Resolvent sampling based Rayleigh–Ritz methods using iterative solvers for the numerical solution of Hermitian linear eigenvalue problems

Project TM

Gerhard Dorn

1 Introduction

The aim of this project is to solve the linear eigenvalue problem of a large Hermitian matrix $A = A^\dagger \in \mathbb{C}^{n\times n}$ in a specified search domain (interval) $\Omega = [a, b]$ for excitation analysis. The goal is to evaluate how many eigenvalues lie within Ω and to determine the eigenpairs with an adjustable accuracy.
The methods discussed in this work are based on the Rayleigh–Ritz projection method, which tries to approximate the projection operator $P = \sum_\iota v_\iota v_\iota^*$ containing the orthonormal eigenvectors v_ι corresponding to the eigenvalues $\lambda_\iota \in [a, b]$ within the search domain Ω.
The approximation of the projection operator is carried out with two methods, namely the contour integral method (CIM) using the residue theorem and the rational interpolation method (RIM) leading to filter functions.
The most cost intensive numerical operation of both methods is to solve shifted linear systems $(z_k\mathbb{1} - A)x_{ki} = y_i$ for N different sampling points z_k (integration resp. interpolation points) and L random vectors y_i. In this project iterative solvers are investigated since the physical problem requires to solve rather large sparse matrices where direct solvers cannot be applied.
Whereas CIM requires to choose sampling points z_k in the complex plane leading to a non–Hermitian system matrix, the introduced symmetrized contour integral method (SCIM) leads to a positive definite and Hermitian matrix $\tilde{A}_k$ at the cost of a larger condition number. Depending on the choice of the contour (eccentricity of the ellipse) either the condition number or the approximation property of the projector becomes worse. RIM unifies both advantages (Hermitian system matrix and not quadratically increased condition number).

Research question: This project is dealing with three research questions:

1) In Sec. 1.7 the contour integral method and the rational interpolation method are compared in terms of filter functions, namely how the chosen contour and its discretization influence the approximation property of the projection operator on the desired eigenspace.

2) In Sec. 2 the condition number of the resulting linear systems and especially the influence of the eccentricity of the chosen ellipse is investigated. In order to make the conjugate gradient method applicable the symmetrized contour integral method is introduced.

3) Sec. 3 is dedicated to the questions of how different iterative solver perform in combination with CIM, SCIM and RIM and how the accuracy of the solution of the linear systems influences the final residual of the calculated eigenpairs.

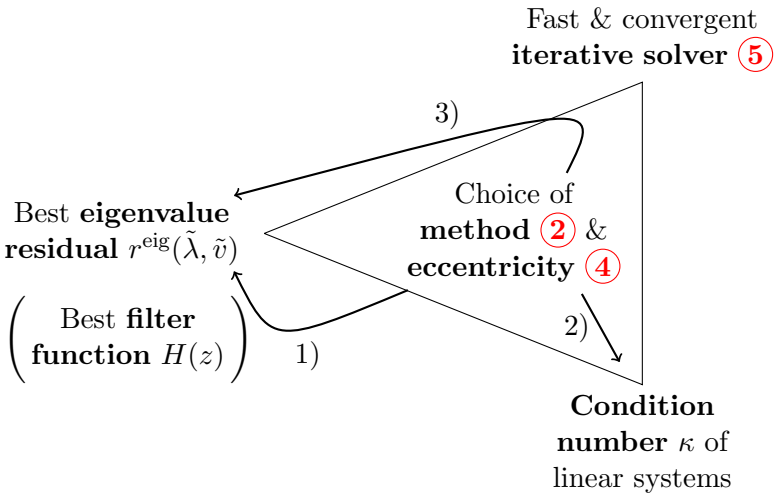


Figure 1: Illustration of the research questions.

Before the three questions are discussed in detail, a short introduction about the physical problem and the used methods is given.

1.1 Physical problem: Spectral analysis of superconducting cuprates

The current superconductors with the highest critical temperature[1] at atmospheric pressure[2] are cuprates - materials comprising a copper oxide layer. It is thus one of the major challenges of theoretical condensed matter physics to develop a theory of superconductivity for these materials. One way to start with is the so-called Hubbard model, an effective model for the electronic degrees of freedom which allows for strong correlation effects [6].
Solving the Hamiltonian derived from the Hubbard model is a nearly infeasible task since the complexity increases exponentially with increasing number of considered atomic orbitals. One way to circumvent this exponential wall is to cluster the orbitals of the material and solve smaller systems which will be reconnected afterwards as done in the so-called cluster perturbation theory (CPT) [3]. Since this theory has a perturbative character the solution becomes better with increased cluster size.
The physical problem underlying this project is to analyze the spectrum of such clusters for specific energy ranges. There are already very efficient methods to determine the groundstate of such clusters but the analysis of the central spectral region is still a challenge.
The corresponding matrices are Hermitian, sparse and may have degenerate eigenvalues. Fig. 2 illustrates a flake of the relevant layer of the examined materials consisting of oxygen (O) and copper (Cu) atoms. The effective Hubbard model comprises four possible states of electron configurations (no electron, $\uparrow$ electron, $\downarrow$ electron, $\uparrow$ and $\downarrow$ electron) for each site, includes kinetic energy terms for the possible hopping of one electron to another site (t, t') and a repulsive Coulomb potential (U, not illustrated in the figure), which describes the repulsive force, when two electrons rest at the same site.

[1] The maximal temperature at which the material becomes superconducting.
[2] Recent papers suggest materials which become superconducting at room temperature when exposed to enormous pressure (approximately the pressure at the center of the earth) [5]

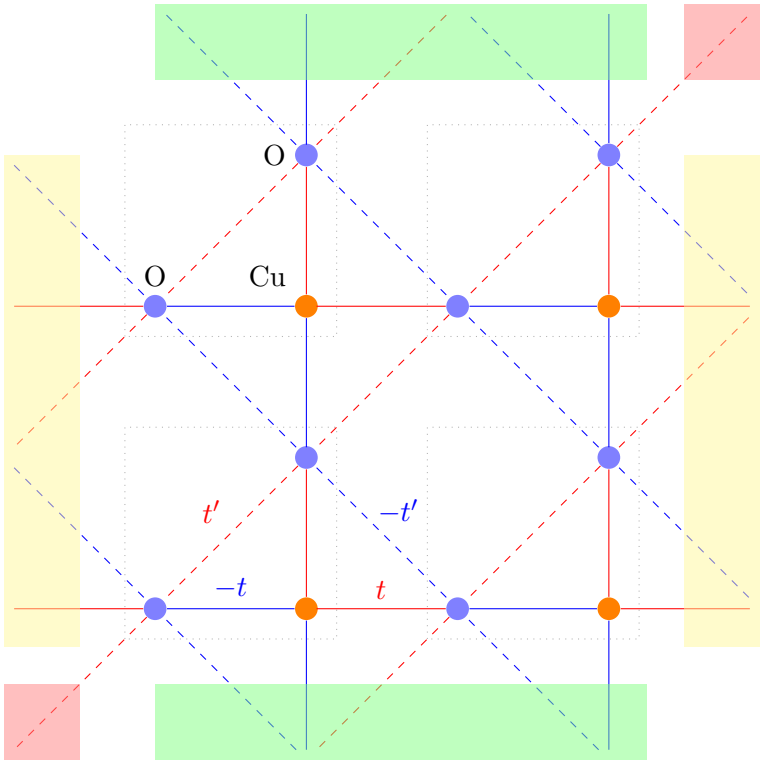


Figure 2: Illustration of the physical model of a copperoxide flake.

The resulting matrix which will be discussed in Sec. 2 and Sec. 3, has at maximally 25 not zero entries in each row and is depicted in Fig. 3.

