Inexact Iterative Projection Methods for Linear and Nonlinear Eigenvalue Problems

Vom Promotionsausschuss der
Technischen Universität Hamburg

zur Erlangung des akademischens Grades

Doktor der Naturwissenschaften (Dr. rer. nat.)

genehmigte Dissertation

von

Nicolai Rehbein

aus Hamburg

2019

Prüfungsausschuss:

Vorsitzender:	Prof. Dr. Norbert Hoffmann
Betreuer:	Prof. Dr. Heinrich Voß
Gutachterin:	Prof. Dr. Sabine Le Borne
Gutachter:	Prof. Dr. Daniel Kressner

Tag der mündlichen Prüfung: 30. August 2019

Danksagungen

Mein besonderer Dank gilt meinem Betreuer Herrn Prof. Dr. Heinrich Voß, der mich den gesamten Weg von der Auswahl des Themas bis zur Diskussion über die Ergebnisse begleitet hat. Sein fundiertes Wissen und seine Erfahrungen waren eine große Bereicherung beim Anfertigen dieser Arbeit.

Des weiteren danke ich Frau Prof. Dr. Sabine Le Borne sowie Herrn Prof. Dr. Daniel Kressner für die Zeit, die sie in die Begutachtung meiner Arbeit investiert haben. Herrn Prof. Dr. Norbert Hoffmann gilt mein Dank für die Übernahme des Vorsitzes des Prüfungsausschusses und für den reibungslosen Ablauf der mündlichen Prüfung.

Bedanken möchte ich mich auch bei Herrn Prof. Dr. Wolfgang Mackens für die Möglichkeit während meiner Promotion als Wissenschaftlicher Mitarbeiter am Institut für Mathematik angestellt gewesen zu sein. Allen Kolleginnen und Kollegen des Instituts für Mathematik gilt ebenfalls mein Dank für die aufregenden Jahre dort, die vielen Diskussionen und nicht zuletzt den Doktor-Hut. Besonders hervorheben möchte ich an dieser Stelle Prof. Dr. Jörg Lampe für die kritische Lektüre meiner Arbeit und die anschließende Diskussion.

Meinen Kolleginnen und Kollegen bei DNV GL danke ich für die anregenden Diskussionen und das Feedback zu meiner Arbeit sowie die Flexibilität während der finalen Prüfungsphase.

Schließlich gilt auch noch großer Dank meiner Familie. Meine Frau, Julia, hat mich während der gesamten Promotionszeit immer unterstützt und mir immer den Rücken frei gehalten. Ohne dich wäre diese Arbeit nicht möglich gewesen. Meinen Eltern gilt großer Dank für die Unterstützung während meiner gesamten Ausbildung, insbesondere auch die Finanzierung meines Studiums.

Das Ergebnis habe ich schon,
jetzt brauche ich nur noch den Weg,
der zu ihm führt.

CARL FRIEDRICH GAUSS

To my son Philipp

Contents

List of Figures

List of Algorithms

List of Tables

Notations

Symbol or Letter	Explanation
$A \dots Z$	usually matrices
$a \dots z$	usually vectors
$\alpha \dots \zeta$	in general scalars
$'$	derivative
$\dot{}$	derivative with respect to the time t
A^T	the transpose of a matrix A
A^H	the complex conjugate of the transpose of the matrix
i	imaginary unit $i = \sqrt{-1}$
$j,\ k$	indices
O	zero matrix
I	identity matrix
$\langle v, w \rangle_B$	denotes the B-inner product for $v, w \in \mathbb{C}^n$, i.e. $\langle v, w \rangle_B := w^H B v$
$\Vert v \Vert$	norm of the vector (unless stated otherwise it means the euclidean norm)
$\Vert v \Vert_B$	the induced norm by $\langle \cdot, \cdot \rangle_B$.
$\Vert A \Vert$	matrix norm (unless stated otherwise it means the euclidean norm)
λ	eigenvalue
$\mathrm{spec}(A)$	the spectrum (set of all eigenvalues) of the matrix A
$\mathrm{spec}(T(\cdot))$	the set of all eigenvalues in the domain of the nonlinear eigenvalue problem
$\sigma_1(A), \dots, \sigma_n(A)$	singular values of an matrix $A \in \mathbb{C}^{m \times n}$ in descending order
$\sigma_{min}(A)$	smallest non-zero singular value
$e^1, \dots, e^n$	the canocial basis of the $\mathbb{R}^n$ or $\mathbb{C}^n$, respectively
$\mathrm{int}\ D$	If D is a closed set, that $\mathrm{int}\ D = D \setminus \partial D$, if D is an open set $\mathrm{int}\ D = D$.
Π_n	Set of polynomials with maximal degree n
$\bar{S}(\lambda, \tau)$	$\{\mu \in \mathbb{C} \mid \lvert \mu - \lambda \rvert \leqslant \tau\}$
$\mathbb{R}_{\geqslant 0}$	all nonnegative real numbers

Chapter 1

Introduction

Linear eigenvalue problems occur in many applications in physics and engineering. When analyzing dynamic systems, the determination of eigenvalues and their eigenvectors is an important task. The location of the eigenvalues in the complex plane gives information about the stability of the system. Furthermore, eigenfrequencies and their eigenmodes are highly relevant for all vibrating systems. It must be guaranteed that no external vibrations are in neighborhoods of the system's eigenfrequencies.

