvalues are contained in the interval \((-1, 1)\). For a given scaling parameter \( S > 1 \) we define a family of ellipses \( \Gamma_S \) as
\[
\Gamma_S = \left\{ \gamma : \gamma = \gamma(\theta) = \frac{Se^{i\theta} + S^{-1}e^{-i\theta}}{S + S^{-1}}, \ \theta \in [0, 2\pi) \right\}.
\]
Note that \( \gamma(\theta) = \cos(\theta) + i\frac{S - S^{-1}}{S + S^{-1}} \sin(\theta) \), hence these ellipses enclose the interval \((-1, 1)\) and pass through the interval endpoints \( \pm1 \). As \( S \to \infty \), the ellipses approach the unit circle. After a straightforward change of variables one can evaluate the integral (1.3) with contour \( \Gamma = \Gamma_S \) via integration over \([0, 2\pi]\) as
\[
f(z) = \frac{1}{2\pi i} \int_0^{2\pi} \frac{\gamma'(\theta)}{\gamma(\theta) - z} d\theta =: \int_0^{2\pi} g_z(\theta) d\theta,
\]
where
\[
g_z(\theta) := \frac{(Se^{i\theta} - S^{-1}e^{-i\theta})/(S + S^{-1})}{2\pi (Se^{i\theta} + S^{-1}e^{-i\theta})/(S + S^{-1}) - z}.
\]
Two different approaches for the numerical approximation of the integral (3.2) have been considered in the context of contour-based eigensolvers.
3.1. Gauss quadrature. It was proposed in [29] to use \( m \) Gauss quadrature nodes \( \tilde{\theta}_j^{(G)} \) \((j = 1, \ldots, m)\) on the interval \([0, \pi]\), and another set of \( m \) Gauss quadrature nodes \( \tilde{\theta}_j^{(G)} \) \((j = m + 1, \ldots, 2m)\) on the interval \([\pi, 2\pi]\) (see, e.g., [8, Section 2.7] for a derivation of Gauss quadrature rules). Denoting the corresponding Gauss weights by \( \omega_j^{(G)} \), we have for (3.2) the quadrature approximation

\[
 f(z) \approx \sum_{j=1}^{2m} \omega_j^{(G)} g_z(\tilde{\theta}_j^{(G)}) =: r_m^{(G)}(z),
\]

with a rational function \( r_m^{(G)} \). Defining the mapped Gauss nodes and weights

\[
 z_j^{(G)} = \frac{S e^{i\tilde{\theta}_j^{(G)}} + S^{-1} e^{-i\tilde{\theta}_j^{(G)}}}{S + S^{-1}}, \quad w_j^{(G)} = \frac{\omega_j^{(G)} S e^{i\tilde{\theta}_j^{(G)}} - S^{-1} e^{-i\tilde{\theta}_j^{(G)}}}{S + S^{-1}}, \quad j = 1, \ldots, 2m,
\]

the function \( r_m^{(G)} \) can be written in the form

\[
 r_m^{(G)}(z) = \sum_{j=1}^{2m} \frac{w_j^{(G)}}{z_j^{(G)} - z}.
\]

This is a rational function of type \((2m - 1, 2m)\). By construction, its \( 2m \) poles have a four-fold symmetry about the origin,

\[
 z_j^{(G)} = -\frac{z_{m+1-j}^{(G)}}{z_{m+1-j}^{(G)}} = -z_{m+j}^{(G)} = z_{2m+1-j}^{(G)} \quad \text{for } j = 1, \ldots, m,
\]

in particular, the poles occur in complex conjugate pairs. If \( A \) and \( B \) are real symmetric this can be computationally convenient because for a real vector \( v \) one has

\[
 (zB - A)^{-1} v = (zB - A)^{-1} w,
\]

and hence the number of linear systems to be solved for computing \( r_m^{(G)} (B^{-1} A) v \) is only \( m \) instead of \( 2m \).

A graphical illustration of \( r_m^{(G)} \) is given in Figure 3.1. When \( S \) decreases this rational function becomes quite “wiggly” on the interval \((-1, 1)\), with the oscillations being caused by the nearby poles and hence becoming larger as the ellipse gets flatter.

3.2. Trapezoid rule. As the integrand \( g_z \) in (3.2) is a \( 2\pi \)-periodic function, it appears most natural to use the trapezoid rule for its integration over \([0, 2\pi]\). Indeed this is the preferred choice of quadrature rule in the moment-based methods (see, e.g., [35, 18]). We refer to [39] for a review of the trapezoid rule and its properties.

Let us take equispaced quadrature nodes \( \tilde{\theta}_j^{(T)} = \pi(j - 1/2)/m \) and equal weights
\[
 \tilde{\omega}_j^{(T)} = \pi/m, \quad j = 1, \ldots, 2m,
\]

and use for (3.2) the trapezoid approximation (cf. [8, Section 2.1])

\[
 f(z) \approx \sum_{j=1}^{2m} \tilde{\omega}_j^{(T)} g_z(\tilde{\theta}_j^{(T)}) =: r_m^{(T)}(z),
\]

with a rational function \( r_m^{(T)} \) of type at most \((2m - 1, 2m)\). Defining the mapped trapezoid nodes and weights

\[
 z_j^{(T)} = \frac{S e^{i\tilde{\theta}_j^{(T)}} + S^{-1} e^{-i\tilde{\theta}_j^{(T)}}}{S + S^{-1}}, \quad w_j^{(T)} = \frac{1}{2m} \frac{S e^{i\tilde{\theta}_j^{(T)}} - S^{-1} e^{-i\tilde{\theta}_j^{(T)}}}{S + S^{-1}}, \quad j = 1, \ldots, 2m,
\]
the rational function $r_m^{(T)}$ can be written in the form
\begin{equation}
    r_m^{(T)}(z) = \sum_{j=1}^{2m} \frac{w_j^{(T)}}{z_j - z}.
\end{equation}

We now show that $r_m^{(T)}$ has a close connection with Chebyshev polynomials.