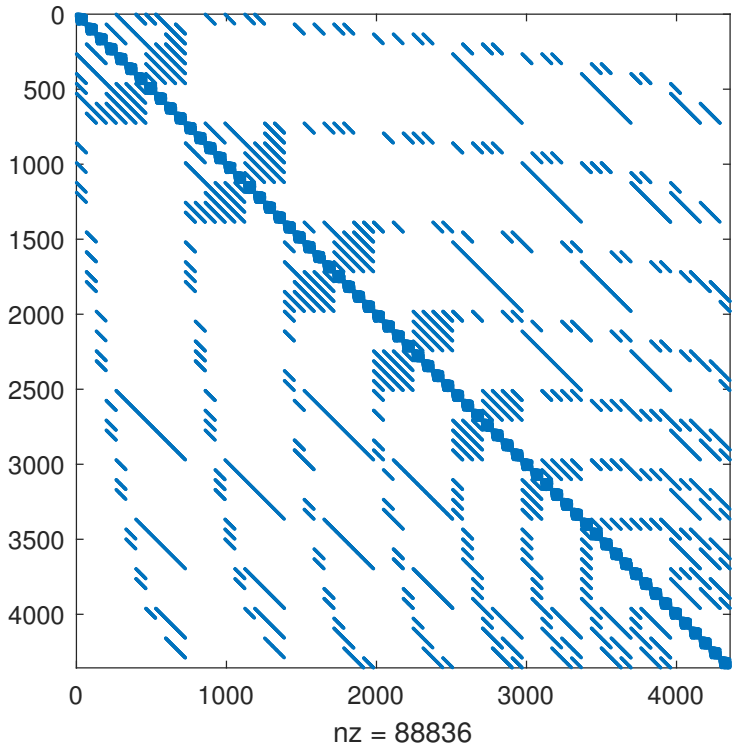


Figure 3: Sparsity structure of the examined matrix A.

1.2 Rayleigh–Ritz projection method

The Rayleigh–Ritz method tries to approximate eigenpairs using a projection operator that spans the space of the corresponding eigenvectors. The Rayleigh–Ritz projection method consists of the following four steps, namely:

1. Compute an orthonormal basis u_ι that approximates the desired eigenspace of interest that corresponds to r eigenvectors of A. These vectors form the projection operator $[u_1, \ldots, u_r] = U \in \mathbb{C}^{n \times r}$.
2. Perform the projection $\mathbb{C}^{r \times r} \ni \tilde{A} := U^* A U$.
3. Solve the eigenproblem of $\tilde{A}$ with corresponding eigenvalues $\tilde{\lambda}_\iota$ and eigenvectors $\tilde{u}_\iota$.
4. Use the Ritz pairs $(\tilde{\lambda}_\iota, \tilde{v}_\iota)$ with $\tilde{v}_\iota = U\tilde{u}_\iota$ as approximation of the desired eigenpairs (λ_ι, v_ι).

Note, that for an exact approximation of the eigenspace, the method yields the exact eigenpairs. The quality of the approximated eigenpairs and thus indirectly the quality of the approximation of the eigenspace can be measured via the **residual of the eigenvalue problem** (eigenvalue residual):

$$r^{\text{eig}}(\tilde{v}, \tilde{\lambda}) = \|A\tilde{v} - \tilde{\lambda}\tilde{v}\|. \tag{1}$$

The Rayleigh–Ritz method using a Krylov subspace as projection operator corresponds to the Arnoldi method.
The following two subsections will introduce two ways to approximate the eigenspace of eigenvalues lying in a certain range based on the resolvent.

1.3 Contour integral method

The residue theorem states that the contour integral of a meromorphic function f with poles z_k yields 2π the sum of weighted residues $\mathrm{Res}(f, z_k)$ that lie within the contour $\mathcal{C}$:

$$\oint_{\mathcal{C}} f(z)dz = 2\pi\mathrm{i}\sum_k \mathrm{Res}(f, z_k)\gamma(\mathcal{C}, z_k).$$

The weight of each residue is the winding number γ that counts how many times the contour winds around the pole z_k and will be one for the poles inside the chosen elliptic contours in this project.
The residue theorem can be extended to matrix-valued functions via the spectral representation or more general via Jordan block representation[3].
The resolvent $R(A, z)$ of a matrix A is given as

$$R(A, z) = \frac{1}{z\mathbb{1} - A},$$

and is meromorphic in the complex plane with poles corresponding to the eigenvalues λ_ι of A. In the case of normal matrices $AA^* = A^*A$ the poles are of order one and the residues are identity operators.
The contour integral of the resolvent of a normal matrix along a Jordan curve representing the boundary $\partial\Omega$ of a simply connected open domain Ω yields the projection P_Λ onto the eigenspace of eigenvectors V_Λ corresponding to the eigenvalues λ that lie within the contour:

$$\frac{1}{2\pi\mathrm{i}}\oint_{\partial\Omega} R(A, z)dz = \frac{1}{2\pi\mathrm{i}}\sum_\iota \oint \frac{1}{z - \lambda_\iota} v_\iota v_\iota^* = \sum_{\lambda_\iota\in\Lambda} v_\iota v_\iota^* =: P_\Lambda, \tag{2}$$

$$\Lambda := \{\lambda \in \sigma(A) : \lambda \in \Omega\}. \tag{3}$$

This result is the basis of the contour integral method which requires the evaluation of the matrix–valued integral. The integral will be discretized using the composite trapezoidal rule with equidistant or uniformly distributed integration points z_k. This discretization of the integral yields exponential convergence [4]. The approximated projection on the eigenspace obtained by discretization of the contour integral is denoted by $\tilde{P}_\Lambda$.
Since the exact evaluation of the resolvent requires a matrix inversion a different approximate approach (resolvent sampling) is discussed in Sec. 1.4).

1.4 Resolvent sampling and use of moments

Since the exact evaluation of the resolvent at points z_k requires a matrix inversion a less elaborate but approximate strategy namely resolvent sampling is pursued. The action of the resolvent is sampled by multiplying random vectors y_i from the right side $x_{ki} := \frac{1}{z_k\mathbb{1}-A}y_i$ yielding the linear system $(z_k\mathbb{1} - A)x_{ki} = y_i$.
In the case of the discretized contour integral this corresponds to projecting the random vectors y_i onto the approximated eigenspace leading to a spanning set

$$S_\Lambda = \left\{ s_i \mid s_i = \tilde{P}_\Lambda y_i,\ i = 1, \ldots, L \right\} = \tilde{P}_\Lambda Y.$$

[3]For more details about the concept of geometric multiplicities and generalized eigenvectors in the non-linear problem can be found in [8].

Moments: Using moments z^j in the contour integral and rational interpolation method can help to enlarge the projection space without having to solve additional linear systems. The residue theorem for resolvents weighted by moments yields a linear combination of projections onto the searched eigenspace:

$$\frac{1}{2\pi \mathrm{i}} \oint_{\partial\Omega} z^j \frac{1}{z\mathbb{1} - A} dz = \sum_{\lambda_\iota \in \Omega} \lambda_\iota^j v_\iota v_\iota^*.$$

The same argument for the enlargement of projection space hold true for the rational interpolation method.
So the combination of resolvent sampling and contour integral method using moments yields also a spanning set S_Λ of the searched eigenspace V_Λ:

$$S_\Lambda = \oint_{\partial\Omega} z^j \underbrace{\frac{1}{z\mathbb{1} - A} Y}_{=X} dz, \tag{4}$$

Since the Rayleigh–Ritz method needs an orthonormal basis of the desired eigenspace, the last step requires a singular value decomposition $S_\Lambda = U \cdot \Sigma \cdot W^*$ with U the orthonormal basis of the approximated eigenspace.
When calculating the spanning set of the desired eigenspace S_Λ using the contour integral method in combination with the resolvent sampling technique, the following approximations are performed in Eq. (4):

- numerical evaluation of the contour integral using the composite trapezoidal formula,
- approximation of the solution of the linear systems using iterative solvers.

1.5 Discretization of the elliptic contour integral

The used elliptic contour is given by the following parametrization:

$$z(\varphi) = \lambda_0 + R\cos(\varphi) + \mathrm{i} \cdot r\sin(\varphi), \qquad \varphi \in [0, 2\pi), \tag{5}$$
$$z'(\varphi) = -R\sin(\varphi) + \mathrm{i} \cdot r\cos(\varphi), \tag{6}$$

with the semi–major and semi–minor axes R and r yielding the eccentricity $e := \sqrt{1 - \frac{r^2}{R^2}}$.
The integral is evaluated numerically using the composite trapezoidal rule. The parameter φ is discretized with φ_k in two ways, namely equidistantly in the arc length and uniformly distributed in the parameter φ. The two different approaches are chosen to discuss the later introduced filter functions in Sec. 1.6.

