Alexander Pichler, BSc

Numerical methods for eigenvalue problems based on the approximation of the poles of the resolvent

MASTERARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

Masterstudium Technomathematik

eingereicht an der

Technischen Universität Graz

Betreuer

Dr. G. Unger

Institut für Numerische Mathematik

Graz, Juli 2016

Acknowledgement

First of all, I would like to thank Dr. Gerhard Unger for introducing me to this interesting topic, for always taking time whenever I ran into trouble or had some questions, and for his excellent supervision of this master thesis in general.

Furthermore, I would like to express my deepest gratitude to my parents, Adelheid and Gerhard Pichler, for their constant support and their encouragement throughout my years of study.

Last but not least, I would like to thank everybody else who supported me during my studies.

Contents

1 Introduction

Many applications in science and engineering lead to nonlinear eigenvalue problems (NEPs). To illustrate this, a collection of some scientifically and practically relevant NEPs is given in [6]. Unfortunately, NEPs have to be treated completely differently in comparison to linear eigenvalues problems since several troubles which are not relevant for linear eigenvalue problems can occur. For instance, it is possible that there are more eigenvalues than the dimension of the problem and that eigenvectors to different eigenvalues are linearly dependent. In particular, the solution of large-scale NEPs, which occur in FEM and BEM applications quite often, is a challenge [20]. In this master thesis we restrict to a certain class of NEPs of the form

$$T(\lambda)v = 0,$$

where $T : \mathcal{D} \to \mathbb{C}^{n\times n}$ is holomorphic in some open domain $\mathcal{D} \subset \mathbb{C}$. Our goal is to compute all eigenvalues (and corresponding eigenvectors) lying within a given contour $\Gamma_{\mathcal{C}} \subset \mathcal{D}$. We will focus on algorithms which make use of the approximation of the poles of the resolvent of T. In literature, there are at least two major ways to do this, one which uses contour integrals [1, 7] and another one based on rational interpolation [32]. Since the presented algorithms in [1] and in [7] are quite similar, we will follow [7] and [32] to derive the corresponding methods, which we will call *contour integral method (CIM)* and *rational interpolation method (RIM)* in this thesis. These numerical methods are very suitable to use in practice since, apart from the holomorphy, there are no other requirements on T such as symmetry, structure, and whether the involved matrices are dense or sparse.

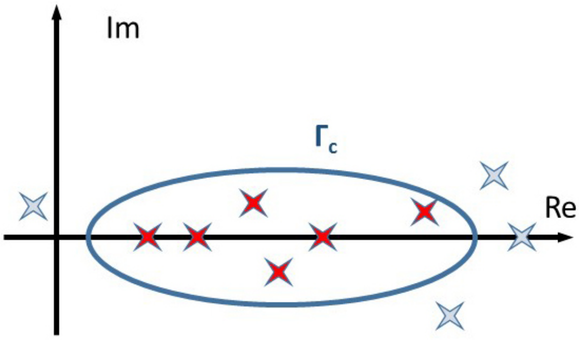


Figure 1.1: An image which shows the contour $\Gamma_{\mathcal{C}}$, the searched eigenvalues inside the contour (red stars) and some eigenvalues outside the contour (light-blue stars).

The thesis is organized in the following way: Roughly speaking, it consists of two parts. The first one involves the Chapters 2-5 and has the aim to introduce the numerical methods mentioned above, and the second one is built by the Chapters 6-8 and has the intention to present boundary integral formulations for acoustic eigenvalue problems and to compare the depicted algorithms for the computation of eigenvalues.

The first part is organized as follows: In Chapter 2, some important definitions related to NEPs are given and the theorem about the representation of the resolvent of T is stated. This theorem shows that the eigenvalues can be characterized as poles of the resolvent and provides the base for the algorithms CIM and RIM. Therefore, these methods can be seen as numerical methods approximating the poles of the resolvent T^{-1}. Our main reference for this chapter is [7]. In Chapter 3, the CIM is derived. This method approximates the poles by using contour integrals. In doing so, the NEP is transformed into a linear matrix eigenvalue problem from which the eigenvalues and eigenvectors can be calculated easily. In Chapter 4, the RIM is derived. This method uses rational interpolation instead of contour integrals. Although the RIM and the CIM do not seem to have much in common at first glance, the RIM can in fact be seen as a generalization of the CIM where the interpolation points are chosen on the contour. In Chapter 5, the key steps of different variants of the *Rayleigh-Ritz procedure* in the context of the CIM and the RIM are summarized [31, 32, 33]. This procedure provides a possibility to transform NEPs with large dimensions, which occur in practice very often, into NEPs with smaller dimensions. The smaller problems can then be solved with comparatively little cost by using the CIM or the RIM.

The second part has the following structure: In Chapter 6, formulations of time-harmonic acoustic eigenvalue problems with Dirichlet and Neumann boundary conditions are given first. Then the boundary integral operators are introduced and boundary integral formulations are derived. Our main reference here is [28]. In Chapter 7, an error estimate for the calculated eigenvalues and eigenvectors of the boundary integral formulations is stated. We will see that we can get a convergence order with respect to the mesh size of up to three for the eigenvalues and up to 3/2 for the eigenvectors in the case, where lowest order Galerkin discretizations are chosen [27, 30]. Finally, in Chapter 8, the different algorithms presented in the Chapters 3-5 are compared in several numerical experiments for interior and exterior eigenvalue problems.

2 Basics of nonlinear eigenvalue problems

In this chapter we would like to introduce the basics of nonlinear eigenvalues problems (NEPs) of the form

$$T(\lambda)v = 0, \quad v \neq 0, \tag{2.1}$$

where $T \in \mathcal{H}(\mathcal{D}, \mathbb{C}^{n\times n})$ is holomorphic in some open domain $\mathcal{D} \subset \mathbb{C}$, i.e., all coefficients of T are holomorphic functions in $\mathcal{D}$. Moreover, we assume that T is regular, i.e., $\det(T)$ does not vanish identically in $\mathcal{D}$. We largely follow [7].

Definition 2.1. *We call $(\lambda, v) \in \mathcal{D} \times \mathbb{C}^n \backslash \{0\}$ an eigenpair of T if $T(\lambda)v = 0$. λ is called eigenvalue and v is the corresponding eigenvector.*

Similarly as for linear eigenvalue problems, one can introduce the geometric multiplicity of an eigenvalue. However, the definition of the algebraic multiplicity of an eigenvalue has to be adapted for NEPs. Therefore, we introduce the concept of root functions.

Definition 2.2. *We call a function $v \in \mathcal{H}(\mathcal{D}, \mathbb{C}^n)$ root function of T at $\lambda \in \mathcal{D}$ if the following condition is satisfied:*

$$T(\lambda)v(\lambda) = 0, \quad v(\lambda) \neq 0.$$

The order of zero of $T(z)v(z)$ at $z = \lambda$ is called multiplicity of v at λ and denoted by ν.

Due to the fact that v is holomorphic, it can be represented by a power series as follows [12, Chapter II, Theorem 4.1.]

$$v(z) = \sum_{j=0}^{\infty} (z-\lambda)^j v_j, \quad v_0 \neq 0.$$

This representation leads to the definition of chains of generalized eigenvectors.

Definition 2.3. *Let*

$$v(z) = \sum_{j=0}^{\infty} (z-\lambda)^j v_j$$

with $v_0 \neq 0$ be a root function of T at $\lambda \in \mathcal{D}$ with multiplicity ν, i.e.,

$$\begin{aligned}
&\frac{d^j}{dz^j}(T(z)v(z))|_{z=\lambda} = 0 \quad \forall j = 0, \dots, \nu - 1, \\
&\frac{d^\nu}{dz^\nu}(T(z)v(z))|_{z=\lambda} \neq 0.
\end{aligned}$$

Then any sequence of vectors

$$(v_0, \dots, v_{\mu-1}), \quad \mu \leqslant \nu,$$

is called a chain of generalized eigenvectors (CGE) or Jordan chain of length μ, v_0 is an eigenvector, and $v_1, \dots, v_{\mu-1}$ are associated vectors for v_0 of T at λ. The set $G_\lambda(T)$ spanned by all the elements of Jordan chains corresponding to λ is called generalized eigenspace of T at λ.

With these definitions in mind, we define the rank of an eigenvector of T and the canonical system of generalized eigenvectors, which finally leads us to a proper definition of the algebraic multiplicity of an eigenvalue of T.

Definition 2.4. *Let $v_0 \in \ker(T(\lambda)), v_0 \neq 0$, be an eigenvector of T at λ. Then*

$$r_\lambda(v_0) := \max\{\nu : v \text{ is a root function of } T \text{ at } \lambda \text{ with multiplicity } \nu \text{ and } v(\lambda) = v_0\}$$

is called the rank of v_0. The number

$$M_\lambda(T) := \max_{v_0 \in \ker T(\lambda) \setminus \{0\}} r_\lambda(v_0)$$

is called maximal length of Jordan chains corresponding to λ.

Note that the definition above is well-defined as $r(v_0)$ is finite [17, Lemma A.8.3].

Definition 2.5. *Let η be the dimension of $\ker(T(\lambda))$. The system of vectors*

$$\tilde{V} = \left(v_j^l, \quad 0 \leqslant j \leqslant m_l - 1, \ 1 \leqslant l \leqslant \eta\right)$$

is called canonical system of generalized eigenvectors (CSGEs) of T at λ if the conditions below are fulfilled:

(1) The vectors $v_0^1, \dots, v_0^\eta$ form a basis of $\ker(T(\lambda))$.

(2) The tuple $(v_0^l, \dots, v_{m_l-1}^l)$ is a CGE of T at λ for $l = 1, \dots, \eta$.

(3) $r_\lambda(v_0^1) = m_1 := \max\{r_\lambda(v_0) : v_0 \in \ker(T(\lambda))\}$.

(4) $r_\lambda(v_0^l) = m_l := \max\{r_\lambda(v_0) : v_0 \in \ker(T(\lambda)) \setminus \operatorname{span}\{v_0^1, \dots, v_0^{l-1}\}\}$ for $l = 2, \dots, \eta$.

The numbers $m_1 \geqslant \dots \geqslant m_\eta$ are called partial multiplicities.

Now the algebraic multiplicity of an eigenvalue of T can be defined as follows.

Definition 2.6. *The number*

$$m_\lambda(T) := \sum_{l=1}^{\eta} m_l$$

is called algebraic mulitplicity of T at λ and η is called geometric mulitplicity. We call the eigenvalue λ simple if its geometric and algebraic multiplicity are equal to one.

The following theorem about the representation of the resolvent $T(z)^{-1}$ [7, Corollary 2.8] is used as base in the construction of the numerical methods presented in [1, 7, 31, 32, 33] for solving NEPs of the form (2.1).

Theorem 2.7. *Let $\mathcal{C} \subset \mathcal{D}$ be a compact subset and let $T \in \mathcal{H}(\mathcal{D}, \mathbb{C}^{n\times n})$. Then $\mathcal{C}$ contains at most finitely many eigenvalues λ_k, $k = 1, \dots, n_\mathcal{C}$, with corresponding CSGEs of T*

$$V_k = \left(v_j^{l,k}, \quad 0 \leqslant j \leqslant m_{l,k} - 1,\ 1 \leqslant l \leqslant \eta_k\right).$$

Further, let the corresponding CSGEs on T^H be given by

$$W_k = \left(w_j^{l,k}, \quad 0 \leqslant j \leqslant m_{l,k} - 1,\ 1 \leqslant l \leqslant \eta_k\right)$$

such that

$$r_{\lambda_k}(w_0^{l,k}) = m_{l,k},$$

and the scaling condition

$$\sum_{\alpha=0}^{j}\sum_{\beta=0}^{m_{\nu,k}} (w_{j-\alpha}^{l,k})^H T_{\alpha+\beta,k} v_{m_{\nu,k}-\beta}^{\nu,k} = \delta_{\nu,l}\delta_{0,j}, \quad 0 \leqslant j \leqslant m_{l,k} - 1,\ 1 \leqslant l, \nu \leqslant \eta_k,$$

where $T_{j,k} = \frac{1}{j!}T^{(j)}(\lambda_k)$, is satisfied. Then there exist a set $\mathcal{C} \subset \mathcal{U} \subset \mathcal{D}$ and a holomorphic function $R \in \mathcal{H}(\mathcal{U}, \mathbb{C}^{n\times n})$ such that

$$T(z)^{-1} = \sum_{k=1}^{n_\mathcal{C}}\sum_{l=1}^{\eta_k}\sum_{j=1}^{m_{l,k}} (z - \lambda_k)^{-j} \sum_{\nu=0}^{m_{l,k}-j} v_\nu^{l,k} (w_{m_{l,k}-j-\nu}^{l,k})^H + R(z) \tag{2.2}$$

holds for all $z \in U \backslash \{\lambda_1, \dots, \lambda_{n_\mathcal{C}}\}$.

Note that the poles of the resolvent $T(z)^{-1}$ are exactly the eigenvalues of T. Therefore, our aim is to extract these poles properly. In the next two chapters we present two ways to approximate them. As already mentioned in the introduction, the first one makes use of contour integrals and the second one of rational interpolation.

3 Contour integral method (CIM)

In this chapter we would like to derive the CIM, which provides a way to approximate the poles of the resolvent $T(z)^{-1}$ by using contour integrals. In doing so, we largely follow [7]. Another algorithm which uses the contour integral approach is the so-called *block SS method*, though this method differs only slightly from the one proposed by Beyn [7]. For a detailed derivation and description of the block SS method we refer to [1]. However, in comparison to [7] where the case of multiple eigenvalues is treated too, in [1] it is assumed for the derivation of the algorithm that all eigenvalues are simple.

3.1 Main idea of the CIM

The central idea of the CIM is to apply the residue theorem [12, Chapter IV, Theorem 3.1.] to the representation of the resolvent (2.2) in Theorem 2.7. This gives the following result [7, Theorem 2.9].

Theorem 3.1. *Let $T \in \mathcal{H}(\mathcal{D}, \mathbb{C}^{n\times n})$ and let $\Gamma_{\mathcal{C}} \subset \mathcal{D}$ be a contour, i.e., a simple closed curve, such that there are no eigenvalues of T on $\Gamma_{\mathcal{C}}$. Let us denote the eigenvalues in the interior* $\operatorname{int}(\Gamma_{\mathcal{C}}) \subset \mathcal{D}$ *by λ_k, $k = 1, \dots, n_{\mathcal{C}}$. Then it holds for every $f \in \mathcal{H}(\mathcal{D}, \mathbb{C})$ with the CSGEs from Theorem 2.7*

$$\frac{1}{2\pi i}\int_{\Gamma_{\mathcal{C}}} f(z)T(z)^{-1}dz = \sum_{k=1}^{n_{\mathcal{C}}}\sum_{l=1}^{\eta_k}\sum_{j=1}^{m_{l,k}} \frac{f^{(j-1)}(\lambda_k)}{(j-1)!}\sum_{\nu=0}^{m_{l,k}-j} v_\nu^{l,k}(w_{m_{l,k}-j-\nu}^{l,k})^H.$$

If we further assume that all eigenvalues are simple, the above formula simplifies to

$$\frac{1}{2\pi i}\int_{\Gamma_{\mathcal{C}}} f(z)T(z)^{-1}dz = \sum_{k=1}^{n_{\mathcal{C}}} f(\lambda_k)v_k w_k^H,$$

where v_k and w_k are left and right eigenvectors corresponding to λ_k which are normalized according to

$$w_k^H T'(\lambda_k)v_k = 1, \quad k = 1, \dots, n_{\mathcal{C}}.$$

Note that the holomorphic term $R(z)$ in the representation of $T(z)^{-1}$ vanishes since by Cauchy's integral theorem [12, Chapter II, Theorem 2.2.] every contour integral for a holomorphic function is equal to zero.

3.2 Derivation of the algorithm

Let $\Gamma_{\mathcal{C}} \subset \mathcal{D}$ be a contour. In this section we summarize the main steps in the derivation of the algorithm for calculating all eigenvalues of $T \in \mathcal{H}(\mathcal{D}, \mathbb{C}^{n\times n})$ which lie in the interior of the contour $\operatorname{int}(\Gamma_{\mathcal{C}}) \subset \mathcal{D}$. Let

$$\kappa = \sum_{k=1}^{n_{\mathcal{C}}} \sum_{l=1}^{\eta_k} m_{l,k} \tag{3.1}$$

be the sum of all algebraic multiplicities of all eigenvalues. For large-scale problems it is necessary to condense $T(z)^{-1}$ so that it is possible to carry out the computations in an affordable time with moderate memory costs. Therefore, we multiply $T(z)^{-1}$ by a random matrix $U \in \mathbb{C}^{n\times l}$, where $l \leqslant n$, from the right. For l we require that it is chosen larger or equal than the maximal algebraic multiplicity of the eigenvalues of T, i.e.,

$$l \geqslant \max_{k=1,\dots,n_{\mathcal{C}}} \left(\sum_{l=1}^{\eta_k} m_{l,k} \right). \tag{3.2}$$

Note that it is always assumed that $\kappa \leqslant n$ because this is typical of large-scale problems. Moreover, we always require that U has full rank. However, this condition is fulfilled in almost all practical situations due to the floating point arithmetic. In the sequel we will need the following matrices.

Definition 3.2. *Let $p \in \mathbb{N}_0$. Then we call A_p given by*

$$A_p := \frac{1}{2\pi i} \int_{\Gamma_{\mathcal{C}}} z^p T(z)^{-1} U dz \in \mathbb{C}^{n\times l} \tag{3.3}$$

contour moment of order p.

The following lemma shows that there exist decompositions of the contour moments A_p into matrices.

Lemma 3.3. *Let A_p by given as in* (3.3) *and let $p \in \mathbb{N}_0$. Then there exists a splitting of A_p in the following way:*

$$A_p = V\Lambda^p W^H U, \tag{3.4}$$

with

$$V = \left(v_j^{l,k}, \quad 0 \leqslant j \leqslant m_{l,k} - 1,\ 1 \leqslant l \leqslant \eta_k, 1 \leqslant k \leqslant n_{\mathcal{C}} \right) \in \mathbb{C}^{n\times\kappa},$$
$$W = \left(w_j^{l,k}, \quad 0 \leqslant j \leqslant m_{l,k} - 1,\ 1 \leqslant l \leqslant \eta_k, 1 \leqslant k \leqslant n_{\mathcal{C}} \right) \in \mathbb{C}^{n\times\kappa},$$

defined as in Theorem 2.7, and Λ *having Jordan normal form*

$$\begin{aligned}\Lambda &= \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_{n_\mathcal{C}} \end{pmatrix}, \quad J_k = \begin{pmatrix} J_{k,1} & & \\ & \ddots & \\ & & J_{k,\eta_k} \end{pmatrix}, \\ J_{k,l} &= \begin{pmatrix} \lambda_k & 1 & & \\ & \ddots & \ddots & \\ & & \lambda_k & 1 \\ & & & \lambda_k \end{pmatrix} \in \mathbb{C}^{m_{l,k} \times m_{l,k}}.\end{aligned} \tag{3.5}$$

Proof. Using Theorem 3.1 we get for every $p \in \mathbb{N}_0$

$$A_p = \frac{1}{2\pi i}\int_{\Gamma_\mathcal{C}} z^p T(z)^{-1} U dz = \sum_{k=1}^{n_\mathcal{C}} \sum_{l=1}^{\eta_k} \sum_{j=1}^{m_{l,k}} \frac{(z^p)^{(j-1)}(\lambda_k)}{(j-1)!} \sum_{\nu=0}^{m_{l,k}-j} v_\nu^{l,k} (w_{m_{l,k}-j-\nu}^{l,k})^H U. \tag{3.6}$$

In order to show the statement, we calculate Λ^p first. Since Λ and J_k have block diagonal form, it holds

$$\Lambda^p = \begin{pmatrix} J_1^p & & \\ & \ddots & \\ & & J_{n_\mathcal{C}}^p \end{pmatrix}, \quad J_k^p = \begin{pmatrix} J_{k,1}^p & & \\ & \ddots & \\ & & J_{k,\eta_k}^p \end{pmatrix}.$$

For $J_{k,l}^p$ we get the following representation

$$J_{k,l}^p = \begin{pmatrix} \lambda_k^p & \frac{p\lambda_k^{p-1}}{1!} & \frac{p(p-1)\lambda_k^{p-2}}{2!} & \dots & \frac{p(p-1)\cdots(p-m_{l,k}+2)\lambda_k^{p-m_{l,k}+1}}{(m_{l,k}-1)!} \\ & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & \frac{p(p-1)\lambda_k^{p-2}}{2!} \\ & & & \ddots & \frac{p\lambda_k^{p-1}}{1!} \\ & & & & \lambda_k^p \end{pmatrix} \in \mathbb{C}^{m_{l,k} \times m_{l,k}}.$$

Note that the entries in this matrix coincide with the terms

$$\frac{(z^p)^{(j-1)}(\lambda_k)}{(j-1)!}$$

in (3.6). Moreover, V can be written in the form

$$V = \left(v_\cdot^{\cdot,1} \dots v_\cdot^{\cdot,n_\mathcal{C}}\right), \quad v_\cdot^{\cdot,k} = \left(v_\cdot^{1,k} \dots v_\cdot^{\eta_k,k}\right), \quad v_\cdot^{l,k} = \left(v_0^{l,k} \dots v_{m_{l,k}-1}^{l,k}\right) \in \mathbb{C}^{n \times m_{l,k}}.$$

Analogously, W^H can be represented as

$$W^H = \begin{pmatrix} (w_\cdot^{\cdot,1})^H \\ \vdots \\ (w_\cdot^{\cdot,n_\mathcal{C}})^H \end{pmatrix}, \quad (w_\cdot^{\cdot,k})^H = \begin{pmatrix} (w_\cdot^{1,k})^H \\ \vdots \\ (w_\cdot^{\eta_k,k})^H \end{pmatrix}, \quad (w_\cdot^{l,k})^H = \begin{pmatrix} (w_0^{l,k})^H \\ \vdots \\ (w_{m_{l,k}-1}^{l,k})^H \end{pmatrix} \in \mathbb{C}^{m_{l,k} \times n}.$$

The statement follows directly by matrix multiplication. □

Let us now choose $K \in \mathbb{N}$ such that $Kl \geqslant \kappa$ and the following rank conditions hold

$$\operatorname{rank}\begin{pmatrix} V \\ \vdots \\ V\Lambda^{K-1} \end{pmatrix} = \operatorname{rank}\begin{pmatrix} W^H U & \cdots & \Lambda^{K-1} W^H U \end{pmatrix} = \kappa. \tag{3.7}$$

If both matrices, V and $W^H U$, have full rank, then $K = 1$ and $l \geqslant \kappa$. In the case where either V or $W^H U$ or both do not have full rank, the following lemma shows that the rank conditions are satisfied if K is chosen larger than the sum of all maximal partial multiplicities at all eigenvalues [7, Lemma 5.1].

Lemma 3.4. *Let the assumptions of Theorem 2.7 be fulfilled. Then the rank conditions* (3.7) *hold for*

$$K \geqslant \sum_{k=1}^{n_{\mathcal{C}}} \max_{1\leqslant l\leqslant \eta_k} m_{l,k}. \tag{3.8}$$

Proof. We restrict to proving the first rank condition since the second rank condition can be shown in a similar way. The proof follows [7]. To start with, from our definition of the partial multiplicities it follows directly

$$m_{1,k} = \max_{1\leqslant l\leqslant \eta_k} m_{l,k}$$

for all $k \in \{1, \dots, n_{\mathcal{C}}\}$. We choose K such that (3.8) is satisfied. In order to prove

$$\operatorname{rank}\begin{pmatrix} V \\ \vdots \\ V\Lambda^{K-1} \end{pmatrix} = \kappa, \tag{3.9}$$

we have to show that for all $j = 0, \dots, K-1$, it follows from $V\Lambda^j x = 0$ for some $x \in \mathbb{C}^\kappa$ that $x = 0$, i.e., we show that the columns of the $Kn \times \kappa$ matrix in (3.9) are linearly independent. We suppose $V\Lambda^j x = 0$ for some $x \in \mathbb{C}^\kappa$ and for all $j = 0, \dots, K-1$. For any arbitrarily chosen $k \in \{1, \dots, n_{\mathcal{C}}\}$ and $0 \leqslant \beta \leqslant m_{1,k} - 1$ we define the polynomial

$$P_{k,\beta}(z) := (z - \lambda_k)^\beta \prod_{r=1, r\neq k}^{n_{\mathcal{C}}} (z - \lambda_r)^{m_{1,r}}.$$

We notice that by our assumption on the choice of K these polynomials have at most degree $K - 1$. Hence, it follows $VP_{k,\beta}(\Lambda)x = 0$. Then we partition V into columns and x into blocks which are compatible with the Jordan structure (3.5), i.e.,

$$V = (V_1 \cdots V_{n_{\mathcal{C}}}), \quad V_k = (V_{1,k} \cdots V_{\eta_k,k}), \quad V_{l,k} = \left(v_0^{l,k} \cdots v_{m_{l,k}-1}^{l,k}\right),$$
$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_{n_{\mathcal{C}}} \end{pmatrix}, \quad x_k = \begin{pmatrix} x_{1,k} \\ \vdots \\ x_{\eta_k,k} \end{pmatrix}, \quad x_{l,k} = \begin{pmatrix} x_0^{l,k} \\ \vdots \\ x_{m_{l,k}-1}^{l,k} \end{pmatrix}.$$

Using this partitioning we get from $VP_{k,\beta}(\Lambda)x = 0$

$$0 = \sum_{j=1}^{n_{\mathcal{C}}} V_j (J_j - \lambda_k)^{\beta} \prod_{r=1, r\neq k}^{n_{\mathcal{C}}} (J_j - \lambda_r)^{m_{1,r}} x_j.$$

Due to $(J_r - \lambda_r)^{m_{1,r}} = 0$, it follows immediately

$$0 = V_k (J_k - \lambda_k)^{\beta} \prod_{r=1, r\neq k}^{n_{\mathcal{C}}} (J_k - \lambda_r)^{m_{1,r}} x_k.$$

After expanding into columns again and using $(J_{k,l} - \lambda_k)^{\beta} = 0$ for $\beta \geqslant m_{l,k}$, we obtain

$$0 = \sum_{\substack{l=1 \\ \beta \leqslant m_{l,k}-1}}^{\eta_k} V_{l,k} \left[\prod_{r=1, r\neq k}^{n_{\mathcal{C}}} (J_{k,l} - \lambda_r)^{m_{1,r}} \right] (J_{k,l} - \lambda_r)^{\beta} x_{l,k}. \tag{3.10}$$

Now we show for fixed $k \in \{1, \dots, n_{\mathcal{C}}\}$ by induction on $\beta = m_{1,k} - 1, \dots, 0$, that

$$x_{\nu}^{l,k} = 0 \tag{3.11}$$

for all $\beta \leqslant \nu \leqslant m_{l,k} - 1$, where $l \in \{1, \dots, \eta_k\}$. For the basis we have to verify that (3.11) holds for $\beta = m_{1,k} - 1$. In fact, (3.10) reads in this case

$$\begin{aligned} 0 &= \sum_{\substack{l=1 \\ m_{l,k}=m_{1,k}}}^{\eta_k} \left(v_0^{l,k} \cdots v_{m_{l,k}-1}^{l,k} \right) \prod_{r=1, r\neq k}^{n_{\mathcal{C}}} (J_{k,l} - \lambda_r)^{m_{1,r}} \begin{pmatrix} 0 & \cdots & 1 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} x_0^{l,k} \\ \vdots \\ x_{m_{l,k}-1}^{l,k} \end{pmatrix} \\ &= \prod_{r=1, r\neq k}^{n_{\mathcal{C}}} (\lambda_k - \lambda_r)^{m_{1,r}} \sum_{\substack{l=1 \\ m_{l,k}=m_{1,k}}}^{\eta_k} v_0^{l,k} x_{m_{l,k}-1}^{l,k}. \end{aligned}$$

Since the vectors $v_0^{l,k}$ are linearly independent by Definition 2.5, $x_{m_{l,k}-1}^{l,k} = 0$ for all $l \in \{1, \dots, \eta_k\}$. For the induction step we assume that (3.11) holds for β (induction hypothesis) and show that it remains true for $\beta - 1$. Taking this hypothesis into consideration, equation (3.10) leads, similarly as above, to

$$\begin{aligned} 0 &= \sum_{\substack{l=1 \\ \beta \leqslant m_{l,k}}}^{\eta_k} \left(v_0^{l,k} \cdots v_{m_{l,k}-1}^{l,k} \right) \prod_{r=1, r\neq k}^{n_{\mathcal{C}}} (J_{k,l} - \lambda_r)^{m_{1,r}} \begin{pmatrix} 0 & \cdots & 1 & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 1 \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} x_0^{l,k} \\ \vdots \\ x_{\beta-1}^{l,k} \\ \vdots \\ 0 \end{pmatrix} \\ &= \prod_{r=1, r\neq k}^{n_{\mathcal{C}}} (\lambda_k - \lambda_r)^{m_{1,r}} \sum_{\substack{l=1 \\ \beta \leqslant m_{l,k}}}^{\eta_k} v_0^{l,k} x_{\beta-1}^{l,k}. \end{aligned}$$

Again, the linear independence of the vectors $v_0^{l,k}$ implies $x_{\beta-1}^{l,k} = 0$ for all $l \in \{1, \dots, \eta_k\}$, and therefore $x = 0$. □

We form the Hankel matrices

$$B_0 := \begin{pmatrix} A_0 & A_1 & \cdots & A_{K-1} \\ A_1 & A_2 & \cdots & A_K \\ \vdots & \vdots & \ddots & \vdots \\ A_{K-1} & A_K & \dots & A_{2K-2} \end{pmatrix}, \quad B_1 := \begin{pmatrix} A_1 & A_2 & \cdots & A_K \\ A_2 & A_3 & \cdots & A_{K+1} \\ \vdots & \vdots & \ddots & \vdots \\ A_K & A_{K+1} & \dots & A_{2K-1} \end{pmatrix}. \tag{3.12}$$

By (3.4) we can represent B_0 and B_1 in the following way

$$B_0 = \begin{pmatrix} V \\ \vdots \\ V\Lambda^{K-1} \end{pmatrix} \begin{pmatrix} W^H U & \cdots & \Lambda^{K-1} W^H U \end{pmatrix} := V_{[K]} W_{[K]}^H \in \mathbb{C}^{Kn \times Kl}$$

and

$$B_1 = \begin{pmatrix} V \\ \vdots \\ V\Lambda^{K-1} \end{pmatrix} \Lambda \begin{pmatrix} W^H U & \cdots & \Lambda^{K-1} W^H U \end{pmatrix} := V_{[K]} \Lambda W_{[K]}^H \in \mathbb{C}^{Kn \times Kl}. \tag{3.13}$$

Then we compute the reduced singular value decomposition (SVD)

$$V_{[K]} W_{[K]}^H = B_0 = V_0 \Sigma_0 W_0^H, \tag{3.14}$$

where $V_0 \in \mathbb{C}^{Kn \times \kappa}$, $V_0^H V_0 = I_\kappa$, $W_0 \in \mathbb{C}^{Kl \times \kappa}$, $W_0^H W_0 = I_\kappa$ and $\Sigma_0 = \operatorname{diag}(\sigma_1, \dots, \sigma_\kappa) \in \mathbb{C}^{\kappa \times \kappa}$. In order to show $\operatorname{rank}(B_0) = \kappa$, we need the following lemma [11, Section 2.5.5].