**Lemma 3.1.** The rational function $r_m^{(T)}$ can be written as fractional transformations of $T_j(z) = \cos(j \arccos(z))$, the first-kind Chebyshev polynomial of degree $j$. More precisely,
\begin{equation}
    r_m^{(T)}(z) = \frac{1}{\alpha + \beta T_{2m}(\frac{S+S^{-1}}{2} z)} = \frac{1}{(\alpha - \beta) + 2\beta T_{m}(\frac{S+S^{-1}}{2} z)^2}
\end{equation}

with
\begin{equation}
    \alpha = \frac{S^{2m} + S^{-2m}}{S^{2m} - S^{-2m}}, \quad \beta = \frac{2}{S^{2m} - S^{-2m}}.
\end{equation}

Therefore, $r_m^{(T)}$ is of exact type $(0, 2m)$. Moreover, it is equioscillating $2m + 1$ times on the interval $[-2/(S+S^{-1}), 2/(S+S^{-1})]$, alternating between the values $(\alpha \pm \beta)^{-1}$.

**Proof.** We only consider the first equality in (3.5), the second following from the relation $T_{2m}(z) = 2T_m(z)^2 - 1$. Let the rational function
\begin{equation}
    w(z) = \frac{1}{\alpha + \beta T_{2m}(\frac{S+S^{-1}}{2} z)}
\end{equation}

be defined with $\alpha$ and $\beta$ as in (3.6). Being clearly of type $(0, 2m)$, it suffices to prove that $w$ has the same poles as $r_m^{(T)}$ defined in (3.4) and the same residues at
these points. We make use of the following formulas for Chebyshev polynomials of a complex variable [24, Section 1.4.2], namely
\[ T_{2m} \left( \frac{z + z^{-1}}{2} \right) = \frac{z^{2m} + z^{-2m}}{2}, \]
\[ U_{2m-1} \left( \frac{z + z^{-1}}{2} \right) = \frac{z^{2m} - z^{-2m}}{z - z^{-1}}, \]
\[ \frac{d}{dz} T_{2m}(z) = 2m U_{2m-1}(z), \]
where \( U_{2m-1} \) is the Chebyshev polynomial of the second kind of degree \( 2m - 1 \).

Defining \( u_j = Se^{i\theta_j(T)} \) and \( S^{2m} e^{2i m \theta_j(T)} = -S^{2m} \),
\[ S + S^{-1} \frac{z_j(T)}{2} = \frac{u_j + u_j^{-1}}{2} \]
and
\[ u_j^{2m} = S^{2m} e^{2i m \theta_j(T)} = -S^{2m}. \]

Therefore,
\[ \alpha + \beta T_{2m} \left( S + S^{-1} \frac{z_j(T)}{2} \right) = \alpha + \beta T_{2m} \left( \frac{u_j + u_j^{-1}}{2} \right) \]
\[ = \alpha + \beta \frac{u_j^{2m} + u_j^{-2m}}{2} \]
\[ = \alpha - \beta \frac{S^{2m} + S^{-2m}}{2}, \]
which gives zero when inserting the values (3.6) for \( \alpha \) and \( \beta \). It remains to show that the residue of \( w \) at \( z_j(T) \) is precisely \( -u_j^{(T)} \). To this end we make use of the fact that the residue at a point \( z \) of a rational function \( p/q \), where \( p \) and \( q \) are polynomials such that \( q \) has a simple root at \( z \) and \( p \) is nonzero there, is given by \( p(z)/q'(z) \). The residue of \( w \) at \( z = z_j(T) \) is hence given by
\[ \frac{1}{\beta m (S + S^{-1}) U_{2m-1} \left( \frac{S + S^{-1}}{2} z_j(T) \right)} \]
\[ = \frac{1}{\beta m (S + S^{-1}) U_{2m-1} \left( \frac{u_j + u_j^{-1}}{2} \right)} \]
\[ = \frac{1}{\beta m (S + S^{-1}) \frac{u_j^{2m} - u_j^{-2m}}{u_j - u_j^{-1}}} \]
\[ = u_j - u_j^{-1}. \]
which indeed agrees with \( -u_j^{(T)} \). The equioscillation property of \( r_m(T) \) follows directly from the equioscillation of \( T_{2m} \). \( \square \)

We learn from Lemma 3.1 that \( r_m(T) \) is precisely a type-1 Chebyshev filter as commonly used in electronic filter design; see, e.g., [17, § 13.5]. A graphical illustration of \( r_m(T) \) is given in Figure 3.2. Note that this rational function is perfectly equioscillating on the interval \([-2/(S + S^{-1}), 2/(S + S^{-1})]\), which becomes wider as the ellipse gets
Fig. 3.2. The contours $\Gamma_S$ along with the associated rational functions $r_m^{(T)}$ obtained from the mapped trapezoid quadrature rule for three different values of $S \in \{\infty, 2, 1.25\}$. The modulus of $r_m^{(T)}$ is plotted for each value of $S$ and for three different values of $m \in \{3, 5, 8\}$ over the interval $[0, 2]$ (it is a symmetric function). For clarity, the poles of $r_m^{(T)}$ are shown for the case $m = 3$ only.

flatter ($S \to 1$). On the other hand, the function values are between $(\alpha \pm \beta)^{-1}$ with $\beta = 2/(S^{2m} - S^{-2m})$, so the oscillations become larger as $S \to 1$. In the other limiting case, when $S \to \infty$, there are no oscillations and

$$r_m^{(T)}(z) = \frac{1}{2m} \sum_{j=1}^{2m} \frac{e^{i\pi(j-1/2)/m}}{e^{i\pi(j-1/2)/m} - z},$$

which is also known as the Butterworth filter; see, e.g., [17, § 12.6] or [3]. This relation between the type-1 Chebyshev and Butterworth filters is well known in the literature (see, e.g., [6, p. 119]). By symmetry considerations one can show that $r_m^{(T)}$ in (3.8) attains the value $1/2$ for $z = \pm 1$. This property is also shared by $r_m^{(G)}$ as we show in the following remark.