Nonlinear eigenvalue problems result from more complex systems. If, for instance, the system has a time delay or a damping part, then the resulting eigenvalue problem becomes nonlinear. To consider these problems in the most general form, nonlinear eigenvalue problems are described by

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

where $T : \mathbb{C} \supset D \to \mathbb{C}^{n\times n}$ is a family of matrices, $\lambda \in D$ is the eigenvalue and $x \in \mathbb{C}^n \setminus \{0\}$ is an eigenvector.

Determining the eigenvalues of a linear $n \times n$ eigenvalue problem is equivalent to finding the roots of a polynomial with degree n. Once the eigenvalue is known, the eigenvector can be obtained by the solution of a linear system. However, for root finding the analytical solution is only possible for special cases or $n \leqslant 4$. Thus, many numerical methods have been developed to obtain approximations for the eigenvalues and eigenvectors. For nonlinear eigenvalue problems, there even exist very low dimensional problems, which cannot be solved analytically.

The bottleneck of many numerical methods is solving a linear system which is required in every iteration step. The power method avoids this expensive operation, but it can only be used to determine the largest eigenvalue in magnitude of a linear eigenvalue problem. Other methods, like the QR-algorithm, require a QR decomposition in every iteration step, which is even more expensive than solving a linear system. Hence, it is suggested to replace the solution of the linear system by an *inexact solution* for large[1] n. This inexact solution is computed by using an iterative method to solve the linear system. The nonlinear Arnoldi method determines a subspace extension by replacing the linear system by simplification where the occurring matrix is replaced by a preconditioner.

An important question in this context is how accurate the inexact solution has to be in order for the eigenvector iteration to converge (more or less) quickly to an

[1] i.e. $n \gg 1000$

eigenpair. This question was discussed by Smit and Paardekooper in [68] for symmetric matrices. Additionally, there are papers by Simoncini and Eldén [64] and Notay [52] discussing the same context. Berns-Müller and Spence present convergence results for non-symmetric problems in [7].

In [27] and [51], Hochstenbach and Notay discuss stopping criteria for inner iterations.

The main problem of these methods is the growing number of necessary inner iterations with the convergence of the outer iteration, i.e. if we are close to the solution, many inner iterations are required to solve the corresponding linear system to obtain a better new iterate.

Procedures to stabilize the outer iteration process by a special preconditioning technique were suggested by Freitag and Spence in [15, 16, 17, 18].

Szyld and Xue analyzed an inexact method for nonlinear eigenvalue problems in [70].

In this thesis, we consider, in general, nonlinear eigenvalue problems; since linear eigenvalue problems are special cases of the nonlinear ones.

Voss developed a theory in [77] about the robustness of search space expansions. Here, the case is considered where a current eigenvector approximation is expanded by a correction to a two-dimensional search space. The search space expansion has to be chosen orthogonal to the current subspace to minimize the influence of perturbations.

A typical method which provides an orthogonal search space expansion is the Jacobi-Davidson method which was introduced by Sleijpen and Van der Vorst in [66]. This method is motivated by Newton's method where the Newton correction is orthogonalized against the current iterate. We regard the Jacobi-Davidson method more generally as a procedure to stabilize an iterative method to determine eigenvalues and eigenvectors of a linear or nonlinear eigenvalue problem.

Based on this theory, we will discuss the question how a perturbed search space influences the convergence behavior of the iterative projection method. Therefore, we do not distinguish between linear and nonlinear eigenvalue problems, but we consider those cases where special structural properties (like minmax characterization or local symmetry) can be exploited.

The good convergence results are reasoned by the use of Rayleigh functionals, if these are available. Their stationarity at eigenvectors results in very good approximation properties. Motivated by excellent numerical results, we analyze a special class of nonlinear eigenvalue problems with two real parameters[2]. These problems are caused by analyzing dynamic systems with time-delay for critical frequencies. We present different kinds of functionals and prove their stationarity at eigenvectors.

This thesis is structured into six chapters, first being this introduction.

In the second chapter we give important preliminaries about the theory of nonlin-

[2] instead of one complex eigenvalue

ear eigenvalue problems from the literature. We then introduce the most common numerical methods to solve these problems. Moreover, inexact methods and their convergence theories are presented. The chapter is finished with a numerical example, which reveals the problems caused by inexact solutions of the linear systems during the iteration.

The third chapter introduces the Jacobi-Davidson method. After briefly describing the history of this method, the Jacobi-Davidson algorithm is given. We discuss the *Jacobi-Davidson method without Subspace Acceleration* and the convergence theory for its inexact version. This has been obtained by Szyld and Xue in [70]. For the robustness against perturbations we regard the theory published by Voss in [77]. Finally, we give a very general version of the Jacobi-Davidson method, which can be constructed based on every vector iteration to determine eigenvalues.

In Chapter 4 we analyze the convergence behavior for two-dimensional search spaces which have been rotated by an angle ε. We present a convergence proof for the smallest eigenvalue of Hermitian problems whose eigenvalues can be minmax characterized. Furthermore, a different approach for the convergence to interior eigenvalues and its eigenvectors is given. Therefore, we answer the question where Ritz vectors are located in the two dimensional search plane after solving the projected problem. Moreover, we consider real non-symmetric matrices where no minmax characterization can be applied and we give a convergence proof for this case. At the end, the numerical example from Chapter 2 is reconsidered. Here, the advantages of the Jacobi-Davidson method are shown. Furthermore, an example for a nonlinear problem is presented.