Lemma 3.5 (Sylvester's rank inequality). *Let $A \in \mathbb{C}^{m \times n}$ and $B \in C^{n \times p}$. Then it holds*

$$\operatorname{rank} A + \operatorname{rank} B - n \leqslant \operatorname{rank} AB \leqslant \min\{\operatorname{rank} A, \operatorname{rank} B\}.$$

Due to the rank conditions (3.7) we know that $V_{[K]} \in \mathbb{C}^{Kn \times \kappa}$ and $W_{[K]}^H \in \mathbb{C}^{\kappa \times Kl}$ have rank κ. Therefore, we obtain by applying Sylvester's rank inequality to B_0

$$\kappa \leqslant \operatorname{rank} B_0 \leqslant \kappa.$$

This implies that $\operatorname{rank}(B_0) = \kappa$. Hence, B_0 has singular values (s. [9, Lemma 4.29])

$$\sigma_1 \geqslant \cdots \geqslant \sigma_\kappa > 0 = \sigma_{\kappa+1} = \cdots = \sigma_{Kl}.$$

Another consequence of the rank conditions (3.7) and of equation (3.14) is

$$\operatorname{ran}(B_0) = \operatorname{ran}(V_{[K]}) = \operatorname{ran}(V_0), \tag{3.15}$$

where ran denotes the range. This follows directly from the fact that the columns of the matrix products $V_{[K]}W_{[K]}^H$ and $V_0\Sigma_0 W_0^H$ are linear combinations of the column vectors of $V_{[K]}$ and V_0 respectively. Since $V_0, V_{[K]} \in \mathbb{C}^{Kn\times\kappa}$ and V_0 has orthonormal columns, we can write

$$V_{[K]} = V_0 S \tag{3.16}$$

with

$$S = V_0^H V_{[K]} \in \mathbb{C}^{\kappa\times\kappa}$$

regular due to (3.15). With (3.14) and (3.16) we obtain

$$V_0 S W_{[K]}^H = V_0 \Sigma_0 W_0^H,$$

and it follows by multiplying this equation with the matrix V_0^H from the left side and using the definition of orthogonal matrices $V_0^H V_0 = I_\kappa$ that

$$S W_{[K]}^H = \Sigma_0 W_0^H.$$

Due to the regularity of S, we can apply S^{-1} to this equation. This leads to

$$W_{[K]}^H = S^{-1}\Sigma_0 W_0^H. \tag{3.17}$$

Plugging (3.16) and (3.17) into (3.13) gives

$$B_1 = V_{[K]}\Lambda W_{[K]}^H = V_0 S\Lambda S^{-1}\Sigma_0 W_0^H,$$

and after multiplication by V_0^H from the left and by $W_0\Sigma_0^{-1}$ from the right, we finally obtain by using $V_0^H V_0 = I_\kappa$ and $W_0^H W_0 = I_\kappa$

$$B := V_0^H B_1 W_0 \Sigma_0^{-1} = S\Lambda S^{-1}. \tag{3.18}$$

Note that this matrix B can be calculated without any knowledge of the eigenvalues and CSGEs, only by calculating the matrices B_0 and B_1 defined as in (3.12) via the contour integrals (3.3) and computing the reduced SVD of B_0. We summarize this result in the following theorem [7, Theorem 5.2].

Theorem 3.6. *Let $T \in \mathcal{H}(\mathcal{D}, \mathbb{C}^{n\times n})$ and let $\Gamma_{\mathcal{C}} \subset \mathcal{D}$ such that there are no eigenvalues of T on $\Gamma_{\mathcal{C}}$. Let us denote the pairwise distinct eigenvalues inside $\Gamma_{\mathcal{C}}$ by λ_k, $k = 1, \ldots, n_{\mathcal{C}}$, and the corresponding partial multiplicities by $m_{1,k} \geqslant \cdots \geqslant m_{\eta_k,k}$. Further, we suppose that the rank conditions* (3.7) *are fulfilled with κ given by* (3.1). *Then the matrix $B \in \mathbb{C}^{\kappa\times\kappa}$ defined in* (3.18) *has Jordan normal form with the same eigenvalues λ_k and corresponding partial multiplicities $m_{l,k}$, $l = 1, \ldots, \eta_k$, $k = 1, \ldots, n_{\mathcal{C}}$, as T. If $s_j^{l,k}$ are corresponding CSGEs for B, one can obtain suitable CSGEs for T via*

$$v_j^{l,k} = V_0^{[1]} s_j^{l,k}, \quad 0 \leqslant j \leqslant m_{l,k} - 1, 1 \leqslant l \leqslant \eta_k, 1 \leqslant k \leqslant n_{\mathcal{C}},$$

where $V_0^{[1]}$ denotes the upper $n \times \kappa$ block in

$$V_0 = \begin{pmatrix} V_0^{[1]} \\ \vdots \\ V_0^{[K]} \end{pmatrix}.$$

3.3 Approximation of B_0 and B_1

The major cost of the algorithms lies in the calculation of the integrals A_p given by (3.3). Since it is generally not possible to calculate A_p explicitly, we approximate these integrals by performing numerical integration. Note that the generation of A_p also requires to compute $T(z)^{-1}U$ in the integrand. To do this, equations of the form

$$T(z)X = U,$$

where $X \in \mathbb{C}^{n\times l}$, have to be solved. Let us assume that there exists a 2π-periodic smooth parametrization of $\Gamma_{\mathcal{C}}$, i.e., there exists $\phi \in C^1(\mathbb{R}, \mathbb{C})$ such that

$$\phi(t + 2\pi) = \phi(t) \quad \forall t \in \mathbb{R}.$$

For the approximation of A_p we use the trapezoidal sum because it can be shown that the approximation $\hat{A}_p$ of A_p converges exponentially to A_p with respect to the number of quadrature nodes N if equidistant nodes $t_j = \frac{2j\pi}{N}$, $j = 0, \dots, N$, are taken as quadrature points [7, Theorem 4.7]. We obtain approximations of the form

$$A_p := \frac{1}{2\pi i} \int_{\Gamma_{\mathcal{C}}} z^p T(z)^{-1} U dz \approx \frac{1}{iN} \sum_{j=0}^{N-1} T(\phi(t_j))^{-1} U \phi(t_j)^p \phi'(t_j) =: \hat{A}_p. \tag{3.19}$$

3.4 Algorithm (CIM)

By summarizing the steps from the previous sections we obtain the following algorithm, which is called *Integral algorithm 2* in [7].

Algorithm 1 CIM

```
1: Fix the contour $\Gamma_{\mathcal{C}}$ and the number $N$ of quadrature points for the trapezoidal rule. Choose $l \leqslant n$ according to (3.2), and $K \in \mathbb{N}$ such that $Kl \geqslant \kappa$ and (3.7) are fulfilled. Note that in most cases the number of eigenvalues $\kappa$ is unknown beforehand. Construct a $n \times l$ random matrix $U$.
2: Calculate the matrices $\hat{A}_p$ given by (3.19) for $p = 0, \dots, 2K - 1$.
3: Form the approximations $\hat{B}_0$ and $\hat{B}_1$ of the block Hankel matrices $B_0$ and $B_1$ given as in (3.12) by using $\hat{A}_p$ instead of $A_p$.
4: Compute the reduced SVD of $\hat{B}_0 = \tilde{V}\Sigma\tilde{W}^H$, where $\tilde{V} \in \mathbb{C}^{Kn\times Kl}$, $\tilde{W} \in \mathbb{C}^{Kl\times Kl}$, $\tilde{V}^H\tilde{V} = \tilde{W}^H\tilde{W} = I_{Kl}$, and $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \dots, \sigma_{Kl})$.
5: Determine $\kappa$ by $\sigma_1 \geqslant \cdots \geqslant \sigma_\kappa > \mathrm{tol}_{\mathrm{rank}} > \sigma_{\kappa+1} \approx \cdots \approx \sigma_{Kl} \approx 0$.
6: Set $V_0 = \tilde{V}(1 : Kn, 1 : \kappa)$, $W_0 = \tilde{W}(1 : Kl, 1 : \kappa)$ and $\Sigma_0 = \mathrm{diag}(\sigma_1, \sigma_2, \dots, \sigma_\kappa)$.
7: Compute the matrix $\hat{B} = V_0^H \hat{B}_1 W_0 \Sigma_0^{-1} \in \mathbb{C}^{\kappa\times\kappa}$ and solve the eigenvalue problem for $\hat{B}$. Let $(\lambda_j, s_j)$ be an eigenpair of $\hat{B}$. Calculate $v_j = V_0^{[1]} s_j$, and accept $\lambda_j$, if $\lambda_j \in \mathrm{Int}(\Gamma_{\mathcal{C}})$ and $\|T(\lambda_j)v_j\|/\|v_j\| \leqslant \mathrm{tol}_{\mathrm{res}}$.
```

Remark (Choice of K, tol_{rank} and tol_{res})**.** *When looking at the above algorithm, the question arises, how to choose the parameters K, tol_{rank} and tol_{res}.*

First of all, the parameter K can be chosen one if the generalized eigenvectors are linearly independent. However, numerical tests have shown that in this case it is even better to choose $K = 2$, but larger values are not recommended due to the special structure of the Hankel matrices. For the case of linearly dependent eigenvectors, Lemma 3.4 has to be taken into consideration.

As far as tol_{rank} is concerned, this parameter describes in fact the numerical rank of a matrix, which is defined as the largest value k such that

$$\sigma_k > \delta \cdot \sigma_1$$

for a given tolerance δ. Clearly, there exists no explicit formula for the proper choice of δ since the singular value distribution of B_0 and B_1 is affected by several factors, e.g. the position of the eigenvalues and their distribution. In our numerical tests we almost always chose $\delta \approx 10^{-14}$ for the cases where dense matrices were used in the calculation of $T(\phi(t_j))$ and a direct solver (LU) was chosen for solving the systems of linear equations. For the cases where these matrices were approximated with tolerance ϵ, i.e.,

$$\|T(\phi(t_j)) - T_\epsilon(\phi(t_j))\|_2 \lesssim \epsilon,$$

and the systems of linear equations were solved using an iterative solver (GMRES) with tolerance $\text{gmres}_{\text{tol}}$, the following approximation formula for δ worked nicely

$$\delta \approx \max\{\epsilon, \text{gmres}_{\text{tol}}\} \cdot 10^{-2}.$$

Finally, the parameter tol_{res} is used to "remove" spurious eigenvalues lying inside of $\Gamma_{\mathcal{C}}$ from the list of all calculated eigenvalues. Typically there is a gap of several orders in the residuals of the real and the spurious eigenvalues. In most cases tol_{res} can be chosen 10^{-4} or even lower.

4 Rational interpolation method (RIM)

In this chapter we would like to present an algorithm for calculating all eigenvalues of $T \in \mathcal{H}(\mathcal{D}, \mathbb{C}^{n\times n})$ lying in the interior of a given contour $\Gamma_{\mathcal{C}} \subset \mathcal{D}$ which uses the rational interpolation approach to approximate the poles of $T(z)^{-1}$. The main ideas are taken from [32], though the algorithm is derived in a bit different way.

4.1 Jacobi rational interpolation algorithm

The central idea of the RIM is to compute an appropriate rational interpolant of $T(z)^{-1}U$ inside $\Gamma_{\mathcal{C}}$, where U is a random matrix chosen as in the CIM, and to calculate the zeros of the denominator then. One possibility to compute the rational interpolant of a function inside $\Gamma_{\mathcal{C}}$ is to use the so-called *Jacobi rational interpolation algorithm.* This algorithm is presented in [10] for scalar functions f. In this section we follow [10] and generalize the algorithm for matrix-valued functions. Let

$$F(z) := T(z)^{-1}U \in \mathbb{C}^{n\times l} \tag{4.1}$$

be matrix-valued, where $U \in \mathbb{C}^{n\times l}$ is a random matrix with $l \leqslant n$ linearly independent columns and l should be chosen according to (3.2), i.e.,

$$l \geqslant \max_{k=1,\dots,n_{\mathcal{C}}} \left(\sum_{l=1}^{\eta_k} m_{l,k} \right).$$

For a given set of N tuples $(z_i, F(z_i))$, $i = 0, \dots, N-1$, we would like to determine a matrix-valued polynomial P_μ of degree μ and a scalar polynomial q_ν of degree ν with $N = \mu + \nu + 1$ such that

$$F(z_i) = \frac{P_\mu(z_i)}{q_\nu(z_i)} \quad \forall 0 \leqslant i \leqslant N-1. \tag{4.2}$$

Note that the coefficients of P_μ are matrices. Now we need some definitions from the barycentric Lagrange interpolation theory [5]. First, we define the *node polynomial*

$$l(z) := (z - z_0)(z - z_1)\cdots(z - z_{N-1})$$

and the *barycentric weights*

$$w_i := \frac{1}{l'(z_i)} = \frac{1}{\prod_{j=0,j\neq i}^{N-1}(z_i - z_j)} \quad \forall 0 \leqslant i \leqslant N-1. \tag{4.3}$$

From (4.2) we obtain for every $j \geqslant 0$

$$w_i z_i^j P_\mu(x_i) = w_i z_i^j F(z_i) q_\nu(z_i) \quad \forall 0 \leqslant i \leqslant N-1,$$

and by summation of these terms we get

$$\sum_{i=0}^{N-1} w_i z_i^j P_\mu(z_i) = \sum_{i=0}^{N-1} w_i z_i^j F(z_i) q_\nu(z_i). \tag{4.4}$$

Before proceeding, we repeat some important definitions from the interpolation theory for scalar-valued functions g.

Definition 4.1. *The divided differences of a function g with respect to x_i, $i = 0, \dots, n$, are defined recursively as follows.*

- *zeroth divided difference:*

$$g[x_i] = g(x_i),$$

- *first divided difference:*

$$g[x_i, x_{i+1}] = \frac{g[x_{i+1}] - [g(x_i)]}{x_{i+1} - x_i},$$

- *k^{th} divided difference:*

$$g[x_i, x_{i+1}, \dots, x_{i+k}] = \frac{g[x_{i+1}, x_{i+2}, \dots, x_{i+k}] - g[x_i, x_{i+1}, \dots, x_{i+k-1}]}{x_{i+k} - x_i}.$$

We denote $g[x_0, \dots, x_k]$ by $[g(x)]_{0,\dots,k}$.

Definition 4.2. *We define the Newton interpolation polynomial of order n of a function g at the interpolation points $x_0, \dots, x_n$ by*

$$p_n(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \cdots + a_n(x - x_0)\cdots(x - x_{n-1}). \tag{4.5}$$

It can be shown via induction that the coefficients a_k are given by the divided differences $[g(x)]_{0,\dots,k}$ (see e.g. [9, Chapter 8] for real-valued g). Hence, it follows immediately $[g(x)]_{0,\dots,n} = 0$ if g is a polynomial of degree smaller than n. In addition, $a_n = [g(x)]_{0,\dots,n}$ is the leading coefficient of $p_n(x)$, i.e., the coefficient of x^n.

Theorem 4.3 (Lagrange interpolation formula)**.** *Let $g(x_0), g(x_1), \dots, g(x_n)$ be given. Then there exists a unique polynomial $p_n \in \Pi_n$ such that*

$$p_n(x_i) = g(x_i) \quad \forall 0 \leqslant i \leqslant n.$$

In particular, p_n can be represented as

$$p_n(x) = \sum_{i=0}^{n} g(x_i) l_i(x), \tag{4.6}$$

where

$$l_i(x) := \prod_{k=0, k \neq i}^{n} \frac{x - x_k}{x_i - x_k} \tag{4.7}$$

are the Lagrange basis polynomials.

Proof. See e.g. [9, Theorem 8.3] for real-valued g. □

Note that the leading coefficient of the Lagrange basis polynomial l_i defined in (4.7) for fixed $i \in \{0, \dots, n\}$ is given by

$$\prod_{k=0, k \neq i}^{n} \frac{1}{x_i - x_k} = w_i.$$

Hence, the coefficient of x^n in $p_n(x)$ defined by (4.6) is

$$\sum_{i=0}^{n} w_i g(x_i).$$

By equating it with the corresponding coefficient in the formula for the Newton interpolation polynomial (4.5)

$$a_n = [g(x)]_{0,\dots,n},$$

we deduce the formula

$$[g(x)]_{0,\dots,n} = \sum_{i=0}^{n} w_i g(x_i). \tag{4.8}$$

Note that the definitions and results above also remain true for matrix-valued functions $G \in \mathbb{C}^{n \times l}$ because they can be written as

$$G(z) := \begin{pmatrix} g_{1,1}(z) & \cdots & g_{1,l}(z) \\ \vdots & & \vdots \\ g_{n,1}(z) & \dots & g_{n,l}(z) \end{pmatrix}$$

and the results can be applied to each component $g_{i,j}(z)$. Clearly, the coefficients $a_0, \ldots, a_n$ in (4.5) are matrices $A_0, \ldots, A_n \in \mathbb{C}^{n \times l}$ when G is used. The corresponding analogue of formula (4.8) for G is

$$[G(x)]_{0,\ldots,n} = \sum_{i=0}^{n} w_i G(x_i).$$

By setting $G(z) = z^j P_\mu(z)$ and $G(z) = z^j F(z) q_\nu(z)$ respectively, equation (4.4) can be represented via the divided differences of order $N-1$ with respect to the interpolation values z_i as

$$[z^j P_\mu(z)]_{0,\ldots,N-1} = [z^j F(z) q_\nu(z)]_{0,\ldots,N-1}.$$

The left term vanishes for $0 \leqslant j \leqslant \nu - 1$ since $N = \mu + \nu + 1$ and

$$\deg(z^j P_\mu(z)) = j + \mu \leqslant \nu - 1 + \mu = N - 2.$$

Hence,

$$[z^j F(z) q_\nu(z)]_{0,\ldots,N-1} = 0 \quad \forall 0 \leqslant j \leqslant \nu - 1. \tag{4.9}$$

This represents a system of ν matrix-equations and $\nu+1$ unknowns, namely the coefficients of $q_\nu(z)$. Therefore, we have one degree of freedom in the choice of the coefficients of $q_\nu(z)$. Since $q_\nu(z)$ is a polynomial of degree ν, it can be represented as

$$q_\nu(z) = \sum_{k=0}^{\nu} d_k z^k,$$

where we choose $d_\nu = 1$. With this ansatz we receive from (4.9) the following linear systems of equations

$$\sum_{i=0}^{N-1} w_i z_i^j F(z_i) q_\nu(z_i) = \sum_{i=0}^{N-1} w_i z_i^j F(z_i) \sum_{k=0}^{\nu} d_k z^k = 0 \quad \forall 0 \leqslant j \leqslant \nu - 1,$$

which can be rewritten as

$$\sum_{k=0}^{\nu} h_{j+k} d_k = 0 \quad \forall 0 \leqslant j \leqslant \nu - 1, \tag{4.10}$$

where

$$h_s := \sum_{i=0}^{N-1} w_i z_i^s F(z_i) \in \mathbb{C}^{n \times l} \quad \forall 0 \leqslant s \leqslant 2\nu - 1. \tag{4.11}$$

Finally, we can put (4.10) into the form

$$\begin{pmatrix} h_0 & h_1 & h_2 & \cdots & h_{\nu-1} \\ h_1 & h_2 & h_3 & \cdots & h_\nu \\ h_2 & h_3 & h_4 & \cdots & h_{\nu+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ h_{\nu-1} & h_\nu & h_{\nu+1} & \dots & h_{2\nu-2} \end{pmatrix} \begin{pmatrix} D_0 \\ D_1 \\ D_2 \\ \vdots \\ D_{\nu-1} \end{pmatrix} = - \begin{pmatrix} h_\nu \\ h_{\nu+1} \\ h_{\nu+2} \\ \vdots \\ h_{2\nu-1} \end{pmatrix}, \tag{4.12}$$

where $D_i = \mathrm{diag}(d_i, \dots, d_i) \in \mathbb{C}^{l \times l}$. This system is called *Yule-Walker system.*

Remark. *Note that for given μ and ν a rational interpolant satisfying* (4.2) *at all points does not have to exist. This is the case, when z_i is a zero of the numerator and the denumerator polynomials at the same time. However, Saff [24] showed the following: Let f be a scalar function. If the degree ν of the monic polynomial q_ν is chosen as the number of poles of f, then there exist rational interpolants p_μ/q_ν which converge uniformly to f in some region. Moreover, the poles of those rational interpolants converge to the poles of f, provided that μ is sufficiently large. We will see that for $\nu = K$, where K is the number of blocks per row/column of the Hankel matrices B_0 and B_1 in the CIM given by* (3.12)*, a rational interpolant of the matrix-valued F defined as in* (4.1) *exists, provided that N is sufficiently large. Clearly, it also has to be assumed that* (3.8) *holds in the case where the generalized eigenspace is rank-deficient. If all generalized eigenvectors are linearly independent, K can be chosen one.*

4.2 Relationship between the zeros of q_ν and a generalized eigenvalue problem

In this section we show that there exists a relationship between the zeros of the polynomial q_ν in the rational interpolant of F and a generalized eigenvalue problem involving Hankel matrices. For this, we need the following lemma.

Lemma 4.4. *Let q_ν be a complex-valued monic polynomial of degree ν with zeros $x_1, \dots, x_\nu$, i.e.,*

$$q_\nu(z) = d_0 + d_1 z + \dots + d_{\nu-1} z^{\nu-1} + z^\nu = (z - x_1) \cdot \dots \cdot (z - x_\nu). \tag{4.13}$$

Then the eigenvalues of the Frobenius companion matrix

$$C := \begin{pmatrix} 0 & 0 & \cdots & 0 & -d_0 \\ 1 & 0 & \cdots & 0 & -d_1 \\ 0 & 1 & \cdots & 0 & -d_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -d_{\nu-1} \end{pmatrix}$$

coincide with the zeros of q_ν.

Proof. " $\Leftarrow$ " Let x_j be a zero of q_ν, i.e., $q_\nu(x_j) = 0$. Then it holds

$$C^T \begin{pmatrix} 1 \\ x_j \\ \vdots \\ x_j^{\nu-2} \\ x_j^{\nu-1} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -d_0 & -d_1 & -d_2 & \cdots & -d_{\nu-1} \end{pmatrix} \begin{pmatrix} 1 \\ x_j \\ \vdots \\ x_j^{\nu-2} \\ x_j^{\nu-1} \end{pmatrix} = \begin{pmatrix} x_j \\ x_j^2 \\ \vdots \\ x_j^{\nu-1} \\ x_j^{\nu} \end{pmatrix} = x_j \begin{pmatrix} 1 \\ x_j \\ \vdots \\ x_j^{\nu-2} \\ x_j^{\nu-1} \end{pmatrix},$$

where

$$-d_0 - d_1 x_j - \cdots - d_{\nu-1} x_j^{\nu-1} = -q_\nu(x_j) + x_j^\nu = x_j^\nu$$

is used. Therefore, x_j is also an eigenvalue of C. " $\Rightarrow$ " Now we assume that x_j is an eigenvalue of C. Then x_j is also an eigenvalue of C^T and hence there exists an eigenvector $(a_0, a_1, \ldots, a_{\nu-1})^T \in \mathbb{C}^\nu \backslash \{0\}$ such that

$$C^T \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_{\nu-2} \\ a_{\nu-1} \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_{\nu-1} \\ -d_0 a_0 - d_1 a_1 - \cdots - d_{\nu-1} a_{\nu-1} \end{pmatrix} = x_j \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_{\nu-2} \\ a_{\nu-1} \end{pmatrix}.$$

If we choose $a_0 = 1$, we obtain $(a_0, a_1, \ldots, a_{\nu-1})^T = (1, x_j, \ldots, x_j^{\nu-1})^T$. The last row implies

$$d_0 + d_1 x_j + \cdots + d_{\nu-1} x_j^{\nu-1} + x_j^\nu = 0.$$

Therefore, x_j is a zero of q_ν. □

A generalization of this result is given by the following corollary.

Corollary 4.5. *Let q_ν be given as in Lemma 4.4. Then the eigenvalues of the matrix*

$$\tilde{C} := \begin{pmatrix} 0 & 0 & \cdots & 0 & -D_0 \\ I_{l\times l} & 0 & \cdots & 0 & -D_1 \\ 0 & I_{l\times l} & \cdots & 0 & -D_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & I_{l\times l} & -D_{\nu-1} \end{pmatrix},$$

where $D_i = \operatorname{diag}(d_i, \ldots, d_i) \in \mathbb{C}^{l\times l}$, coincide with the zeros of q_ν (each zero occurs l times).

Proof. Analogously to the proof of Lemma 4.4 with eigenvector $(I_{1\times l}, x_j I_{1\times l}, \ldots, x_j^{\nu-1} I_{1\times l})^T$. □

Using this corollary, we are now able to prove the following theorem which describes the relationship between the zeros of q_ν and a generalized eigenvalue problem using Hankel matrices, mentioned in the introduction of this section.