**Remark 1.** Assume that the poles and corresponding weights have a four-fold symmetry about the origin, i.e., if $(w, z)$ is a weight–pole pair, then also $(-w, -z)$, $(w, -z)$, and $(\bar{w}, \bar{z})$ are weight–pole pairs. In this case one can verify that

$$r_m(\pm 1) = \sum_{j=1}^{2m} \frac{w_j}{z_j - 1} = \sum_{j=1}^{m/2} \frac{\omega_j}{\pi} \frac{S^4 - 1}{S^8 - 2 \cos(2\theta_j)S^4 + 1}.$$  

Whatever the values $\theta_j$, we have

$$\frac{1}{\pi} \frac{S^4 - 1}{S^4 + 1} \sum_{j=1}^{m/2} \omega_j < r_m(\pm 1) < \frac{1}{\pi} \frac{S^4 + 1}{S^4 - 1} \sum_{j=1}^{m/2} \omega_j.$$  

For $S \to \infty$ we obtain $r_m(\pm 1) \to \pi^{-1} \sum_{j=1}^{m/2} \omega_j$. Therefore for both the Gauss and the
trapezoid quadrature rules we have \( \sum_{j=1}^{m/2} \omega_j^{(G)}(T) = \pi/2 \), hence
\[
\lim_{s \to \infty} r_m^{(G), (T)}(\pm 1) = \frac{1}{2}.
\]

4. A method based on Zolotarev approximants. Both quadrature rules in Section 3 achieve a small approximation error for (1.3) throughout the complex plane, except when \( z \) is close to the contour \( \Gamma \). Assume again that the wanted eigenvalues of \( M \) are contained in the interval \((-1, 1)\). Then for a fast convergence of FEAST, in view of Theorem 2.2, our main concern should be the accuracy of \( r_m(M) \) as an approximation to the indicator function \( \text{ind}_{[-G, G]}(M) \), where
\[
\text{ind}_{[-G, G]}(z) = \begin{cases} 
1 & \text{if } z \in [-G, G] \\
0 & \text{otherwise},
\end{cases}
\]
with some \( G < 1 \). We will refer to \( G \) as the gap parameter; because it is related to the gap between the wanted and unwanted eigenvalues. The smaller the value of \( G \), the larger the gap. Since \( M \) has real eigenvalues, it seems natural to concentrate all of \( r_m's \) “approximation power" to the real line. In other words, we are looking for a rational function \( r_m \) of degree \( 2m \) such that \( r_m \) is closest to 1 on a largest possible interval \([-G, G] \subset (-1, 1)\), and closest to 0 on a largest possible subset of the complement. Such a rational function is explicitly known due to an ingenious construction of Zolotarev [44] and in the filter design literature typically referred to as band-pass Cauer filter or elliptic filter (see, e.g., [6, § 3.7.4] or [40, § 13.6]). The construction of this filter makes use of elliptic functions.

Let the Jacobi elliptic function \( sn(w; \kappa) = x \) be defined by
\[
w = \int_0^x \frac{1}{\sqrt{(1-t^2)(1-\kappa^2 t^2)}} dt,
\]
and let the complete elliptic integral for the modulus \( \kappa \) be denoted by
\[
K(\kappa) = \int_0^1 \frac{1}{\sqrt{(1-t^2)(1-\kappa^2 t^2)}} dt.
\]
The following well-known theorem summarizes one of Zolotarev’s findings; we use a formulation given by Akhiezer [2, Chapter 9].

**Theorem 4.1 (Zolotarev, 1877).** The best uniform rational approximant of type \((2m - 1, 2m)\) for the signum function \( \text{sgn}(x) \) on the set \([-R, -1] \cup [1, R] \), \( R > 1 \), is given by
\[
s_m(x) = x D \prod_{j=1}^{m-1} (x^2 + c_{2j}) \prod_{j=1}^{m} (x^2 + c_{2j-1}) \quad \text{with} \quad c_j = \frac{\text{sn}^2(j K(\kappa)/(2m); \kappa)}{1 - \text{sn}^2(j K(\kappa)/(2m); \kappa)},
\]
where \( \kappa = \sqrt{1-1/R^2} \) and the constant \( D \) is uniquely determined by the condition
\[
\min_{x \in [-R, -1]} s_m(x) + 1 = \max_{x \in [1, R]} -s_m(x) + 1.
\]

1. The definition of elliptic functions is not consistent in the literature. We stick to the definitions used in [2, §25]. For example, in MATLAB one would type \( \text{sn} = \text{ellipj}(w, \text{kappa}^2) \) and \( K = \text{ellipke(kappa}^2) \) to obtain the values of \( \text{sn}(w; \kappa) \) and \( K(\kappa) \), respectively.
Fig. 4.1. Transformed rational function $r_m^{(Z)}(z)$ based on Zolotarev’s approximant. The parameters are $m = 3$ and $R \in \{1e2, 1e4, 1e6\}$ is varied. The interval of equioscillation about the value 1 is $[-G, G]$, where $G = \sqrt{R^2 - 1}/(\sqrt{R} + 1)$; see formula (4.4). As stated in Corollary 4.2 all poles lie on the unit circle and appear in complex conjugate pairs. The upper right plot for $R = 1e6$ contains a zoomed-in version to show the proximity of the poles to $z = \pm 1$.

The last normalization condition in Theorem 4.1 ensures that $s_m(x)$ is equioscillating about the value $-1$ on $[-R, -1]$, and equioscillating about the value 1 on $[1, R]$. In fact, it is known that there is a number of $4m + 2$ equioscillation points, a number that clearly has to be even due to the symmetry $s_m(-x) = -s_m(x)$. This is one equioscillation point more than required by Chebyshev’s characterization theorem for uniform best rational approximations (see, e.g., [28, §2.2]), which states that a rational function of type $(\mu, \nu)$ with $\mu + \nu + 2$ equioscillation points is a unique best approximant.

Let us briefly highlight some properties of $s_m(x)$. First of all, $s_m(0) = 0$ due to the symmetry, and $s_m(\infty) = 0$ as $s_m$ is a rational function of type $(2m - 1, 2m)$. Let us define by $E_m$ the maximal modulus of the error function $e_m(x) := \text{sgn}(x) - s_m(x)$ over the set $[-R, -1] \cup [1, R]$, i.e.,

$$E_m := \max_{x \in [-R, -1] \cup [1, R]} |e_m(x)| = \max_{x \in [-R, -1] \cup [1, R]} |\text{sgn}(x) - s_m(x)|.$$ 

Then $|e_m(x)|$ takes on its maximum $E_m$ at the points $x \in \{-R, -1, 1, R\}$. In [25, eq. (3.17)] lower and upper bounds on $E_m$ have been given as

$$\frac{4\rho^m}{1 + \rho^m} \leq E_m \leq 4\rho^m,$$

\[2\text{Note that Chebyshev’s classical equioscillation criterion is typically stated for a single closed interval and does not strictly apply in the case of two intervals. However, looking closer at Zolotarev’s construction [44] we find that it is based on a weighted best rational approximant for } 1/\sqrt{x}\text{ on the single interval } [1, R^2], \text{ on which the equioscillation criterion holds. Zolotarev then uses the relation } \text{sgn}(x) = x/\sqrt{x^2} \text{ to find } s_m(x).\]
where
\[ \mu = \left( \frac{\sqrt{R} - 1}{\sqrt{R} + 1} \right)^2, \quad \mu' = \sqrt{1 - \mu^2}, \quad \rho = \rho(\mu) = \exp \left( -\frac{\pi K(\mu')}{2K(\mu)} \right). \]

Our aim is to determine a Moebius transformation
\[ x = t(z) = \frac{a + bz}{c + dz}, \]
such that the rational function
\[ r_m(z) := \frac{s_m(t(z)) + 1}{2} \]
is an approximation of the indicator function \( \text{ind}_{[-G,G]}(z) \) on some interval \([-G,G] \subset (-1,1)\) with gap parameter \( G < 1 \).