Equidistant arc length: The parameter φ is discretized in such a way that the grid points $z_k = z(\varphi_k^{\mathrm{arc}})$ are equidistantly spaced along the contour.
The number of discretization points N shall be divisible by four and the points shall be aligned symmetrically along the real axis.
The circumference of the ellipse is given by $4RE(\pi/2, e)$, with the elliptic integral of the second kind[4] $E(\varphi, e) = \int_0^\varphi \sqrt{1 - e^2\cos^2\phi} d\phi$.
The angles φ_k^{arc} parametrizing the equidistant points on the elliptic curve are given by the inverse of the (incomplete) elliptic integral of the second kind via

$$\frac{4RE(\pi/2, e)}{N}(k - \frac{1}{2}) = rE(\varphi_k^{\mathrm{arc}}, e), \quad k = \{1, \ldots, N\}. \tag{7}$$

[4]The angle φ used in this definition starts from the positive real axis as used in the parametrization in contrast to the usual definition of the elliptic integral using a sin.

Uniformly distributed parameters: The second variant of discretization uses uniformly distributed parameters in the interval $[0, 2\pi]$:

$$\varphi_k^{\text{uni}} = \frac{2\pi}{N}(k - \frac{1}{2}), \quad k = \{1, \ldots, N\}, \tag{8}$$

Trapezoidal rule: The discretized points and the discretized derivatives of the parametrization are given by

$$z_k = \lambda_0 + R\cos(\varphi_k) + \mathrm{i} \cdot r\sin(\varphi_k), \quad k = \{1, \ldots, N\}, \tag{9}$$
$$z'_k = -R\sin(\varphi_k) + \mathrm{i} \cdot r\cos(\varphi_k). \tag{10}$$

The discretization of the integral [Eq. (4)] using the composite trapezoidal formula yields:

$$S_\Lambda = \frac{1}{2\pi\mathrm{i}} \int_0^{2\pi} z(\varphi)^j \left[z(\varphi)I - A\right]^{-1} Y z'(\varphi) d\varphi \approx \frac{1}{2\pi\mathrm{i}} \sum_{k=1}^{N} z_k^j \left[z_k I - A\right]^{-1} Y z'_k \tilde{\omega}_k, \tag{11}$$

with the integration weights ω_k:

$$2\tilde{\omega}_k = \begin{cases} \varphi_2 - \varphi_N + 2\pi & k = 1 \\ \varphi_{k+1} - \varphi_{k-1} & k > 1 \wedge k < N \\ \varphi_1 - \varphi_{N-1} + 2\pi & k = N \end{cases} \tag{12}$$

Combining the discretized differential and the prefactor in a general summation weight $\omega_k := \frac{1}{2\pi\mathrm{i}} z'_k \tilde{\omega}_k$ yields the standard form of the used algorithm in this project yielding $N \cdot M$ vectors $S^{(ij)}$ approximately spanning the searched eigenspace:

$$S^{(ij)} = \sum_k \omega_k z_k^j \frac{1}{z_k \mathbb{1} - A} y_i. \tag{13}$$

For the uniformly distributed parameter discretization, the integration weights are $\tilde{\omega}_k^{\text{uni}} = \frac{2\pi}{N}$ yielding the general summation weight $\omega_k^{\text{uni}} = \frac{1}{\mathrm{i}N} z'_k$.

1.6 Rational interpolation method

A second method to approximate the projection operator for the Rayleigh–Ritz approach is the so–called rational interpolation method (RIM).
The original idea in terms of search of eigenvalues stems from interpolating the resolvent with two polynomials $\frac{p(z)}{q(z)} \approx \frac{1}{z\mathbb{1}-A}$ so that the roots of the denominator polynomial $q(z)$ match the poles of the resolvent and thus correspond to the eigenvalues. The coefficients of the polynoms of degree μ_p and μ_q are determined by enforcing the interpolation equation $p(z_k) = \frac{1}{z_k\mathbb{1}-A} q(z_k)$ at interpolation points z_k. These interpolation points z_k are chosen to be in the search domain of the spectrum, so that the interpolation of poles works well for the eigenvalues in this domain of interest.
Since this straightforward approach suffers from numerical instabilities (see an example in Chapter 3 of [2]) the actual RIM method uses **rational filters**.
The idea is that the interpolation of the resolvent using interpolation points z_k in the search domain Ω amplifies the eigenvectors corresponding to those eigenvalues within Ω. This so–called **rational filter function** $H(z)$ has the general form:

$$H(z) = \frac{1}{\prod_k (z_k - z)} = \sum_k \frac{w_k}{z_k - z}, \tag{14}$$

with the barycentric weights $w_k = c \cdot \left(\prod_{j\neq k}(z_j - z_k)\right)^{-1}$. The normalization c is given by the product of the deviations from the mean value of the interpolation points $\bar{z} = \frac{\sum_k z_k}{N}$ (for a symmetrically setup $\bar{z} = \frac{b+a}{2}$):

$$c = \prod_k (\bar{z} - z_k)\,.$$

Applying the filter function to the normal matrix A amplifies the spectral projectors $v_\iota v_\iota^*$ corresponding to eigenvalues λ_ι close to the interpolations points z_k:

$$H(A) = \sum_k \frac{w_k}{z_k \mathbb{1} - A} = \sum_{\iota k} \frac{w_k}{z_k - \lambda_\iota} v_\iota v_\iota^*. \tag{15}$$

When applying the resolvent sampling technique via multiplication with a random vector y_i one obtains a spanning set of the amplified eigenspace. Adding the trick of moments z^j to save the computation of linear systems (see Sec. 1.4) one formally ends up with the same generic formula as derived after the numerical evaluation of the contour integral [see Eq. 13] having barycentric weights w_k instead of the general weights ω_k stemming from the contour integral discretization. One can show in fact that the rational interpolation method is a generalization of the contour integral method using resolvent sampling and a discretized integral [2].
The strong advantage of RIM is that the interpolation points don't have to lie on a contour in the complex plane but can be chosen also on the real axis yielding a symmetric linear system which is much easier to solve. I use two different discretizations of the real axis namely equidistantly distributed interpolation points along the real axis, labelled linear nodes:

$$z_k^{\text{lin}} = a + \frac{k - \frac{1}{2}}{N - 1}(b - a), \quad k = \{1, \ldots, N\}, \tag{16}$$

and Chebyshev nodes - the roots of the Chebyshev polynomials $T_N(x)$ rescaled for the interval $[a, b]$:

$$x_k = \cos\left((k - \frac{1}{2})\frac{\pi}{N}\right), \quad k = \{1, \ldots, N\}, \tag{17}$$

$$z_k^{\text{cheby}} = \frac{x_k + 1}{2} \cdot (b - a) + a, \quad k = \{1, \ldots, N\}. \tag{18}$$

Note, that the Chebyshev nodes correspond exactly to the real values of the discretization of the ellipse using the uniformly distributed parameters φ_k^{uni} with twice the number of integration points N, see Eq. (8) and Eq. (9). This way, by using a uniformly distributed parameter discretization with $2N$ integration points, the filter function of CIM converges for $e \to 1$ to the filter function of RIM based on Chebyshev nodes (compare both middle panels in Fig. 4).
The barycentric weights of the Chebyshev nodes for the interval $[-1, 1]$ are given by

$$w_k = \frac{1}{N} T_{N-1}(z_k^{\text{cheby}}) = \frac{1}{N} \cos\left((k - \frac{1}{2})\frac{(N-1)\pi}{N}\right).$$

1.7 Analysis of the resulting filter functions

The question remains which interpolation points yield the best filter function. There are two quality criteria which are relevant for a filter function, namely:

1. Homogeneity (homogeneous amplification) of the filter function in the search interval $[a, b]$
2. Amplification of the eigenspace inside the search interval in relation to the suppression of eigenvectors outside the search interval. This amplification ratio is visible at the steep descent of the filter function at the boundaries of the search interval

Four different discretizations for interpolation points to gain filter functions of type Eq. (14) are examined in Fig. 4, namely

a) equidistantly distributed along the real axis (linear RIM),

b) roots of the Chebyshev polynom as interpolation points on the real axis (Chebyshev RIM),

c) equidistantly distributed with respect to the arc length on an elliptic contour for different eccentricities [φ_k^{arc} of Eq. (7) used with Eq. (9)],

d) uniformly distributed with respect to the parameter φ on an elliptic contour for different eccentricities [φ_k^{uni} of Eq. (8) used with Eq. (9)],

for an interval $[-2, 2]$ and compared to the filter functions retrieved from the discretized CIM. Complex eccentricities are those ellipses whose minor axis is larger than their major axis. The upper panels of Fig. 7 show the discretization d) for the discussed eccentricities.
It turns out that using the filter functions obtained from discretization type d) (integration points z_k^{uni}) using CIM and RIM are equivalent up to a constant factor c.
The different choices of discretization lead to filter functions with different homogeneity and amplification ratio. Whereas discretization b) and d) yield a homogeneous filter behavior in the desired search interval, a) and c) tend to amplify the center of the interval more strongly. The amplification ratio seem to become better the closer the contour is next to the real axis ($e \approx 1$).