Theorem 4.6. *Let P_μ/q_ν be the rational interpolant of a matrix-valued function F in the interpolation points $z_1, \dots, z_N$, where $F(z) \in \mathbb{C}^{n \times l}$, P_μ is a matrix-valued polynomial of degree μ and q_ν is a monic polynomial of degree ν represented as in* (4.13)*. Denote the poles of P_μ/q_ν by $x_1, \dots, x_\nu$. Then $x_1, \dots, x_\nu$ are eigenvalues of the eigenvalue problem*

$$(H^< - \lambda H)y = 0, \tag{4.14}$$

where

$$H := \begin{pmatrix} h_0 & h_1 & \cdots & h_{\nu-1} \\ h_1 & h_2 & \cdots & h_\nu \\ \vdots & \vdots & \ddots & \vdots \\ h_{\nu-1} & h_\nu & \cdots & h_{2\nu-2} \end{pmatrix}, \quad H^< := \begin{pmatrix} h_1 & h_2 & \cdots & h_\nu \\ h_2 & h_3 & \cdots & h_{\nu+1} \\ \vdots & \vdots & \ddots & \vdots \\ h_\nu & h_{\nu+1} & \cdots & h_{2\nu-1} \end{pmatrix}, \tag{4.15}$$

and h_s is defined as in (4.11) *for $0 \leqslant s \leqslant 2\nu - 1$.*

Proof. The proof is similar to the proof in [18, Theorem 4] where the statement is shown for a scalar f. By Corollary 4.5 the zeros $x_1, \dots, x_\nu$ of q_ν are eigenvalues of the matrix $\tilde{C}$. Let x_j be an eigenvalue of $\tilde{C}$ with corresponding eigenvector v, i.e.,

$$\tilde{C}v = x_j v.$$

Then it follows by multiplication with H from the left side

$$H\tilde{C}v = x_j H v.$$

Using (4.12), one can easily show that $H\tilde{C} = H^<$. This proves the statement. □

4.3 Relationship between the RIM and the CIM

In this section we show that the CIM can be seen as a special case of the RIM in some way. In Section 4.1 we have already seen that the RIM leads to the computation of

$$h_p = \sum_{j=0}^{N-1} w_j z_j^p F(z_j), \quad p \in \mathbb{N}_0, \tag{4.16}$$

where z_j are interpolation points, and w_j are the corresponding barycentric weights given by (4.3). In the CIM we have to compute

$$\hat{A}_p = \frac{1}{iN} \sum_{j=0}^{N-1} T(\phi(t_j))^{-1} U \phi(t_j)^p \phi'(t_j) = \sum_{j=0}^{N-1} \tilde{w}_j \tilde{z}_j^p F(\tilde{z}_j), \quad p \in \mathbb{N}_0, \tag{4.17}$$

where

$$\tilde{z}_j := \phi(t_j)$$

are quadrature points on $\Gamma_{\mathcal{C}}$,

$$\tilde{w}_j := \frac{1}{iN}\phi'(t_j) \tag{4.18}$$

are the corresponding weights with $t_j = \frac{2j\pi}{N}$, and ϕ is a 2π-periodic smooth parametrization of $\Gamma_{\mathcal{C}}$. We show that the two methods are equivalent in the case, where the eigenvalues within the unit disc are searched and it is assumed that the N roots of unity are chosen as quadrature and interpolation points, i.e.,

$$z_j = \tilde{z}_j = \exp\left(\frac{2\pi ij}{N}\right), \quad j = 0, \dots, N-1.$$

This relationship has already been noticed for one-dimensional problems in [2] and for a symmetric generalized eigenvalue problem in [3].

Under the assumptions that $z_0, \dots, z_{N-1}$ are the roots of unity, it holds

$$l(z) = \prod_{j=0}^{N-1}(z - z_j) = z^N - 1.$$

Hence, we obtain for every $j = 0, \dots, N-1$

$$w_j = \frac{1}{l'(z_j)} = \frac{1}{Nz_j^{N-1}} = \frac{1}{N}z_j.$$

Due to

$$\tilde{z}_j = \phi(t_j) = \exp\left(\frac{2\pi ij}{N}\right),$$

it follows $\phi(t) = \exp(it)$ and thus $\phi'(t) = i\exp(it)$. Finally, we get by (4.18)

$$\tilde{w}_j = \frac{1}{iN}\phi'(t_j) = \frac{1}{N}\exp(it_j) = \frac{1}{N}z_j = w_j.$$

It can be shown in a similar way that the weights $\tilde{w}_j$ and w_j are the same up to common factors in the case, where the N roots of unity are multiplied by a constant $\alpha > 0$ and shifted by $\beta \in \mathbb{C}$. Furthermore, this remains true for the case, where the same points are chosen as quadrature and interpolation points, and the contours are ellipses. It seems reasonable to suppose that these results can also be transferred to the case, where the contours are even more arbitrary parametrizable C^1-contours. Therefore, the CIM can be seen as a special case of the RIM, if the same points are taken as quadrature points and interpolation points, and ν is chosen equal to K defined in the CIM. Provided that all generalized eigenvectors are linearly independent, K can be chosen one, otherwise K has to be chosen according to Lemma 3.4. We want to note that in [32] the RIM is derived in a bit different way and there one obtains for interpolation points which are chosen arbitrarily inside the contour $\Gamma_{\mathcal{C}}$ the same dimensions for the Hankel system as for the one in the CIM.

After having calculated the matrices H and $H^<$ defined in (4.15), the next steps to solve the generalized eigenvalue problem (4.14) are the same as for the CIM. These involve the calculation of the reduced SVD of H and the transformation of (4.14) into a linear matrix eigenvalue problem.

4.4 Algorithm (RIM)

Summarizing the steps described above leads to the following algorithm [32, Section 3.3].

Algorithm 2 RIM

```
1: Fix the contour $\Gamma_{\mathcal{C}}$, the number of interpolation points $N$, and the interpolation points $z_j$, $j = 0, 1, \ldots, N-1$, on or within the contour. Calculate the corresponding weights $\omega_j$. Choose $l \leqslant n$ and $K \in \mathbb{N}$ such that $Kl \geqslant \kappa$, and the same requirements for $l$ and $K$ as in the CIM are satisfied. Construct a $n \times l$ random matrix $U$.
2: Calculate the matrices $h_s$ given by (4.11) for $s = 0, \ldots, 2K-1$.
3: Form the two block Hankel matrices $H$ and $H^{<}$.
4: Compute the reduced SVD of $H = \tilde{V}\Sigma\tilde{W}^H$, where $\tilde{V} \in \mathbb{C}^{Kn\times Kl}$, $\tilde{W} \in \mathbb{C}^{Kl\times Kl}$, $\tilde{V}^H\tilde{V} = \tilde{W}^H\tilde{W} = I_{Kl}$, and $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_{Kl})$.
5: Determine $\kappa$ by $\sigma_1 \geqslant \cdots \geqslant \sigma_\kappa > \mathrm{tol}_{\mathrm{rank}} > \sigma_{\kappa+1} \approx \cdots \approx \sigma_{Kl} \approx 0$.
6: Set $V_0 = \tilde{V}(1 : Kn, 1 : \kappa)$, $W_0 = \tilde{W}(1 : Kl, 1 : \kappa)$ and $\Sigma_0 = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_\kappa)$.
7: Compute $B = V_0^H H^{<} W_0 \Sigma_0^{-1} \in \mathbb{C}^{\kappa\times\kappa}$ and solve the eigenvalue problem for $B$. Let $(\lambda_j, s_j)$ be an eigenpair of $B$. Compute $v_j = V_0^{[1]} s_j$ and accept $\lambda_j$, if $\lambda_j \in \mathrm{Int}(\Gamma_{\mathcal{C}})$ and $\|T(\lambda_j)v_j\|/\|v_j\| \leqslant \mathrm{tol}_{\mathrm{res}}$.
```

This algorithm differs only slightly from the CIM, yet it guarantees a more general choice of the interpolation points.

5 Rayleigh-Ritz procedure

In practical applications the occurring NEPs often have large dimensions. In order to circumvent solving large problems, the Rayleigh-Ritz procedure can be used. The idea is to project the original NEP onto a NEP with a smaller dimension and then solve this smaller NEP by using the CIM (Algorithm 1) or the RIM (Algorithm 2). The two following steps are the key steps of the classical Rayleigh-Ritz procedure.

(1) Build a proper search space $\mathcal{S}$ as a good approximation of a part of the generalized eigenspace. Let Q be an orthogonal basis of $\mathcal{S}$. The number of columns $k_{\mathcal{S}}$ of Q is in general much smaller than the dimension of the original NEP, denoted by n.

(2) Calculate approximate eigenpairs (λ, v) which satisfy the Galerkin condition

$$v \in \mathcal{S} \text{ and } T(\lambda)v \perp \mathcal{S}.$$

This is equivalent to determining eigenpairs (λ, g) of the NEP with smaller dimension

$$T_Q(\lambda)g = 0, \tag{5.1}$$

where $T_Q(z) = Q^H T(z) Q \in \mathbb{C}^{k_s \times k_s}$. If (λ, g) is an eigenpair of (5.1), an approximation of the corresponding eigenpair of (2.1) is (λ, Qg).

In this chapter we follow [31] and [32] where two possibilities to construct a proper search space $\mathcal{S}$, called the *moment scheme* and the *sampling scheme*, are presented.

5.1 Rayleigh-Ritz procedure based on the contour integral approach

In this section we summarize the the major steps of the moment scheme and the sampling scheme based on the contour integral approach.

5.1.1 Moment scheme

This scheme was first proposed in [33]. The main idea is to generate $\mathcal{S}$ by calculating contour moments of higher order defined as in Definition 3.2. To start with, we choose a random matrix $\hat{U} \in \mathbb{C}^{n \times L}$, where L should be larger or equal than the maximal algebraic multiplicity of all eigenvalues of T in $\Gamma_{\mathcal{C}}$, i.e.,

$$L \geqslant \max_{k=1,\dots,n_{\mathcal{C}}} \left(\sum_{l=1}^{\eta_k} m_{l,k} \right). \tag{5.2}$$

Then we define

$$M := (A_0, A_1, \dots, A_{K^{RR}-1}),$$

where A_p is defined as in (3.3) for $p \in \mathbb{N}_0$ and the splitting (3.4) is needed

$$A_p = \frac{1}{2\pi i} \int_{\Gamma_C} z^p T(z)^{-1} \hat{U} dz = V \Lambda^p W^H \hat{U}.$$

If K^{RR} is chosen such that

$$\operatorname{rank}(M) = \operatorname{rank}(V), \tag{5.3}$$

then it follows

$$\operatorname{span}(M) = \operatorname{span}(V).$$

This relationship can be seen as follows: Firstly, M can be written in the way

$$M = V(W^H \hat{U}, \Lambda W^H \hat{U}, \dots, \Lambda^{K^{RR}-1} W^H \hat{U}). \tag{5.4}$$

Secondly, we can represent V in terms of the column vectors v_i as

$$V = (v_1, \dots, v_\kappa).$$

The matrix multiplication in (5.4) gives a new matrix where each column is a linear combination of the vectors $v_1, \dots, v_\kappa$. Then the desired equality follows directly with (5.3). Therefore, we could choose $\mathcal{S} = \operatorname{span}(M)$ as approximation space of the eigenspace $\operatorname{span}(V)$. However, the large problem is that A_p can not be computed exactly in practice. In order to compute A_p, we have to use numerical quadrature rules like the trapezoid rule. This leads to the computation of the approximated moments

$$A_p \approx \hat{A}_p = \sum_{j=0}^{N^{RR}-1} \tilde{w_j}^{RR} (\tilde{z_j}^{RR})^p T^{-1}(\tilde{z_j}^{RR}) \hat{U}$$

with $\tilde{z_j}^{RR}$ and $\tilde{w_j}^{RR}$ defined as for (4.17). For the approximation of M we obtain

$$M \approx \hat{M} = (\hat{A}_0, \hat{A}_1, \dots, \hat{A}_{K^{RR}-1}). \tag{5.5}$$

If K^{RR} and N^{RR} are chosen such that the rank condition

$$\operatorname{rank}(\hat{M}) \geqslant \operatorname{rank}(V)$$

is satisfied, it holds

$$\operatorname{span}(\hat{M}) \approx \operatorname{span}(V).$$

We choose $\mathcal{S} = \operatorname{span}(\hat{M})$, what is also proposed in [33]. In practice, an orthogonal basis Q of $\hat{M}$ is used. This orthogonal basis can be computed easily by calculating the reduced SVD of $\hat{M}$ with tolerance δ^{RR}, i.e.,

$$\hat{M} \approx Q\Sigma^{RR}V^H,$$

where Σ^{RR} contains only singular values which are larger than $\delta^{RR} \cdot \sigma_1^{RR}$ with σ_1^{RR} denoting the largest singular value. The number of these singular values is called the numerical rank of $\hat{M}$ and denoted by $k_{\mathcal{S}}$. In [32] it is proposed to set $\delta^{RR} = 10^{-14}$. For the proper approximation of all poles of $T(z)^{-1}$ inside $\Gamma_{\mathcal{C}}$ it is necessary to guarantee that the condition $k_{\mathcal{S}} \geqslant \operatorname{rank}(V)$ is satisfied. As $\kappa \geqslant \operatorname{rank}(V)$, it is more convenient to use

$$K^{RR}L \geqslant k_{\mathcal{S}} \geqslant \kappa. \tag{5.6}$$

To summarize, K^{RR} and L have to be chosen such that the above condition is fulfilled, where there is also the restriction (5.2) on the choice of L. These considerations lead to the following algorithm which is quite similar to the CIRR(S) algorithm in [31] and the RSRR algorithm in [32].

Algorithm 3 CIM-RRm algorithm

```
1: Fix the contour $\Gamma_{\mathcal{C}}$ and the number $N^{RR}$ of quadrature points $\tilde{z}_j^{RR}$ for the trapezoidal
   rule. Choose $K^{RR}$ and $L$ such that (5.2) holds and $K^{RR}L \geqslant k_{\mathcal{S}} \geqslant \kappa$ is satisfied, where
   $\kappa$, defined as in (3.1), is unknown. Construct a $n \times L$ random matrix $\hat{U}$.
2: Build $\hat{M}$ as in (5.5).
3: Compute the reduced SVD of $\hat{M} = \tilde{Q}\Sigma^{RR}\tilde{W}^H$, where $\tilde{Q} \in \mathbb{C}^{n\times r}$, $\tilde{W} \in \mathbb{C}^{K^{RR}L\times r}$,
   $\tilde{Q}^H\tilde{Q} = \tilde{W}^H\tilde{W} = I_r$, $r = \min\{n, K^{RR}L\}$, and $\Sigma^{RR} = \operatorname{diag}(\sigma_1^{RR}, \sigma_2^{RR}, \dots, \sigma_r^{RR})$.
   Determine the number $k_{\mathcal{S}}$ of singular values which are larger than $\delta^{RR} \cdot \sigma_1^{RR}$. Set
   $Q = \tilde{V}(1 : n, 1 : k_{\mathcal{S}})$.
4: Calculate $T_Q(z) = Q^HT(z)Q$. Solve the projected NEP (5.1) by using the CIM (Algo-
   rithm 1) or the RIM (Algorithm 2) with choosing $U = I_{k_{\mathcal{S}}}$. This gives the eigenpairs
   $(\lambda_j, g_j)$, $j = 1, \dots, \kappa$.
5: Calculate the eigenpairs of the original NEP (2.1) by $(\lambda_j, Qg_j)$, $j = 1, \dots, \kappa$, and check
   the residuals $\|T(\lambda_j)v_j\|_2/\|v_j\|_2$, where $v_j = Qg_j$.
```

Remark. *Note that in Step 4 the random matrix U in the CIM or in the RIM can be chosen $I_{k_{\mathcal{S}}}$ since the projected NEP $T_Q(\lambda)g = 0$ usually has a small dimension. Furthermore, the interpolation points or quadrature points which are chosen for solving the projected NEP can be completely different than the quadrature points chosen in Step 1 of the algorithm above.*

The big disadvantage of this scheme is that for growing numbers of K, which could be needed to satisfy the rank condition (5.6), the columns of $\hat{M}$ become more and more linearly dependent. As a consequence, the accuracy of computed eigenvalues can be low, spurious eigenvalues can occur and there can even be a loss of eigenvalues (see example in [31, Section 3.1]). Finally, it is also possible that (5.6) can not be satisfied accurately any more. Hence, the moment generated eigenspace is not always reliable.

5.1.2 Sampling scheme

In this section we would like to summarize the key steps in the construction of a more reliable scheme, known as the sampling scheme. This can be seen as a modification of the moment scheme in some way. Similarly as in the moment scheme, we also choose a random matrix $\hat{U} \in \mathbb{C}^{n \times L}$, where L should be chosen such that (5.2) holds again, first. However, instead of $\hat{M}$ we construct

$$\hat{S} = \Big(T(\tilde{z}_0^{RR})^{-1}\hat{U}, T(\tilde{z}_1^{RR})^{-1}\hat{U}, \ldots, T(\tilde{z}_{N^{RR}-1}^{RR})^{-1}\hat{U} \Big) \in \mathbb{C}^{n \times N^{RR}L}. \tag{5.7}$$

It turns out that $\operatorname{span}(\hat{S})$ is an appropriate approximation of the generalized eigenspace $\operatorname{span} V$ because each $\hat{A}_p$ can be represented as a linear combination of $T(\tilde{z}_j)^{-1}\hat{U}$, i.e.,

$$\hat{A}_p = \sum_{j=0}^{N^{RR}-1} \tilde{w}_j^{RR} (\tilde{z}_j^{RR})^p T^{-1}(\tilde{z}_j^{RR})\hat{U}, \quad p \in \mathbb{N}_0,$$

and hence it follows directly

$$\operatorname{span}(V) \approx \operatorname{span}(\hat{M}) \subset \operatorname{span}(\hat{S}).$$

We set $\mathcal{S} = \operatorname{span}(\hat{S})$. For the proper approximation of all poles of $T(z)^{-1}$ inside $\Gamma_{\mathcal{C}}$ it is necessary to guarantee that the condition $k_{\mathcal{S}} \geqslant \operatorname{rank}(V)$ is satisfied. As $\kappa \geqslant \operatorname{rank}(V)$, it is more convenient to use

$$N^{RR}L \geqslant k_{\mathcal{S}} \geqslant \kappa. \tag{5.8}$$

To summarize, N^{RR} and L have to be chosen such that the above condition is fulfilled, where there is also the restriction (5.2) on the choice of L. As for the moment scheme, the next steps are to compute an orthogonal basis via performing a reduced SVD with tolerance δ^{RR} on $\operatorname{span}(\hat{S})$ and solving the projected NEP with the CIM or the RIM. The algorithm looks similar to Algorithm 3. This algorithm is called *CIRR(S)* in [31].

Algorithm 4 CIM-RRs algorithm

```
Fix the contour $\Gamma_{\mathcal{C}}$ and the number $N^{RR}$ of quadrature points $\tilde{z}_j^{RR}$ for the trapezoidal rule. Choose $L$ such that (5.2) holds and $N^{RR}L \geqslant k_{\mathcal{S}} \geqslant \kappa$ is satisfied, where $\kappa$, defined as in (3.1), is unknown. Construct a $n \times L$ random matrix $\hat{U}$.
Build $\hat{S}$ as in (5.7).
Compute the reduced SVD of $\hat{S} = \tilde{Q}\Sigma^{RR}\tilde{W}^H$, where $\tilde{Q} \in \mathbb{C}^{n \times r}$, $\tilde{W} \in \mathbb{C}^{N^{RR}L \times r}$, $\tilde{Q}^H\tilde{Q} = \tilde{W}^H\tilde{W} = I_r$, $r = \min\{n, N^{RR}L\}$, and $\Sigma^{RR} = \operatorname{diag}(\sigma_1^{RR}, \sigma_2^{RR}, \ldots, \sigma_r^{RR})$. Determine the number $k_{\mathcal{S}}$ of singular values which are larger than $\delta^{RR} \cdot \sigma_1^{RR}$. Set $Q = \tilde{V}(1:n, 1:k_{\mathcal{S}})$.
Calculate $T_Q(z) = Q^H T(z) Q$. Solve the projected NEP (5.1) by using the CIM (Algorithm 1) or the RIM (Algorithm 2) with choosing $U = I_{k_{\mathcal{S}}}$. This gives the eigenpairs $(\lambda_j, g_j)$, $j = 1, \ldots, \kappa$.
Calculate the eigenpairs of the original NEP (2.1) via $(\lambda_j, Qg_j)$, $j = 1, \ldots, \kappa$, and check the residuals $\|T(\lambda_j)v_j\|_2/\|v_j\|_2$, where $v_j = Qg_j$.
```

5.2 Rayleigh-Ritz procedure based on the rational interpolation approach

In this section we state the algorithms which correspond to Algorithm 3 and Algorithm 4 when the rational interpolation approach is used instead of the contour integral approach. The derivation of the algorithms works quite similarly as for the CIM. A detailed description is given in [31, 32].

5.2.1 Moment scheme

Analogously as above, we define the matrix

$$\hat{M} = (h_0, h_1, \dots, h_{K^{RR}-1}), \tag{5.9}$$

where h_p is given as in (4.16)

$$h_p = \sum_{j=0}^{N^{RR}-1} w_j^{RR} (z_j^{RR})^p F(z_j)$$

for $p \in \{0, \dots, K^{RR} - 1\}$ with interpolation points z_j^{RR} on or within Γ_C and corresponding weights w_j^{RR}. Then the algorithms based on the moment scheme reads as follows.

Algorithm 5 RIM-RRm algorithm

```
Fix the contour $\Gamma_\mathcal{C}$, the number of interpolation points $N^{RR}$, and the interpolation points $z_j^{RR}$, $j = 0, 1, \dots, N^{RR}-1$, on or within the contour. Compute the corresponding weights $\omega_j^{RR}$. Choose $K^{RR}$ and $L$ such that (5.2) holds and $K^{RR}L \geqslant k_\mathcal{S} \geqslant \kappa$ is satisfied, where $\kappa$ defined as in (3.1), is unknown. Construct a $n \times l$ random matrix $\hat{U}$.
Build $\hat{M}$ as in (5.9).
Compute the reduced SVD of $\hat{M} = \tilde{Q}\Sigma^{RR}\tilde{W}^H$, where $\tilde{Q} \in \mathbb{C}^{n\times r}$, $\tilde{W} \in \mathbb{C}^{K^{RR}L\times r}$, $\tilde{Q}^H\tilde{Q} = \tilde{W}^H\tilde{W} = I_r$, $r = \min\{n, K^{RR}L\}$, and $\Sigma^{RR} = \operatorname{diag}(\sigma_1^{RR}, \sigma_2^{RR}, \dots, \sigma_r^{RR})$. Determine the number $k_\mathcal{S}$ of singular values which are larger than $\delta^{RR} \cdot \sigma_1^{RR}$. Set $Q = \tilde{V}(1 : n, 1 : k_\mathcal{S})$.
Calculate $T_Q(z) = Q^H T(z) Q$. Solve the projected NEP (5.1) by using the CIM (Algorithm 1) or the RIM (Algorithm 2) with choosing $U = I_{k_\mathcal{S}}$. This gives the eigenpairs $(\lambda_j, g_j)$, $j = 1, \dots, \kappa$.
Calculate the eigenpairs of the original NEP (2.1) by $(\lambda_j, Qg_j)$, $j = 1, \dots, \kappa$, and check the residuals $\|T(\lambda_j)v_j\|_2/\|v_j\|_2$, where $v_j = Qg_j$.
```

5.2.2 Sampling scheme

Let $\hat{S}$ be defined analogously as above by

$$\hat{S} = \Big(T(z_0^{RR})^{-1}\hat{U}, T(z_1^{RR})^{-1}\hat{U}, \dots, T(z_{N^{RR}-1}^{RR})^{-1}\hat{U}\Big) \in \mathbb{C}^{n\times N^{RR}L}. \tag{5.10}$$

Then the corresponding algorithm, which is called *RSRR* in [32], is given below.

Algorithm 6 RIM-RRs algorithm

```
Fix the contour $\Gamma_{\mathcal{C}}$, the number $N^{RR}$ of interpolation points $z_j^{RR}$, and the interpolation points $z_j^{RR}$, $j = 0, 1, \ldots, N^{RR} - 1$ on or within the contour. Choose $L$ such that (5.2) holds and $N^{RR}L \geqslant k_{\mathcal{S}} \geqslant \kappa$ is satisfied, where $\kappa$, defined as in (3.1), is unknown. Construct a $n \times L$ random matrix $\hat{U}$.
Build $\hat{S}$ as in (5.10).
Compute the reduced SVD of $\hat{S} = \tilde{Q}\Sigma^{RR}\tilde{W}^H$, where $\tilde{Q} \in \mathbb{C}^{n\times r}$, $\tilde{W} \in \mathbb{C}^{N^{RR}L\times r}$, $\tilde{Q}^H\tilde{Q} = \tilde{W}^H\tilde{W} = I_r$, $r = \min\{n, N^{RR}L\}$, and $\Sigma^{RR} = \mathrm{diag}(\sigma_1^{RR}, \sigma_2^{RR}, \ldots, \sigma_r^{RR})$. Determine the number $k_{\mathcal{S}}$ of singular values which are larger than $\delta^{RR} \cdot \sigma_1^{RR}$. Set $Q = \tilde{V}(1 : n, 1 : k_{\mathcal{S}})$.
Calculate $T_Q(z) = Q^H T(z) Q$. Solve the projected NEP (5.1) by using the CIM (Algorithm 1) or the RIM (Algorithm 2) with choosing $U = I_{k_{\mathcal{S}}}$. This gives the eigenpairs $(\lambda_j, g_j)$, $j = 1, \ldots, \kappa$.
Calculate the eigenpairs of the original NEP (2.1) by $(\lambda_j, Qg_j)$, $j = 1, \ldots, \kappa$, and check the residuals $\|T(\lambda_j)v_j\|_2/\|v_j\|_2$, where $v_j = Qg_j$.
```

6 Boundary element method for eigenvalue problems in acoustics

In this chapter we formulate interior and exterior eigenvalue problems for the Laplace operator first, then we derive corresponding boundary integral formulations, and finally we give a characterization of the eigenpairs. The material of this chapter is based on [28].