In the fifth chapter, the two-parameter eigenvalue problem is considered. We introduce methods to solve this kind of nonlinear eigenvalue problems for small dense problems as well as large sparse matrices. Finally, we present different variants of Rayleigh functionals and analyze their properties.

Chapter 6 includes a summary and the conclusions of this thesis. We will review the most important results and state some open questions. Lastly, we give an outlook on possible projects in the future.

The appendix gives deeper information about the three dimensional geometric approach to prove the convergence in Chapter 4.

Chapter 2

Nonlinear Eigenvalue Problems

In this chapter we give an introduction to the field of nonlinear eigenvalue problems. Nonlinear eigenvalue problems can be seen as a generalization of linear eigenvalue problems. Further information about the theory of linear eigenvalue problems can be found, for instance, in [3], [20], [57], and [83].

To begin with, nonlinear eigenvalue problems are introduced and the variational characterization of eigenvalues is explained. Afterward, the Rayleigh functional is considered with respect to its approximation properties. Additionally, error measures for numerical methods seeking one or more eigenpairs are presented. Finally, different numerical methods to compute eigenvalues and eigenvectors, their advantages and disadvantages, as well as their costs, are discussed.

2.1 Problem Description

The nonlinear eigenvalue problem is given by the following definition.

Definition 2.1. Let

$$T : \left\{ \begin{array}{rcl} \mathbb{C} \supset D & \to & \mathbb{C}^{n\times n} \\ \lambda & \mapsto & T(\lambda) \end{array} \right.$$

be a continuous function, which maps a parameter λ onto a family of matrices. Every $\lambda \in D$ satisfying that

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

has a nontrivial solution is an *eigenvalue.* Every nontrivial solution $v \in \mathbb{C}^n$ of (2.1) is called *(right) eigenvector*. Analogously, each nontrivial solution $w \in \mathbb{C}^n$ of

$$w^H T(\lambda) = 0^H \tag{2.2}$$

is called *left eigenvector*.

The hereby defined eigenvalue problem is the most general representation. Possible cases are shown in the following remark.

Remark 2.2. Possible examples for T are:

- the linear eigenvalue problem: $T(\lambda) = \lambda I - A$,
- the generalized linear eigenvalue problem: $T(\lambda) = \lambda B - A$,
- the quadratic eigenvalue problem: $T(\lambda) = \lambda^2 C + \lambda B - A$,
- the polynomial eigenvalue problem: $T(\lambda) = \lambda^n A_n + \ldots + \lambda A_1 + A_0$,
- the delay eigenvalue problem: $T(\lambda) = A - \lambda B + \exp(\lambda\tau)C$,
- the rational eigenvalue problem: $T(\lambda) = -K + \lambda M + \sum_{j=1}^{k} \frac{\lambda}{\sigma_j - \lambda} C_j$.

A good overview about nonlinear eigenvalue problems can be found in [19], [25], [46], and [81]. Further theory about the quadratic eigenvalue problem is given in [72].

Similar to the linear eigenvalue problems, the eigenvalues of (2.1) have also algebraic and geometric multiplicities. A simple *eigenvalue* is defined as follows.

Definition 2.3. Let $\hat{\lambda}$ be an eigenvalue of (2.1) and $\hat{v}$ a corresponding right eigenvector. The eigenvalue $\hat{\lambda}$ is called *(algebraically) simple*, if and only if

$$\frac{d}{d\lambda} \det T(\lambda)\Big|_{\lambda=\hat{\lambda}} \neq 0.$$

The geometric multiplicity is defined analog to the linear eigenvalue problems:

Definition 2.4. Let $\hat{\lambda}$ be an eigenvalue of (2.1) and $\hat{v}$ a corresponding right eigenvector. The geometric multiplicity is defined as for the linear eigenvalue problems by

$$\gamma(\hat{\lambda}) = n - \text{rank } (T(\hat{\lambda}))$$

The following two theorems proven by Schreiber in [62] provide the geometric multiplicity and one further property.

Theorem 2.5 ([62]).
Let $\hat{\lambda}$ be an algebraically simple eigenvalue of (2.1). *Then, the eigenvalue $\hat{\lambda}$ is geometrically simple, i.e. rank* $(T(\hat{\lambda})) = n - 1$.

Theorem 2.6 ([62]).
Let $\hat{\lambda}$ be a geometrically simple eigenvalue of (2.1), *i.e.,* $\det(T(\hat{\lambda})) = 0$ *and dim ker* $T(\hat{\lambda}) = 1$, *ker* $T(\hat{\lambda}) = span\{\hat{v}\}$, *ker* $T(\hat{\lambda})^H = span\{\hat{w}\}$,
with $\|\hat{v}\|_2 = \|\hat{w}\|_2 = 1$, *and let T be differentiable. Then we have*

$$\hat{\lambda} \textit{ is algebraically simple } \Leftrightarrow \hat{w}^H T'(\hat{\lambda})\hat{v} \neq 0.$$

2.2 Variational Characterization of Eigenvalues

Variational characterization is known for self-adjoint linear operators on a Hilbert space $\mathcal{H}$. By using it, the eigenvalues can be localized within a bounded interval. Furthermore, variational characterization is a good tool to compare eigenvalues, interlacing eigenvalues and to prove the convergence of numerical methods. In this section, we will briefly introduce the generalization of variational characterization for nonlinear eigenvalue problems. In Chapter 4 we will use this tool for convergence proofs. We restrict this introduction to $\mathcal{H} = \mathbb{C}^n$.