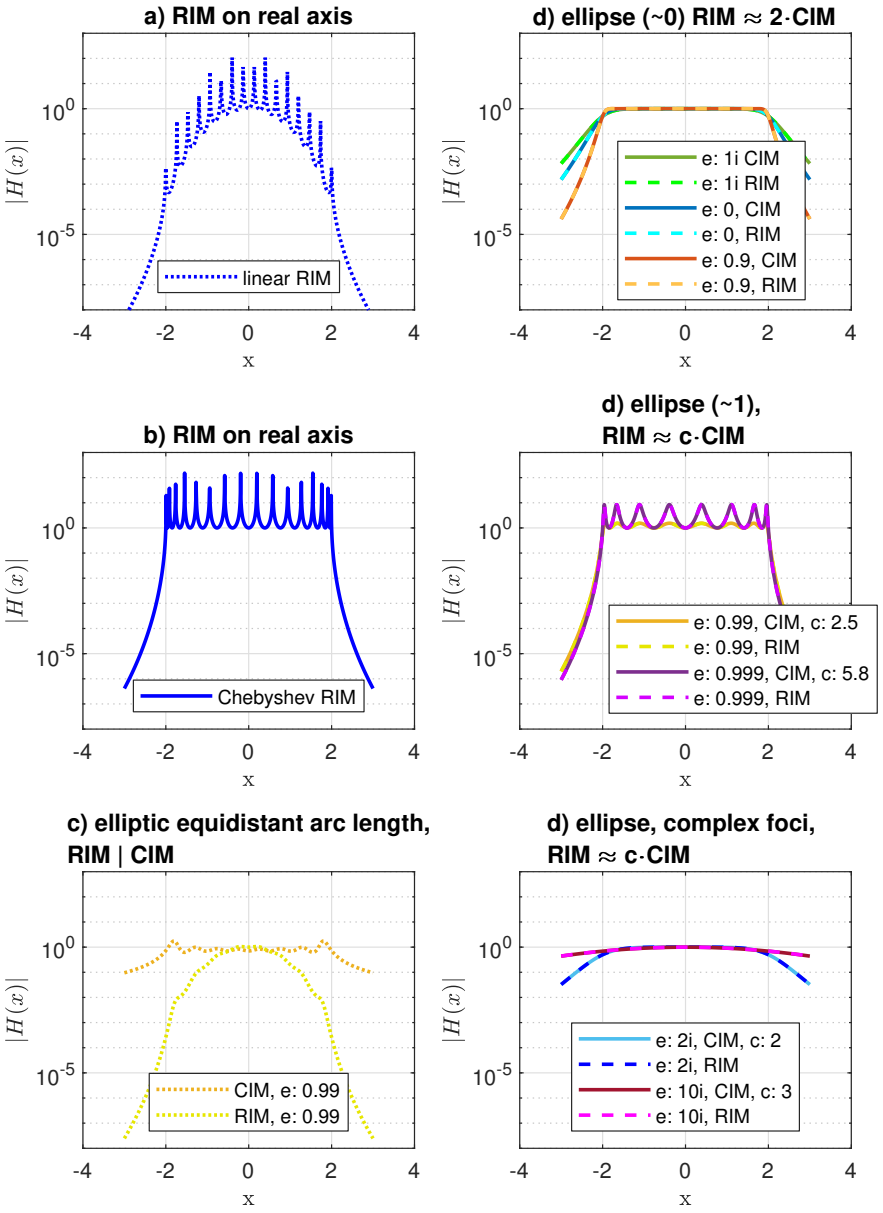


Figure 4: Comparison of the filter functions obtained from RIM and CIM for different choices of the discretization of the search domain containing the search interval $[-2, 2]$ using $N = 16$ discretization points z_k. The absolute value of the filter functions $|H(x)|$ is plotted on a logarithmic scale to visualize the decay behavior and the homogeneity of the filter functions.
Depending on the method used the weights w_k of the filter function are given either by the barycentric weights (RIM) or the integration weights (CIM). The right panels show, that the filter functions obtained from CIM and RIM using a uniformly distributed parameter discretization d) are approximately the same up to a factor of c.

1.8 Configuration steps for the developed eigenvalue solver RESARARI using Rayleigh–Ritz projection technique in combination with RIM and CIM

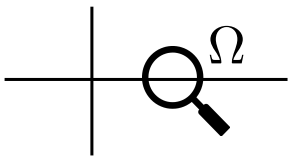
This subsection shall illustrate and visualize the configuration steps of the developed eigensolver **RESARARI** (for **RE**solvent **SA**mpling **RA**yleigh– **RI**tz)

1. Choice of the domain of interest (e.g. interval $[a, b]$)
2. Choice of the method used to approximate the eigenspace: RIM or CIM
3. Choice of the number N of interpolation points / integration points z_k, the number M of used moments and the number L of random vectors y_i for the resolvent sampling
4. Choice of the actual contour (eccentricity) and discretization scheme (linear, Chebyshev nodes resp. equidistant arc lengths, uniformly distributed parameters)
5. Choice of the parameters to solve the linear systems: solver, preconditioner, start vector, number of iteration, symmetrized version, accuracy (relative linear residual)
6. Choice of the tolerance of the singular values in the Rayleigh–Ritz projection technique

See also Fig. 5

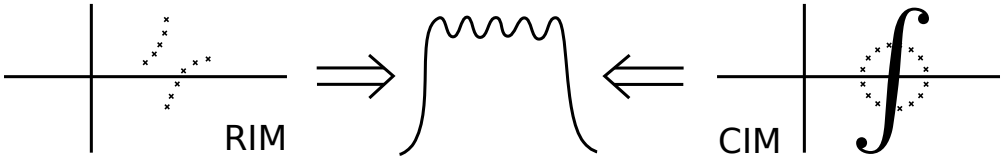
Resolvent sampling Rayleigh-Ritz

1) Search domain



2) Method

Rational filter function



Rational interpolation method — Contour integral method

3) Number of

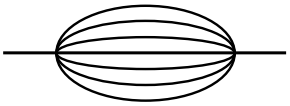
Integration points	Moments	Random vectors for resolvent sampling
N	M	L

4) Discretization

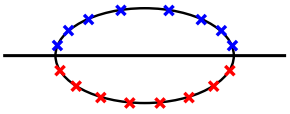
linear

chebyshev

eccentricity of ellipse



parametrization

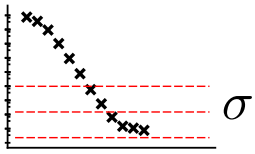


5) Linear System

CHOOSE:
- solver
- preconditioner
- start vector

- max iterations
- symmetrized CIM
- accuracy of linear system solution

6) SVD Tolerance



Result: $Av_i \approx \lambda_i v_i, \quad \lambda_i \in \Omega$

Figure 5: Configuration steps of the resolvent sampling Rayleigh–Ritz solver available on GITHUB

2 Influence of the projector approximation method on the condition number of the linear systems

2.1 Fixed configuration parameters of all numerical experiments

Throughout this and the next section the same sparse matrix A stemming from the physical problem introduced in Sec. 1.1 is analyzed that has a dimension of $n = 4356$ and a condition number of $\kappa(A) = 1.88$. The full spectrum can be evaluated numerically within Matlab and is plotted in Fig. 6.
The following configuration parameters of RESARARI (see Fig. 5) are fixed throughout all calculations:

(1) The domain of interest is chosen to be $\Omega = [121.0, 121.2]$ which comprises five eigenvalues (marked in green in Fig. 6).

(3) The number of discretization points and random vectors is fixed with $N = M = 16$ and the number of used moments is $M = 4$.

(5) No preconditioner is used throughout all calculations. The start vector $x_{i1}^{(0)}$ is zero for the first linear system and then uses the solution of the last linear system $x_{ik}^{(0)} = x_{ik-1}^{m*}$ to speed up convergence. The maximum number of iterations is fixed with maxit $= 8000$.

(6) The tolerance of the singular value decomposition is chosen dynamically from the set $\sigma \in \{10^k, k \in \{-17, \ldots, -10\}\}$ in such a way that the sum of eigenvalue residuals is minimized.

In the next subsection the influence of the chosen method (2) (and contour form (4)) on the condition number κ of the arising linear systems is examined.