6.1 Formulation of eigenvalue problems for the Laplace operator

Our starting point is the Helmholtz equation in $\mathbb{R}^3$

$$-\Delta u(x) - k^2 u(x) = 0,$$

where u is a complex, scalar-valued function and k is the wave number. This equation describes the propagation of a time-harmonic sound wave in a homogeneous, isotropic, and friction free medium. Note that the Helmholtz equation follows directly from the acoustic wave equation in $\mathbb{R}^3$

$$\frac{1}{c^2}\frac{\partial^2}{\partial t^2}U(x,t) = \Delta U(x,t),$$

where c denotes the speed of sound, by plugging in the ansatz for time-harmonic sound waves with frequency ω

$$U(x,t) = \mathrm{Re}\{u(x)e^{-i\omega t}\},$$

and defining $k := \omega/c$. In the sequel we will assume that $\mathrm{Re}(k) \geqslant 0$. Thus, we define

$$\mathbb{C}^+ := \{z \in \mathbb{C} : \mathrm{Re}(z) \geqslant 0\}$$

as the space of complex numbers with real part larger or equal than zero. Before stating the eigenvalue problems, we need some general definitions and assumptions. First of all, let us assume that $\Omega^- \subset \mathbb{R}^3$ is a bounded Lipschitz domain with closed, piecewise smooth boundary, and that $\Omega^+ := \mathbb{R}^3 \backslash \overline{\Omega^-}$ is simply connected. We denote the boundary of Ω^- by Γ and define the space of test functions in $\Omega^\pm$ by

$$\mathcal{D}_\infty := C_0^\infty(\Omega^\pm),$$

where $C_0^\infty(\Omega^\pm)$ is the space of infinitely differentiable functions with compact support in $\Omega^\pm$. Furthermore, we introduce the space

$$L^2(\Omega^-) := \left\{ u : \Omega^- \to \mathbb{C} \text{ measurable } : \int_{\Omega^-} |u(x)|^2 dx < \infty \right\}.$$

The Sobolev space $H^1(\Omega^-)$ is given by

$$H^1(\Omega^-) := \{u \in L^2(\Omega^-) : \exists D_j u \in L^2(\Omega^-) \text{ for all } j = 1, 2, 3\},$$

where $D_j u$ is the weak derivative of u with respect to x_j, i.e.,

$$\int_{\Omega^-} u(x) \frac{\partial \phi}{\partial x_j} dx = - \int_{\Omega^-} D_j u(x) \phi(x) dx$$

for all $\phi \in \mathcal{D}_\infty(\Omega^-)$. Moreover, we introduce the space

$$\begin{aligned} H_0^1(\Omega^-) : &= \overline{\mathcal{D}_\infty(\Omega^-)}^{H^1(\Omega^-)} \\ &= \left\{ u \in H^1(\Omega^-) : \exists (\phi_n)_n \subset \mathcal{D}_\infty(\Omega^-) \text{ such that } \phi_n \to u \text{ in } H^1(\Omega^-) \right\}. \end{aligned}$$

Then we define the spaces

$$L^2_{\text{loc}}(\Omega^+) := \left\{ f \in L^2(\Omega^+) : \int_U |u(x)|^2 dx < \infty \quad \forall U \Subset \Omega^+ \right\},$$

where $U \Subset \Omega^+$ means that u is open, bounded, and $\overline{U} \subset \Omega^+$, and the Sobolev space

$$H^1_{\text{loc}}(\Omega^+) := \left\{ u \in L^2_{\text{loc}}(\Omega^+) : \exists D_j u \in L^2_{\text{loc}}(\Omega^+) \text{ for all } j = 1, 2, 3 \right\}.$$

For a more detailed description of these spaces and their properties we refer to [19]. In the following we will also need the spaces

$$\begin{aligned} H^1(\Delta, \Omega^-) &:= \{u \in H^1(\Omega^-) : \Delta u \in L^2(\Omega^-)\}, \\ H^1_{\text{loc}}(\Delta, \Omega^+) &:= \{u \in H^1_{\text{loc}}(\Omega^+) : \Delta u \in L^2_{\text{loc}}(\Omega^+)\}. \end{aligned}$$

Now we consider the Dirichlet and Neumann trace operators. In the case of smooth functions $u^\pm \in \mathbb{C}^\infty(\overline{\Omega^\pm})$ the Dirichlet trace operators for Ω^+ and Ω^- respectively are given by

$$\gamma_0^\pm u^\pm(x) = \lim_{\Omega^\pm \ni \tilde{x}^\pm \to x} u(\tilde{x}^\pm)$$

and the Neumann trace operators are defined by

$$\gamma_1^\pm u^\pm(x) = \lim_{\Omega^\pm \ni \tilde{x}^\pm \to x} \nabla u^\pm(\tilde{x}^\pm)|_\Gamma \cdot n(x)$$

for $x \in \Gamma$. It can be shown that for these operators there exist unique extensions

$$\begin{aligned}
\gamma_0^+ &: H^1_{\text{loc}}(\Omega^+) \to H^{1/2}(\Gamma), & \gamma_0^- &: H^1(\Omega^-) \to H^{1/2}(\Gamma), \\
\gamma_1^+ &: H^1_{\text{loc}}(\Delta, \Omega^+) \to H^{-1/2}(\Gamma), & \gamma_1^- &: H^1(\Delta, \Omega^-) \to H^{-1/2}(\Gamma),
\end{aligned}$$

see [19, Theorem 3.37], [19, Lemma 4.3] and [8, Lemma 3.2]. The spaces $H^{1/2}(\Gamma)$ and $H^{-1/2}(\Gamma)$ are called trace spaces. These spaces can be defined as follows: The former one is given by

$$H^{1/2}(\Gamma) := \overline{C^0(\Gamma)}^{\|\cdot\|_{H^{1/2}(\Gamma)}},$$

where $C^0(\Gamma)$ denotes the space of continuous functions on Γ and

$$\|u\|_{H^{1/2}(\Gamma)} := \left(\int_\Gamma |u(x)|^2 ds_x + \int_\Gamma \int_\Gamma \frac{|u(x) - u(y)|^2}{|x-y|^3} ds_x ds_y \right)^{1/2}.$$

The latter one is the dual space of $H^{1/2}(\Gamma)$, i.e.,

$$H^{-1/2}(\Gamma) := \left(H^{1/2}(\Gamma)\right)' := \{v : H^{1/2}(\Gamma) \to \mathbb{C} \text{ linear and continuous}\},$$

where the corresponding norm is given by

$$\|v\|_{H^{-1/2}(\Gamma)} := \sup_{0 \neq w \in H^{1/2}(\Gamma)} \frac{|\langle v, w \rangle_\Gamma|}{\|w\|_{H^{1/2}(\Gamma)}}$$

with the duality pairing

$$\langle v, w \rangle_\Gamma := v(w),$$

which can be represented as

$$\langle v, w \rangle_\Gamma := \int_\Gamma v(x) w(x) ds_x,$$

if $v \in L^2(\Gamma) \subset H^{-1/2}(\Gamma)$. Now we formulate interior and exterior eigenvalue problems for the Laplacian with Dirichlet and Neumann boundary conditions respectively.

Interior eigenvalue problems:

(D) *Find pairs* $(k, u) \in \mathbb{C}^+ \times H_0^1(\Omega^-) \backslash \{0\}$ *such that*

$$-\Delta u - k^2 u = 0 \text{ in } \Omega^- \tag{6.1}$$

is fulfilled in a weak sense.

(N) *Find pairs* $(k,u) \in \mathbb{C}^+ \times H^1(\Omega^-)\backslash\{0\}$ *such that*

$$\begin{aligned} -\Delta u - k^2 u &= 0 \text{ in } \Omega^-, \\ \gamma_1^- u &= 0 \text{ on } \Gamma \end{aligned} \tag{6.2}$$

is fulfilled in a weak sense.

Exterior eigenvalue problems:

(D) *Find pairs* $(k,u) \in \mathbb{C}^+ \times H^1_{loc}(\Omega^+)\backslash\{0\}$ *such that*

$$\begin{aligned} -\Delta u - k^2 u &= 0 \text{ in } \Omega^+, \\ \gamma_0^+ u &= 0 \text{ on } \Gamma, \\ u &\text{ satisfies a radiation condition} \end{aligned} \tag{6.3}$$

is fulfilled in a weak sense.

(N) *Find pairs* $(k,u) \in \mathbb{C}^+ \times H^1_{loc}(\Omega^+)\backslash\{0\}$ *such that*

$$\begin{aligned} -\Delta u - k^2 u &= 0 \text{ in } \Omega^+, \\ \gamma_1^+ u &= 0 \text{ on } \Gamma, \\ u &\text{ satisfies a radiation condition} \end{aligned} \tag{6.4}$$

is fulfilled in a weak sense.

Definition 6.1. *We call pairs* (k,u) *which are solutions of the above exterior eigenvalue problems for the Laplacian with Dirichlet and Neumann boundary conditions respectively, scattering-resonance pairs. The number* k *is called resonance and* u *is called corresponding resonance function.*

We require for the radiation condition in (6.3) and (6.4) that u can be expanded as

$$u(x) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} a_{n,m} h_n^{(1)}(kr) Y_n^m \left(\frac{x}{|x|} \right) \text{ for } r = |x| > r_0, \tag{6.5}$$

where r_0 is chosen such that $\Omega^- \subset B_{r_0} := \{x : |x| < r_0\}$, $a_{n,m}$ are constants, $h_n^{(1)}$ are the spherical Hankel functions of the first kind, and Y_n^m are the spherical harmonics [22]. Solutions with an expansion of this form are called *outgoing solutions.* For $k \in \mathbb{R}_0^+$ this condition matches with the Sommerfeld radiation condition

$$\lim_{r\to\infty} r \left(\frac{\partial u}{\partial r}(x) - iku(x) \right) = 0, \tag{6.6}$$

where $r = |x|$ and $\frac{\partial}{\partial r} u(x) = \frac{x}{|x|} \cdot \nabla u(x)$ [16, Remark 2.1]. In addition, for $k \in \mathbb{C}^+$ with $\operatorname{Im}(k) \geqslant 0$ one can show that a solution of the above scattering-resonance problems (6.3) and (6.4) which satisfies the Sommerfeld radiation condition (6.6) also fulfils the radiation condition (6.5). Note that this can even be done for $k \in \mathbb{C}$ with $0 \leqslant \arg(k) < \pi$ (s. [19, Chapter 9]).

6.2 Representation formula and boundary integral operators

In this section we state the representation formulas for solutions of the interior and exterior eigenvalue problems and introduce the boundary integral operators for the Helmholtz equation. To start with, the fundamental solution of the Helmholtz equation in $\mathbb{R}^3$ is given by [26, Section 5.4]

$$U_k^*(x,y) = \frac{1}{4\pi}\frac{e^{ik|x-y|}}{|x-y|}.$$

We define for given $k \in \mathbb{C}$ the single layer potential $\mathrm{SL}(k) : H^{-1/2}(\Gamma) \to H^1(\Delta,\Omega^-) \times H^1_{\mathrm{loc}}(\Delta,\Omega^+)$ by

$$(\mathrm{SL}(k)\psi)(x) = \int_\Gamma U_k^*(x,y)\psi(y)ds_y, \quad x \in \mathbb{R}^3\backslash\Gamma,$$

and the double layer potential $\mathrm{DL}(k) : H^{1/2}(\Gamma) \to H^1(\Delta,\Omega^-) \times H^1_{\mathrm{loc}}(\Delta,\Omega^+)$ by

$$(\mathrm{DL}(k)\phi)(x) = \int_\Gamma \frac{\partial}{\partial n_y}U_k^*(x,y)\phi(y)ds_y, \quad x \in \mathbb{R}^3\backslash\Gamma.$$

The representation formulas for solutions of the interior and exterior eigenvalue problems are given by the following theorem.

Theorem 6.2. *Let $k \in \mathbb{C}$. Then the following assertions about the representation of a solution of the Helmholtz equation hold true:*

(i) Every solution $u \in H^1(\Omega^-)$ of the Helmholtz equation in Ω^- has the representation

$$u(x) = (\mathrm{SL}(k)\gamma_1^-u)(x) - (\mathrm{DL}(k)\gamma_0^-u)(x) \text{ for } x \in \Omega^-. \tag{6.7}$$

(ii) Every solution $u \in H^1_{\mathrm{loc}}(\Omega^+)$ of the Helmholtz equation in Ω^+ has the representation

$$u(x) = (-\mathrm{SL}(k)\gamma_1^+u)(x) + (\mathrm{DL}(k)\gamma_0^+u)(x) \text{ for } x \in \Omega^+, \tag{6.8}$$

provided that u satisfies the radiation condition (6.5) for $k \neq 0$.

Proof. For (i) see [19, Theorem 6.10], and for (ii) [28, Corollary 6.5]. □

The lemma below summarizes some properties of the single layer potential and the double layer potential.

Lemma 6.3. *Let $k \in \mathbb{C}$. Then the mappings $\mathrm{SL}(k)$ and $\mathrm{DL}(k)$ are linear and continuous. Moreover, these potentials satisfy the jump relations*

$$\gamma_0^+\,\mathrm{SL}(k)\psi - \gamma_0^-\,\mathrm{SL}(k)\psi = 0, \qquad \gamma_1^+\,\mathrm{SL}(k)\psi - \gamma_1^-\,\mathrm{SL}(k)\psi = -\psi, \tag{6.9}$$

$$\gamma_0^+\,\mathrm{DL}(k)\phi - \gamma_0^-\,\mathrm{DL}(k)\phi = \phi, \qquad \gamma_1^+\,\mathrm{DL}(k)\phi - \gamma_1^-\,\mathrm{DL}(k)\phi = 0, \tag{6.10}$$

for $\psi \in H^{-1/2}(\Gamma)$ and $\phi \in H^{1/2}(\Gamma)$.

Proof. See [19, Theorem 6.11]. □

In addition, we need another property.

Theorem 6.4. *Let $k \in \mathbb{C}$. We define the second-order differential operator $\mathcal{P} := -\Delta - k^2$. Then it holds*

$$\mathcal{P}\,\mathrm{SL}(k)\psi = 0 = \mathcal{P}\,\mathrm{DL}(k)\phi$$

for $\psi \in H^{-1/2}(\Gamma)$ and $\phi \in H^{1/2}(\Gamma)$. Hence, the single layer potential and the double layer potential are solutions of the Helmholtz equation in $\Omega^{\pm}$. In addition, they satisfy (6.5) *for $k \neq 0$.*

Proof. See [19, p. 202] for the first and [28, Theorem 6.4] for the second statement. □

By applying the trace operators to the single layer potential and the double layer potential, we can define the following boundary integral operators:

- single layer boundary integral operator $V(k) : H^{-1/2}(\Gamma) \to H^{1/2}(\Gamma)$

$$V(k) := \frac{1}{2}[\gamma_0^+ \,\mathrm{SL}(k) + \gamma_0^- \,\mathrm{SL}(k)],$$

- adjoint double layer boundary integral operator $K'(k) : H^{-1/2}(\Gamma) \to H^{-1/2}(\Gamma)$

$$K'(k) := \frac{1}{2}[\gamma_1^+ \,\mathrm{SL}(k) + \gamma_1^- \,\mathrm{SL}(k)],$$

- double layer boundary integral operator $K(k) : H^{1/2}(\Gamma) \to H^{1/2}(\Gamma)$

$$K(k) := \frac{1}{2}[\gamma_0^+ \,\mathrm{DL}(k) + \gamma_0^- \,\mathrm{DL}(k)],$$

- hypersingular boundary integral operator $D(k) : H^{1/2}(\Gamma) \to H^{-1/2}(\Gamma)$

$$-D(k) := \frac{1}{2}[\gamma_1^+ \,\mathrm{DL}(k) + \gamma_1^- \,\mathrm{DL}(k)].$$

Some of the properties of the above defined boundary integral operators are summarized in the following lemma.

Lemma 6.5. *Let $k \in \mathbb{C}$. Then the mappings $V(k)$, $K'(k)$, $K(k)$, and $D(k)$ are linear and continuous. Moreover, the following operators are compact:*

$$\begin{aligned} V(k) - V(0) : &\quad H^{-1/2}(\Gamma) \to H^{1/2}(\Gamma), \\ K'(k) - K'(0) : &\quad H^{-1/2}(\Gamma) \to H^{-1/2}(\Gamma), \\ K(k) - K(0) : &\quad H^{1/2}(\Gamma) \to H^{1/2}(\Gamma), \\ D(k) - D(0) : &\quad H^{1/2}(\Gamma) \to H^{-1/2}(\Gamma). \end{aligned}$$

Proof. For the first assertion see [19, Theorem 6.11] and for the second [25, Lemma 3.9.8]. Note that in [25] the assertion is only showed for real k, but according to [30] it can be generalized for complex k. □

Furthermore, the following property of $V(k)$ and $D(k)$ will be important later.

Lemma 6.6. *The following two assertions hold true:*

(i) The mapping $\mathbb{C} \ni k \mapsto \mathrm{V}(k) \in \mathcal{L}(H^{-1/2}(\Gamma), H^{1/2}(\Gamma))$ *is holomorphic, i.e., for every* $k \in \mathbb{C}$ *there exists an element* $V_{*,k} \in \mathcal{L}(H^{-1/2}(\Gamma), H^{1/2}(\Gamma))$ *such that*

$$\lim_{h\to 0} \left\| \frac{V(k+h) - V(k)}{h} - V_{*,k} \right\|_{\mathcal{L}(H^{-1/2}(\Gamma),H^{1/2}(\Gamma))} = 0,$$

where $\|\cdot\|_{\mathcal{L}(H^{-1/2}(\Gamma),H^{1/2}(\Gamma))}$ *is the operator norm.*

(ii) The mapping $\mathbb{C} \ni k \mapsto \mathrm{D}(k) \in \mathcal{L}(H^{1/2}(\Gamma), H^{-1/2}(\Gamma))$ *is holomorphic, i.e., for every* $k \in \mathbb{C}$ *there exists an element* $D_{*,k} \in \mathcal{L}(H^{1/2}(\Gamma), H^{-1/2}(\Gamma))$ *such that*

$$\lim_{h\to 0} \left\| \frac{D(k+h) - D(k)}{h} - D_{*,k} \right\|_{\mathcal{L}(H^{1/2}(\Gamma),H^{-1/2}(\Gamma))} = 0.$$

Proof. See [30, Lemma 5.1.1] for (i) and [30, Lemma 5.1.2] for (ii). □

By inserting the jump conditions (6.9) and (6.10) into the definitions of the boundary integral operators, we obtain the following identities:

$$\gamma_0^{\pm}\,\mathrm{SL}(k)\psi = V(k)\psi, \qquad \gamma_1^{\pm}\,\mathrm{SL}(k)\psi = \left[\mp\frac{1}{2}I + K'(k)\right]\psi, \tag{6.11}$$

$$\gamma_0^{\pm}\,\mathrm{DL}(k)\phi = \left[\pm\frac{1}{2}I + K(k)\right]\phi, \qquad \gamma_1^{\pm}\,\mathrm{DL}(k)\phi = -D(k)\phi. \tag{6.12}$$

6.3 Boundary integral formulations for the Dirichlet eigenvalue problems

In this section a direct and an indirect boundary integral formulation of the eigenvalue problems (6.1) and (6.3) are given. In the sequel we will only focus on one of the obtained boundary integral formulations, and we will show that the eigenvalues with negative imaginary part are the resonances of the scattering-resonance problem (6.3) and the eigenvalues with non-negative imaginary part are the eigenvalues of (6.1).

Direct ansatz based on the representation formula:

If we apply γ_0^- and γ_1^- to the representation formula (6.7), we obtain

$$\begin{pmatrix} \gamma_0^- u \\ \gamma_1^- u \end{pmatrix} = \begin{pmatrix} \frac{1}{2}I - K(k) & V(k) \\ D(k) & \frac{1}{2}I + K'(k) \end{pmatrix} \begin{pmatrix} \gamma_0^- u \\ \gamma_1^- u \end{pmatrix},$$

and by applying γ_0^+ and γ_1^+ to the representation formula (6.8), we get

$$\begin{pmatrix} \gamma_0^+ u \\ \gamma_1^+ u \end{pmatrix} = \begin{pmatrix} \frac{1}{2}I + K(k) & -V(k) \\ -D(k) & \frac{1}{2}I - K'(k) \end{pmatrix} \begin{pmatrix} \gamma_0^+ u \\ \gamma_1^+ u \end{pmatrix}.$$

Remark. *Note that that these identities are called Calderon identities.*

With $\gamma_0^- u = 0 = \gamma_0^+ u$, it follows that every eigenfunction u of (6.1) or (6.3) can be represented as

$$u(x) = \begin{cases} (\mathrm{SL}(k)\gamma_1^- u)(x), & \text{for } x \in \Omega^-, \\ (-\mathrm{SL}(k)\gamma_1^+ u)(x), & \text{for } x \in \Omega^+, \end{cases}$$

and we obtain with $\psi^\pm := \gamma_1^\pm u \in H^{-1/2}(\Gamma)$ the boundary integral formulations

$$\begin{aligned} V(k)\psi^\pm &= 0, \\ \left[\pm\frac{1}{2}I + K'(k)\right]\psi^\pm &= 0. \end{aligned} \tag{6.13}$$

Indirect ansatz based on the double layer potential ansatz:

We make the following ansatz for the solution u by using the double layer potential

$$u(x) = (\mathrm{DL}(k)\phi)(x) \text{ for } x \in \Omega^- \cup \Omega^+,$$

where $\phi \in H^{1/2}(\Gamma)\backslash\{0\}$. By applying the Dirichlet traces $\gamma_0^\pm$ to this ansatz, we obtain the following boundary integral formulation

$$\left[\pm\frac{1}{2}I + K(k)\right]\phi = 0.$$

Note that it is also possible to choose an indirect ansatz for u based on the single layer potential. However, this ansatz coincides with the representation formula for u.

Characterization of the eigenvalues of (6.13)**:**

Since (6.13) is most appropriate for a conforming Galerkin approximation, we will only focus on this boundary integral formulation in the sequel.

Due to Lemma 6.6, we know that $k \mapsto V(k)$ is a holomorphic map from $\mathbb{C}$ into the space $\mathcal{L}(H^{-1/2}(\Gamma), H^{1/2}(\Gamma))$. Therefore, (6.13) has the form (2.1) in a "continuous level", i.e.,

$$T(\lambda)v = 0,$$

where $T \in \mathcal{H}(\mathcal{D}, \mathcal{L}(X, Y))$ for Hilbert spaces X, Y. By discretizing T we get a NEP of exactly the form (2.1).

For the characterization of the eigenpairs of (6.13) we need the following two statements. The first one concerns the eigenvalues of the interior eigenvalue problems (6.1) and (6.2).

Lemma 6.7. *Let $k \in \mathbb{C}^+$ be an eigenvalue of* (6.1) *or* (6.2)*. Then it follows $k \in \mathbb{R}^+$.*

Proof. Since $k \in \mathbb{C}^+$ is an eigenvalue of either (6.1) or (6.2), there exists a corresponding eigenvector $u \in H_0^1(\Omega^-)\backslash\{0\}$ or $u \in H^1(\Omega^-)\backslash\{0\}$ respectively, such that (6.1) or (6.2) is fulfilled in the weak sense, i.e.,

$$\int_{\Omega^-} \nabla u \cdot \overline{\nabla v} dx - k^2 \int_{\Omega^-} u\overline{v} dx = 0$$

for all $\overline{v} \in H^1(\Omega^-)\backslash\{0\}$ or $\overline{v} \in H_0^1(\Omega^-)\backslash\{0\}$ respectively. If we set $v = u$, it follows directly $k^2 > 0$ and due to $k \in \mathbb{C}^+$, we get $k \in \mathbb{R}^+$. □

The second statement is the following uniqueness theorem for the solutions of homogeneous exterior boundary value problems for the Helmholtz equation [19, Theorem 9.10].

Theorem 6.8. *Let $k \in \mathbb{C}^+$ with $\operatorname{Im}(k) \geqslant 0$ and let $u \in H^1_{\text{loc}}(\Omega^+)$ be a solution of the homogeneous exterior mixed boundary value problem with $\Gamma = \Gamma_D \cup \Gamma_N$*

$$\begin{aligned} -\Delta u - k^2 u &= 0 \text{ in } \Omega^+, \\ \gamma_0^+ u &= 0 \text{ on } \Gamma_D, \\ \gamma_1^+ u &= 0 \text{ on } \Gamma_N. \end{aligned}$$

If u satisfies the Sommerfeld radiation condition (6.6)*, then $u = 0$ in Ω^+.*

Note that this theorem is formulated in [19] even for $k \in \mathbb{C}$ with $0 \leqslant \arg(k) < \pi$. The next theorem characterizes the eigenpairs of the boundary integral formulation (6.13).

Theorem 6.9. *Let $(k, \psi) \in \mathbb{C}^+ \times H^{-1/2}(\Gamma)\backslash\{0\}$ be an eigenpair of*

$$V(k)\psi = 0.$$

Then the following statements hold true:

(i) If $\operatorname{Im}(k) < 0$, then (k, u) with $u = (-\operatorname{SL}(k)\psi)|_{\Omega^+}$ is a scattering-resonance pair of (6.3).

(ii) If $\operatorname{Im}(k) \geqslant 0$, then $k \in \mathbb{R}_0^+$, and (k, u) with $u = (\operatorname{SL}(k)\psi)|_{\Omega^-}$ is an eigenpair of (6.1)*. Furthermore, it holds $\operatorname{SL}(k)\psi = 0$ in Ω^+.*

Proof. (i): We suppose that $\operatorname{Im}(k) < 0$. Firstly, we define $v := -\operatorname{SL}(k)\psi$ in $\Omega^- \cup \Omega^+$. Then it follows from (6.11) that $\gamma_0^{\pm} v = -V(k)\psi = 0$. This implies that $v = 0$ in Ω^- because otherwise (k, v) would be a solution of (6.1) due to Theorem 6.4. However, it follows $k \in \mathbb{R}_0^+$ for this problem by Lemma 6.7, which is a contradiction to $\operatorname{Im}(k) < 0$. From $v = 0$ in Ω^- we obtain $\gamma_1^- v = 0$, and further $\gamma_1^+ v = \psi \neq 0$ by (6.9). Therefore, $v \neq 0$ in Ω^+. The statement follows with $u := v|_{\Omega^+}$.

(ii) Now we assume that $\operatorname{Im}(k) \geqslant 0$. We define $v := \operatorname{SL}(k)\psi$ in $\Omega^- \cup \Omega^+$. By (6.11) it follows $\gamma_0^{\pm} v = V(k)\psi = 0$. Due to Theorem 6.4, we know that the single layer potential is a solution of the Helmholtz equation, and we get $v = 0$ in Ω^+ by Theorem 6.8. This implies $\gamma_1^+ v = 0$. With the jump condition (6.9) we obtain $\gamma_1^- v = \psi \neq 0$, and therefore $v \neq 0$ in Ω^-. This implies that $(k, (\operatorname{SL}(k)\psi)|_{\Omega^-})$ is an eigenpair of (6.1). By Lemma 6.7 we obtain $k \in \mathbb{R}_0^+$. □

6.4 Boundary integral formulations for the Neumann eigenvalue problems

In this section we derive boundary integral formulations for the eigenvalue problems (6.2) and (6.4) with Neumann boundary conditions. Similarly as in the previous section, we will focus on one of the obtained formulations and characterize the eigenpairs.

Direct ansatz based on the representation formula:

With $\gamma_1^- u = 0 = \gamma_1^+ u$, it follows that every eigenfunction u of (6.2) or (6.4) can be represented as

$$u(x) = \begin{cases} (-\operatorname{DL}(k)\gamma_0^- u)(x), & \text{for } x \in \Omega^-, \\ (\operatorname{DL}(k)\gamma_0^+ u)(x), & \text{for } x \in \Omega^+, \end{cases}$$

and we obtain with $\phi^\pm := \gamma_0^\pm u \in H^{1/2}(\Gamma)$ the boundary integral formulations

$$\begin{aligned} \left[\mp\frac{1}{2}I + K(k)\right]\phi^\pm &= 0, \\ D(k)\phi^\pm &= 0. \end{aligned} \tag{6.14}$$

Indirect ansatz based on the single layer potential ansatz:

We make the following ansatz for the solution u by using the single layer potential

$$u(x) = (\operatorname{SL}(k)\psi)(x) \text{ for } x \in \Omega^- \cup \Omega^+,$$

where $\psi \in H^{-1/2}(\Gamma)\backslash\{0\}$. By applying the Neumann traces $\gamma_1^\pm$ to this ansatz, we obtain the following boundary integral equations

$$\left[\mp\frac{1}{2}I + K(k)\right]\phi = 0.$$

Note that it is also possible to choose an indirect ansatz for u based on the double layer potential. However, this ansatz coincides with the representation formula for u. The next theorem [28, Theorem 2.1, Proposition 2.2] characterizes the eigenpairs of the boundary integral equation (6.14).

Theorem 6.10. *Let* $(k, \phi) \in \mathbb{C}^+ \times H^{1/2}(\Gamma)\backslash\{0\}$ *be an eigenpair of*

$$D(k)\phi = 0.$$

Then the following statements hold true:

(i) *If* $\operatorname{Im}(k) < 0$*, then* (k, u) *with* $u = (\mathrm{DL}(k)\phi)|_{\Omega^+}$ *is a scattering-resonance pair of* (6.4).

(ii) *If* $\operatorname{Im}(k) \geqslant 0$*, then* $k \in \mathbb{R}_0^+$*, and* (k, u) *with* $u = (\mathrm{DL}(k)\phi)|_{\Omega^-}$ *is an eigenpair of* (6.2). *Furthermore, it holds* $\mathrm{DL}(k)\phi = 0$ *in* Ω^+*.*

Proof. The proof works analogously to the proof in Theorem 6.9 by defining $v := \mathrm{DL}(k)\phi$ in $\Omega^- \cup \Omega^+$ and interchanging $\gamma_0^\pm$ and $\gamma_1^\pm$. For a more detailed proof see [28]. □

7 Error estimates for the Galerkin approximation of boundary integral formulations of Laplacian eigenvalue problems

In this chapter we consider the Galerkin approximation of the boundary integral formulations (6.13) and (6.14), i.e.,

$$V(k)\psi = 0, \quad D(K)\phi = 0,$$

and we give error estimates for the approximated eigenvalues and the corresponding eigenvectors. These eigenvalue problems are considered as eigenvalue problems for holomorphic Fredholm operator-valued functions. For the Galerkin approximation of such eigenvalue problems there exists a convergence theory [14, 15].

In the first section we will introduce Fredholm operators in general, then we will state convergence results for the Galerkin approximation of eigenvalue problems for holomorphic Fredholm operator-valued functions, and finally we will show that these results can also be applied to V and D. Our main references for this chapter are [27] and [30] .

7.1 Fredholm operators

In this section we would like to give the general definition of Fredholm operators and state an important property concerning compact perturbations of Fredholm operators.

Definition 7.1. *Let X, Y be Banach spaces. Then a bounded, linear operator $A \in \mathcal{L}(X, Y)$ is called Fredholm operator, if $\dim(\ker A) < \infty$ and $\operatorname{codim}(\operatorname{ran} A, Y) < \infty$. The number*

$$\operatorname{ind} A = \dim(\ker A) - \operatorname{codim}(\operatorname{ran} A, Y)$$

is called Fredholm index of A.

Remark. *The codimension $\operatorname{codim}(\operatorname{ran} A, Y)$ of $\operatorname{ran} A$ in Y is defined as the dimension of the factor space $\dim(Y/\operatorname{ran} A)$.*

There holds the following property for Fredholm operators [19, Theorem 2.26]:

Theorem 7.2. *Let X, Y be Banach spaces and let $A \in \mathcal{L}(X, Y)$ be a Fredholm operator. Further, let $C \in \mathcal{L}(X, Y)$ be a compact operator. Then $A + C$ is a Fredholm operator and* $\operatorname{ind}(A + C) = \operatorname{ind} A$.