Assumption 2.7 ([82]). *Let $J \subset \mathbb{R}$ be an open interval and $T(\lambda) = T^H(\lambda)$ for all $\lambda \in J$. Define the function*

$$f(\lambda, x) := x^H T(\lambda) x, \quad \lambda \in J, \quad x \in \mathbb{C}^n. \tag{2.3}$$

We assume, that for every fixed $x \in \mathbb{C}^n \setminus \{0\}$ the real scalar equation

$$f(\lambda, x) \overset{!}{=} 0 \tag{2.4}$$

has, at most, one solution $\lambda =: p(x) \in J$.

This defines the Rayleigh functional

$$p : \left\{ \begin{array}{rcl} D(p) \subset \mathbb{C}^n & \to & J \\ x & \mapsto & p(x) \end{array} \right. .$$

The domain $D(p)$ is a subset of $\mathbb{C}^n$ satisfying that:

- $0 \notin D(p)$,
- (2.4) has exactly one solution $\lambda \in J$ for every $x \in D(p)$.

For generalized linear eigenvalue problems, ($Av = \lambda Bv$), minmax characterization of the eigenvalues is only possible for Hermitian matrices A and Hermitian positive definite matrices B (cf. [57]). The definiteness of B is generalized for nonlinear problems in the following assumption.

Assumption 2.8 ([79]). *Let f be defined as in* (2.3)*. We assume, that for every $x \in D$ and every $\lambda \in J$ with $\lambda \neq p(x)$,*

$$(\lambda - p(x))\, f(\lambda, x) > 0 \tag{2.5}$$

holds.

Firstly, we consider and define overdamped problems and secondly, an approach for non-overdamped eigenvalue problems is presented.

2.2.1 Overdamped Nonlinear Eigenvalue Problems

Definition 2.9. A nonlinear eigenvalue problem is called *overdamped* if $D(p) = \mathbb{C}^n \setminus \{0\}$. Otherwise, the problem is called *non-overdamped.*

The name originates from the quadratic eigenvalue problem

$$T(\lambda) = \lambda^2 M + \lambda C + K, \quad M, C, K \in \mathbb{R}^{n\times n}, \text{symmetric positive definite.}$$

The function f, defined in (2.3), is obtained by

$$f(\lambda, x) = \lambda^2 x^H M x + \lambda x^H C x + x^H K x.$$

Taking the roots of this polynomial yields

$$p_\pm(x) = -\frac{1}{2x^H M x}\left(x^H C x \pm \sqrt{(x^H C x)^2 - 4x^H K x \cdot x^H M x}\right).$$

Let J_+ and J_- be the images of the two functionals p_+ and p_-.

If

$$\left(x^H C x\right)^2 - 4x^H K x \cdot x^H M x > 0, \quad \forall x \in \mathbb{R}^n \setminus \{0\}, \tag{2.7}$$

J_+ and J_- are real disjoint intervals, containing all eigenvalues of $T(\cdot)$. This implies only real and negative eigenvalues which yields descending solutions without any oscillations for the corresponding system of ordinary differential equations, $M\ddot{q} + D\dot{q} + Kq = 0$.

Duffin [12] shows Poincaré's minmax characterization for overdamped quadratic eigenvalue problems and Rogers [58] for general nonlinear overdamped eigenvalue problems. Here it is suggested to enumerate the eigenvalues in their usual order beginning with the smallest eigenvalue.

2.2.2 Non-Overdamped Nonlinear Eigenvalue Problems

For non-overdamped problems, Voss and Werner introduced the minmax characterization in [82]. With J from Assumption 2.7 they defined for $j \in \mathbb{N}$ and $\lambda \in J$:

$$\mu_j(\lambda) := \sup_{\dim(V)=j} \min_{\substack{v\in V \\ \|v\|_2=1}} v^H T(\lambda) v, \tag{2.8}$$

which yields that $\mu_j(\lambda)$ denotes the j-th largest eigenvalue of the linear eigenvalue problem

$$T(\lambda)v(\lambda) = \mu(\lambda)v(\lambda). \tag{2.9}$$

To be able to define a minmax characterization for nonlinear eigenvalue problems we define the following enumeration of eigenvalues

Definition 2.10. $\lambda \in J$ is a *k-th eigenvalue* of $T(\cdot)$ if $\mu_k(\lambda) = 0$ for $1 \leqslant k \leqslant n$.

The following lemma was proved in [82] for differentiable $T(\cdot)$ and in [79] for the non-differentiable case.

Lemma 2.11. *Under the conditions of Assumptions 2.7 and 2.8, let $\lambda \in J$, and assume that V is a subspace of $\mathbb{C}^n$ satisfying $V \cap D(p) \neq \emptyset$. Then*

$$\lambda \begin{Bmatrix} < \\ = \\ > \end{Bmatrix} \sup_{v \in V \cap D(p)} p(v) \Leftrightarrow \min_{v \in V \setminus \{0\}} \frac{v^H T(\lambda) v}{v^H v} \begin{Bmatrix} < \\ = \\ > \end{Bmatrix} 0$$

Using Lemma 2.11, Voss and Werner presented in [82] the minmax characterization for non-overdamped nonlinear eigenvalue problems. But they required $T(\cdot)$ to be differentiable. In [79] Voss presented a proof where $T(\cdot)$ does not have to be differentiable with respect to λ anymore. This is presented in the following theorem.