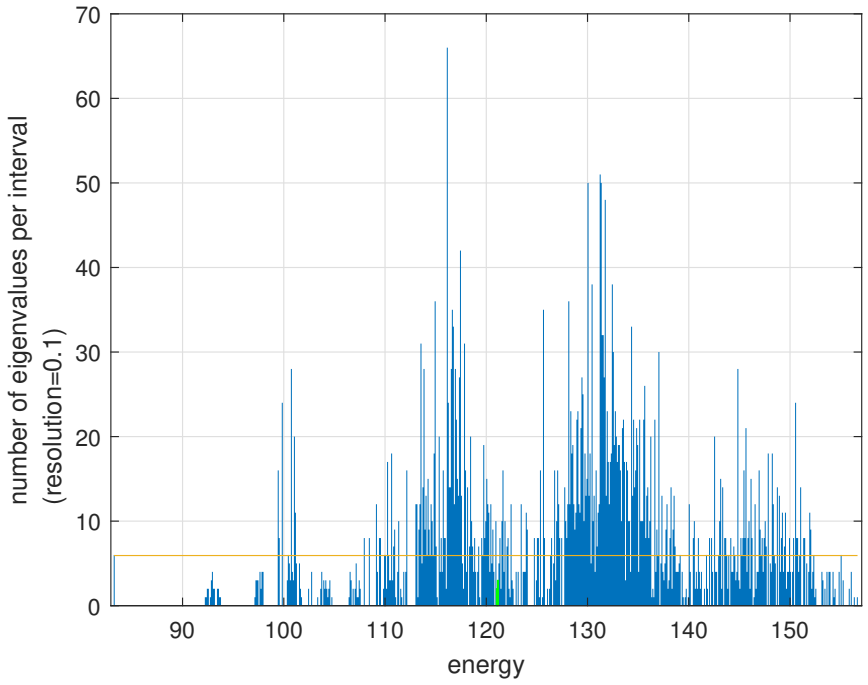


Figure 6: Spectral density of the test matrix A. The domain of interest $\Omega = [121.0, 121.2]$ used in the numerical tests is depicted in green and comprises five eigenvalues.

2.2 Condition number of linear systems arising in CIM for different shapes of the contour

Condition number: The condition number κ of the linear system provides in many cases an estimate for the approximate number of iterations needed to reach a certain accuracy when solving linear systems with an iterative solver. The error $e^{(m)} = x^{(m)} - x^*$ of an iterative method after m iterations is given by the deviation of the iterative solution $x^{(m)}$ from the exact solution x^* of the linear system $Ax^* = y$. For the conjugate gradient method there exists the following convergence relation

$$\|e^{(m)}\|_A \leq 2 \left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \right)^m \|e^{(0)}\|_A, \tag{19}$$

with the condition number $\kappa(A)$ and the A-norm $\|e\|_A = \|A^{1/2}e\|_2$. This relation will be used in order to estimate the number of iterations needed to reach a certain accuracy when solving the linear systems with a positive definite and Hermitian system matrix.

Numerical evaluation of the condition number: The condition number is evaluated for different shapes of the ellipse embracing the search interval. The discretization of the ellipses is given by Eq. (9) using uniformly distributed parameters φ_k^{uni} according to Eq. (8).
The $N = 16$ integration points z_k for the different eccentricities and the resulting condition numbers are depicted in the lower left panel of Fig. 7.

Estimation of the condition number using Gershgorin discs for different eccentricities: For normal matrices A the condition number $\kappa(z_k\mathbb{1} - A)$ can be estimated using Gershgorin discs. The maximal eigenvalue of $z_k\mathbb{1} - A$ in term of modulus can be estimated by

$$|\lambda_{\max}| \leq \max_\iota |z_k - A_{\iota\iota}| + r_\iota, \quad r_\iota := \max \left(\sum_{\iota \neq j} |A_{\iota j}|, \sum_{\iota \neq j} |A_{j\iota}| \right). \tag{20}$$

Due to the shift of the spectrum into the complex plane by the imaginary part of the integration point z_k, the minimal eigenvalue of $z_k\mathbb{1} - A$ can be bounded from below by

$$|\lambda_{\min}| \geq |\text{Im}\{z_k\}|.$$

Thus we gain the estimated condition number:

$$\kappa(z_k\mathbb{1} - A) \leq \frac{\max_\iota |z_k - A_{\iota\iota}| + r_\iota}{|\text{Im}\{z_k\}|}. \tag{21}$$

2.3 Symmetrization of contour integral method (SCIM)

In this subsection the symmetrized contour integral method (SCIM) is derived in order to get Hermitian, positive definite linear systems for which also the conjugate gradient (CG) method is applicable.
Using the symmetry properties of the chosen discretization [Eq. (9)], $\cos(\varphi_k) = \cos(\varphi_{N+1-k})$ and $\sin(\varphi_k) = -\sin(\varphi_{N+1-k})$, we have

$$\begin{aligned} z_{N+1-k} &= \overline{z_k}, \\ z'_{N+1-k} &= -\overline{z'_k}, \\ \tilde{\omega}_{N+1-k} &= \overline{\tilde{\omega}_k}, \\ \omega_{N+1-k} &= \overline{\omega_k}, \end{aligned}$$

for $k = 1, \ldots, N/2$.
The sum [Eq. (13)] thus can be split into

$$S = \sum_{k=1}^{N/2} \left[\omega_k z_k^j \left(z_k \mathbb{1} - A\right)^{-1} - \overline{\omega_k z_k}^j \left(\overline{z_k}\mathbb{1} - A\right)^{-1} \right] Y =$$
$$= \sum_{k=1}^{N/2} \underbrace{\left[\omega_k z_k^j (\overline{z_k}\mathbb{1} - A) - \overline{\omega_k z_k}^j (z_k \mathbb{1} - A) \right]}_{=:\omega_{kj}} \left[\underbrace{(z_k \mathbb{1} - A)(\overline{z_k}\mathbb{1} - A)}_{=:\tilde{A}_k} \right]^{-1} Y. \tag{22}$$

By decomposing the discretization points $z_k = z_k^{(\mathrm{r})} + \mathrm{i} z_k^{(\mathrm{i})}$, the matrix can be written as:

$$\tilde{A}_k = (z_k^{(\mathrm{r})}\mathbb{1} - A + z_k^{(\mathrm{i})})(z_k^{(\mathrm{r})}\mathbb{1} - A - z_k^{(\mathrm{i})}) = (z_k^{(\mathrm{r})}\mathbb{1} - A)^2 + {z_k^{(\mathrm{i})}}^2 \mathbb{1}. \tag{23}$$

Using the hermiticity of the matrix $A^* = A$, the resulting linear systems to solve are Hermitian since

$$\tilde{A}_k^* = \left[{z_k^{(\mathrm{r})}}^2 \mathbb{1} - 2 z_k^{(\mathrm{r})} A + A^2 + {z_k^{(\mathrm{i})}}^2 \mathbb{1} \right]^* = \tilde{A}_k,$$

and also positive definite since the non-zero imaginary part $z_k^{(\mathrm{i})} \neq 0$ always gives a positive contribution to the spectrum in Eq. (23).
Thus the conjugate gradient algorithm can be applied which is rather robust in contrast to the algorithms apt for non-Hermitian problems.
The condition number of the resulting linear problem may become large since $\tilde{A}_k$ comprises the square of A but $\kappa(\tilde{A}_k)$ can be tuned by the choice of the semi-minor axis r.
One can estimate the condition number of $\tilde{A}_k$ using Gershgorin discs [see Eq. (20)]:

$$\kappa(\tilde{A}_k) \leq \frac{\left(\max_\iota |z_k^{(\mathrm{r})} - A_{\iota\iota}| + r_\iota \right)^2 + {z_k^{(\mathrm{i})}}^2}{{z_k^{(\mathrm{i})}}^2}, \tag{24}$$

If the matrix A is sparse the computation of $\tilde{A}_k$ would require a rather large storage. When applying iterative algorithms this computation can be avoided by splitting the multiplication with $\tilde{A}_k$ in two multiplications with $(\mathrm{Re}\{z_k\}\mathbb{1} - A)$.