Remark. *Recall that an operator $C \in \mathcal{L}(X, Y)$ is called compact if for each bounded sequence $(x_n)_n \subset X$ the sequence $(Cx_n)_n \subset Y$ has a convergent subsequence.*

7.2 Convergence results for the Galerkin approximation of eigenvalue problems for holomorphic Fredholm operator-valued functions

In the sequel let X be a Hilbert space over $\mathbb{C}$ with inner product $(\cdot, \cdot)_X$, and let $\mathcal{D} \subset \mathbb{C}$ be some open and connected subset of $\mathbb{C}$. Furthermore, we assume that $S \in \mathcal{H}(\mathcal{D}, \mathcal{L}(X, X))$ is a Fredholm operator function with index zero, i.e., $S(\lambda)$ is a Fredholm operator of index zero for all fixed $\lambda \in \mathcal{D}$. We consider the following eigenvalue problem: *Find $\lambda \in \mathcal{D}$ and $v \in X \backslash \{0\}$ such that*

$$S(\lambda)v = 0. \tag{7.1}$$

In order to compute approximated eigenpairs of (7.1), we use the Galerkin approximation. For the Galerkin variational formulation we choose a sequence $\{X_n\}_n \subset X$ of conforming finite-dimensional subspaces of X. The variational formulation reads as follows: *Find $\lambda_n \in \mathcal{D}$ and $v_n \in X_n \backslash \{0\}$ such that*

$$(S(\lambda_n)v_n, x_n)_X = 0 \quad \forall x_n \in X_n. \tag{7.2}$$

Next we consider the orthogonal projection $\Pi_n : X \to X_n$. Then the orthogonality relation

$$(S(\lambda_n)v_n - \Pi_n S(\lambda_n)v_n, x_n)_X = 0 \quad \forall x_n \in X_n$$

implies that the variational problem (7.2) is equivalent to the variational problem: *Find $\lambda_n \in \mathcal{D}$ and $v_n \in X_n \backslash \{0\}$ such that*

$$(\Pi_n S(\lambda_n)v_n, x_n)_X = 0 \quad \forall x_n \in X_n,$$

and therefore to the problem

$$\Pi_n S(\lambda_n)v_n = 0. \tag{7.3}$$

Concerning the eigenvalues of the operator S, one can show the following important property [30, Theorem 3.2.2]:

Lemma 7.3. *Let S be defined as above and let us assume that for the resolvent set of S there holds*

$$\rho(S) := \{\lambda \in \mathcal{D} : \exists S(\lambda)^{-1} \in \mathcal{L}(X, X)\} \neq \emptyset.$$

Then the spectrum

$$\sigma(S) := \mathcal{D} \backslash \rho(S)$$

does not contain any cluster points and every $\lambda \in \sigma(S)$ is an eigenvalue.

The following theorem provides the desired convergence result [27, Theorem 4.2].

Theorem 7.4. *Let S be a Fredholm operator function with index zero. Further we assume that for every $\lambda \in \mathcal{D}$ the operator $S(\lambda)$ has a splitting into a X-elliptic operator $S_0 \in \mathcal{L}(X, X)$ and a compact operator $C(\lambda) \in \mathcal{L}(X, X)$, i.e.,*

$$S(\lambda) = S_0 + C(\lambda).$$

Moreover, let $\{X_n\}_n \subset X$ be a sequence of conforming finite-dimensional subspaces such that the condition

$$\lim_{n\to\infty} \inf_{x_n \in X_n} \|x - x_n\|_X = 0 \quad \forall x \in X \tag{7.4}$$

is satisfied. Then the following assertions hold:

(i) Let $\lambda_0 \in \sigma(S)$ be an eigenvalue. Then there exists a sequence $(\lambda_n)_n$ of eigenvalues of the Galerkin eigenvalue problem (7.3) *such that*

$$\lim_{n\to\infty} \lambda_n = \lambda_0.$$

(ii) Let $\{(\lambda_n, v_n)\}_n$ be a sequence of eigenpairs of the Galerkin eigenvalue problem (7.3) *with $\|v_n\|_X = 1$. Then it holds*

$$\lim_{n\to\infty} \lambda_n = \lambda_0 \in \sigma(S).$$

Moreover, there exists a subsequence of $(v_n)_n$ which convergences to an eigenvector corresponding to λ_0.

An error estimate for an approximated eigenpair is given by the following theorem [27, Theorem 4.4].

Theorem 7.5. *Let the assumptions of Theorem 7.4 hold true. Furthermore, let $\mathcal{D}_0 \subset \mathcal{D}$ be a compact subset such that there are no eigenvalues on the boundary of this subset, i.e., $\partial\mathcal{D}_0 \subset \rho(S)$, and $\mathcal{D}_0 \cap \sigma(S) = \{\lambda_0\}$. Then there exist constants $C > 0$ and $N \in \mathbb{N}$ such that for all $n \geqslant N$ and $\lambda_n \in \sigma(\Pi_n S) \cap \mathcal{D}_0$*

$$|\lambda_n - \lambda_0| \leqslant C(\delta_n \delta_n^*)^{1/M_{\lambda_0}(S)},$$

where

$$\delta_n := \max_{v_0 \in G_{\lambda_0}(S), \|v_0\|_X \leqslant 1} \inf_{v_n \in X_n} \|v_0 - v_n\|_X, \quad \delta_n^* := \max_{w_0 \in G_{\overline{\lambda_0}}(S^*), \|w_0\|_X \leqslant 1} \inf_{v_n \in X_n} \|w_0 - v_n\|_X,$$

and $G_{\overline{\lambda_0}}(S^*)$ is the generalized eigenspace corresponding to the adjoint eigenvalue problem

$$S^*(\overline{\lambda_0}) w_0 := [S(\lambda_0)]^* \, w_0 = 0.$$

Moreover, we get the following estimate for the eigenvectors v_n corresponding to λ_n with $\|v_n\|_X = 1$: There exists a constant $c > 0$ such that for all $n \geqslant N$

$$\inf_{v_0 \in \ker S(\lambda_0)} \|v_n - v_0\|_X \leqslant c \left(|\lambda_n - \lambda_0| + \max_{z_0 \in \ker S(\lambda_0), \|z_0\|_X \leqslant 1} \inf_{z_n \in X_n} \|z_0 - z_n\|_X \right).$$

7.3 Application of the results to V and D

In this section we show how Theorem 7.4 and Theorem 7.5 can be applied to the boundary integral formulations corresponding to V and D. For that, we need the following lemma.

Lemma 7.6. *Let $A \in \mathcal{H}(\mathcal{D}, \mathcal{L}(X, X))$, where X and $\mathcal{D}$ are defined as in the previous section. Let us assume that $A(\lambda)$ has a splitting into a X-elliptic operator $A_0 \in \mathcal{L}(X, X)$ and a compact operator $C(\lambda) \in \mathcal{L}(X, X)$, i.e.,*

$$A(\lambda) = A_0 + C(\lambda),$$

for all $\lambda \in \mathcal{D}$. Then each $A(\lambda)$ is a Fredholm operator with index zero and thus A is a Fredholm operator function with index zero.

Proof. Since A_0 is X-elliptic, there exists a constant $c_{A_0} > 0$ such that

$$(A_0 x, x)_X \geqslant c_{A_0} \|x\|_X^2$$

for all $x \in X$. Due to the Lemma of Lax-Milgram [25, Lemma 2.1.51], A_0 is invertible. Hence, $\dim(\ker A_0) = 0$. From $\dim(X/X) = 0$ it follows that A_0 is a Fredholm operator with index zero. With Theorem 7.2 we obtain that $A(\lambda)$ is Fredholm with index zero for every $\lambda \in \mathcal{D}$, and thus A is a Fredholm operator function with index zero. □

As already mentioned above, our goal is to apply the convergence results of Section 7.2 to the boundary integral formulations

$$V(k)\psi = 0, \quad D(k)\phi = 0. \tag{7.5}$$

However, the problem which occurs has to do with the mapping properties

$$V \in \mathcal{H}(\mathcal{D}, \mathcal{L}(H^{-1/2}(\Gamma), H^{1/2}(\Gamma))), \quad D \in \mathcal{H}(\mathcal{D}, \mathcal{L}(H^{1/2}(\Gamma), H^{-1/2}(\Gamma))),$$

since we have assumed for S defined in the previous section that $S \in \mathcal{H}(\mathcal{D}, \mathcal{L}(X, X))$. In the sequel a formulation for (7.5) is derived which allows the direct application of the results in Section 7.2. A variational formulation of (7.5) reads as follows: *Find $k \in \mathcal{D}$, $\psi \in H^{-1/2}(\Gamma)$ and $\phi \in H^{1/2}(\Gamma)$ respectively, such that*

$$\langle \overline{u}, V(k)\psi \rangle_{H^{-1/2}(\Gamma)\times H^{1/2}(\Gamma)} = 0 \quad \forall u \in H^{-1/2}(\Gamma), \tag{7.6}$$

$$\langle \overline{D(k)\phi}, v \rangle_{H^{-1/2}(\Gamma)\times H^{1/2}(\Gamma)} = 0 \quad \forall v \in H^{1/2}(\Gamma). \tag{7.7}$$

Note that

$$(\cdot, \cdot)_{H^{-1/2}(\Gamma)\times H^{1/2}(\Gamma)} := \langle \bar{\cdot}, \cdot \rangle_{H^{-1/2}(\Gamma)\times H^{1/2}(\Gamma)}$$

is a sequilinear form, i.e., a form which is antilinear in the first and linear in the second argument.

7.3.1 Representation of sesquilinear forms as inner products in Hilbert spaces

In this subsection we show how sesquilinear forms can be represented as inner products in Hilbert spaces, which we will utilize for an appropriate formulation of the eigenvalue problems (7.5). Let X be a Hilbert space over $\mathbb{C}$ with inner product $(\cdot, \cdot)_X$. Remember that an inner product on X is a complex-valued form which is antilinear in the first argument, linear in the second argument, and positive definite, i.e., the following properties are fulfilled for all $x, y, z \in X$ and $\lambda \in \mathbb{C}$:

- $(\lambda x + y, z)_X = \overline{\lambda}(x, z)_X + (y, z)_X$,
- $(x, \lambda y + z)_X = \lambda(x, y)_X + (x, z)_X$,
- $(x, y)_X = \overline{(y, x)}_X$,
- $(x, x)_X \geqslant 0$ with equality if and only if $x = 0$.

First, we recap some general definitions. The dual space of X is defined as

$$X' := \{f : X \to \mathbb{C} \text{ linear and continuous}\}.$$

For $f \in X'$ and $x \in X$ the duality pairing is given by

$$\langle f, x \rangle_{X' \times X} := f(x),$$

with the related norm

$$\|f\|_{X'} := \sup_{0 \neq x \in X} \frac{|\langle f, x \rangle_{X' \times X}|}{\|x\|_X}.$$

If the bidual space X'' is identified with X, then we set

$$\langle x, f \rangle_{X \times X'} = \langle f, x \rangle_{X' \times X}. \tag{7.8}$$

Moreover, we define the complex conjugate of a linear and bounded functional $f \in X'$ by

$$\langle \overline{f}, x \rangle_{X' \times X} := \overline{\langle f, \overline{x} \rangle}_{X' \times X}. \tag{7.9}$$

Let $u \in X$. Then we consider the linear and bounded functional $Ju \in X'$ given by

$$\langle Ju, v \rangle_{X' \times X} = (u, v)_X \quad \forall v \in X. \tag{7.10}$$

It is obvious that this functional is linear. The boundedness follows by using the Cauchy-Schwarz inequality

$$|\langle Ju, v \rangle_{X' \times X}| = |(u, v)_X| \leqslant \|u\|_X \|v\|_X.$$

Note that the map $J : X \to X'$ is conjugate-linear due to the definition of the inner product, and surjective. From the following Riesz representation theorem [19, Theorem 2.30] it follows that J is also injective and isometric, i.e., $\|Ju\|_{X'} = \|u\|_X$.

Theorem 7.7 (Riesz representation theorem)**.** *Let X be a Hilbert space and let $f \in X'$. Then there exists a uniquely determined $u \in X$ such that*

$$\langle f, v \rangle_{X' \times X} = (u, v)_X \quad \forall v \in X.$$

Moreover it holds $\|f\|_{X'} = \|u\|_X$.

We define the mapping $\iota_X : X \to X'$ by

$$\iota_X u := \overline{Ju}.$$

Note that this mapping is an isomorphism. It follows directly from (7.10) that

$$\langle \overline{\iota_X u}, v \rangle_{X' \times X} = \langle Ju, v \rangle_{X' \times X} = (u, v)_X \quad \forall u, v \in X.$$

Further, we obtain with $u = \iota_X^{-1} w$

$$(\iota_X^{-1} w, v)_X = \langle \overline{w}, v \rangle_{X' \times X} \quad \forall w \in X', \forall v \in X,$$

and

$$(w, (\iota_X^{-1})^* v)_{X'} = \langle \overline{w}, v \rangle_{X' \times X} \quad \forall w \in X', \forall v \in X.$$

If we now define

$$\mathcal{I} := (\iota_X^{-1})^* : X \to X',$$

we get the desired relationship between the inner product in X and the sesquilinear forms

$$(w, \mathcal{I} v)_{X'} = \langle \overline{w}, v \rangle_{X' \times X} \quad \forall w \in X', \forall v \in X, \tag{7.11}$$
$$(\mathcal{I}^* w, v)_X = \langle \overline{w}, v \rangle_{X' \times X} \quad \forall w \in X', \forall v \in X. \tag{7.12}$$

7.3.2 Application of the derived relationship to V and D

In this subsection we first rewrite the equations (7.6) and (7.7) by using the relationships (7.11) and (7.12), and then we show that the results of Section 7.2 can be applied. To start with, by (7.11) we can rewrite (7.6) with $X = H^{1/2}(\Gamma)$, $X' = H^{-1/2}(\Gamma)$, $w = u \in H^{-1/2}(\Gamma)$, and $v = V(k)\psi \in H^{1/2}(\Gamma)$ in the following way

$$(u, \mathcal{I}V(k)\psi)_{H^{-1/2}(\Gamma)} = \langle \overline{u}, V(k)\psi \rangle_{H^{-1/2}(\Gamma)\times H^{1/2}(\Gamma)} = 0 \quad \forall u \in H^{-1/2}(\Gamma). \tag{7.13}$$

Similarly, we get with (7.12) by setting $w = D(k)\phi \in H^{-1/2}(\Gamma)$

$$(\mathcal{I}^* D(k)\phi, v)_{H^{1/2}(\Gamma)} = \langle \overline{D(k)\phi}, v \rangle_{H^{-1/2}(\Gamma)\times H^{1/2}(\Gamma)} = 0 \quad \forall v \in H^{1/2}(\Gamma). \tag{7.14}$$

Due to the fact that $\mathcal{I}$ and $\mathcal{I}^*$ are isomorphisms, the following theorem follows directly [30, Theorem 5.1.3 ii), Theorem 5.1.4 ii)].

Theorem 7.8. *Let $\mathcal{I} : H^{1/2}(\Gamma) \to H^{-1/2}(\Gamma)$ and $\mathcal{I}^* : H^{-1/2}(\Gamma) \to H^{1/2}(\Gamma)$ be given as above. Then it holds:*

(i) The spectra of V and $\mathcal{I}V$ coincide. Moreover, the kernels of $V(k)$ and $\mathcal{I}V(k)$ coincide for every arbitrarily chosen $k \in \mathbb{C}$. For every $k \in \sigma(V)$ the maximal length of Jordan chains corresponding to k, and the algebraic multiplicity are the same for V and $\mathcal{I}V$.

(ii) The spectra of D and $\mathcal{I}^*D$ coincide. Moreover, the kernels of $D(k)$ and $\mathcal{I}^*D(k)$ coincide for every arbitrarily chosen $k \in \mathbb{C}$. For every $k \in \sigma(D)$ the maximal length of Jordan chains corresponding to k, and the algebraic multiplicity are the same for D and $\mathcal{I}^*D$.

The next theorem [30, Theorem 5.1.3 i), Theorem 5.1.4 i)] shows that $\mathcal{I}V$ and $\mathcal{I}^*D$ are Fredholm operator functions with index zero.

Theorem 7.9. *Let $\mathcal{I} : H^{1/2}(\Gamma) \to H^{-1/2}(\Gamma)$ and $\mathcal{I}^* : H^{-1/2}(\Gamma) \to H^{1/2}(\Gamma)$ be given as above. Then it holds:*

(i) The operator function

$$\begin{aligned} \mathcal{I}V : \mathbb{C} &\to \mathcal{L}(H^{-1/2}(\Gamma), H^{-1/2}(\Gamma)), \\ k &\mapsto \mathcal{I}V(k), \end{aligned}$$

is holomorphic and defines a Fredholm operator function of index zero.

(ii) The operator function

$$\begin{aligned} \mathcal{I}^*D : \mathbb{C} &\to \mathcal{L}(H^{1/2}(\Gamma), H^{1/2}(\Gamma)), \\ k &\mapsto \mathcal{I}^*D(k), \end{aligned}$$

is holomorphic and defines a Fredholm operator function of index zero.

Proof. We follow the proofs in [30].
(i) The holomorphy of $\mathcal{I}V$ is a direct consequence of the holomorphy of V. For every arbitrarily chosen $k \in \mathbb{C}$ we can write $\mathcal{I}V(k)$ in the following way

$$\mathcal{I}V(k) = \mathcal{I}V(0) + \mathcal{I}(V(k) - V(0)).$$

One can show that $V(0)$ is $H^{-1/2}(\Gamma)$-elliptic, i.e., there exists a constant c_{V_0} such that

$$\langle \overline{w}, V(0)w \rangle_{H^{-1/2}(\Gamma)\times H^{1/2}(\Gamma)} \geqslant c_{V_0} \|w\|^2_{H^{-1/2}(\Gamma)} \quad \forall w \in H^{-1/2}(\Gamma)$$

[19, Corollary 8.13]. Then it follows with (7.11) that $\mathcal{I}V(0)$ is $H^{-1/2}(\Gamma)$-elliptic. Furthermore, $V(k) - V(0)$ is compact by Lemma 6.5. This implies that $\mathcal{I}(V(k) - V(0))$ is compact, too. Lemma 7.6 shows that $\mathcal{I}V(k)$ is a Fredholm operator with index zero and since k was chosen arbitrarily, $\mathcal{I}V$ is a Fredholm operator function with index zero.
(ii) The holomorphy of $\mathcal{I}^*D$ is a direct consequence of the holomorphy of D. For every arbitrarily chosen $k \in \mathbb{C}$ we can write $\mathcal{I}^*D(k)$ in the following way

$$\mathcal{I}^*D(k) = \mathcal{I}^*\tilde{D}(0) + \mathcal{I}^*(D(k) - \tilde{D}(0)),$$

where $\tilde{D}(0)$ is given by

$$\tilde{D}(0) := D(0) + \alpha\langle \overline{1}_\Gamma, \cdot \rangle_{H^{-1/2}(\Gamma)\times H^{1/2}(\Gamma)} \overline{1}_\Gamma,$$

with $\alpha \in \mathbb{R}^+$, and $\overline{1}_\Gamma \in H^{-1/2}(\Gamma)$ is a functional defined by

$$\tilde{1}_\Gamma(v) := \int_\Gamma v(x) ds_x$$

for all $v \in H^{1/2}(\Gamma)$. Note that in contrast to $D(0)$, the operator $\tilde{D}(0)$ is $H^{1/2}(\Gamma)$-elliptic [26, p. 177]. The reason why the operator $D(0)$ is not $H^{1/2}(\Gamma)$-elliptic is that the constants lie in the kernel of this operator. However, with (7.12) it follows that $\mathcal{I}^*D(0)$ is $H^{1/2}(\Gamma)$-elliptic, too. The operator $D(k) - \tilde{D}(0)$ is compact, since $D(k) - D(0)$ is compact by Lemma 6.5 and the operator given by $v \mapsto \langle \overline{1}_\Gamma, \cdot \rangle_{H^{-1/2}(\Gamma)\times H^{1/2}(\Gamma)} \overline{1}_\Gamma$ is also compact. It is known that the sum of two compact operators is again compact. Hence, $\mathcal{I}^*(D(k) - \tilde{D}(0))$ is also compact and with Theorem 7.6 the assertion follows. □

This theorem allows us to apply Theorem 7.4 and Theorem 7.5 to $\mathcal{I}V$ and $\mathcal{I}^*D$. In the sequel we restrict to the special case where the maximal length of Jordan chains is equal to one. The following relationship between δ_n and δ_n^* holds:

Theorem 7.10. *Let $\lambda_0 \in \mathcal{D}_0 \cap \sigma(\mathcal{IV})$, where $\mathcal{D}_0$ is defined as in Theorem 7.5. Further, we assume that $M_{\lambda_0}(\mathcal{I}V) = 1$. Then it holds*

$$\delta_n = \delta_n^*$$

in Theorem 7.5. The same holds true for $\mathcal{I}^*D$ instead of $\mathcal{I}V$.

Remark. *For real eigenvalues λ_0 one can show that the maximal length of any Jordan chain is equal to one [27, Lemma 5.3].*

In order to prove this theorem, we need the following lemma.

Lemma 7.11. *Let $k \in \mathbb{C}$. Then the following two assertions hold:*

(i) $[\mathcal{I}V(k)]^* = \mathcal{I}V(-\overline{k})$.

(ii) (k, t) *with* $t \in H^{-1/2}(\Gamma)\backslash\{0\}$ *is an eigenpair of* V*, i.e.,* $V(k)t = 0$*, if and only if* $(-\overline{k}, \overline{t})$ *is an eigenpair of* V*, i.e.,* $V(-\overline{k})\overline{t} = 0$.

Proof. (i) The proof of this statement can also be found in [27, Lemma 5.2]. By [25, p. 120] the single layer boundary integral operator admits a representation as weakly singular boundary integral of the form

$$(V(k)w)(x) = \int_\Gamma U_k^*(x,y)w(y)ds_y, \quad \forall x \in \mathbb{R}^3,$$

for all $w \in L^\infty(\Gamma)$. Let now $t, u \in H^{-1/2}(\Gamma)$. Then it holds with (7.13)

$$\begin{aligned}(u, \mathcal{I}V(k)t)_{H^{-1/2}(\Gamma)} &= \langle \overline{u}, V(k)t\rangle_{H^{-1/2}(\Gamma)\times H^{1/2}(\Gamma)} = \frac{1}{4\pi}\int_\Gamma\int_\Gamma \frac{e^{ik|x-y|}}{|x-y|}t(y)ds_y\overline{u(x)}ds_x \\ &= \frac{1}{4\pi}\int_\Gamma\int_\Gamma \overline{\frac{e^{-i\overline{k}|x-y|}}{|x-y|}}\,\overline{u(x)}ds_x t(y)ds_y = \overline{\frac{1}{4\pi}\int_\Gamma\int_\Gamma \frac{e^{-i\overline{k}|x-y|}}{|x-y|}u(x)ds_x\overline{t(y)}ds_y} \\ &= \overline{\langle \overline{t}, V(-\overline{k})u\rangle_{H^{-1/2}(\Gamma)\times H^{1/2}(\Gamma)}} = \overline{(t, \mathcal{I}V(-\overline{k})u)_{H^{-1/2}(\Gamma)}} \\ &= (\mathcal{I}V(-\overline{k})u, t)_{H^{-1/2}(\Gamma)}.\end{aligned}$$

This proves the first assertion.
(ii) Let $(k,t) \in \mathbb{C} \times H^{-1/2}(\Gamma)\backslash\{0\}$ be an eigenpair of V, i.e., $V(k)t = 0$. By taking the complex conjugate we get $\overline{V(k)t} = 0$. We can rewrite $\overline{V(k)t}$ in the following way

$$(\overline{V(k)t})(x) = \overline{\int_\Gamma U_k^*(x,y)t(y)ds_y} = \int_\Gamma \overline{U_k^*(x,y)t(y)}ds_y = \int_\Gamma U^*_{-\overline{k}}(x,y)\overline{t(y)}ds_y = (V(-\overline{k})\overline{t})(x)$$

for every $x \in \mathbb{R}^3$, where

$$\overline{U_k^*(x,y)} = U^*_{-\overline{k}}(x,y) \tag{7.15}$$

is used. Hence, the statement follows. □

An analogous result holds true for $\mathcal{I}^*D$ instead of $\mathcal{I}V$:

Lemma 7.12. *Let $k \in \mathbb{C}$. Then the following two assertions hold:*

(i) $[\mathcal{I}^*D(k)]^* = \mathcal{I}^*D(-\overline{k})$.

(ii) (k, u) *with* $u \in H^{1/2}(\Gamma)\backslash\{0\}$ *is an eigenpair of* D*, i.e.,* $D(k)u = 0$*, if and only if* $(-\overline{k}, \overline{u})$ *is an eigenpair of* D*, i.e.,* $D(-\overline{k})\overline{u} = 0$.

Proof. (i) The proof of this statement can also be found in [30, Lemma 5.1.6] for real k. By [25, Corollary 3.3.24] the hypersingular boundary integral operator admits for all $u, v \in H^{1/2}(\Gamma)$ a representation of the form

$$\begin{aligned}\langle D(k)u, v\rangle_{H^{-1/2}(\Gamma)\times H^{1/2}(\Gamma)} \\ = \int_\Gamma \int_\Gamma U_k^*(x, y)\left\{(\underline{\operatorname{curl}}_\Gamma u(y), \underline{\operatorname{curl}}_\Gamma v(x)) - k^2 u(y)v(x)(\underline{n}(x), \underline{n}(y))\right\} ds_y ds_x,\end{aligned}$$

where $\underline{n}$ is the unit normal vector pointing from the domain Ω^- to Ω^+ and $\underline{\operatorname{curl}}_\Gamma$ is the surface rotation given by

$$\underline{\operatorname{curl}}_\Gamma u(x) = \underline{n}(x) \times \nabla\tilde{u}(x), \quad x \in \Gamma,$$

with a local extension $\tilde{u}$ of u. Let now $u, v \in H^{1/2}(\Gamma)$. Then it holds with (7.14), (7.15), (7.9), and (7.8)

$$\begin{aligned}(\mathcal{I}^* D(k)u, v)_{H^{1/2}(\Gamma)} &= \langle\overline{D(k)u}, v\rangle_{H^{-1/2}(\Gamma)\times H^{1/2}(\Gamma)} = \langle\overline{u}, D(-\overline{k})v\rangle_{H^{1/2}(\Gamma)\times H^{-1/2}(\Gamma)} \\ &= \overline{\langle u, \overline{D(-\overline{k})v}\rangle}_{H^{1/2}(\Gamma)\times H^{-1/2}(\Gamma)} = \overline{\langle\overline{D(-\overline{k})v}, u\rangle}_{H^{-1/2}(\Gamma)\times H^{1/2}(\Gamma)} \\ &= \overline{(\mathcal{I}^* D(-\overline{k})v, u)}_{H^{1/2}(\Gamma)} = (u, \mathcal{I}^* D(-\overline{k})v)_{H^{1/2}(\Gamma)}.\end{aligned}$$

This proves the first assertion.
(ii) Let $(k, u) \in \mathbb{C} \times H^{1/2}(\Gamma)\backslash\{0\}$ be an eigenpair of D, i.e., $D(k)u = 0$. By taking the complex conjugate we get $\overline{D(k)u} = 0$. Due to (7.15), we can rewrite $\overline{D(k)u}$ as

$$\langle\overline{D(k)u}, v\rangle_{H^{-1/2}(\Gamma)\times H^{1/2}(\Gamma)} = \langle D(-\overline{k})\overline{u}, v\rangle_{H^{-1/2}(\Gamma)\times H^{1/2}(\Gamma)}$$

for all $v \in H^{1/2}(\Gamma)$. Hence, the statement follows. □

We are now able to prove Theorem 7.10.