Due to the symmetry of the indicator function, it is natural to demand that
\[ r_m(z) = r_m(-z) \]
for real \( z \), and we will also prescribe \( r_m(-1) = r_m(1) = 1/2 \).

This yields the following conditions for the transformation \( t \):
\[ t(-1) = 0, \quad t(-G) = 1, \quad t(G) = R, \quad t(1) = \infty. \]

From these conditions the transformation \( t \) and \( G \) are readily determined as
\[ x = t(z) = \sqrt{R} \frac{1 + z}{1 - z}, \quad G = \frac{\sqrt{R} - 1}{\sqrt{R} + 1}. \]

By construction, the rational function \( r_m(z) \) is indeed equioscillating about the value 1 for \( z \in [-G,G] \), and equioscillating about the value 0 for \( z \in [-\infty,-G^{-1}] \) and \( z \in [G^{-1},+\infty] \). The number of \( 4m + 2 \) equioscillation points of \( s_m(x) \) is inherited by \( r_m(z) \). Note that a rational transformation of type \((1,1)\) inserted into a rational function of type \((2m - 1,2m)\) in general gives a rational function of type \((2m,2m)\).

For a visual example see Figure 4.1.

The following corollary summarizes the above findings.

**Corollary 4.2.** The rational function \( r_m(z) \) given by (4.3) and (4.4) is the best uniform rational approximant of type \((2m,2m)\) of the indicator function \( \text{ind}_{[-G,G]}(z) \) on
\[ [-G,G] \quad \text{and} \quad [-\infty,-G^{-1}] \cup [G^{-1},+\infty]. \]
The error curve \( e'_m(z) := \text{ind}_{[-G,G]}(z) - r_m(z) \) equioscillates on these sets with error \( E'_m := \max_{z \in [-G,G]} |e'_m(z)| \) bounded by
\[ \frac{2\rho^m}{1 + \rho^m} \leq E'_m \leq 2\rho^m, \]
where
\[ \mu = G^2, \quad \mu' = \sqrt{1 - \mu^2}, \quad \rho = \rho(\mu) = \exp \left( -\frac{\pi K(\mu')}{2K(\mu)} \right). \]
Moreover, all \( 2m \) poles of \( r_m(z) \) lie on the unit circle and appear in complex conjugate pairs.
Proof. The first statement follows from the fact that \( r_m^{(z)} \) defined in (4.3) has been obtained from \( s_m(x) \) by the bijective transformation \( x = t(z) \) in (4.2).

The error inequalities follow from (4.3) and (4.1), and the fact that \( \mu = G^2 \).

Finally, from (4.3), we find that \( z_j \) is a pole of \( r_m^{(Z)} \) if and only if \( t(z_j) \) is a pole of \( s_m \) defined in Theorem 4.1. Inverting the relation \( x = t(z) \) and using the fact that the poles of \( s_m \) are \( \pm i \sqrt{c_{2j-1}} \), \( j = 1, \ldots, m \), the poles of \( r_m^{(Z)} \) are found to be

\[
\pm i \sqrt{c_{2j-1}} - \sqrt{R} = \frac{c_{2j-1} - R}{c_{2j-1} + R} \pm \frac{2\sqrt{c_{2j-1}R}}{c_{2j-1} + R}, \quad j = 1, \ldots, m,
\]

which are complex conjugate and of modulus 1. \( \square \)

Remark 2. When \( G \) (and hence \( \mu \)) is sufficiently close to 1, it is possible to use [1, (17.3.11) and (17.3.26)] to derive the asymptotically sharp estimates \( K(\mu) \approx \log(4/\mu) \) and \( K(\mu^2) \approx \pi/2 \) (see also [25]), and thereby give the estimate

\[
E_m^{(z)} \approx 2 \exp \left( -m \frac{\pi^2}{4 \log(4/\mu^2)} \right) = 2 \exp \left( -m \frac{\pi^2}{2 \log(16/(1-G^2))} \right)
\]

in terms of elementary functions.

Remark 3. It may be instructive to study the simplest Zolotarev function \( r_m^{(Z)} \) for \( m = 1 \), that is, a rational function of type \((2m,2m) = (2,2)\). Let a gap parameter \( G < 1 \) be given. As the poles of \( r_m^{(Z)} \) lie on the unit circle, and due to symmetry must be \( \pm i \), this rational function is of the form

\[
r_1^{(Z)}(z) = \gamma + \frac{2\delta}{z^2 + 1},
\]

with real numbers \( \gamma \) and \( \delta \). Due to the equioscillation property we have \( r_1^{(Z)}(\infty) = \gamma = 1 - r_1^{(Z)}(0) \), from which we find that \( 2\delta = 1 - 2\gamma \). Also due to equioscillation we have \( r_1^{(Z)}(G) = 1 + \gamma \), from which we then find \( \gamma = -G^2/2 \), i.e.,

\[
r_1^{(Z)}(z) = -\frac{G^2}{2} + \frac{1 + G^2}{z^2 + 1} = -\frac{G^2}{2} + \frac{(i + i G^2)/2}{z + i} - \frac{(i + i G^2)/2}{z - i}.
\]

5. Comparison of the three quadrature rules. We are now in the position to assess the three discussed rational functions \( r_m^{(G)} \), \( r_m^{(T)} \), and \( r_m^{(Z)} \) in view of their performance within the FEAST method for computing eigenpairs. A main tool will be Theorem 2.2, which allows us to characterize the convergence of FEAST in terms of the underlying rational function. Again assume that all eigenvalues are ordered such that for a given rational function \( r_m \) we have (2.1). We also assume that a number of \( \ell \leq n \) wanted eigenvalues \( \lambda_1, \lambda_2, \ldots, \lambda_{\ell} \) of \( M \) are scaled and shifted to be contained in the interval \([-G, G] \subset (-1,1)\) for some gap parameter \( G < 1 \). Finally, assume that the eigenvalues \( \lambda_{n+1}, \lambda_{n+2}, \ldots, \lambda_N \), which are those outside the search interval not “covered” by the \( n \)-dimensional search space, are contained in the set \([-\infty, -G^{-1}] \cup [G^{-1}, +\infty)\). Such a situation can always be achieved for an appropriately chosen \( G \) provided that \( r_m \) can separate \( \lambda_1 \) and \( \lambda_{n+1} \), i.e., \( |r_m(\lambda_{n+1})| > |r_m(\lambda_1)| \). We can then compute for each quadrature rule and parameter \( m \) the shape parameter \( S > 1 \) (for the Gauss and trapezoid rules), or parameter \( R > 1 \) (for the Zolotarev case), so that
the worst-case convergence factor

\[ \text{factor}_{\text{worst}}(m, G) = \frac{\max_{z \in (-\infty, -G^{-1}] \cup [G^{-1}, +\infty)} |r_m(z)|}{\min_{z \in [-G, G]} |r_m(z)|} \]  

is smallest possible.