Theorem 2.12 (Minmax Characterization).
Let Assumptions 2.7 and 2.8 be satisfied. Then for every $m \in \mathbb{N}$ there exists at most one m-th eigenvalue in J, and the following characterization holds:

$$\lambda_m = \min_{\substack{V \in S_m \\ D(p) \cap V \neq \emptyset}} \sup_{\substack{v \in V \cap D(p) \\ v \neq 0}} p(v),$$

and

$$\exists \lambda_m,\ \lambda_k,\ k < m \Rightarrow \exists \lambda_j,\ k \leqslant j \leqslant m,$$

where S_m denotes the set of all m-dimensional subspaces of $\mathcal{H}$.

The maxmin characterization from Courant, Fischer and Weyl is generalized in [73]. Similarly to the minmax characterization, a proof for maxmin characterization was presented by Voss in [79], where $T(\cdot)$ does not have to be differentiable with respect to λ anymore.

Theorem 2.13 (Maxmin characterization).
Assume that the Assumptions 2.7 and 2.8 are satisfied. If there is an m-th eigenvalue $\lambda_m \in J$ of problem (2.1)*, then*

$$\lambda_m = \max_{\substack{V \in S_{m-1} \\ V^\perp \cap D(p) \neq \emptyset}} \inf_{\substack{v \in V^\perp \cap D(p) \\ v \neq 0}} p(v),$$

and the maximum is attained by $W := span\{u^1, \dots, u^{m-1}\}$ where u^j denotes an eigenvector corresponding to the j-th largest eigenvalue $\mu_j(\lambda_m)$ of $T(\lambda_m)$.

The numerical method *Safeguarded Iteration* was developed from this characterization. It is described in Section 2.5.

2.3 Rayleigh Functional

In the previous section, we introduced the Rayleigh functional to characterize the eigenvalues (cf. Theorem 2.12 and Theorem 2.13). However, this is not the only useful property of a Rayleigh functional. Since evaluating the Rayleigh functional at an eigenvector yields the eigenvalue, the Rayleigh functional can be used to determine a good approximation for the eigenvalue, if an approximation for the eigenvector is available.

2.3.1 Generalized Definition

In Section 2.2 we defined the Rayleigh functional, if Assumptions 2.7 and 2.8 are fulfilled. In this section a broader definition by Schwetlick and Schreiber ([63]) is discussed such that the Rayleigh functional can be applied, even if the eigenvalues cannot be characterized by minmax characterization. The domain of the Rayleigh functional is reduced to a n-dimensional cone around the sought eigenvector.

Before we can define the Rayleigh functional the following assumption for $T(\cdot)$ is needed.

Assumption 2.14. *Let $\hat{\lambda} \in \mathbb{C}$ be a simple eigenvalue of* (2.1) *and let $T(\cdot) : \mathbb{C} \supset D \to \mathbb{C}^{n\times n}$ be a matrix valued function. Suppose that there is a radius $\hat{\tau} > 0$ such that the disc $\bar{S}(\hat{\lambda}, \hat{\tau}) := \{\lambda \in \mathbb{C} \mid |\lambda - \hat{\lambda}| \leqslant \hat{\tau}\}$ is contained in int D and that T is holomorphic on $\bar{S}(\hat{\lambda}, \hat{\tau})$.*

If T and $\hat{\lambda}$ are real, we can consider $T : \mathbb{R} \supset D \to \mathbb{R}^{n\times n}$ instead, and $\bar{S}(\hat{\lambda}, \hat{\tau})$ is an open interval. In this case, we assume that T' exists and that T' is Lipschitz continuous in $\bar{S}(\hat{\lambda}, \hat{\tau})$ with some constant $L > 0$, i.e. that

$$\|T'(\lambda) - T'(\mu)\| \leqslant L\,|\lambda - \mu| \quad \textit{for all} \quad \lambda, \mu \in \bar{S}(\hat{\lambda}, \hat{\tau}).$$

Using this assumption we can now define the Rayleigh functional.

Definition 2.15 ([63])**.** Let Assumption 2.14 be satisfied for some $\hat{\tau} > 0$, $\mathcal{K}_\varepsilon(\hat{v}) := \{x \in \mathbb{C}^n \mid \measuredangle(x, \hat{v}) \leqslant \varepsilon\}$ and $0 < \varepsilon < \frac{\pi}{2}$, then a functional $p : \mathcal{K}_\varepsilon(\hat{v}) \to \bar{S}(\hat{\lambda}, \hat{\tau}) \subset \mathbb{C}$ fulfilling for all $x \in \mathcal{K}_\varepsilon(\hat{v})$

(i) $p(\alpha x) = p(x), \quad \forall \alpha \in \mathbb{C} \setminus \{0\}$,

(ii) $x^H T(p(x))\, x = 0$,

(iii) $x^H T'(p(x))\, x \neq 0$,

defines a *Rayleigh functional.*

Remark 2.16. Let p be a Rayleigh functional as defined in Definition 2.15, then for every eigenpair (λ_j, v^j),

$$p(v^j) = \lambda_j$$

is satisfied.

2.3.2 Stationarity

For this subsection we assume, in addition to Assumption 2.14, that

$$p : \mathcal{K}_\varepsilon(\hat{v}) \to J \subset \mathbb{R} \quad \text{and} \quad T(\lambda) = T^H(\lambda) \quad \forall \lambda \in J$$

holds.