Proof of Theorem 7.10. We restrict to showing the proof for $\mathcal{I}V$. The proof for $\mathcal{I}^*D$ works analogously. To start with, we would like to recall the definitions of δ_n and δ_n^*

$$\delta_n := \max_{v_0\in G_{\lambda_0}(\mathcal{I}V), \|v_0\|_X\leqslant 1} \inf_{v_n\in X_n} \|v_0 - v_n\|_X, \quad \delta_n^* := \max_{w_0\in G_{\overline{\lambda_0}}([\mathcal{I}V]^*), \|w_0\|_X\leqslant 1} \inf_{v_n\in X_n} \|w_0 - v_n\|_X.$$

Due to the definition of the adjoint eigenvalue problem and Lemma 7.11(i), the following relationship holds

$$(\mathcal{I}V)^*(\overline{\lambda_0}) = [\mathcal{I}V(\lambda_0)]^* = \mathcal{I}V(-\overline{\lambda_0}),$$

and therefore it follows for the corresponding generalized eigenspace

$$G_{\overline{\lambda_0}}([\mathcal{I}V]^*) = G_{-\overline{\lambda_0}}(\mathcal{I}V).$$

Since we have assumed that $M_{\lambda_0}(\mathcal{I}V) = 1$, the generalized eigenspaces $G_{\lambda_0}(\mathcal{I}V)$ and $G_{\overline{-\lambda_0}}(\mathcal{I}V)$ coincide with $\ker \mathcal{I}V(\lambda_0)$ and $\ker \mathcal{I}V(-\overline{\lambda_0})$ by definition. It remains to show that

$$\max_{v_0 \in \ker \mathcal{I}V(\lambda_0), \|v_0\|_X \leqslant 1} \inf_{v_n \in X_n} \|v_0 - v_n\|_X = \max_{w_0 \in \ker \mathcal{I}V(-\overline{\lambda_0}), \|w_0\|_X \leqslant 1} \inf_{v_n \in X_n} \|w_0 - v_n\|_X.$$

In order to show this, we fix $v_0 \in \ker \mathcal{I}V(\lambda_0)$ with $\|v_0\|_X \leqslant 1$. By using Lemma 7.11(ii) $\overline{v_0} \in \ker \mathcal{I}V(-\overline{\lambda_0})$. Since X_n is finite-dimensional, there exists a finite basis $\{\phi_1, \dots, \phi_n\}$. Therefore, every element $v_n \in X_n$ can be written as

$$v_n = \sum_{j=1}^{n} \alpha_j \phi_j$$

with constants $\alpha_j \in \mathbb{C}$. As X_n is a subspace, the element $\overline{v_n}$ is also contained in X_n. The assertion follows now directly from

$$\|v_0 - v_n\|_X^2 = \|\mathrm{Re}(v_0 - v_n)\|_X^2 + \|\mathrm{Im}(v_0 - v_n)\|_X^2 = \|\overline{v_0} - \overline{v_n}\|_X^2.$$

□

7.3.3 Error estimate for lowest order Galerkin approximation

In this subsection we additionally assume that Ω^- has a polygonal boundary. Our aim is to state the error estimates for lowest order Galerkin approximations of the boundary integral formulations. As approximation spaces we consider $X_n = S_h^0(\Gamma) \subset H^{-1/2}(\Gamma)$ and $X_n = S_h^1(\Gamma) \subset H^{1/2}(\Gamma)$. The former space denotes the space of piecewise constant functions, whereas the latter one is the space of piecewise linear functions. These spaces are defined as follows:

Definition 7.13. *Let Γ_h be a valid decompositions of Γ, i.e.,*

(i) Γ_h is a finite set of non-degenerate triangular elements $\{\tau_1, \dots, \tau_{m_h}\}$,

(ii) $\Gamma_h = \bigcup_{l=1}^{m_h} \overline{\tau_l}$,

(iii) two neighbouring elements of Γ_h either share a node or an edge.

Let n_h denote the number of vertices $\{x_l\}_{l=1}^{n_h}$ of Γ_h. Then

$$S_h^0(\Gamma) := \left\{ v_n = \sum_{k=1}^{m_h} \beta_k \psi_k,\ \beta_k \in \mathbb{C} \right\}, \quad S_h^1(\Gamma) := \left\{ v_n = \sum_{k=1}^{n_h} \gamma_k \phi_k,\ \gamma_k \in \mathbb{C} \right\},$$

where the basis functions ψ_k and ϕ_k are given by

$$\psi_k(x) = \begin{cases} 1, & x \in \tau_k, \\ 0, & x \notin \tau_k, \end{cases} \qquad \phi_k(x) = \begin{cases} 1, & x = x_k, \\ 0, & x = x_l \neq x_k, \\ \textit{piecewise linear}, & \textit{elsewhere}. \end{cases}$$

Let $\{\Gamma_h\}_h$ be a family of valid decompositions of Γ, which are uniformly shape regular, i.e., there exists a constant $c > 0$ independent of the decomposition, such that

$$d_l \leqslant c h_l, \quad \forall l = 1, \ldots, m_h,$$

where d_l and h_l are the diameter and the local mesh size of τ_l

$$d_l := \sup_{x,y\in\tau_l} |x-y|, \quad h_l := \left(\int_{\tau_l} ds_x \right)^{1/2}.$$

Moreover, we assume that $\{\Gamma_h\}_h$ is globally quasi-uniform, i.e., there exists a constant $c_G \geqslant 1$ independent of m_h such that

$$\frac{h_{\max}}{h_{\min}} := \frac{\max\limits_{l=1,\ldots,m_h} h_l}{\min\limits_{l=1,\ldots,m_h} h_l} \leqslant c_G.$$

The following lemma points out that the assumption on the choice of the finite-dimensional subspace X_n in Theorem 7.4 is satisfied by $S_h^0(\Gamma)$ and $S_h^1(\Gamma)$.

Lemma 7.14. *Let $\{\Gamma_h\}_h$ be given as above. Then the spaces $S_h^0(\Gamma)$ and $S_h^1(\Gamma)$ fulfil the condition* (7.4)*.*

Proof. See [25, Corollary 4.1.28] for $S_h^0(\Gamma)$. The proof for $S_h^1(\Gamma)$ can be done analogously. □

Before stating the error estimates for the eigenvalues and eigenvectors, we need the following properties of $S_h^0(\Gamma)$ and $S_h^1(\Gamma)$ [30, Eq. (5.18)]

$$\inf_{v_h\in S_h^\eta(\Gamma)} \|v - v_h\|_{H^{\eta-1/2}(\Gamma)} \leqslant c h^{s-\eta+1/2} \|v\|_{H^s_{\mathrm{pw}}(\Gamma)} \quad \forall v \in H^s_{\mathrm{pw}}(\Gamma),$$

where $\eta \in \{0,1\}$, $s \in [\eta - 1/2, \eta + 1]$, and $H^s_{\mathrm{pw}}(\Gamma)$ is defined as in [26]. Under the assumption that the maximal length of any Jordan chain is equal to one, we obtain from Theorem 7.5 for $\eta \in \{0,1\}$ the following error estimates for the eigenvalues

$$|\lambda_n - \lambda_0| \leqslant \tilde{C} h^{2s-2\eta+1} \max_{v_0\in\ker A^\eta(\lambda_0), \|v_0\|_{H^{\eta-1/2}(\Gamma)}\leqslant 1} \|v_0\|^2_{H^s_{\mathrm{pw}}(\Gamma)} \sim \mathcal{O}(h^{2s-2\eta+1}), \tag{7.16}$$

and for the eigenvectors

$$\begin{aligned} \inf_{v_0\in\ker A^\eta(\lambda_0)} \|v_n - v_0\|_{H^{\eta-1/2}(\Gamma)} \leqslant \hat{C} \Bigg(& h^{2s-2\eta+1} \max_{v_0\in\ker A^\eta(\lambda_0), \|v_0\|_{H^{\eta-1/2}(\Gamma)}\leqslant 1} \|v_0\|^2_{H^s_{\mathrm{pw}}(\Gamma)} \\ & + h^{s-\eta+1/2} \max_{z_0\in\ker A^\eta(\lambda_0), \|z_0\|_{H^{\eta-1/2}(\Gamma)}\leqslant 1} \|z_0\|_{H^s_{\mathrm{pw}}(\Gamma)} \Bigg) \\ \sim \mathcal{O}(h^{s-\eta+1/2}) & \end{aligned}$$

where $A^0 := \mathcal{I}V$, $A^1 := \mathcal{I}^*D$, and $\ker A^\eta(\lambda_0) \subset H^s_{\mathrm{pw}}(\Gamma)$ for some $s \in [\eta - 1/2, \eta + 1]$. Therefore, the highest orders of convergence using lowest order approximation spaces are $\mathcal{O}(h^3)$ for the eigenvalues and $\mathcal{O}(h^{3/2})$ for the eigenvectors, provided that $\ker \mathcal{I}V(\lambda_0) \subset H^1_{\mathrm{pw}}(\Gamma)$ and $\ker \mathcal{I}^*D(\lambda_0) \subset H^2_{\mathrm{pw}}(\Gamma)$.

8 Numerical examples

In this chapter we would like to compare the presented numerical methods by discussing the results of several numerical experiments.

8.1 Interior eigenvalue problem for the Laplace operator on the unit cube

In this section we focus on the computation of eigenvalues of the interior Dirichlet Laplace eigenvalue problem

$$-\Delta u - k^2 u = 0 \text{ in } \Omega^-,$$
$$\gamma_0^- u = 0 \text{ on } \Gamma,$$

where $\Gamma := \partial\Omega^-$ and $\Omega^- = (0,1)^3$ is the unit cube in $\mathbb{R}^3$. Our aim is to find all eigenvalues k which lie in the interval $[1.0, 19.0]$. By Theorem 6.9 we have seen that the eigenvalues of the above eigenvalue problem correspond to the numbers $k \in \mathbb{R}_0^+$ of the boundary integral formulation

$$\mathrm{V}(k)t = 0, \tag{8.1}$$

where $t \in H^{-1/2}(\Gamma)\backslash\{0\}$. Note again that equation (8.1) describes a NEP since it is nonlinear in the parameter k.

Discretization and Galerkin approximation:

Let $S_h^0(\Gamma) := \{\psi_j^h\}_{j=1}^{m_h} \subset H^{-1/2}(\Gamma)$ be the space of piecewise constant functions. Then the Galerkin variational formulation leads to the following eigenvalue problem

$$V_h(k_h)\underline{t} = 0, \tag{8.2}$$

where

$$V_h(k_h)[i,j] = \int_\Gamma \int_\Gamma \frac{1}{4\pi} \frac{e^{ik_h|x-y|}}{|x-y|} \psi_j^h(y) ds_y \overline{\psi_i^h(x)} ds_x.$$

Note that (8.2) has exactly the required form (2.1).

In order to solve (8.2) we choose a discretization of Γ in triangles with mesh size $h = 0.1$. This leads to 1,468 degrees of freedom (DOFs). Figure 8.1 was created with Gmsh [13] and shows the discretization.

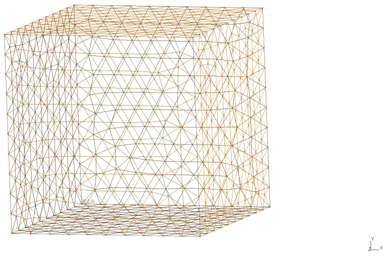


Figure 8.1: Discretization of Γ in triangles with mesh size $h = 0.1$.

For computing the BEM matrices we used the open-source Galerkin boundary element library BEM++ 3.0.0 [21]. Note that in this example only dense matrices were used. The calculations were performed in IPython [23]. In BEM++ the boundary integrals are, by default, computed by using Gauß quadrature rules with 12 symmetric integration points per triangle and Duffy type transformations. For solving the systems of linear equations for the matrices $\hat{A}_p$ and h_p, defined by (3.19) and (4.11), we used the command *numpy.linalg.solve*, which calls the LAPACK routine *_gesv*. This routine solves a well-determined, i.e. full rank, linear matrix equation $AX = B$ by computing the LU decomposition of A.

Exact eigenvalues and their distribution:

The exact eigenvalues of (8.2) are given by the formula [29, Section 10.1, Example 1]

$$k_h = \pi\sqrt{k_1^2 + k_2^2 + k_3^2}, \quad k_1, k_2, k_3 \in \mathbb{N}. \tag{8.3}$$

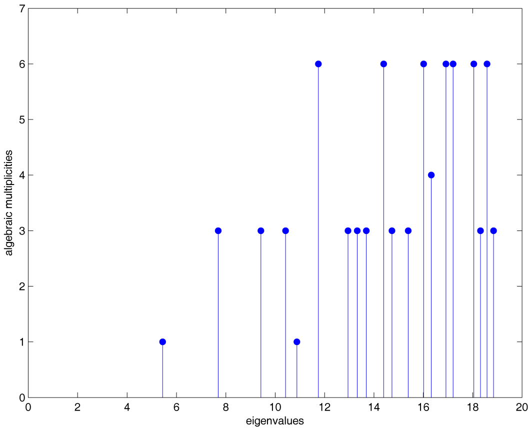


Figure 8.2: Eigenvalues in $[1.0, 19.0]$ with their corresponding algebraic multiplicities.

One can verify that there are 78 eigenvalues in the interval $[1.0, 19.0]$, when counting them according to their multiplicities. Figure 8.2 shows their distribution and their corresponding algebraic multiplicities. We observe that there are several eigenvalues with algebraic multiplicity six and only two eigenvalues with algebraic multiplicity one. In addition, the eigenvalues with algebraic multiplicity six are all in the right half of the interval.

Comparison of the CIM and the RIM:

Note that for all our tests we used the same contour $\Gamma_\mathcal{C}$, namely an ellipse with center in $(x_0, y_0) = (10.0, 0.0)$, semi-major axis $a = 9.0$, and semi-minor axis $b = 0.1$.

First of all, we would like to compare the CIM with the RIM for different values of K and l, where K denotes the number of blocks of the Hankel matrices B_0, B_1 in the CIM and H, $H^<$ in the RIM, and l is the number of columns of the random matrix U. For our tests we fixed the following parameters:

- $N = 150$ equidistant quadrature points on $\Gamma_\mathcal{C}$ for the CIM;
- $N = 150$ Chebyshev points of the first kind in the interval $[1.0, 19.0]$ for the RIM;
- $\delta = 10^{-14}$, which is used for the determination of the numerical rank $\sigma_k > \delta \cdot \sigma_1$, for both algorithms;
- $\text{tol}_{\text{res}} = 10^{-3}$, which is used in the residual tests, for both algorithms.

In order to facilitate the comparison, we first sorted the calculated eigenvalues in ascending order according to their real values, then we assigned them indices according to their position, and finally we plotted the residuals $\|T(\lambda_j)v_j\|_2/\|v_j\|_2$ against these indices. In Figure 8.3, our results for different values of K and l are illustrated.

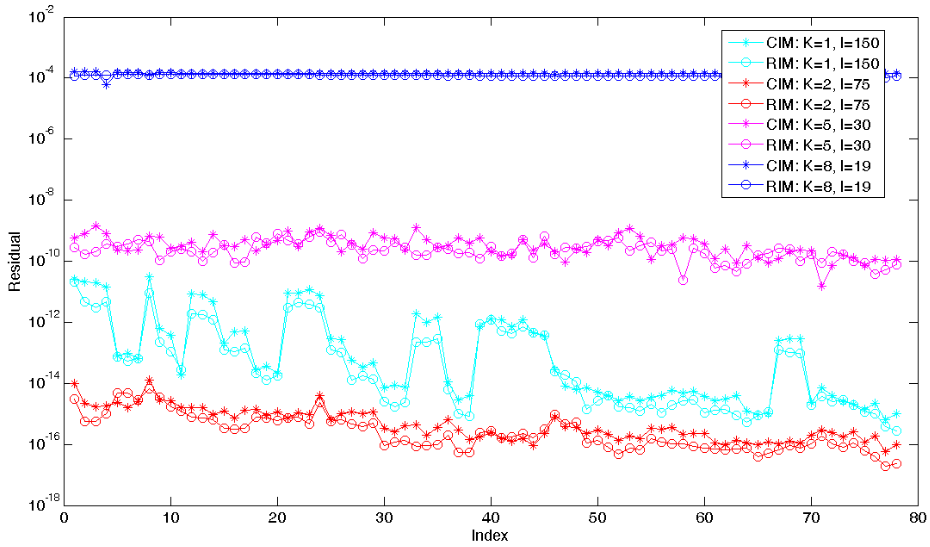


Figure 8.3: Comparison of the CIM and the CIM for different values of K and l.

One can notice that for fixed values of K and l the residuals $\|T(\lambda_j)v_j\|_2/\|v_j\|_2$ are nearly the same for both algorithms. Furthermore, the residuals increase with increasing values of K, except for $K = 1$ and $K = 2$. For $K = 1$ the residuals are not only a few orders higher than for $K = 2$, but they also show higher oscillations. At some point both algorithms start to fail completely. We noticed that when trying the cases $K = 15$, $l = 10$ and $K = 21$, $l = 7$. In the former case both algorithms only found 50 eigenvalues and in the latter one 35. Besides that, some calculated eigenvalues were totally wrong, although (3.2) and $Kl \geqslant \kappa$ were fulfilled. This observation could be explained as follows. The larger the value of l is, the more information of $T(z)^{-1}$ can be acquired, which leads to a higher accuracy. For increasing values of K though, the Hankel matrices B_0, B_1, H, and $H^<$ become more and more ill-conditioned due to their special structure. Hence, it is better to choose small values of K and larger values of l. The problem which occurs is that with larger values of l the number of linear systems which have to be solved becomes larger, too.

Before moving on to the results of the CIM with Rayleigh-Ritz procedure, we would like to have a look at the discretization errors of our calculated eigenvalues with respect to the exact eigenvalues given by formula (8.3). Figure 8.4 shows these discretization errors in the case, where the RIM with $K = 2$ and $l = 75$ was used.

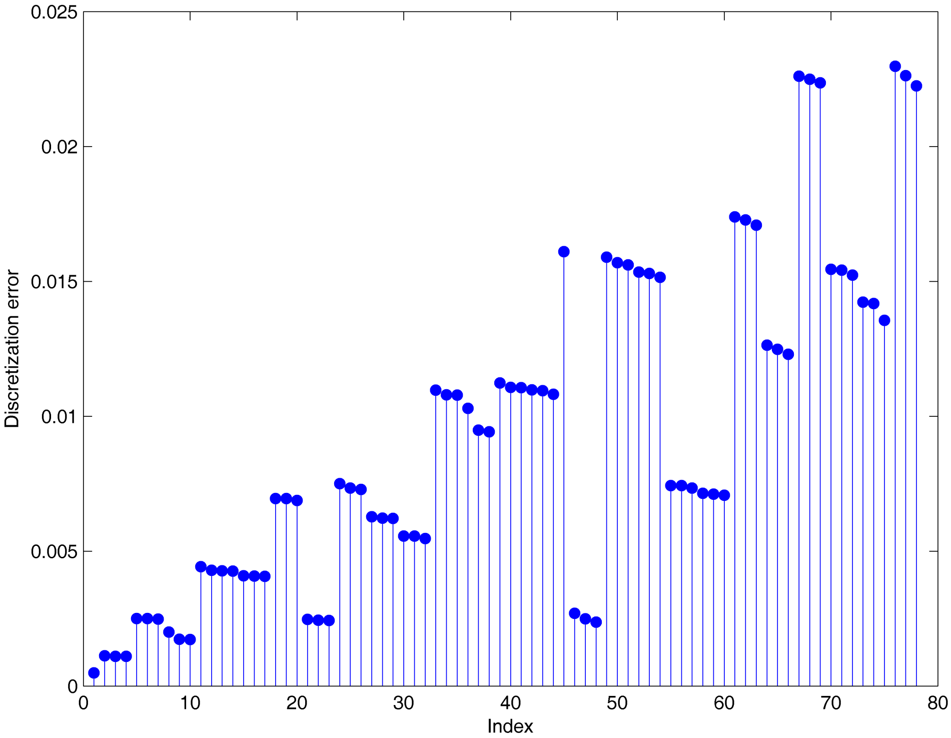


Figure 8.4: Discretization errors using the RIM with $K = 2$ and $l = 75$.

Comparison of the CIM-RRm and the CIM-RRs:

Now we would like to compare the CIM using the Rayleigh-Ritz procedure based on the moment scheme with the corresponding method based on the sampling scheme. Since we have seen before that the results for the CIM and the RIM are nearly the same, we only focus on the CIM here. Note that we also solved the projected NEP $T_Q(\lambda)g = 0$ by using the CIM. For our tests the following parameters were fixed:

- $N^{RR} = 150$ equidistant quadrature points on the contour $\Gamma_{\mathcal{C}}$ for the Rayleigh-Ritz procedure;
- $\delta^{RR} = 10^{-14}$, which is used for the determination of the numerical rank $k_{\mathcal{S}}$ in the moment and the sampling scheme;
- same quadrature points for the projected NEP ($N = N^{RR}$);
- $K = 2$ blocks in each row and column of the Hankel matrices $\hat{B}_0$ and $\hat{B}_1$ assembled when solving the projected NEP;
- $\delta = 10^{-14}$, which is used for the determination of the numerical rank of $\hat{B}_0$ in the projected NEP;
- $\text{tol}_{\text{res}} = 10^{-3}$ for the residual tests.

We performed the following tests:

- CIM-RRs with $L = 15$ and $L = 6$;
- CIM-RRm with $L = 15$ and $K^{RR} = 10, 20$;
- CIM-RRm with $L = 6$ and $K^{RR} = 25, 50$;

In Table 8.1, the calculated numerical ranks for $\hat{S}$ and $\hat{M}$ are summarized.

Method	Parameters	Numerical rank $k_{\mathcal{S}}$
CIM-RRs	$L = 15$	480
CIM-RRs	$L = 6$	272
CIM-RRm	$L = 15$, $K^{RR} = 10$	75
CIM-RRm	$L = 15$, $K^{RR} = 20$	73
CIM-RRm	$L = 6$, $K^{RR} = 25$	32
CIM-RRm	$L = 6$, $K^{RR} = 50$	30

Table 8.1: Numerical ranks $k_{\mathcal{S}}$ for $\hat{S}$ and $\hat{M}$.

We can notice that for the CIM-RRs the necessary condition

$$N^{RR}L \geqslant k_{\mathcal{S}} \geqslant \kappa = 78$$

for the proper approximation of the generalized eigenspace corresponding to the eigenvalues inside $\Gamma_{\mathcal{C}}$ is fulfilled for both choices of L, whereas for the CIM-RRm the necessary condition

$$K^{RR}L \geqslant k_{\mathcal{S}} \geqslant \kappa = 78$$

is not satisfied for any case. This usually has the effect that some of the calculated eigenvalues are totally wrong and others are missing. However, in the cases $L = 15$, $K^{RR} = 10$, and $L = 15$, $K^{RR} = 20$, for the CIM-RRm, the determined numerical rank is only slightly smaller than the real rank (78), and with the help of $K = 2$ all 78 eigenvalues were calculated. Though, the residuals of these calculated eigenvalues are much smaller than those when the CIM-RRs is used, as one can see in Figure 8.5. In the cases $L = 6$, $K^{RR} = 25$, and $L = 6$, $K^{RR} = 50$, only 17 and 30 eigenvalues respectively were computed.

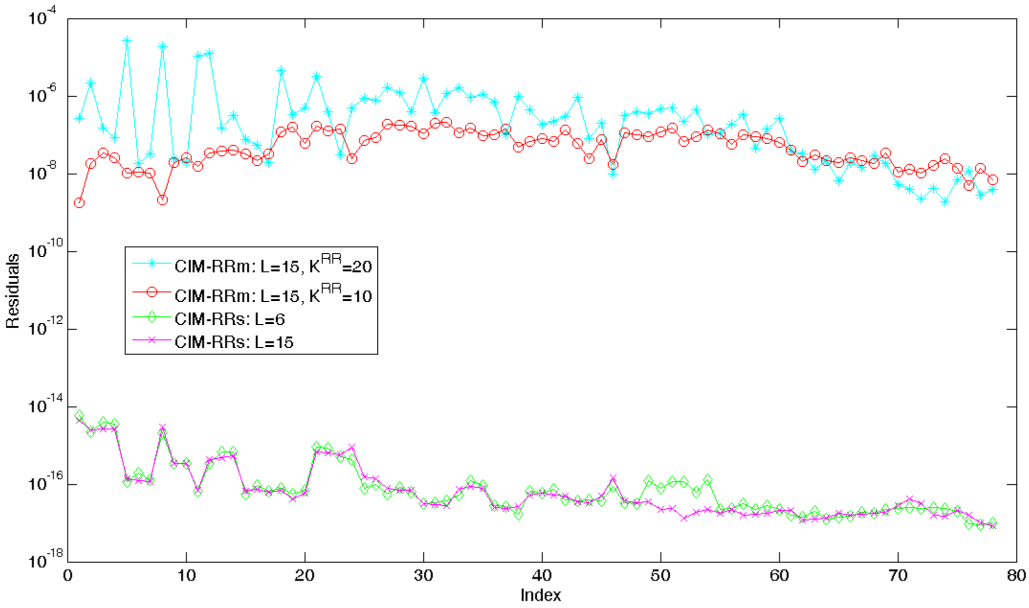


Figure 8.5: Residuals using the Rayleigh-Ritz procedure.

The different behaviour of the residuals of the calculated eigenvalues can be explained by the different quality of the approximated generalized eigenspaces which are obtained from the matrices $\hat{M}$ and $\hat{S}$. In order to fulfil the rank condition (5.6)

$$K^{RR}L \geqslant k_{\mathcal{S}} \geqslant \kappa,$$

the choice of small values of L implies that K^{RR} has to be chosen large. The problem which occurs is that by using higher moments in $\hat{M}$, the columns of this matrix become more and more linearly dependent.

The different qualities of the approximated eigenspaces can also be observed by looking at the singular values of $\hat{M}$ and $\hat{S}$, which are shown in Figure 8.6. Note that in order to be able to compare the singular value distributions for $\hat{M}$ and $\hat{S}$ more effectively, we first scaled the singular values such that the first one is always one and we only consider the first 150 singular values of $\hat{M}$ and $\hat{S}$, which normally have $\min\{n, K^{RR}L\}$ and $\min\{n, N^{RR}L\}$ singular values respectively.

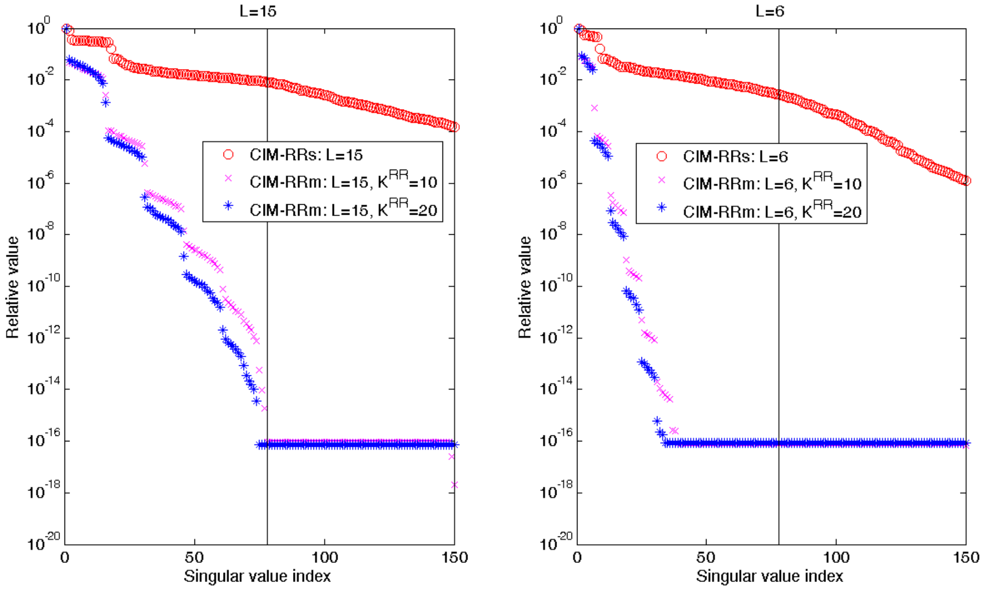


Figure 8.6: Singular values of $\hat{M}$ and $\hat{S}$ for $L = 15$ (left) and $L = 6$ (right).

We can notice that the singular values decrease much faster when using the moment scheme. Whereas the 78^{th} singular value, which is the first eigenvalue to the left of the vertical line in the figures, is only around two orders lower than the first one in $\hat{S}$, the ratio is around 16 orders for $\hat{M}$. Since we chose $\delta^{RR} = 10^{-14}$ for the moment scheme, the eigenvalues can not be extracted properly. Figure 8.7 even shows that the numerical rank never reaches 78 no matter how large we choose K^{RR}. To conclude, we have seen that the sampling scheme generates far more reliable and better approximations of the eigenspaces.