In Table 5.1 we show a comparison of the worst-case convergence factors for various values of \(G\) and \(m\). In practice, the gap parameter \(G\) is of course unknown so that we better consider a whole range of this parameter. As can be seen for all gap parameters \(G\) listed in Table 5.1, the optimal worst-case convergence factors of the Zolotarev rule consistently outperform those obtained via trapezoid and Gauss quadrature (with Gauss being slightly better than trapezoid). Let us discuss this table in some more detail.

**Trapezoid rule.** For the trapezoid rule (3.4), a “natural” choice of the parameter \(S\) (in Table 5.1 denoted as “\(S = \text{nat}^t\)”) is to achieve equioscillation on \([-G, G]\), and by Lemma 3.1 this means that \(2/(S + S^{-1}) = G\) should be satisfied. Due to the strictly monotone decay of \(r_m^{(T)}\) outside the interval \([-G, G]\) of equioscillation, the maximum in (5.1) is always attained at \(z = \pm G^{-1}\) and the worst-case convergence factor is given by

\[ \text{factor}^{(T)}_{\text{worst}}(m, G) = \frac{r_m^{(T)}(G^{-1})}{r_m^{(T)}(G)} = \frac{\alpha + \beta}{\alpha + \beta T_{2m}(\frac{\alpha + \beta T_{2m}(G^{-2})}{G^{-1}})} \]

with \(\alpha\) and \(\beta\) defined in Lemma 3.1.

However, this “natural” choice does not necessarily minimize (5.1), see also Table 5.1. We observed numerically that (5.1) decreases monotonically when \(S \to 1\). However, taking \(S\) very close to 1 may be problematic from a numerical point of view because it means that the ellipse \(\Gamma_S\) given by (3.1) degenerates to an interval. This means that the poles of \(r_m^{(T)}\), which lie on \(\Gamma_S\), come potentially close to the wanted eigenvalues, rendering the shifted linear systems in FEAST ill-conditioned or even singular. In our numerical minimization of (5.1) for finding \(S\) we have therefore enforced the constraint \(S \geq 1.01\). In most cases reported in Table 5.1 the optimum for (5.1) was attained for \(S = 1.01\) (or \(S\) being very close to this value).

**Gauss rule.** Due to the irregular behaviour of \(r_m^{(G)}\) defined in (3.3) it appears difficult to make a direct link between the gap parameter \(G\) and the optimal shape parameter \(S\) in the case of Gauss quadrature. For given \(m\) and \(G\) we have therefore computed the optimal parameter \(S\) by minimizing (5.1) numerically (in Table 5.1 denoted as “\(S = \text{opt}^t\)”). Again, similar to the case for the trapezoid rule, we find that the optimal value for \(S\) is very close to 1, causing the ellipse \(\Gamma_S\) given by (3.1) to be very close to the search interval.

**Zolotarev rational function.** Corollary 4.2 tells us that the interval of equioscillation of \(r_m^{(Z)}\) about the value 1 is \([-G, G]\) when \(R\) is chosen such that \(G = (\sqrt{R} - 1)/(\sqrt{R} + 1)\) (see also (4.4)). In Table 5.1 this choice is denoted as “\(R = \text{opt}^t\)”. Moreover, using the error bounds in that same corollary, the worst-case convergence factor can be above from above as

\[ \text{factor}^{(Z)}_{\text{worst}}(m, G) = \frac{r_m^{(Z)}(G^{-1})}{r_m^{(Z)}(G)} = \frac{E_m'}{1 - E_m'} \leq \frac{2\rho^m}{1 - 2\rho^m}. \]

**Remark 4.** It should be emphasized that the Gauss and trapezoid rational approximants decay for \(|z| \to \infty\), whereas the Zolotarev approximant equioscillates by
By dropping the absolute term in the partial fraction expansion of \( z_m^{(2)} \), which is precisely of modulus \( E_m \), this rational function could also be forced to decay for \( |z| \to \infty \). However, we will not explore such a modification in this paper.

---

**Table 5.1**

Worst-case convergence factors (5.1) for various parameter gaps \( G \) and (half-) degrees \( m \).

<table>
<thead>
<tr>
<th>( G )</th>
<th>( m )</th>
<th>( S = \infty )</th>
<th>( S = \text{opt} )</th>
<th>( S = \infty )</th>
<th>( S = \text{opt} )</th>
<th>( G )</th>
<th>( G )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.98</td>
<td>3</td>
<td>8.86e-1</td>
<td>6.01e-1</td>
<td>6.02e-1</td>
<td>1.22</td>
<td>1.36e-1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>7.85e-1</td>
<td>3.15e-1</td>
<td>3.00e-1</td>
<td>1.02</td>
<td>7.46e-3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>6.95e-1</td>
<td>1.89e-1</td>
<td>1.01e-1</td>
<td>1.01</td>
<td>4.51e-4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>6.16e-1</td>
<td>1.18e-1</td>
<td>3.14e-2</td>
<td>1.01</td>
<td>2.74e-5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>5.45e-1</td>
<td>7.38e-2</td>
<td>9.52e-3</td>
<td>1.01</td>
<td>1.67e-6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>2.98e-1</td>
<td>6.24e-3</td>
<td>2.39e-5</td>
<td>1.01</td>
<td>9.73e-13</td>
<td></td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>1.99e-1</td>
<td>1.16e-3</td>
<td>4.50e-7</td>
<td>1.01</td>
<td>1.23e-16</td>
<td></td>
</tr>
</tbody>
</table>