Real and symmetric: In the case of having a real symmetric matrix A and real right hand sides y_i, the discretization points and the solution vector can be decomposed in real and imaginary parts $z_k = z_k^{(\mathrm{r})} + \mathrm{i} z_k^{(\mathrm{i})}$, $x_{ki} = x_{ki}^{(\mathrm{r})} + \mathrm{i} x_{ki}^{(\mathrm{i})}$ so that the linear problem $(z_k \mathbb{1} - A) x_{ki} = y_i$ reduces to a real valued problem:

$$\left(\frac{1}{z_k^{(\mathrm{i})}} (z_k^{(\mathrm{r})}\mathbb{1} - A)^2 + z_k^{(\mathrm{i})} \mathbb{1} \right) x_{ki}^{(\mathrm{i})} = -y_i, \tag{25}$$

$$x_{ki}^{(\mathrm{r})} = -\frac{1}{z_k^{(\mathrm{i})}} (z_k^{(\mathrm{r})}\mathbb{1} - A) x_{ki}^{(\mathrm{i})}. \tag{26}$$

Resulting condition numbers of SCIM and RIM The lower right panel of Fig. 7 depicts the condition numbers of the resulting linear systems using the symmetrized contour integral method for different eccentricities of the ellipses as well as the condition numbers of the rational interpolation method that uses the roots of Chebyshev polynomials as interpolation points on the real axis. A

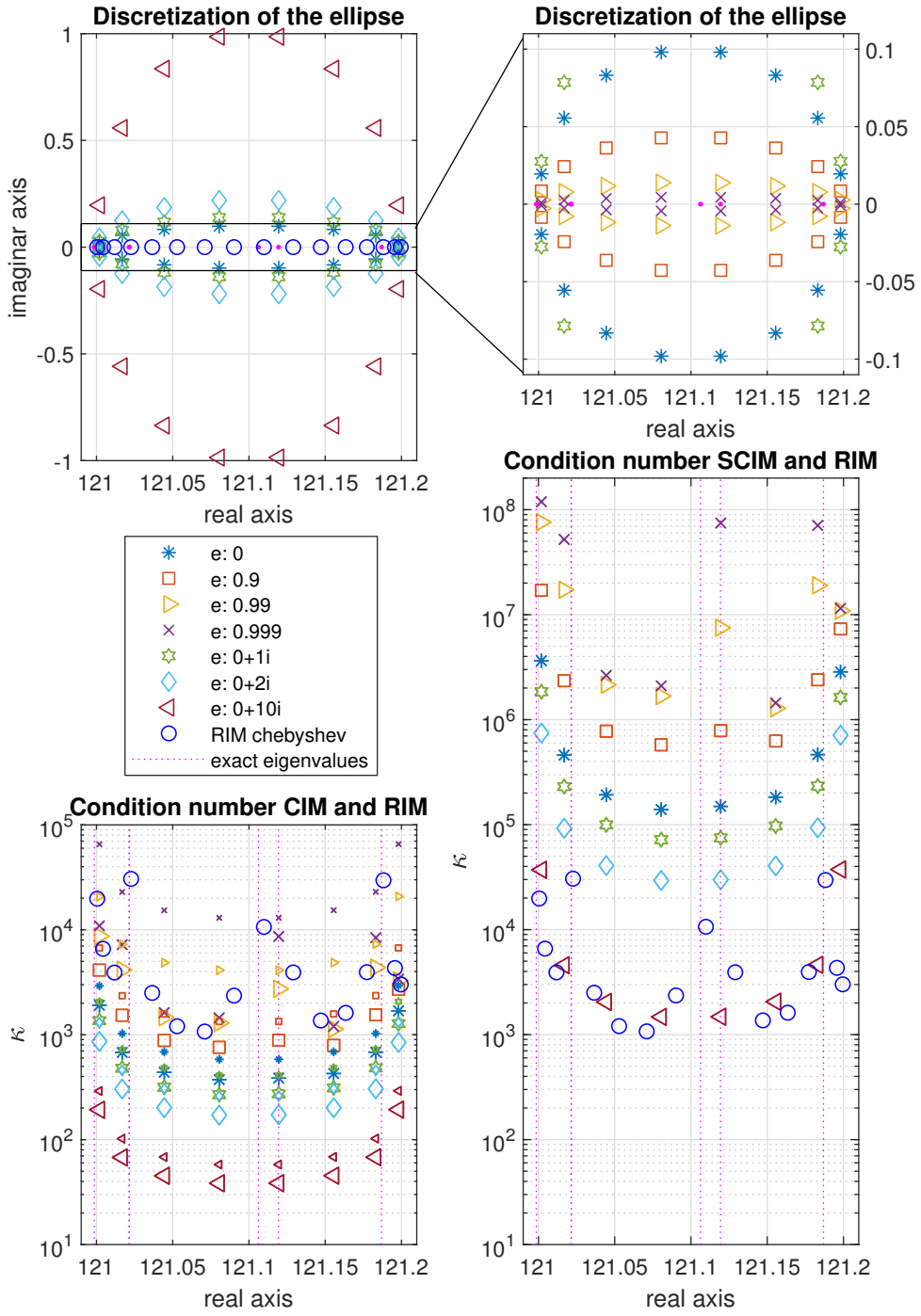


Figure 7: Upper panels: Discretization points z_k of ellipses with different eccentricities e. Lower panels: Condition number $\kappa(\tilde{A}_k)$ of the linear systems using CIM (left panel) and SCIM (right panel) with a uniformly distributed parameter discretization z_k^{uni} and condition number of RIM (both panels) method using Chebyshev interpolation points depicted as a function of the real value of the according sampling point z_k. The small symbols in the left lower panels are the estimated condition numbers of CIM. The exact eigenvalues that influence the exact condition number are depicted with dotted magenta lines.

2.4 Influence of the chosen eccentricity of the ellipses on the eigenvalue residual using a direct solver

This subsection relates the discussed filter function qualities to the final residuals of the eigenproblem. The occurring linear systems for the three methods CIM, symmetric CIM and RIM are solved using Matlab direct solvers (reached relative linear residual in the order of 10^{-13}) and the eigenvalue residuals are plotted in Fig. 8.

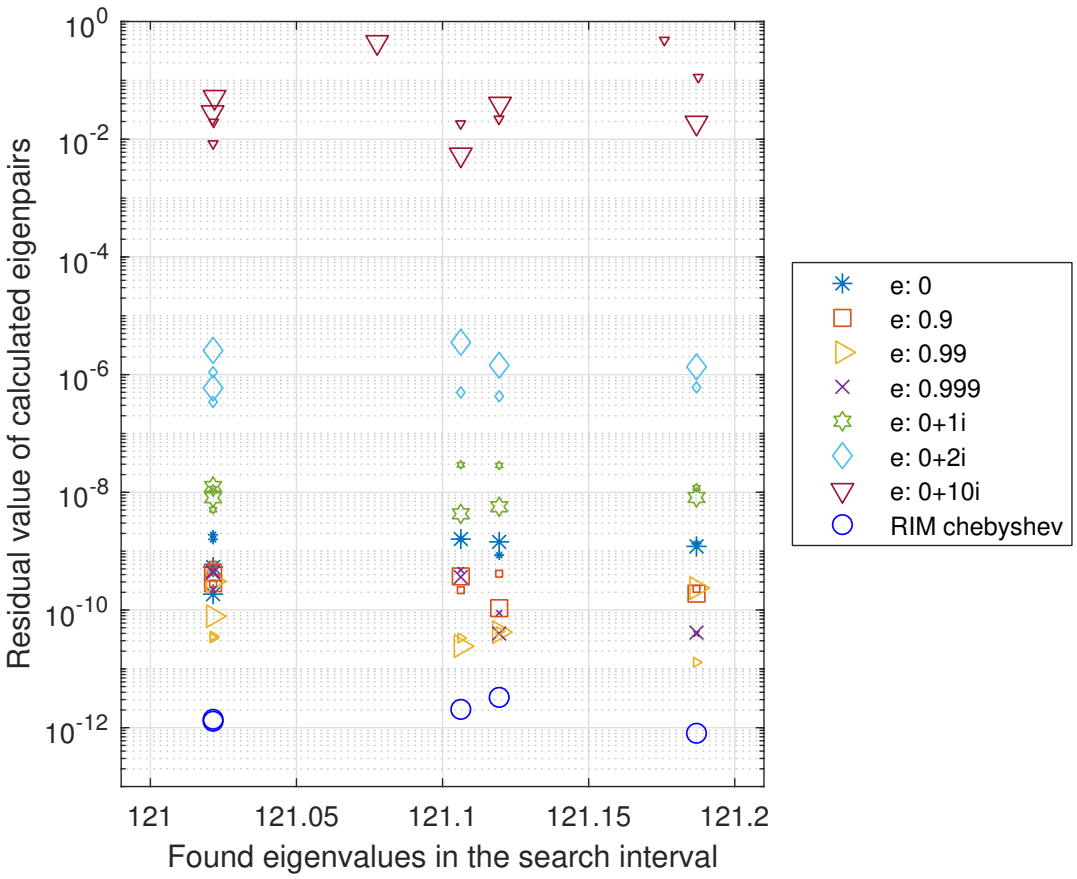


Figure 8: Residuals of the eigenpairs using a direct solver (MATLAB) in combination with the three approaches RIM using Chebyshev nodes (circles), CIM (other large symbols) and SCIM (small symbols) for different choices of the contour indicated by the eccentricity e.