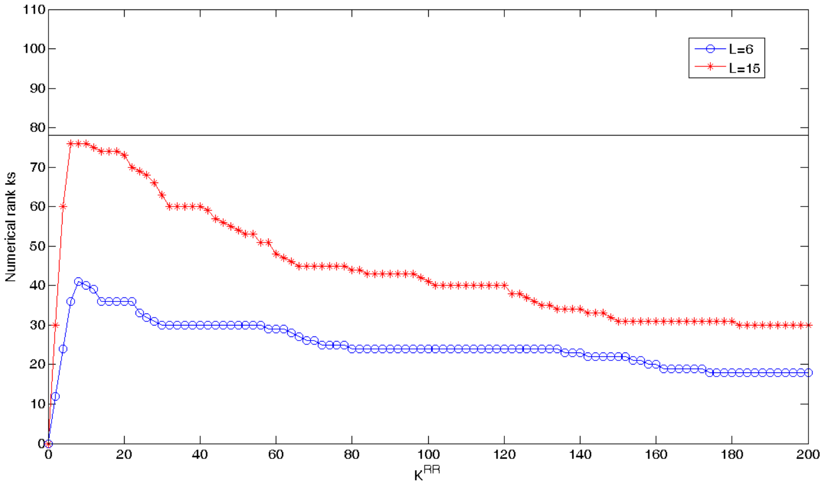


Figure 8.7: Numerical rank k_S in dependence of the number of moments K^{RR}.

Comparison of the CIM and the RIM from a viewpoint of reducing the number of quadrature points N:

Since the assembly of BEM matrices is expensive, we are interested in finding out how small the number of quadrature and interpolation points N can be chosen in the CIM and the RIM such that accurate results are still obtained. Figure 8.8 shows the residuals against N for the cases $K = 1$, $l = 100$, and $K = 2$, $l = 50$. As before, the N quadrature points for the CIM were chosen equidistantly on the contour, and as interpolation points for the RIM N Chebyshev points of the first kind in the interval $[1.0, 19.0]$ were taken.

For the tests we fixed the following parameters:

- $\delta = 10^{-14}$ for the determination of the numerical rank;
- $\text{tol}_{\text{res}} = 10^{-3}$ for the residual tests.

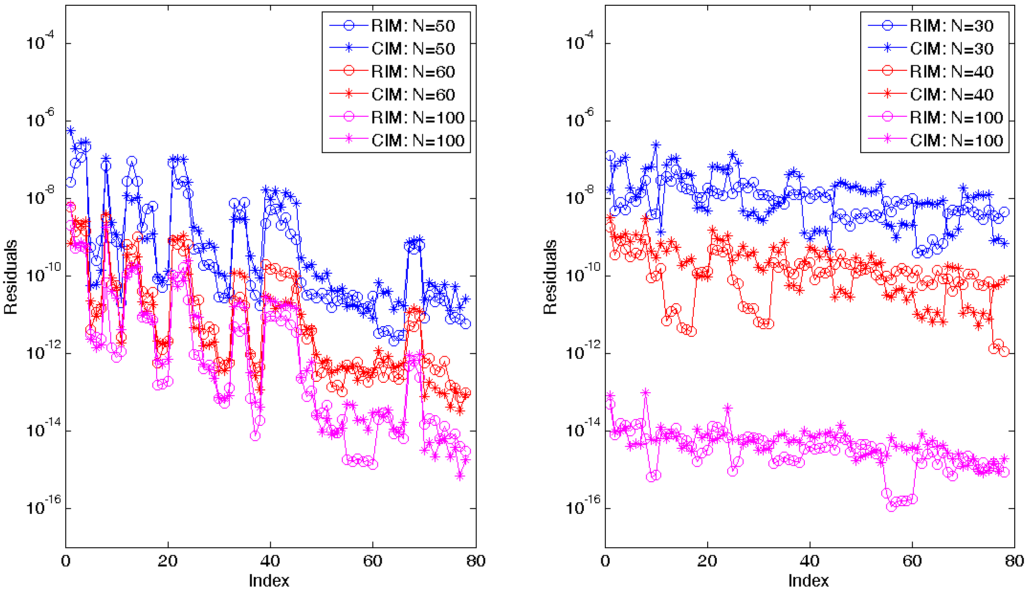


Figure 8.8: Residuals for $K = 1$, $l = 100$ (left) and $K = 2$, $l = 50$ (right).

From Figure 8.8 we can make two observations: Firstly, we can notice again that the residuals are quite the same for both methods and that they show higher oscillations in the case $K = 1$, $l = 100$, than for $K = 2$, $l = 50$. Secondly, whereas in the former case N should be chosen larger or equal than 50 to get accurate results, N should be chosen at least 30 in the latter one. Much smaller values than these are not recommended to be chosen.

Comparison of the CIM and the CIM-RRs from a viewpoint of reducing the number of quadrature points N:

Similarly as before, we would like to compare the CIM and the CIM-RRs from a viewpoint of reducing the number of quadrature points N. For the numerical tests we chose the following parameters for the CIM:

- $\delta = 10^{-14}$ for the determination of the numerical rank;
- $K = 2$ blocks in each row and column of the Hankel matrices $\hat{B}_0$ and $\hat{B}_1$;
- $\mathrm{tol}_{\mathrm{res}} = 10^{-3}$ for the residual test.

For the CIM-RRs we fixed:

- $\delta^{RR} = \delta = 10^{-14}$ for the determination of the numerical ranks of $\hat{S}$ and $\hat{B}_0$;
- $N^{RR} = N$ equidistant quadrature points for $\hat{S}$;
- $\mathrm{tol}_{\mathrm{res}} = 10^{-3}$ for the residual test.

Figure 8.9 shows the residuals for different values of N. Note that for $N = 20$ the CIM does not work any more since some eigenvalues are not found.

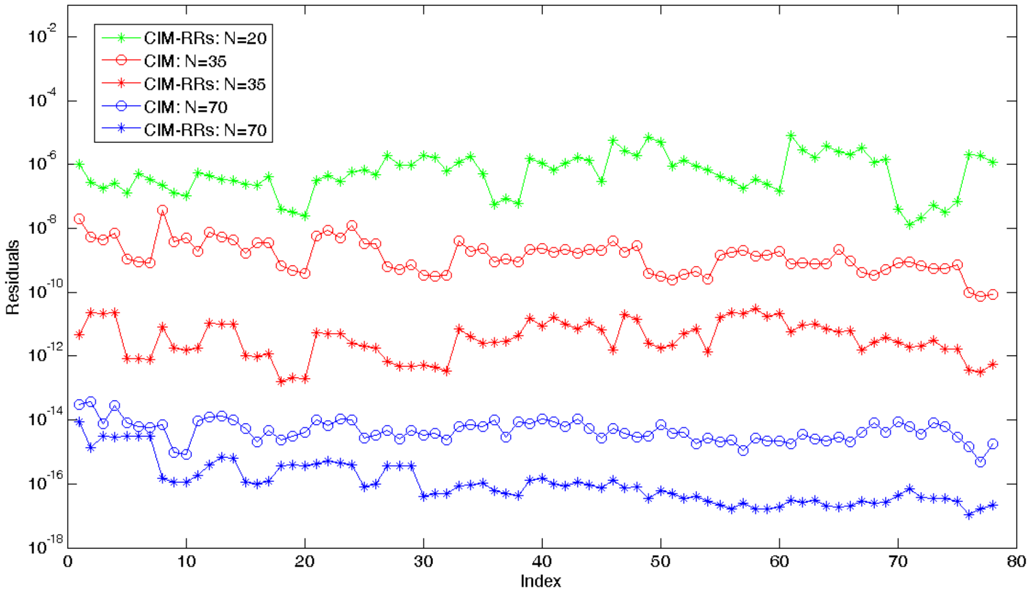


Figure 8.9: Comparison of the residuals when using the CIM and the CIM-RRs.

We can notice that for the same N the residuals are always smaller when the CIM-RRs is used instead of the CIM. Therefore, it really make sense to use the CIM-RRs, provided that the matrices $T(z_1^{RR}), \dots, T(z_{N^{RR}}^{RR})$, calculated for $\hat{S}$, can be stored. Otherwise, a value of N means for the CIM-RRs that in fact $2N$ matrices have to be assembled. In this case it could be better to choose the CIM, depending on the costs for solving the linear systems.

Improvement of the residuals in the CIM-RRs for constant N:

We are now interested in the effects of the choice of L, which describes the number of columns of the random matrix $\hat{U}$, on the residuals for a constant number of equidistant quadrature points on the contour. We performed the tests for $N = 20, 35$ with the following parameters:

- $\delta^{RR} = \delta = 10^{-14}$ for the determination of the numerical ranks of $\hat{S}$ and $\hat{B}_0$;
- $K = 2$ blocks in each row and column of the Hankel matrices $\hat{B}_0$ and $\hat{B}_1$ in the projected NEP;
- $\mathrm{tol}_{\mathrm{res}} = 10^{-3}$ for the residual test.

The plot on the left-hand side of Figure 8.10 shows the residuals against L for $N = 20$, and the other one on the right-hand side of Figure 8.10 for $N = 35$.

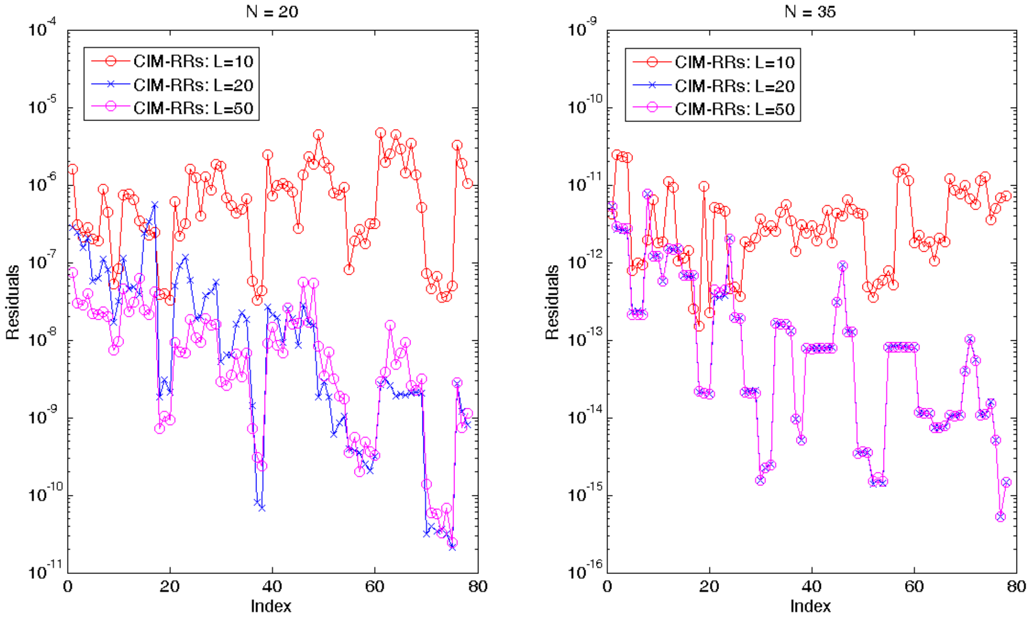


Figure 8.10: Residuals when using the CIM-RRs for fixed values of $N = 20$ and $N = 35$ with variable L.

By condition (5.2)

$$L \geqslant \max_{k=1,\ldots,n_{\mathcal{C}}} \left(\sum_{l=1}^{\eta_k} m_{l,k} \right),$$

L has to be chosen at least six. We can observe that a change from ten to 20 has a much larger effect on the residuals than a change from 20 to 50.

Improvement of the residuals in the CIM for constant N:

As before, we would like to find out more about the relationship between l, which denotes the number of columns of the random matrix U, and the residuals for constant $N = 20, 35$. We fixed the following parameters:

- $\delta = 10^{-14}$ for the determination of the numerical rank;
- $K = 2$ blocks in each row and column of the Hankel matrices $\hat{B}_0$ and $\hat{B}_1$;
- $\mathrm{tol}_{\mathrm{res}} = 10^{-3}$ for the residual test.

Note that the CIM does not work for $N = 20$, $l = 50$, since some eigenvalues are not found. However, by increasing l more information is available which has the consequence that all eigenvalues are detected now.

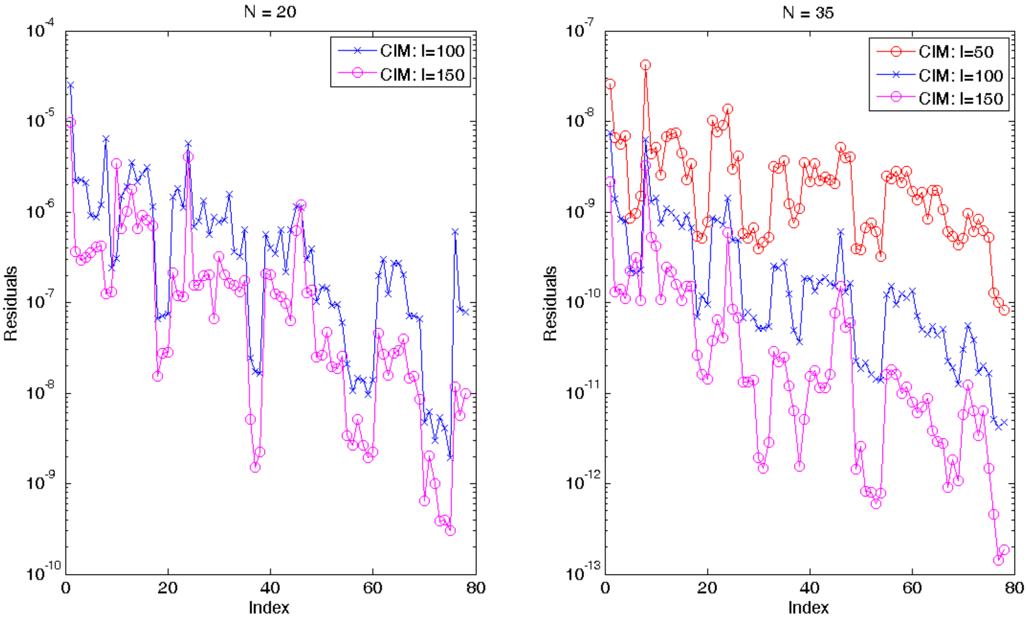


Figure 8.11: Residuals when using the CIM for fixed values of $N = 20$ and $N = 35$ with variable l.

By the necessary condition

$$Kl \geqslant \kappa = 78,$$

l has to be chosen at least 39 for $K = 2$. In the right plot of Figure 8.11 we can see that a change of l from 100 to 150 still has a large effect on the residuals.

Experimental order of convergence (EOC):

In general, the experimental order of convergence (EOC) can be computed by the following formula:

$$\text{eoc} = \frac{\log|\lambda - \lambda_{h_{l-1}}| - \log|\lambda - \lambda_{h_l}|}{\log h_{l-1} - \log h_l},$$

where λ denotes the exact eigenvalue, $\lambda_{h_{l-1}}$ is the calculated approximation of λ based on a coarser level of discretization, and λ_{h_l} is the approximated eigenvalue based on the current level of discretization.

We used the following meshes:

mesh size h	number of nodes
0.1	1468
0.05	5668
0.035	11928

In Table 8.2, the EOC is calculated for some eigenvalues. For the calculation we used the CIM with $N = 35$ equidistant quadrature points on the contour, $K = 2$, $l = 50$, $\delta = 10^{-10}$, and $\text{tol}_{\text{res}} = 10^{-3}$. Note that for the eigenvalues with algebraic multiplicity larger than one we always took the calculated eigenvalue closest to the exact eigenvalue.

h	exact eigenvalue	approximated eigenvalue	absolute error	eoc
0.1	10.88279619	10.87838690-0.00011426j	0.00441076	
0.05		10.88232402-3.48820227e-07j	0.00047216	3.223
0.035		10.88267076+7.88826624e-07j	0.00012542	3.716
0.1	14.39658614	14.39115293-0.00037488j	0.00544612	
0.05		14.39597381-9.38423554e-06j	0.00061239	3.152
0.035		14.39641940+5.19538255e-08j	0.00016673	3.647
0.1	17.20721163	17.20016987-0.00088348j	0.00709696	
0.05		17.20635512-3.84305336e-05j	0.00085737	3.049
0.035		17.20698510-1.96376116e-06j	0.00022653	3.731
0.1	18.84955592	18.82721087-0.00112661j	0.02237343	
0.05		18.84718760-5.79938118e-05j	0.00236902	3.239
0.035		18.84885440-4.52768175e-06j	0.00070153	3.411

Table 8.2: Experimental order of convergence for different eigenvalues.

In all cases we have a nearly cubic convergence. This matches with formula (7.16), which gives an error estimate for the eigenvalues.

8.2 Neumann eigenvalues of a pipe

In this example we are interested in finding all eigenvalues of the boundary integral formulation

$$D(k)\phi = 0$$

which lie close to the real axis and have a real part in the interval [-0.1,2.1]. The domain Ω^- is a pipe with length 5, outer radius 0.5, and inner radius 0.25.

By Theorem 6.10 we know that the eigenvalues of the above problem with negative imaginary part are the resonances of (6.4) and the ones with non-negative imaginary part are the eigenvalues of (6.2).

Discretization and Galerkin approximation:

Similarly as in the previous example, discretization by using the space of piecewise linear functions $S_h^1(\Gamma) := \{\phi_j^h\}_{j=1}^{n_h} \subset H^{1/2}(\Gamma)$ leads to the following eigenvalue problem

$$D_h(k_h)\underline{u} = 0,$$

where

$$D_h(k_h)[i,j] = \langle \overline{D(k)\phi_i}, \phi_j \rangle_{H^{-1/2}(\Gamma)\times H^{1/2}(\Gamma)}.$$

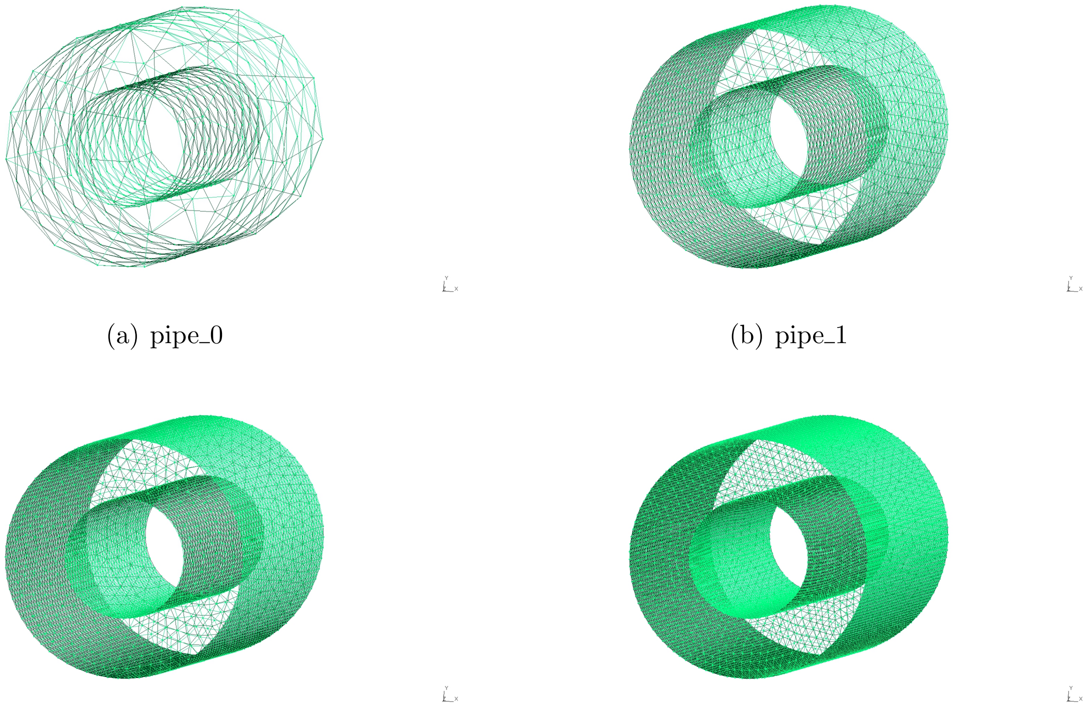
In the sequel we will use the following discretizations of the pipe.

Mesh	Number of nodes	Number of elements
pipe_0	484	968
pipe_1	3784	7568
pipe_2	7307	14614
pipe_3	15136	30272

Table 8.3: Meshes with the numbers of corresponding nodes and elements.

Figure 8.12 illustrates the different meshes. For the pictures we used Gmsh [13], and for the assembling of the BEM and Fast BEM matrices the open-source Galerkin boundary element library BEM++ 3.0.3 [21]. All the other calculations were performed with Python2.7.

Since we are only interested in eigenvalues and resonances which lie close to the real axis, we always used an ellipse with center in $(x_0, y_0) = (1.0, 0.0)$, semi-major axis $a = 1.1$, and semi-minor axis $b = 0.2$ as contour.



(a) pipe_0

(b) pipe_1

(c) pipe_2

(d) pipe_3

Figure 8.12: Meshes showing the discretized pipe.

Our first aim again is to investigate whether it is more efficient to use the CIM or the CIM with the Rayleigh Ritz procedure based on the sampling scheme (CIM-RRs) for the computation of the eigenvalues in the case where dense matrices are used.

As a second point, we would like to find out more about the effects on the results when $\mathcal{H}$-matrices are used instead of dense matrices. Note that the term $\mathcal{H}$-matrix refers to the so-called "Hierarchical matrices", which are used in the Fast BEM. These matrices provide approximations of the dense matrices used in the BEM. The big advantage is that the overall complexity of the assembly of the matrices linked to the discretized boundary integral operators can be reduced from $\mathcal{O}(n^2)$ for dense matrices to $\mathcal{O}(n \log n)$ for $\mathcal{H}$-matrices, where n denotes the number of nodes in the corresponding grid. In BEM++ the ACA (adaptive cross approximation) is used for generating the low-rank approximations. A detailed description of Hierarchical matrices can, for instance, be found in [4].

Roughly speaking, we know that the main cost of the CIM is based on the computation of the matrices

$$\hat{A}_p = \sum_{j=0}^{N-1} \tilde{w}_j T(\tilde{z}_j)^{-1} U \in \mathbb{C}^{n \times l}, \quad p = 0, \dots, 2K-2,$$

and the residual tests. However, the cost for the CIM-RRs is mainly determined by the assembly of the approximation of the generalized eigenspace

$$\hat{S} = \left(T(\tilde{z}_0^{RR})^{-1}\hat{U}, T(\tilde{z}_1^{RR})^{-1}\hat{U}, \ldots, T(\tilde{z}_{N^{RR}-1}^{RR})^{-1}\hat{U}\right) \in \mathbb{C}^{n\times N^{RR}L},$$

the computation of the matrices

$$\hat{\hat{A}}_p = \sum_{j=0}^{N-1} \tilde{w}_j T_Q(\tilde{z}_j)^{-1} I \in \mathbb{C}^{k_{\mathcal{S}}\times k_{\mathcal{S}}}, \quad p = 0, \ldots, 2K-2,$$

and the residual tests. Now we would like to have a closer look at the following operations:

- Matrix assemblies: If N denotes the number of equidistant quadrature points on the contour, we have to assemble N BEM or Fast BEM matrices for the computation of $\hat{A}_p$, $p = 0, \ldots, 2K-2$, in the CIM. However, for the CIM-RRs N^{RR} matrices have to be assembled for $\hat{S}$, plus N matrices for $\hat{\hat{A}}_p$. For simplicity reasons we always assumed that $N = N^{RR}$.

  The number of operations for the assembly of a BEM matrix is $\mathcal{O}(n^2)$, whereas a Fast BEM matrix can be assembled with a cost of $\mathcal{O}(n \log n)$. In both cases, n denotes the number of nodes of the mesh. In the case where we can store the matrices assembled for $\hat{S}$, we can reuse them for the CIM which is applied to the projected NEP, and hence the number of matrices which have to be assembled is the same for the CIM and the CIM-RRs in this special case. In our tests we did not reuse the matrices due to memory requirements.
- Solving of linear systems: For the assembly of $\hat{A}_p$ in the CIM we have to solve N equations of the form

  $$T(\tilde{z}_j)X = U, \tag{8.4}$$

  where $T(\tilde{z}_j) \in \mathbb{C}^{n\times n}$ and $U \in \mathbb{C}^{n\times l}$ with at least $l \geqslant \kappa/2$ for $K = 2$. In comparison to that, for the CIM-RRs we have to solve $N^{RR} = N$ equations

  $$T(\tilde{z}_j)X = \hat{U}, \tag{8.5}$$

  where $\hat{U} \in \mathbb{C}^{n\times L}$ with

  $$L \geqslant \max_{k=1,\ldots,n_{\mathcal{C}}} \left(\sum_{l=1}^{\eta_k} m_{l,k}\right).$$

  We can assume that l and L are chosen such that $L \leqslant l$. Moreover, the equations

  $$T_Q(\tilde{z}_j)X = I,$$

  where $T_Q(\tilde{z}_j), I \in \mathbb{C}^{k_{\mathcal{S}}\times k_{\mathcal{S}}}$, have to be solved, but due to $k_{\mathcal{S}} \ll n$ in most technical applications we neglect the cost to solve them. For dense matrices we solved the equations (8.4) and (8.5) by calculating the LU decomposition of $T(\tilde{z}_j)$, which has a cost of $\mathcal{O}(n^3)$, once and by computing the solution column-wise $\mathcal{O}(n^2)$. For $\mathcal{H}$-matrices we used the GMRES.

- Residual tests: The number of residual tests depends on the choice of the parameter δ, which is used for the determination of the numerical rank. For each test a BEM or Fast BEM matrix has to be assembled. Therefore, it is desirable to choose δ in such a way that the numerical rank coincides with the number of eigenvalues.

  For the CIM with RRs one could circumvent the problem of performing expensive residual tests by storing the matrices $T_Q(z_1), \dots, T_Q(z_N)$ (s. Algorithm 4) and then computing an interpolant of T_Q as in (4.6)

$$T_Q^I(z) = \sum_{i=1}^{N} T_Q(z_i) l_i(z).$$

  Note that since the matrices $T_Q(z_1), \dots, T_Q(z_N)$ have a small dimension, namely $k_{\mathcal{S}} \times k_{\mathcal{S}}$, the storage of these matrices is no problem at all. Instead of $\|T(\lambda_j)v_j\|_2/\|v_j\|_2$ we could compute $\|T_Q^I(\lambda_j)g_j\|_2/\|g_j\|_2$ now, where v_j and g_j are defined as in Algorithm 4. Another possible improvement for the CIM-RRs from the viewpoint of reducing the computational cost could be to calculate an interpolant T_Q^I given as above and then apply the CIM to this interpolant in the second step of the CIM-RRs algorithm. As the evaluation of T_Q^I is cheap, there can be used even more quadrature points in the CIM for the calculation of the eigenpairs of the interpolated projected NEP. To summarize, we performed the following four variants of the CIM-RRs, where the notations are the same as in Chapter 5:

Variant	Description
1	Choose N^{RR} quadrature points $\tilde{z}_j^{RR}$, build $\hat{S}$ and compute Q. Then apply the CIM to $T_Q(z) = Q^H T(z) Q$ with N quadrature points $\tilde{z}_j$. Compute the residuals $\|T(\lambda_j)v_j\|_2/\|v_j\|_2$.
2	Choose N^{RR} quadrature points $\tilde{z}_j^{RR}$, build $\hat{S}$ and compute Q. Then apply the CIM to $T_Q(z) = Q^H T(z) Q$ with N quadrature points $\tilde{z}_j$. Store the matrices $T_Q(\tilde{z}_1), \dots, T_Q(\tilde{z}_N)$ computed for the CIM and assemble the interpolation polynomial $T_Q^I(z)$. Compute the residuals $\|T_Q^I(\lambda_j)g_j\|_2/\|g_j\|_2$.
3	Choose N^{RR} quadrature points $\tilde{z}_j^{RR}$, build $\hat{S}$ and compute Q. Then choose N quadrature points z_j and calculate the matrices $T_Q(z_1), \dots, T_Q(z_N)$. Use these matrices to assemble the interpolant $T_Q^I(z)$. Apply the CIM with $4N$ quadrature points to $T_Q^I(z)$. Compute the residuals $\|T(\lambda_j)v_j\|_2/\|v_j\|_2$.
4	Choose N^{RR} quadrature points $\tilde{z}_j^{RR}$, build $\hat{S}$ and compute Q. Then choose N quadrature points z_j and calculate the matrices $T_Q(z_1), \dots, T_Q(z_N)$. Use these matrices to assemble the interpolant $T_Q^I(z)$. Apply the CIM with $4N$ quadrature points to $T_Q^I(z)$. Compute the residuals $\|T_Q^I(\lambda_j)g_j\|_2/\|g_j\|_2$.