Similar to the Rayleigh quotient, the Rayleigh functional provides a good approximation for the eigenvector, if a vector close to a corresponding eigenvector is given.

This approximation quality of the Rayleigh functional is improved, if it is additionally stationary at eigenvectors. (cf. [62]) Therefore, we will consider the gradient and Hessian of real Rayleigh functionals and refer to the definition of stationarity for the complex case. The following lemma describes the gradient of the Rayleigh functional.

Lemma 2.17. *Let $T : \mathbb{R} \supset J \to \mathbb{R}^{n\times n}$, $p : \mathcal{K}(\hat{v}) \to \mathbb{R}$ be the corresponding Rayleigh functional and $T(\lambda) = T^T(\lambda)$, $\forall \lambda \in \mathbb{R}$. Furthermore, let $T(\cdot)$ be differentiable with respect to λ. Then the gradient of this Rayleigh functional is given by*

$$\nabla p(x) = -\frac{2}{x^T T'(p(x))\, x} T(p(x))\, x.$$

Proof. We consider

$$x^T T(p(x))\, x = 0.$$

Computing the derivative with respect to x yields

$$2T(p(x))\, x + \nabla p(x) \left(x^T T'(p(x))\, x \right) = 0 \tag{2.10}$$

$$\Leftrightarrow \nabla p(x) = -\frac{2}{x^T T'(p(x))\, x} T(p(x))\, x. \tag{2.11}$$

□

Corollary 2.18. *The eigenvectors v^j are stationary points of the Rayleigh functional p.*

Since the expression $x^H T(\lambda) x$ is not holomorphic with respect to x, the stationarity for complex vectors cannot be proven by showing that the gradient is zero as it was done in Lemma 2.17 and Remark 2.18. Schwetlick and Schreiber showed in [63] and [62] the stationarity of a Rayleigh functional at eigenvectors. Therefore, they defined stationarity as follows.

Definition 2.19. A complex function $f : \mathbb{C}^n \supset D \to \mathbb{C}$ is called stationary at $z \in \operatorname{int} D$ if

$$f(z + \Delta z) - f(z) = o(\|\Delta z\|).$$

The Hessian of the Rayleigh functional in real arithmetic will be of interest, too. Therefore, we introduce a further lemma.

Lemma 2.20. *Let $T : \mathbb{R} \supset J \to \mathbb{R}^{n\times n}$, $p : \mathcal{K}(\hat{v}) \to \mathbb{R}$ be the corresponding Rayleigh functional and $T(\lambda) = T^T(\lambda)$, $\forall \lambda \in \mathbb{R}$. Furthermore, let $T(\cdot)$ be twice differentiable with respect to λ. Then the Hessian of the Rayleigh functional is given by*

$$\begin{aligned}\nabla^2 p(x) = -\frac{1}{x^T T'(p(x)) x} \Bigg(& 2T(p(x)) + 2T'(p(x))\, x\, \nabla p(x)^T \\ & + 2\, \nabla p(x) x^T T'(p(x)) + \left(x^T T''(p(x))\, x \right)\, \nabla p(x)\, \nabla p(x)^T \Bigg).\end{aligned} \tag{2.12}$$

Proof. Differentiating (2.10) with respect to x yields

$$\begin{aligned}&2T\left(p(x)\right) + 2T'\left(p(x)\right) x\, \nabla p(x)^T \\ &+ \nabla p(x) \left(x^T T''\left(p(x)\right) x\, \nabla p(x)^T + 2x^T T'\left(p(x)\right)\right) \\ &+ x^T T'\left(p(x)\right) x\, \nabla^2 p(x) = 0. \end{aligned} \tag{2.13}$$

Exploiting $x^T T'\left(p(x)\right) x \neq 0$ we can solve this equation for $\nabla^2 p(x)$ which yields the result. □

Remark 2.21. The Hessian evaluated at an eigenvector $x = \hat{v}$ yields

$$\nabla^2 p\left(\hat{v}\right) = -\frac{2T\left(p\left(\hat{v}\right)\right)}{\hat{v}^T T'\left(p(\hat{v})\right) \hat{v}}.$$

At the end of this chapter, numerical methods are presented which exploit this property of the Rayleigh functional.

2.3.3 Approximation Properties

The stationarity of Rayleigh functionals at eigenvectors improves the approximation property for the eigenvalue. This is discussed in detail by Schwetlick and Schreiber in [63, Cor. 18 and Th. 21]. They prove that

$$|p(u) - \hat{\lambda}| = \begin{cases} \mathcal{O}\left(\tan^2\left(\measuredangle(u,\hat{v})\right)\right) & \text{if } T(\cdot) \text{ is Hermitian} \\ \mathcal{O}\left(\tan\left(\measuredangle(u,\hat{v})\right)\right) & \text{else} \end{cases} \tag{2.14}$$

holds, if $T(\cdot)$ is holomorphic and u is sufficiently close to the eigenvector $\hat{v}$.

2.3.4 Generalized Rayleigh Functional

Evaluating the Rayleigh functional requires the solution of a scalar nonlinear equation. This might be expensive. Therefore, this step can be replaced by one Newton step applied to the nonlinear equation

$$\left(x^{(k+1)}\right)^H T(\lambda) x^{(k+1)} \overset{!}{=} 0,$$

thus

$$\lambda^{(k+1)} = \lambda^{(k)} - \frac{\left(x^{(k+1)}\right)^H T(\lambda^{(k)}) x^{(k+1)}}{(x^{(k+1)})^H T'(\lambda^{(k)}) x^{(k+1)}}.$$

This simplification is called *generalized Rayleigh functional* ([40, 60]).