| 0.998  | 3      | 9.88e-1     | 9.33e-1     | 9.33e-1     | 1.07        | 3.58e-1 |
|        | 6      | 9.76e-1     | 7.84e-1     | 7.84e-1     | 1.07        | 4.23e-2 |
|        | 9      | 9.65e-1     | 6.29e-1     | 6.29e-1     | 1.07        | 5.83e-3 |
|        | 12     | 9.53e-1     | 5.03e-1     | 5.04e-1     | 1.07        | 8.26e-4 |
|        | 15     | 9.42e-1     | 4.09e-1     | 4.09e-1     | 1.07        | 1.85e-4 |
|        | 30     | 8.87e-1     | 1.79e-1     | 8.89e-2     | 1.01        | 6.87e-9 |
|        | 40     | 8.52e-1     | 1.10e-1     | 2.73e-2     | 1.01        | 1.05e-11|

| 0.9998 | 3      | 9.99e-1     | 9.93e-1     | 9.93e-1     | 1.02        | 6.32e-1 |
|        | 6      | 9.98e-1     | 9.72e-1     | 9.72e-1     | 1.02        | 3.81e-2 |
|        | 9      | 9.96e-1     | 9.40e-1     | 9.40e-1     | 1.02        | 2.53e-2 |
|        | 12     | 9.95e-1     | 8.98e-1     | 8.99e-1     | 1.02        | 5.09e-3 |
|        | 15     | 9.94e-1     | 8.51e-1     | 8.52e-1     | 1.02        | 1.14e-3 |
|        | 30     | 9.88e-1     | 6.07e-1     | 6.10e-1     | 1.02        | 6.44e-7 |
|        | 40     | 9.84e-1     | 4.81e-1     | 4.83e-1     | 1.02        | 4.41e-9 |

| 0.9998 | 3      | 1.00        | 9.99e-1     | 9.99e-1     | 1.01        | 1.00      |
|        | 6      | 1.00        | 9.97e-1     | 9.97e-1     | 1.01        | 2.15e-1   |
|        | 9      | 1.00        | 9.94e-1     | 9.94e-1     | 1.01        | 5.55e-2   |
|        | 12     | 1.00        | 9.89e-1     | 9.89e-1     | 1.01        | 1.59e-2   |
|        | 15     | 9.99e-1     | 9.82e-1     | 9.83e-1     | 1.01        | 4.67e-3   |
|        | 30     | 9.99e-1     | 9.34e-1     | 9.38e-1     | 1.01        | 1.08e-5   |
|        | 40     | 9.98e-1     | 8.89e-1     | 8.99e-1     | 1.01        | 1.90e-7   |
In practice, however, it is very difficult to get a hand on \( G \) and \( G_{\text{eff}} \), and even the number of wanted eigenvalues, in the first place. It is therefore problematic to rely on the faster convergence that FEAST could exhibit using the trapezoid or Gauss quadrature rules. Moreover, the parameter \( n \) for which faster convergence is achieved with these rules could be unfeasibly large (e.g., when many eigenvalues are clustered at the endpoints of the search interval). An exceptional case is when storage and communication are not an issue and \( n \) can be made (much) larger than the number of wanted eigenvalues \( \ell \).

The Zolotarev rule, on the other hand, makes FEAST robust in the sense that the worst-case convergence factor is independent of \( G_{\text{eff}} \) and explicitly known by (5.2). In the following we will discuss this property in view of the load balancing problem.

6. Load balancing over interval partitions. In order to achieve perfect load balancing when using FEAST in parallel over multiple search intervals one would need to know in advance the number of wanted eigenvalues in each search interval, and then distribute the available parallel resources accordingly. This problem requires information about the eigenvalue distribution over the whole spectrum of interest, and this information is often not available (though we also mention the possibility to use stochastic estimates; see, e.g., [11]). Instead we assume here that the location of the search intervals as well as an estimate for the number of eigenvalues in each interval are given. Our goal is then to obtain fast convergence within approximately the same number of FEAST iterations on each search interval. Some problems related to dissecting the FEAST method and choosing the subspaces appropriately have been discussed in [13, 21].

In the original FEAST publication [29] it was suggested to use a subspace size of \( \times 1.5 \) the estimated number of eigenvalues \( \ell \) inside a given search interval (i.e., \( n = 1.5\ell \)) to obtain a fast convergence using at least 8 nodes for the Gauss quadrature rule along a semi-circle. While these choices for \( n \) and \( m \) worked well for a large number of examples, it is now well understood from the discussions in the previous sections that they are also far from optimal. In particular, we note the following two limiting cases for the choice of \( n \) when the number of contour points \( m \) stays fixed:

- If \( n \) is chosen too small (but still \( n \geq \ell \)), \( \lambda_{n+1} \) could be located very close to the edges of the search interval. This situation is likely to occur, e.g., if the eigenvalues are not evenly distributed and particularly dense at the edges of the interval. By Theorem 2.2 the convergence is expected to be poor in particular when \( |r_m(z)| \) does not decay quickly enough near the search interval and this is the case with the Gauss and trapezoid quadrature rules; see Figure 6.1. The dependence of convergence on the value \( |r_m(\lambda_{n+1})| \) can cause difficulties for achieving load balancing when FEAST runs on several search intervals in parallel. While on some search intervals the method may converge in, say, 2 to 3 iterations because \( |r_m(\lambda_{n+1})| \leq 10^{-5} \), other intervals could require much longer if the subspace size \( n \) is not sufficiently large.

- If \( n \) is chosen too large, \( |r_m(\lambda_{n+1})| \) is likely to be very small and hence convergence would be rapid (see Figure 6.1 in the case of \( m = 8 \)). However, the location of \( \lambda_{n+1} \) is not known a priori and using a very large search space leads to three major problems: (i) a considerable increase in computation time since \( n \) represents the number of right-hand sides to solve for each shifted linear system, (ii) an increase in communication cost as \( n \) vectors need to be communicated to a master processor at each iteration, and (iii) a possibly highly rank-deficient search subspace \( Z_k \) which may prevent the reduced
pencil to be constructed stably without explicit orthogonalization of $Z_k$.