For our computations we used all four meshes described in Table 8.3. Since the results, as far as the residuals are concerned, are quite similar for the different meshes, we would like to present the results for *pipe_3*. As already mentioned above, we always chose $N^{RR} = N$ in the CIM-RRs. For all tests the following parameters were fixed:

CIM	$N = 15,\ K = 2,\ l = 20,\ \delta = 10^{-7}$
CIM-RRs	$N^{RR} = N = 15,\ L = 6,\ K = 2,\ \delta^{RR} = 10^{-8},\ \delta = 10^{-7}$

Note that several additional tests suggested that it is not advisable to choose N much lower than 15 in this example. In the Fast BEM we used $\text{hmat}_{\text{eps}} = 10^{-5}$ as approximation accuracy for the $\mathcal{H}$ matrices. The equations

$$T(z)X = U$$

were solved by using the LU decomposition in the case of dense matrices and by using the GMRES method with a tolerance $\text{gmres}_{\text{tol}} = 10^{-8}$ in the case of $\mathcal{H}$-matrices. The following plots show the calculated eigenvalues and their corresponding residuals by using the CIM and the different variants of the CIM-RRs for dense matrices and $\mathcal{H}$-matrices. The meaning of the characters used in the plots is explained by the legend to the right of this text.

○	1e-03
◇	1e-04
×	1e-05
✳	1e-06
○	1e-07
◇	1e-08
×	1e-09
✳	1e-10
○	1e-11
◇	1e-12
×	1e-13
✳	1e-14
○	1e-15
◇	1e-16

BEM (dense matrices):

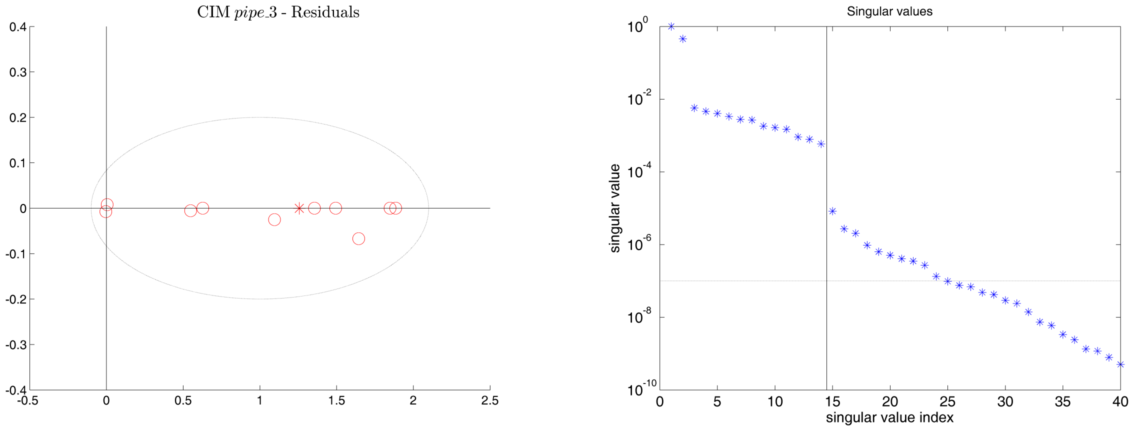


Figure 8.13: CIM, N=15, *pipe_3*.

When looking at the CIM in Figure 8.13, we can notice that all 14 eigenvalues were found by the algorithm and that their residuals have an order of around 1e-06 or 1e-07. Actually, we can only see eleven eigenvalues since the ones at 1.848, 1.493, and 1.355 have algebraic multiplicity two. Moreover, no spurious eigenvalues occur. This is typical of dense matrices if N is sufficiently large. Note that in the plot of the singular values (in all plots they were scaled such that the first one always is one) there are two gaps, one at around 10^{-1} and the other at 10^{-4}. If we had therefore set $\delta = 10^{-4}$, we would not have computed other eigenvalues lying outside of $\Gamma_{\mathcal{C}}$ either. However, for the choice $\delta = 10^{-1}$, we would have got completely wrong results. Hence, it is not advisable to use the criteria

$$\kappa = \underset{j=1,\ldots,Kl-1}{\operatorname{Argmax}} (\sigma_j/\sigma_{j+1})$$

for the determination of the numerical rank, although this is proposed in some articles.

The numerical results for the different variants of the CIM-RRs are presented below.

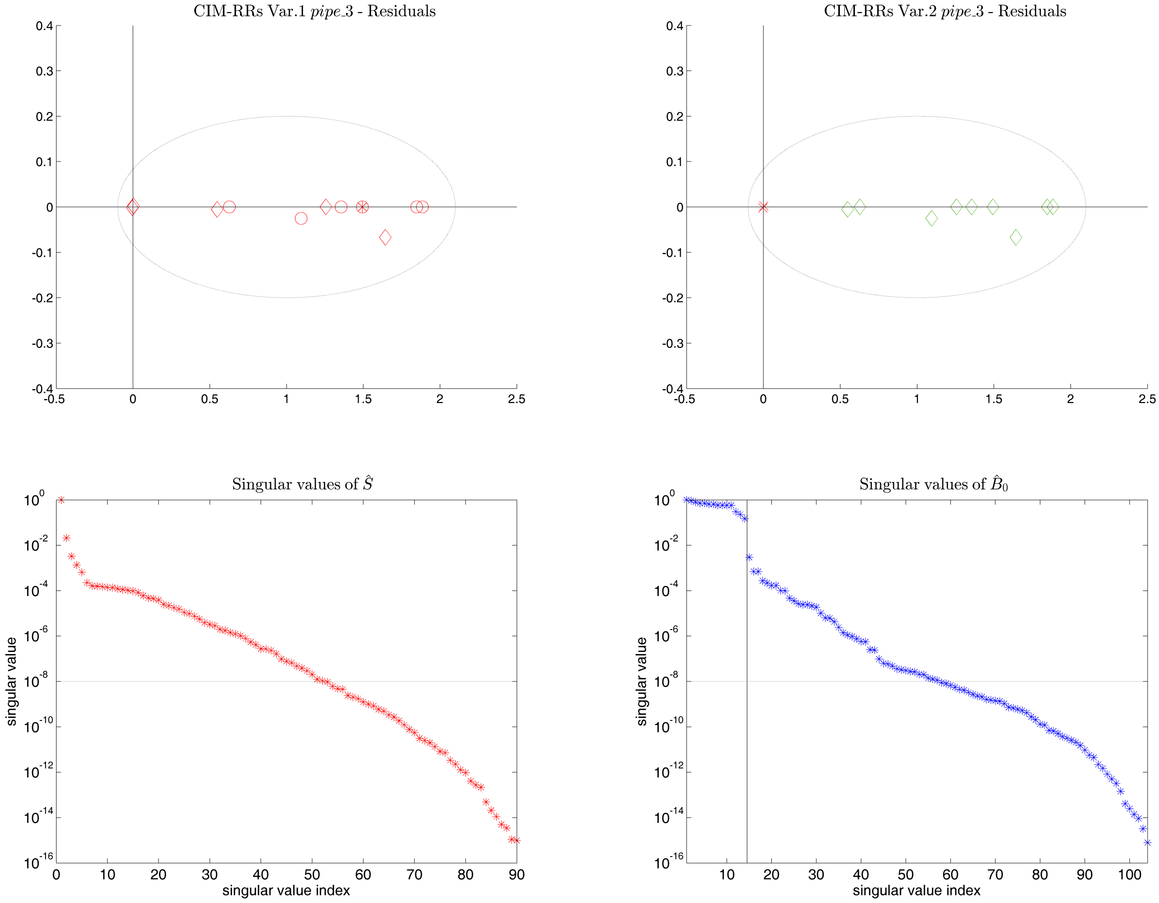
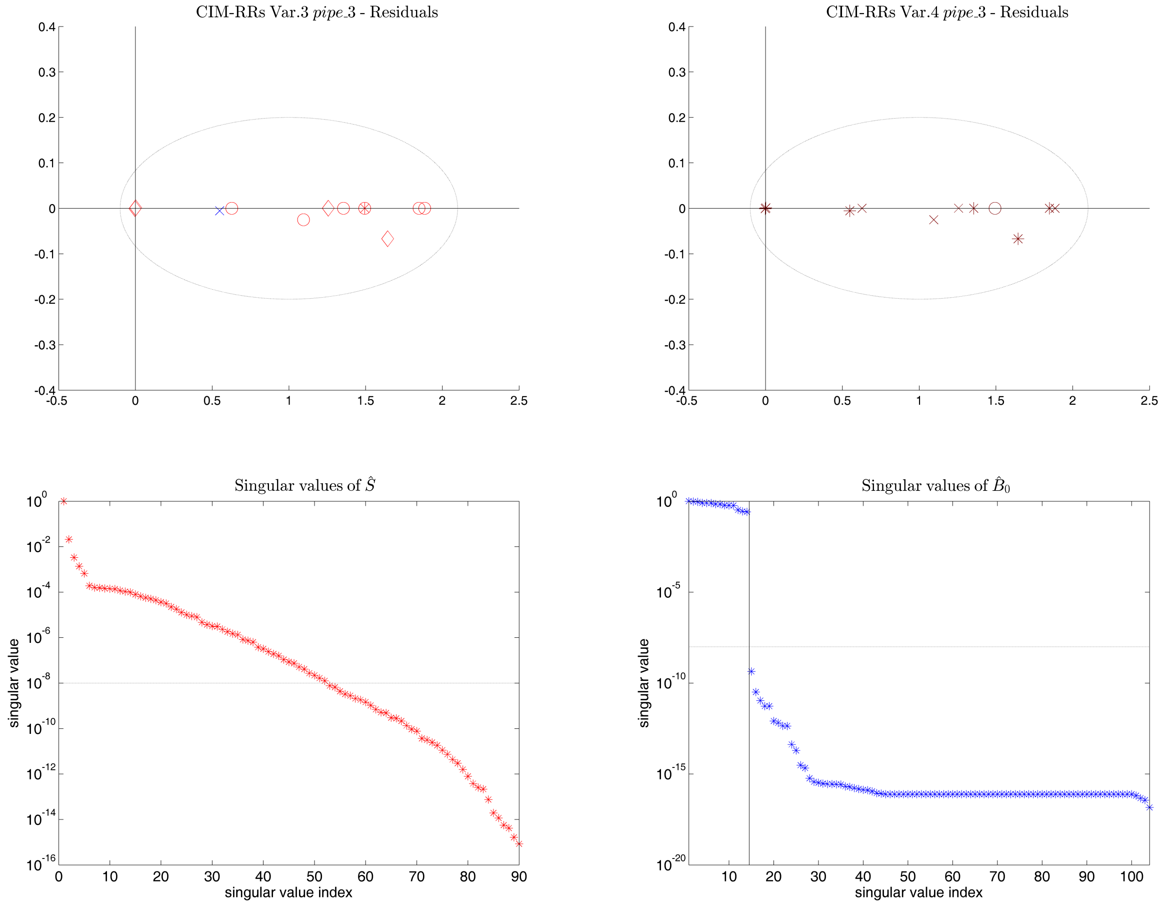


Figure 8.14: CIM-RRs Variants 1 and 2, N=15, *pipe_3*.


CIM-RRs Var.3 *pipe_3* - Residuals

CIM-RRs Var.4 *pipe_3* - Residuals

Singular values of $\hat{S}$

Singular values of $\hat{B}_0$

Figure 8.15: CIM-RRs Variants 3 and 4, N=15, *pipe_3*.

We observe that the results for Variant 1 in Figure 8.14 are nearly the same as for the CIM. Again, all eigenvalues inside the contour are found, no spurious ones occur, and the residuals nearly coincide. In contrast, the residuals in Variant 2 are all a few orders lower than those in Variant 1. Note that in the singular values of the matrix $\hat{B}_0$ we can observe a gap of one order. This gap comes directly after the 14^{th} singular value.

In Figure 8.15, where the CIM is applied to the interpolated projected NEP, we can notice that the gap in the singular values of $\hat{B}_0$ is around eight orders. Since δ was chosen 10^{-7}, the numerical rank now coincides with the real number of eigenvalues which has the effect that no other eigenvalues lying outside the ellipse are calculated. Furthermore, the residuals in Variant 4 are extremely low. Let us now turn to the Fast BEM.

Fast BEM:

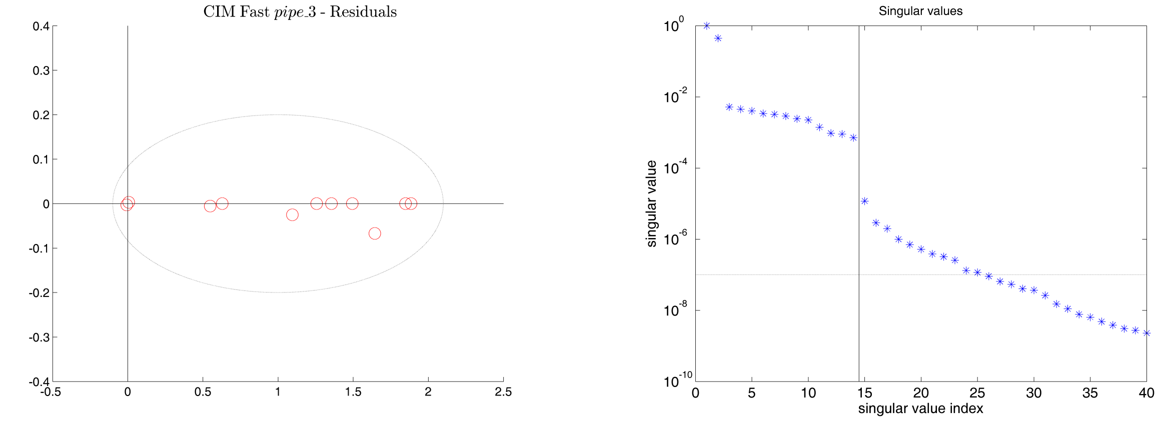


Figure 8.16: CIM, Fast BEM, N=15, *pipe_3*.

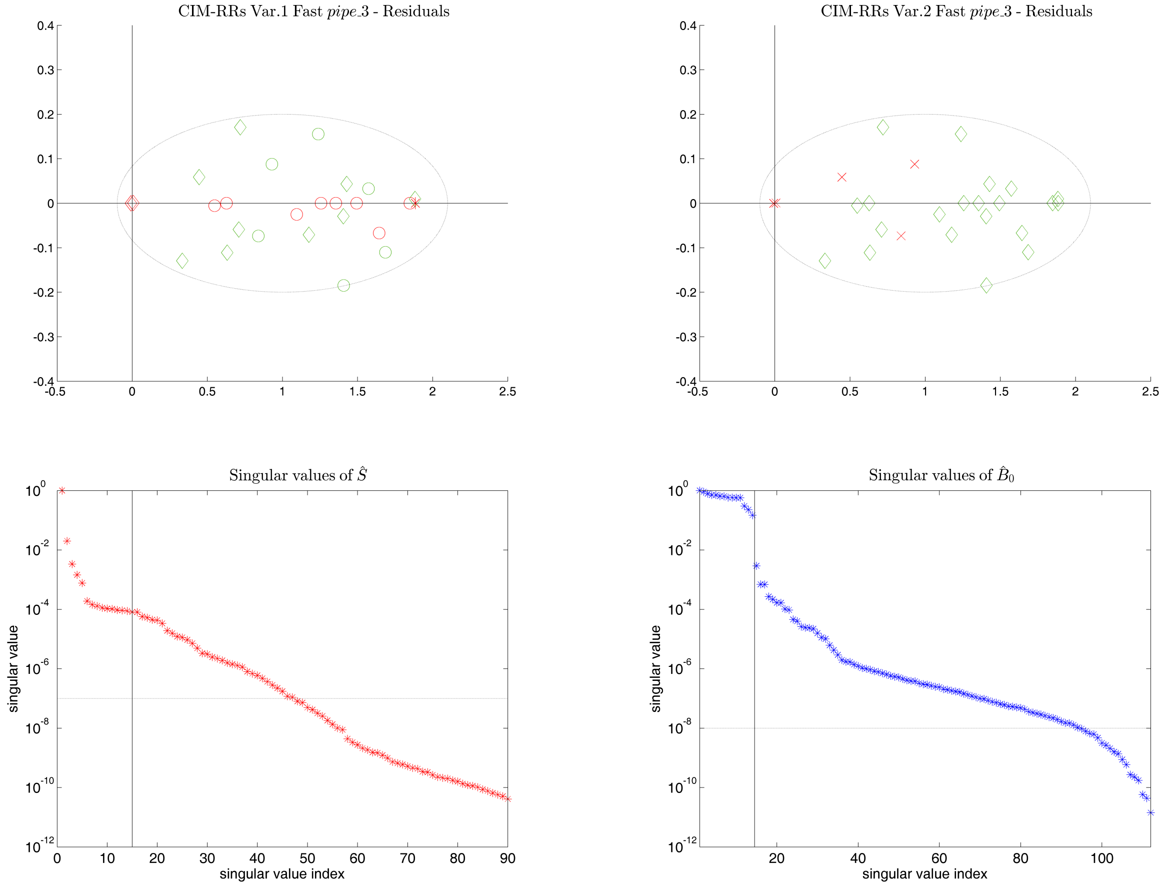


Figure 8.17: CIM-RRs Variants 1 and 2, Fast BEM, N=15, *pipe_3*.

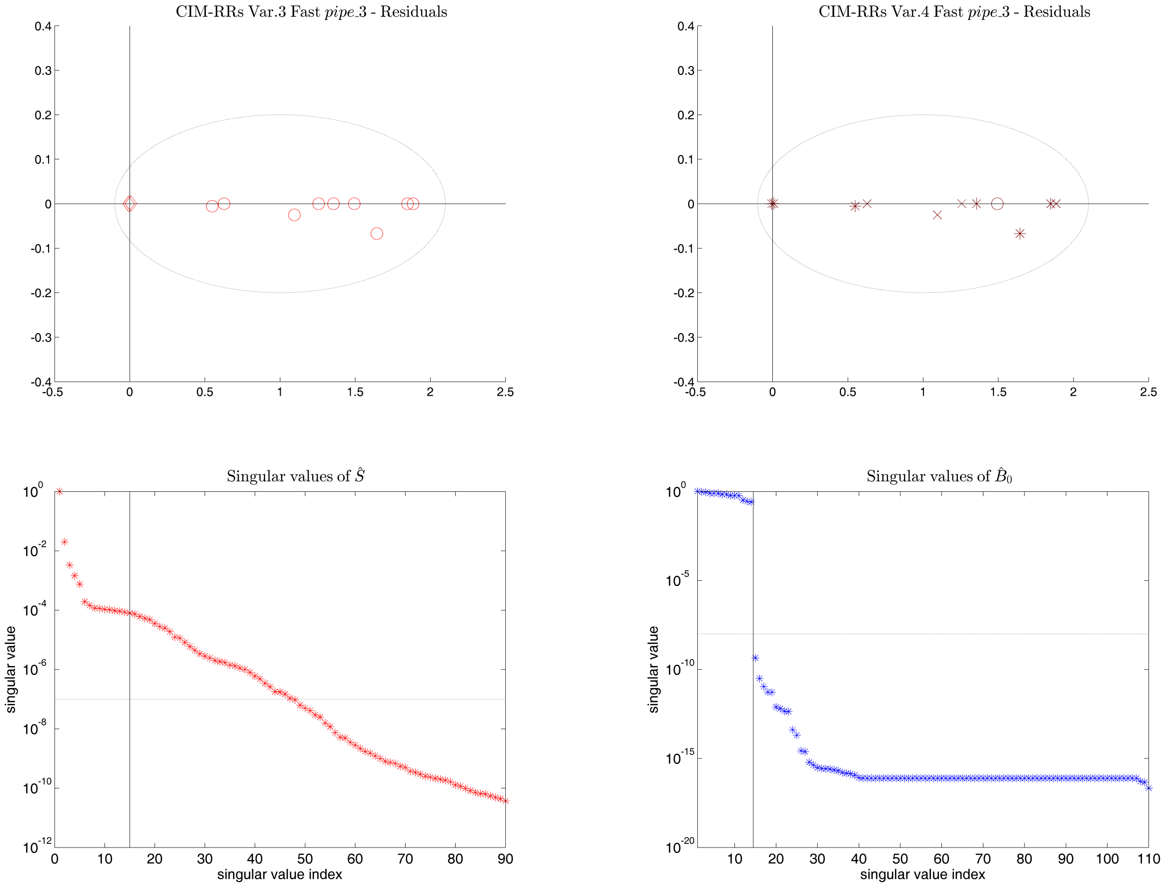


Figure 8.18: CIM-RRs Variants 3 and 4, Fast BEM, N=15, *pipe_3*.

When we compare the above plots for the Fast BEM with those from the BEM, we can notice that the residuals for the CIM (Figure 8.16, Figure 8.13) are quite the same. However, in Figure 8.17 spurious eigenvalues lying inside the contour occur. For Variant 1 it is no problem to distinguish the real eigenvalues and the spurious ones since there is a difference of a least three orders in the residuals, though in Variant 2 such a distinction is not possible any more. Hence, Variant 2 should be avoided completely. The results for Variant 3 and Variant 4 in Figure 8.18 nearly coincide with those from the BEM.

To conclude, due to the large gap in the singular values of $\hat{B}_0$, Variant 3 or 4 seem to be most suitable for the computation of eigenvalues. Moreover, these variants also have much smaller computational times than the CIM at least for the case when $\mathcal{H}$-matrices are used. In the case of dense matrices, Variant 3 and 4 certainly have a large potential provided that the matrices which have to be assembled for $\hat{S}$ can be stored.

Computational times:

CIM:

Mesh	Time in seconds (BEM)	Time in seconds (Fast BEM)
pipe_0	15.3	50.8
pipe_1	463.0	1500.7
pipe_2	2501.9	5519.4
pipe_3	18883.6	22383.0

CIM-RRs Variant 1:

Mesh	Time in seconds (BEM)	Time in seconds (Fast BEM)
pipe_0	22.9	54.8
pipe_1	638.9	1462.8
pipe_2	3704.6	3798.5
pipe_3	23504.9	11700.0

CIM-RRs Variant 2:

Mesh	Time in seconds (BEM)	Time in seconds (Fast BEM)
pipe_0	15.9	38.6
pipe_1	489.7	904.1
pipe_2	2958.3	2688.3
pipe_3	19517.4	8960.6

CIM-RRs Variant 3:

Mesh	Time in seconds (BEM)	Time in seconds (Fast BEM)
pipe_0	23.1	53.6
pipe_1	643.2	1194.6
pipe_2	3626.3	3387.7
pipe_3	22589.3	10541.8

CIM-RRs Variant 4:

Mesh	Time in seconds (BEM)	Time in seconds (Fast BEM)
pipe_0	16.1	38.5
pipe_1	497.8	906.6
pipe_2	2862.0	2744.0
pipe_3	18929.0	9118.7

Bibliography

[1] J. Asakura, T. Sakurai, H. Tadano, T. Ikegami, and K. Kimura. A numerical method for nonlinear eigenvalue problems using contour integrals. *JSIAM Letters*, 1(0):52–55, 2009.

[2] A. Austin, P. Kravanja, and L.N. Trefethen. Numerical algorithms based on analytic function values at roots of unity. *SIAM Journal on Numerical Analysis*, 52(4):1795–1821, 2014.

[3] A. 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.

[4] M. Bebendorf. *Hierarchical Matrices: A Means to Efficiently Solve Elliptic Boundary Value Problems*. Springer, Berlin Heidelberg, 2008.

[5] J.-P. Berrut and L.N. Trefethen. Barycentric Lagrange Interpolation. *SIAM Review*, 46(3):501–517, 2004.

[6] T. Betcke, N. J. Higham, V. Mehrmann, C. Schröder, and F. Tisseur. NLVEP: A Collection of Nonlinear Eigenvalue Problems. *MIMS EPrint 2011.116*, 2011.

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

[8] M. Costabel. Boundary integral operators on Lipschitz domains: elementary results. *SIAM J. Math. Anal.*, 19(3):613–626, 1988.

[9] W. Dahmen and A. Reusken. *Numerik für Ingenieure und Naturwissenschaftler*. Springer, Berlin Heidelberg, 2008. 2. Auflage.

[10] Ö. Eğecioğlu and Ç.K. Koç. A Fast Algorithm for Rational Interpolation Via Orthogonal Polynomials. *Mathematics of Computation*, 53(187):249–264, 1989.

[11] G. Fischer. *Lineare Algebra - Eine Einführung für Studienanfänger*. Springer Spektrum, Wiesbaden, 2014. 18. Auflage.

[12] W. Fischer and I. Lieb. *Einführung in die Komplexe Analysis*. Vieweg+Teubner, Wiesbaden, 2010.

[13] C. Geuzaine and J.-F. Remacle. Gmsh: a three-dimensional finite element mesh generator with built-in pre- and post-processing facilities. *International Journal for Numerical Methods in Engineering*, 79(11):1309–1331, 2009.

[14] O. Karma. Approximation in eigenvalue problems for holomorphic Fredholm operator functions I. *Numer. Funct. Anal. Optim.*, 17(3-4):365–387, 1996.

[15] O. Karma. Approximation in eigenvalue problems for holomorphic Fredholm operator functions II (Convergence rate). *Numer. Funct. Anal. Optim.*, 17(3-4):389–408, 1996.

[16] S. Kim and J.E. Pasciak. The computation of resonances in open systems using a perfectly matched layer. *Mathematics of Computation*, 78(267):1375–1398, 2009.

[17] V. Kozlov and V. Maz'ya. *Differential Equations with Operator Coefficients with Applications to Boundary Value Problems for Partial Differential Equations.* Springer, Berlin Heidelberg, 1999.

[18] P. Kravanja. *On Computing Zeros of Analytic Functions and Related Problems in Structured Numerical Linear Algebra.* PhD thesis, Katholieke Universiteit Leuven, 1999.

[19] W. McLean. *Strongly Elliptic Systems and Boundary Integral Equations.* Cambridge University Press, Cambridge, 2000.

[20] V. Mehrmann and H. Voss. Nonlinear eigenvalue problems: a challenge for modern eigenvalue methods. *GAMM-Mitteilungen*, 27(2):121–152, 2004.

[21] W. Śmigaj, S. Arridge, T. Betcke, J. Phillips, and M. Schweiger. Solving Boundary Integral Problems with BEM++. *ACM Trans. Math. Software 41*, 6:1–40, 2015.

[22] F.W.J. Olver, D.W. Lozier, R.F.Boisvert, and Ch.W.Clark. *NIST Handbook of Mathematical Functions.* U.S. Department of Commerce National Institute of Standards and Technology, Washington, DC, 2010.

[23] Fernando Pérez and Brian E. Granger. IPython: a system for interactive scientific computing. *Computing in Science and Engineering*, 9(3):21–29, May 2007. doi:10.1109/MCSE.2007.53.

[24] E.B. Saff. An extension of Montessus de Ballore's Theorem on the Convergence of Interpolating Rational Functions. *Journal of Approximation Theory*, 6(1):63–67, 1972.

[25] S.A. Sauter and Ch. Schwab. *Boundary Element Methods.* Springer-Verlag, Berlin Heidelberg, 2011.

[26] O. Steinbach. *Numerical Approximation Methods for Elliptic Boundary Value Problems. Finite and Boundary Elements.* Springer, New York, 2008.

[27] O. Steinbach and G. Unger. Convergence analysis of a Galerkin boundary element method for the Dirichlet Laplacian eigenvalue problem. *SIAM J. Numer. Anal.*, 50(2):710–728, 2012. doi:10.1137/100801986.

[28] O. Steinbach and G. Unger. Combined boundary integral equations for acoustic scattering-resonance problems. *Mathematical Methods in the Applied Sciences*, 2016. doi:10.1002/mma.4075.

[29] W.A. Strauss. *Partial Differential Equations: An Introduction.* John Wiley and Sons, 2nd edition, 2008.

[30] G. Unger. *Analysis of Boundary Element Methods for Laplacian Eigenvalue Problems*, volume 6. Verlag der Technischen Universität Graz, 2009.

[31] J. Xiao, C. Zhang, T.-M. Huang, and T. Sakurai. Contour integral based Rayleigh-Ritz method for large-scale nonlinear eigenvalue problems. *arXiv:1510.07522v2*, 2015.

[32] J. Xiao, C. Zhang, T.-M. Huang, and T. Sakurai. Solving large-scale nonlinear eigenvalue problems by rational interpolation approach and resolvent sampling based Rayleigh-Ritz method. *arXiv:1605.07951v1*, 2016.

[33] S. Yokota and T. Sakurai. A projection method for nonlinear eigenvalue problems using contour integrals. *JSIAM Letters*, 5(0):41–44, 2013.

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Masterarbeit identisch.

Datum

Unterschrift