It turns out that the eigensolver yields bad results if the semi-minor axis increases too much. This corresponds to the observation in Sec. 1.7 that the filter functions for discretization points further away from the search domain suffer from the amplification ratio although the resulting linear systems have in general a lower condition number. So a compromise between condition number and optimal filter function (contour) has to be found which will be the topic of the next section where the performance of iterative solver and the influence of the accuracy of the solution of the arising linear systems on the eigenvalue residual is benchmarked.

3 Influence of the accuracy of the iteratively solved linear systems on the residuals of the eigenproblem

There is a variety of iterative solver available using Krylov space and projection techniques. In this project I will restrict the numerical tests to the following iterative solvers which have been chosen according to the recommendations available on the Matlab documentation on iterative solver [1]:

1. GMRES(20) for CIM,
2. BiCGSTAB for CIM,
3. MINRES for RIM,
4. SYMMLQ for RIM,
5. PCG for SCIM.

The iterative solvers were used with a maximum number of iterations of `maxit` = 8000. In order to qualify the solution $x^{(m)}$ at iteration m and make a decision on convergence the relative residual r^{lin} of the solution $x^{(m)}$ of the linear system is used:

$$r^{\text{lin}}(x^{(m)}) = \frac{\|Ax^{(m)} - y\|}{\|y\|},$$

which serves as measure of accuracy of the solution of the linear system.
The numerical experiments were performed using the before mentioned iterative solvers in combination with varying the following parameters:

1. Requested accuracy r^{lin} of the linear systems,
2. Eccentricity e of the ellipse (for CIM and SCIM).

The following criteria are used to examine the performance of the iterative solver in the context of the eigensolver:

1. Percentage of not converged linear systems,
2. Mean number of iterations needed,
3. Total time for the calculation,
4. Mean reached relative residual r^{lin} of the solution $x^{(m)}$ of the linear systems,
5. Number of found eigenvalues in the domain (should be five),
6. Mean eigenvalue residual r^{eig} of the five best solutions,

and are depicted in Fig. 9 and Fig. 10 for the before mentioned numerical experiments. The color scheme is identical for all plots, with blue indicating good results and yellow indicating poor results.

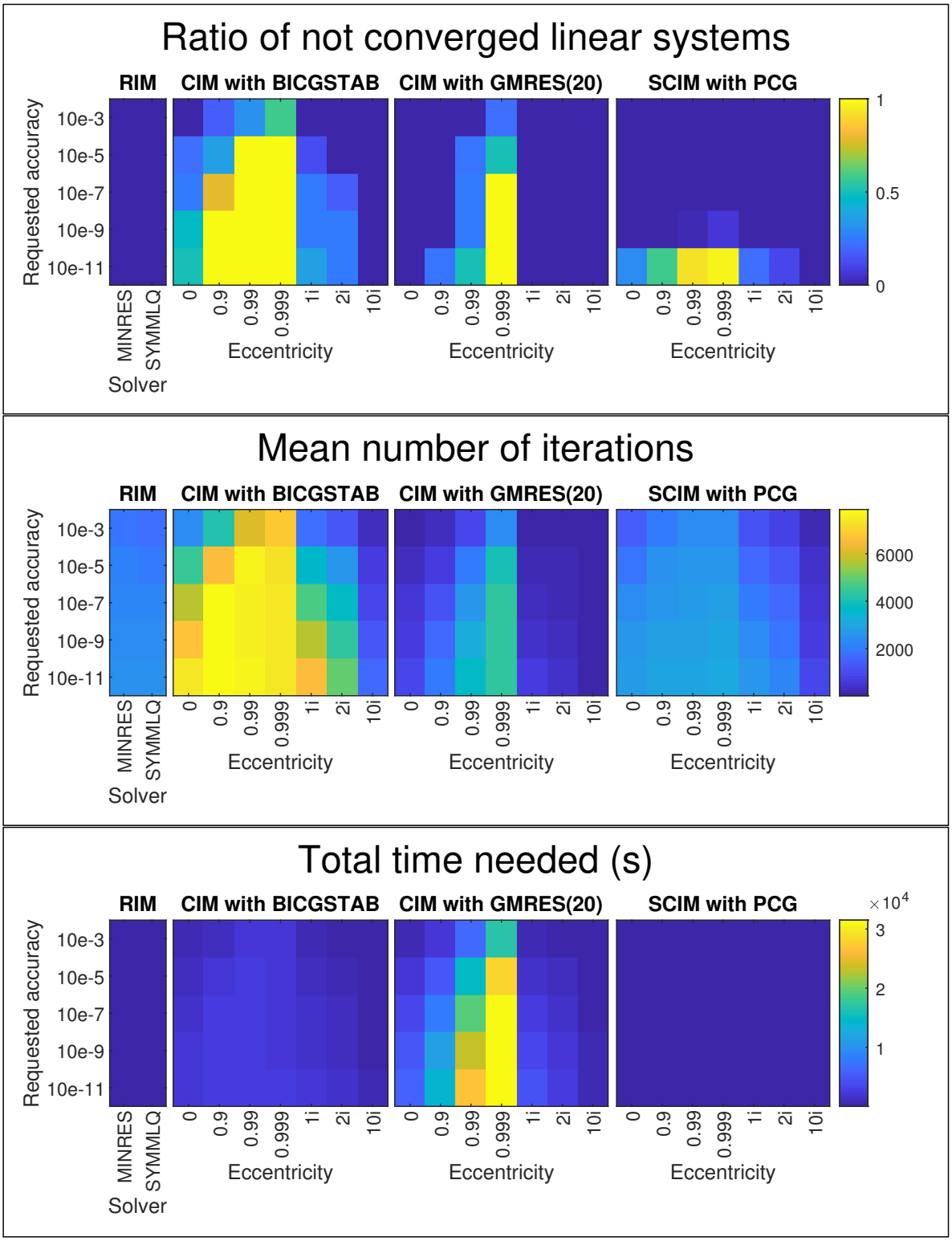


Figure 9: Results from the numerical experiments using a variation of methods (RIM, CIM, SCIM), linear solvers (MINRES; SYMMLQ, BICGSTAB, GMRES; PCG), requested accuracy r^{lin} and used eccentricity e of the ellipse. The color codes are the same for all plots with blue indicating good and yellow poor results. The panels show from top to bottom the ratio of converged linear systems, the mean number of iterations of the iterative solver and the total time needed for solving the linear systems.