Figure 6.1 shows that the Zolotarev rational function (using $R = 1e6$) has a much steeper slope near the edges of the search interval than the Gauss and trapezoid rules (see, in particular, the magnified plot). As a result, the a-priori known convergence factor for Zolotarev can often be obtained using a very small subspace size $n$ ($n \geq \ell$ with $n \approx \ell$). Moreover, the convergence factor will stay almost constant if $n$ is increased further. In this sense the convergence of FEAST with the Zolotarev rule is predictable and robust. In fact, the a priori knowledge of the convergence factor can be useful for detecting whether the subspace size is chosen large enough: after a few FEAST iterations one calculates an approximate convergence factor from the decrease of the eigenvector residuals; if the approximate convergence factor is much worse than expected from Theorem 4.2 (see also Table 5.1) then $n$ should be increased.

In summary, the delicate choice of $n$ to achieve load balancing and computational efficiency is greatly simplified with the Zolotarev approach. In practice one can achieve a uniform convergence behaviour over multiple search intervals by simply covering the region of interest by translated Zolotarev rational functions $r_m(Z)(z + t)$, possibly with a small overlap. An example of three concatenated Zolotarev functions on the interval $[-3, 3]$ is given in Figure 6.2. The expected convergence factor for all three FEAST runs will be the same (provided that $n$ is sufficiently large) and can be calculated from $G = 0.98$ and $m = 6$ using Theorem 4.2 (in this example we read off from Table 5.1 that the expected convergence factor is $7.46e-3$).

7. Numerical experiments. In this section we discuss three numerical experiments stemming from electronic structure calculations and aiming to compare the robustness and efficiency of FEAST running with Gauss and trapezoid quadrature, as well as the Zolotarev rational function, respectively. We will use three examples to visualize the FEAST convergence obtained with these different quadrature rules and discuss the dependency on the dimension of the search space. In the last experiment we will also show the suitability of the Zolotarev rule for achieving load balance.
All results have been obtained using the sparse solver interface of FEAST v2.1 [12], which has been modified for this article to integrate the Zolotarev nodes and weights for the fixed parameter $R = 1e6$ (cf. Section 4). The numerical quadratures with the Gauss and trapezoid rules have been performed along a semi-circle (i.e., $S = \infty$ is fixed; cf. Section 3), taking advantage of the eigenvalue counting approach introduced in FEAST v2.1 [37] (this approach requires the value of $r_m$ to be $1/2$ at the interval endpoints, see also Remark 1). We decided to fix these (suboptimal) values for $R$ and $S$ in practice it is very difficult (or impossible) to choose them optimally for a given matrix without knowing the eigenvalues in advance. With different parameters $R$ and $S$ the FEAST convergence rates and iteration numbers will be different, but the qualitative behavior discussed in this section will be the same. Finally, all spectral values $\lambda$ (including the edges of the search interval) are implicitly stated in the physical unit of electron Volt (for consistency with the unscaled matrix data all numerical values should be multiplied by the electron charge $q = 1.602176 \times 10^{-19}$).

7.1. Example I. Let us first consider the cnt matrix which was presented in [29] and can be found in the FEAST package [12, 30]. This matrix stems from a 2D FEM discretization of the DFT/Kohn–Sham equations at a cross-section of a (13,0) Carbon nanotube (CNT) [43]. The corresponding eigenproblem takes the generalized form (1.1) with $A$ real symmetric and $B$ symmetric positive definite. The size of both matrices is $N = 12,450$ and their sparsity patterns are identical with a number of $\text{nnz} = 86,808$ nonzero entries. We are looking for the $\ell = 100$ eigenvalues contained in the search interval $[\lambda_{\text{min}} = -65, \lambda_{\text{max}} = 4.96]$. 

Figure 7.1 shows the residual norms (more precisely, the maximum among all residual norms of all approximate eigenpairs in the search interval) at each FEAST iteration using three ($m = 3$) and eight ($m = 8$) integration nodes for the Gauss, trapezoid, and Zolotarev quadrature rules. We observe poor convergence with Gauss and trapezoid rules using a subspace of small dimension $n = 102$, i.e., $n = \ell + 2$. Zolotarev, on the other hand, converges robustly even with $n = 102$ but remains to converge at the same rate when $n$ is increased further. As explained in Remark 4, the convergence with Gauss and trapezoid rules systematically improves when the subspace size $n$ is increased.

A more detailed comparison between Gauss and Zolotarev for $m = 8$ is provided in Figure 7.2. Clearly $n = 102$ is insufficient for the Gauss rule to achieve a small value $|r_8^{(G)}(\lambda_{n+1})|$, but for larger subspace sizes this value decreases and hence Gauss converges faster. As $|r_8^{(Z)}(\lambda_{103})|$ is sufficiently small, Zolotarev attains its theoretical convergence factor of $1.12 \times 10^{-2}$ (calculated using the results of Section 5) for $n = 102$. 

Fig. 6.2. Translated rational functions $r_m^{(Z)}(z+2jG)$, $j \in \{-1,0,1\}$, covering a larger interval $[-3G,3G]$. In this example we have chosen $G = 0.98$ and $m = 6$. 

Fig. 7.1. Residual norms (more precisely, the maximum among all residual norms of all approximate eigenpairs in the search interval) at each FEAST iteration using three ($m = 3$) and eight ($m = 8$) integration nodes for the Gauss, trapezoid, and Zolotarev quadrature rules.
7.2. Example II. We now present a case where even a large subspace size of $n = 1.5\ell$ is not enough to yield satisfactory FEAST convergence with Gauss quadrature. This generalized eigenproblem, Caffeinep2, is obtained from a 3D quadratic FEM discretization of the Caffeine molecule ($C_8H_{10}N_4O_2$), using an all-electron DFT/Kohn–Sham/LDA model [23, 15]. The size of both matrices $A$ and $B$ is $N = 176,622$ and their sparsity patterns are identical with $nnz = 2,636,091$ nonzero entries. The eigenvalues can be classified into so-called core, valence, and extended/conduction electron states. We are here searching for the first $\ell = 57$ eigenvalues contained in the interval $[\lambda_{\min} = -711, \lambda_{\max} = -0.19]$ covering the three physical state regions.

Figure 7.3 shows the moduli of the Gauss and Zolotarev rational functions with $m = 8$. For Gauss, the choice of two subspace sizes $n = 71$ (i.e., $n \approx 1.25\ell$) and $n = 85$ (i.e., $n \approx 1.5\ell$) are highlighted with their corresponding values $|r_m^{(G)}(\lambda_{n+1})|$. Both values make us expect very poor convergence factors for FEAST, and indeed after 41 iterations the residual norms are found to have decreased only to $2.4 \times 10^{-5}$ ($n = 71$) and $4.8 \times 10^{-6}$ ($n = 85$), respectively. For Zolotarev the figure indicate that the theoretical converge rate is already attained using a subspace size of $n = 71$, in which case FEAST converges within 9 iterations to a residual norm of $7.8 \times 10^{-14}$. With $n = 59$ Zolotarev-FEAST converges to about the same residual norm in 23 iterations.