2.3.5 Two-sided Rayleigh Functional

Based on Ostrowski's two-sided Rayleigh quotient in [55, 56], there also exists a two-sided Rayleigh functional, which can be used if $T(\lambda)$ is not Hermitian for $\lambda \in D$. This was defined by Schreiber and Schwetlick in [63]:

Definition 2.22. Let $\hat{\lambda}$ be an simple eigenvalue of (2.1). Let $\hat{v}$ and $\hat{w}$ be a right and a left eigenvector corresponding to $\hat{\lambda}$. The map

$$p : \mathcal{K}_\varepsilon(\hat{v}) \times K_\varepsilon(\hat{w}) \to p(x, y) \in S \subset \mathbb{C}$$

defines a two sided-Rayleigh functional if, and only if, the following properties are fulfilled:

(i) $p(\alpha x, \beta y) = p(x, y), \quad \forall \alpha, \beta \in \mathbb{C} \setminus \{0\}$,

(ii) $y^H T\left(p(x, y)\right) x = 0$,

(iii) $y^H T'\left(p(x, y)\right) x \neq 0$.

Schreiber and Schwetlick proved the existence and stationarity of two-sided Rayleigh functionals at eigenvectors. For the two-sided Rayleigh functional,

$$|p(x, y) - \hat{\lambda}| \leqslant C \, \tan \measuredangle (x, \hat{v}) \tan \measuredangle (y, \hat{w}) \tag{2.15}$$

holds.

2.4 Error Measure

When analyzing the numerical method for its convergence rate, it is necessary to have a measure for the error.[1] The first idea is to use the norm of the error vector, which is used, for instance, to measure the error of nonlinear systems of equations. But for eigenvalue problems, all nonzero multiples of an eigenvector are eigenvectors of the problem as well. Therefore, the distance between the vectors is not a suitable measure for the error.

2.4.1 Angles vs Distances

Measuring the approximation error of an eigenpair by angle is usually much more effective than measuring by distance.

Definition 2.23. The angle between two nonzero vectors with respect to an inner product $\langle \cdot, \cdot \rangle$ is defined by

$$\measuredangle(u, v) := \arccos \left(\frac{\langle u, v \rangle}{\|u\| \, \|v\|} \right), \tag{2.16}$$

where $\|\cdot\|$ is the norm induced by $\langle \cdot, \cdot \rangle$.

[1] Measuring the error requires that the exact solution of the (nonlinear) eigenvalue problem is known. Therefore, the content of this section is only used for theoretical considerations (e.g. How fast the error tends to zero in the iteration process?).

We usually use the standard inner product. If the eigenvectors can be chosen B-orthogonal for a given Hermitian positive definite matrix B, the corresponding inner product, $\langle \cdot, \cdot \rangle_B$, is chosen for the angle as well. Additionally, the induced B-norm is used here.

The angle between $x^{(k)}$ and $\hat{v}$ can be used if a simple eigenvalue is sought. For the linear eigenvalue problem

$$Av = \lambda v,$$

where A is normal[2], the eigenvectors can be chosen to be orthogonal. Then the current eigenvector approximation $x^{(k)}$ with norm 1 can be decomposed into a component in the direction of the eigenvector and a direction w, which is orthogonal to the eigenvector.

$$x^{(k)} = \cos(\phi^{(k)})\hat{v} + \sin(\phi^{(k)})w, \tag{2.17}$$

where $\phi_k := \measuredangle(x^{(k)}, \hat{v})$, $\left\| x^{(k)} \right\| = \|\hat{v}\| = \|w\| = 1$ and $\hat{v} \perp w$. Here w is a linear combination of all other eigenvectors except $\hat{v}$.

For the generalized Hermitian linear eigenvalue problem

$$Av = \lambda Bv, \qquad \text{with} \quad A = A^H \quad \text{and} \quad B = B^H \text{ h.p.d.}^3$$

the current approximation $x^{(k)}$ is decomposed as in (2.17), but the angle is chosen with respect to the B-inner product. This choice will be used later for convergence proofs.

Unfortunately, eigenvectors cannot always be chosen to be orthogonal. This occurs for non-Hermitian linear problems and usually for nonlinear eigenvalue problems. Hence, there exists alternative decompositions to (2.17), which result in a generalized sine, cosine and tangent. Those decompositions can be found, for instance, in [70] and will be introduced next.

2.4.2 Generalized Angles

For linear Hermitian eigenvalue problems the eigenvectors can be chosen to be orthogonal. With this and the decomposition of the current iterate $x^{(k)}$ in (2.17) the convergence behavior can then be analyzed. But for non-Hermitian matrices or for nonlinear eigenvalue problems this property cannot be exploited any longer. Therefore, an alternative decomposition is required. Such a decomposition of the current eigenvector approximation corresponding to a simple eigenvalue $\hat{\lambda}$ is given by Szyld and Xue in [70]. While they showed it only for $C = T'(\hat{\lambda})$, we give a more general decomposition for any non-singular matrix $C \in \mathbb{C}^{n \times n}$ satisfying

$$w^H C \hat{v} \neq 0$$

where w denotes now the left eigenvector corresponding to $\hat{\lambda}$. A vector $x \in \mathbb{C}^n$ can be decomposed similarly to (2.17) into a component of the direction of an eigenvector $\hat{v}$ and another direction g, by