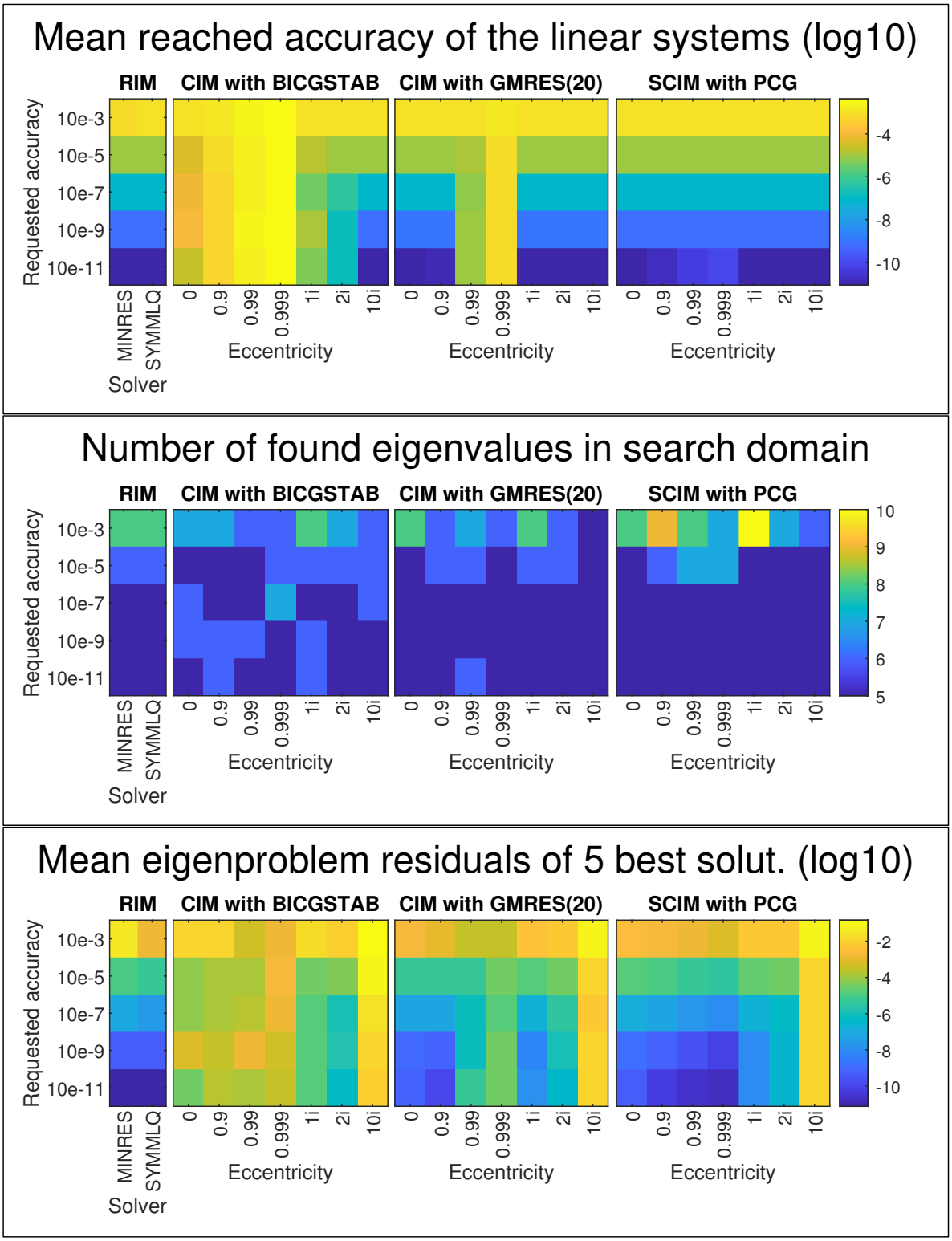


Figure 10: Results from the numerical experiments using a variation of methods (RIM, CIM, SCIM), linear solvers (MINRES; SYMMLQ, BICGSTAB, GMRES; PCG), requested accuracy r^{lin} and used eccentricity e of the ellipse. The color codes are the same for all plots with blue indicating good and yellow poor results. The panels show from top to bottom mean reached relative residual of the linear systems r^{lin} (accuracy), the number of found eigenvalues and mean eigenproblem residual r^{eig} of the five best solutions.

The numerical experiments show that the non-Hermitian linear systems arising in the CIM method and treated with GMRES and especially BICGSTAB solver did hardly converge. Interestingly GMRES did not converge for those linear systems which provide the best filter functions, namely those with a eccentricity close to one. The symmetrized contour integral method did not perform well for large semi–minor axis which can be explained via the amplification ratio of the corresponding filter functions.
The best (smallest) eigenvalue residuals using iterative solvers could be reached with RIM (MINRES and SYMMLQ) and SCIM (PCG) using small semi–minor axes. Although the condition number arising in the corresponding linear systems was quite high, the non–Hermitian linear systems arising in CIM with lower condition numbers suffer from convergence issues.

4 Conclusion and outlook

4.1 Findings from the comparison of RIM and CIM using filter functions

- The discretized contour integral method and the rational interpolation method to approximate an eigenspace projector can be analyzed and compared with filter functions.
- The absolute value of the filter functions $H(z)$ derived from RIM using Chebyshev nodes (barycentric weights w_k) and CIM using a uniformly distributed parameter discretization (summation weights ω_k) are identical up to a constant factor c, which is $c \approx 2$ for eccentricities $e \approx 0$.
- Chebyshev nodes (RIM) and uniformly distributed parameter discretization points (CIM) have the same real parts and coincide for an eccentricity going to one. Both show a rather homogeneous filter function behavior.
- Linear nodes (RIM) and equidistant arc length discretization points (CIM) lead to inhomogeneous filter functions that amplify mostly the center of the search domain.

4.2 Findings and observations from the analysis of the condition number of the resulting linear systems using different methods and shapes of the contour

- The condition number κ for linear systems resulting from using CIM or SCIM can be estimated, see Eq. (21) and Eq. (24).
- The condition number becomes worse when using eccentricities close to one (small semi-minor axis) and rather small for large complex eccentricities (large semi-minor axis).
- The symmetrized contour integral method (SCIM) yields positive definite linear systems which have a quadratically worse condition number than linear systems obtained from using CIM.
- The condition number of RIM is roughly in between those resulting from CIM and SCIM.

4.3 Findings from the anaylsis of the contour shape on the eigenvalue residuals using a direct solver

- The larger the semi–minor axis within CIM and SCIM the worse the reached accuracy of eigenvalues. This can be understood from the shape of the filter functions.
- CIM and SCIM yield approximately the same reached accuracy of eigenvalues (eigenvalue residual r^{eig}) depending on the eccentricity e when using a direct solver.
- SCIM has the advantage of halving the number of linear systems to solve compared to CIM.

4.4 Findings from the comparison of different iterative solver varying the required accuracy of the linear systems

- BICGSTAB and GMRES show convergence issues.
- RIM in combination with SYMMLQ or MINRES shows the best results in terms of reached final eigenvalue residuals and total time needed.
- SCIM in combination with PCG can achieve results compareable with RIM provided the used contour has a small semi–minor axis (eccentricity close to one).
- CIM in combination with GMRES can be used to obtain moderate results provided that the semi–minor axis is not too small otherwise convergence issues arise.
- All-in-all it is best to deal with symmetric linear systems when dealing with iterative solvers even on the cost of an increasing condition number.

4.5 Outlook

Krylov space recycling: The most involved numerical task in the discussed contour and rational interpolation methods is the solution of several shifted linear systems. In order to minimize the number of linear systems to be solved, several methods use recycling of Krylov subspaces which are invariant under linear shifts. For a comprehensive discussion of those methods see [7]. Methods like SQMRCGSTAB could make it possible to solve only L linear systems which would make this approach comparable to a Lanczos method.

Preconditioning: Preconditioning was not a topic of this project, but can help improve on the convergence issues arising in the non–Hermitian linear systems

Randomizing the first solution: In order to save the number of linear systems to be solved, one can obtain the right hand sides y_i by choosing the first solution x_{1i} randomly set $y_i := (z_1\mathbb{1} - A)x_{1i}$.

Successive improvements of the eigensolution: Since the here presented method has a lot of parameters to tune, it would be interesting to define a protocol that improves on the solution while minimizing the necessary matrix vector multiplications. The measures available in improving an eigensolution are

- Adding new sampling points z_k,
- Adding new moments z_k^j,
- Adding new right hand sides y_i,
- Iterate on the solution of linear systems.

References

[1] Matlab documentation on iterative solvers. `https://de.mathworks.com/help/matlab/math/iterative-methods-for-linear-systems.html`. Accessed: 2012-07-22.

[2] A. P. Austin and L. N. Trefethen. Computing eigenvalues of real symmetric matrices with rational filters in real arithmetic. *SIAM Journal on Scientific Computing*, 37(3):A1365–A1387, 2015.

[3] M. Balzer and M. Potthoff. Nonequilibrium cluster perturbation theory. *Phys. Rev. B*, 83:195132, May 2011.

[4] W.-J. Beyn. An integral method for solving nonlinear eigenvalue problems. *Linear Algebra and Its Applications*, 436(10):3839–3863, 2012.

[5] A. Drozdov, M. Eremets, I. Troyan, V. Ksenofontov, and S. Shylin. Conventional superconductivity at 203 kelvin at high pressures in the sulfur hydride system. *Nature*, 525(7567):73, 2015.

[6] V. Emery. Theory of high-t c superconductivity in oxides. *Physical Review Letters*, 58(26):2794, 1987.

[7] J. Meng, P.-Y. Zhu, and H.-B. Li. Qmrcgstab algorithm for families of shifted linear systems. In *2013 Ninth International Conference on Computational Intelligence and Security*, pages 272–276. IEEE, 2013.

[8] A. Pichler. Numerical methods for eigenvalue problems based on the approximation of the poles of the resolvent. Master's thesis, Graz University of Technology, 2016.