From the results in Examples I and II we conclude that the suggested choice of $n = 1.5\ell$ for Gauss (see [29]) is capable of providing good convergence rates but it lacks robustness. The initial choice of a subspace size $n = 1.5\ell$ with Zolotarev will typically be safer in practice. Note that the subspace size can easily be truncated after the first few FEAST iterations without affecting the theoretical convergence factor.

7.3. Example III. With the matrix Caffeinep2 from Example II we now evaluate the efficiency of Zolotarev in terms of load balancing when using two search inter-
Fig. 7.2. Moduli of the rational functions \( r_8^{(G)} \) and \( r_8^{(Z)} \) at the eigenvalues of the CNT matrix in the search interval \( [\lambda_{\min} = -65, \lambda_{\max} = 4.96] \). The moduli of the rational functions are given by the height of the vertical lines, and the \( \lambda \)-position indicates the eigenvalue. There are exactly 100 eigenvalues located in the search interval. The horizontal lines provide information about the moduli of the rational functions evaluated at \( \lambda_{n+1} \) for various subspace sizes \( n \in \{102, 125, 150\} \).

Fig. 7.3. Moduli of the rational functions \( r_8^{(G)} \) and \( r_8^{(Z)} \) at the eigenvalues of the Caffeinep2 matrix in the search interval \( [\lambda_{\min} = -711, \lambda_{\max} = -0.19] \). The moduli of the rational functions are given by the height of the vertical lines, and the \( \lambda \)-position indicates the eigenvalue. For visual clarity we have removed from the plots the large gap between the core eigenvalues (part on the left) and valence/conduction eigenvalues (part on the right), where no eigenvalues are found. The function values for Gauss (in the top) are plotted in linear scale while the ones for Zolotarev (in the bottom) are plotted in logarithmic scale. There are 57 eigenvalues located in the search interval, and the horizontal lines provides information about the moduli of the rational functions evaluated at \( \lambda_{n+1} \) for various subspace sizes \( n \), namely \( n \in \{71, 85\} \) for Gauss and \( n \in \{59, 71\} \) for Zolotarev.
vals. The first one, $[-711, -4]$, captures the $\ell_1 = 51$ core and valence electron states while the second one, $[-4, 1.995]$, captures the first $\ell_2 = 55$ extended/conduction electron states.

Table 7.1 reports the number of FEAST iterations needed to converge to a residual norm\(^4\) below $10^{-13}$ using both Gauss and Zolotarev with a subspace size of $n_j \simeq 1.5 \ell_j$ for the two intervals, i.e., $n_1 = 76$ and $n_2 = 83$. The results indicate that the number of FEAST iterations required on different search intervals can differ significantly using Gauss, whereas Zolotarev is capable of providing reliable load balancing. This is consistent with the discussions in Section 6.

<table>
<thead>
<tr>
<th>Intervals</th>
<th>Gauss $m = 8$</th>
<th>Gauss $m = 16$</th>
<th>Gauss $m = 32$</th>
<th>Zolotarev $m = 8$</th>
<th>Zolotarev $m = 16$</th>
<th>Zolotarev $m = 32$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$[-711, -4]$</td>
<td>39</td>
<td>9</td>
<td>5</td>
<td>8</td>
<td>4</td>
<td>$^*$3</td>
</tr>
<tr>
<td>$[-4, 1.995]$</td>
<td>5</td>
<td>3</td>
<td>$^*$3</td>
<td>9</td>
<td>4</td>
<td>$^*$2</td>
</tr>
</tbody>
</table>

**Summary and future work.** We have studied Zolotarev rational functions as filters in the FEAST eigensolver. We have quantified the expected Zolotarev convergence factor and compared it analytically and numerically with the convergence factors obtained via trapezoid and Gauss quadrature. The Zolotarev rational functions possess a very steep slope at the interval endpoints which often allows for a decrease of the search space dimension. Moreover, these functions do not decay towards infinity which causes FEAST to converge at a predictable, and analytically known, rate (for a sufficiently large search space dimension). We discussed the implications in view of load balancing. The new Zolotarev rules will be part of the next FEAST release, version 3.

Several questions remain open for future work. First of all, some of the poles of the Zolotarev rational functions move very close to the real line. The same is true for the mapped Gauss rule, and even the trapezoid rule when a flat ellipse is used as the contour. It is not clear what is the effect of these poles nearby the search interval on the accuracy of the linear system solver. We have observed numerically that the weights are approximately proportional to the imaginary parts of their associated poles so that, possibly, inaccuracies in the linear system solves are damped out. The use of iterative solvers within FEAST has been considered in [14]. An interesting research question is the numerical stability of implicit $B$-orthogonalization, although the FEAST basis could be re-orthogonalized explicitly if required. Our numerical experiments did not indicate any problems with instability.

\(^4\)The actual value of the residual tolerance is not important when assessing the FEAST speedup obtained with improved quadrature rules. As FEAST converges with a geometric rate by Theorem 2.2, the speedup when converging to a fixed residual tolerance $\text{tol}$ is only dependent on the convergence factor, not the residual tolerance itself. For example, if two quadrature rules yield convergence factors $F_1$ and $F_2$, say, then FEAST will require about $I_1 = \log(\text{tol})/\log(F_1)$ and $I_2 = \log(\text{tol})/\log(F_2)$ iterations, respectively. The quotient $I_1/I_2 = \log(F_2)/\log(F_1)$, which corresponds to the speedup, is indeed independent of $\text{tol}$. 
Another question is how the Zolotarev “quadrature rules” generalize to moments of higher order. We have numerically observed that the Zolotarev rules integrate higher-order moments quite accurately when a polynomial weight function is introduced in (1.2). For quadrature of higher-order moments on the unit circle one can argue that the trapezoid rule is optimal, but for the interval case this problem is open. Also it may be beneficial to distribute the number of equioscillation points of the Zolotarev rational function differently, for example, placing more equioscillation points outside the search interval than inside. Such a rational function can easily be constructed and analyzed using the two-interval Zolotarev approach in [10].

Acknowledgement. We are grateful to Anthony Austin, Daniel Kressner, Lukas Krämer, Bruno Lang, Yuji Nakatsukasa, and Nick Trefethen for useful discussions and comments. We also thank the two anonymous referees for insightful comments.

REFERENCES