$$x = \gamma \left(c\hat{v} + sg \right), \tag{2.18}$$

[2] i.e. $AA^H = A^H A$

[3] Hermitian positive definite

with

$$\gamma := \left\| \begin{pmatrix} w^H \\ W_{n-1}^H \end{pmatrix} Cx \right\|_2, \tag{2.19a}$$

$$s := \frac{\left\| W_{n-1}^H Cx \right\|_2}{\gamma}, \tag{2.19b}$$

$$c := \frac{w^H Cx}{\gamma}, \tag{2.19c}$$

$$g = \frac{1}{s} \left(\frac{1}{\gamma} x - c\hat{v} \right), \tag{2.19d}$$

where w is scaled to satisfy

$$w^H C\hat{v} = 1,$$

and $W_{n-1} \in \mathbb{C}^{n\times(n-1)}$ is a matrix consisting of an orthonormal basis of the orthogonal complement of $\operatorname{span}\{C\hat{v}\}$.

The following lemma elucidates important properties regarding this decomposition.

Lemma 2.24. *Let x be decomposed as in* (2.18) *and $C \in \mathbb{C}^{n\times n}$ non-singular. Furthermore, γ, s, c and g are defined as in* (2.19)*. Then the following is satisfied:*

(i) $w^H Cg = 0$,

(ii) $\left\| W_{n-1}^H Cg \right\|_2 = 1$,

(iii) $c^2 + s^2 = 1$.

Proof. All three parts can be easily proven by inserting the definitions from (2.19). Note that $\gamma \geqslant 0$ and $s \geqslant 0$ are implied by the definition. □

The value γ in (2.19a) is a norm on $\mathbb{C}^n$:

$$\|x\|_\gamma := \left\| \begin{pmatrix} w^H \\ W_{n-1}^H \end{pmatrix} Cx \right\|_2 .$$

Similar to the decomposition (2.17), c can be interpreted as the generalized cosine of the angle between x and $\hat{v}$ and as s as its generalized sine. Alternatively, the generalized tangent is considered to measure the rate of convergence, which we denote by

$$t_g := \frac{s}{c}, \qquad \text{if } c \neq 0.$$

We will now summarize how the generalized angle is related to the angle induced by the Euclidean inner product.

Szyld and Xue additionally showed in [70] that

$$\sin \measuredangle (x, \hat{v}) \leqslant \frac{\|sg\|_2}{\|c\hat{v}\|_2} = t_g \frac{\|g\|_2}{\|\hat{v}\|_2}.$$

$\|g\|_2$ can be bounded by

$$\frac{1}{\sigma_1(W_{n-1}^H C)} \geqslant \|g\|_2 \geqslant \frac{1}{\sigma_{min}(W_{n-1}^H C)}, \tag{2.20}$$

where $\sigma_j(W_{n-1}^H C)$ denotes the j-th and $\sigma_{min}(W_{n-1}^H C)$ denotes the smallest nonzero singular value of $W_{n-1}^H C$. Inequality (2.20) shows that $\|g\|_2$ can be bounded independently of s and c.

Finally, if $|s| \ll |c|$, then the tangent of the angle between $x^{(k)}$ and $\hat{v}$ can be approximated by

$$\tan \measuredangle (x, \hat{v}) = \frac{\|sg_\perp\|_2}{\left\|c\hat{v} + sg_\|\right\|_2} = \frac{\|(s\sin(\chi))g\|_2}{\|c\hat{v} + (s\cos(\chi))g\|_2} \approx t_g \sin(\chi).$$

Here g is decomposed into $g = g_\| + g_\perp$ where $g_\|$ is the orthogonal projection of g onto span$\{\hat{v}\}$, $g_\perp$, that part of g which is orthogonal to $\hat{v}$, and χ denotes the angle between g and $\hat{v}$. If $\chi > 0$, then there exist constants C_1 and C_2, such that

$$C_1 t_g \leqslant \tan \measuredangle (x, \hat{v}) \leqslant C_2 t_g \tag{2.21}$$

holds. For $\chi = \frac{\pi}{2}$ the generalized tangent and the real tangent are the same.

Corollary 2.25. *For the special linear eigenvalue problem where*

$$T(\lambda) = \lambda I - A \qquad \text{with} \quad A = A^H$$

and $C = I$, the generalized sine, cosine, and tangent are identical to the sine, cosine and tangent of the angle according to (2.16).

Proof. Obvious. □

In the following section and especially in Chapter 4, the convergence behavior of the methods is analyzed. We have to distinguish between two different cases:

1. Hermitian linear problems like $Ax = \lambda Bx$, where $A^H = A$ and $B = B^H$ is positive definite. Here the eigenvectors can be chosen B-orthogonal, and, thus, the decomposition (2.17) with $\langle \cdot, \cdot \rangle_B$ can be used.
   Here, $C = B$ yield that the decompositions in (2.17) and (2.19) are identical.
2. All other problems (non-Hermitian linear problems and nonlinear problems). Here the generalized decomposition (2.18) is used.

The following corollary combines the generalized angles with the approximation properties of the Rayleigh functional.

Corollary 2.26. *With the help of* (2.21)*, the approximation property of* (2.14) *can be extended to the generalized tangent. Thus, if $\chi = \measuredangle(\hat{v}, g)$ is not too small in magnitude, the error between evaluated Rayleigh functional and the sought eigenvalue behaves as*

$$\begin{aligned} |p(u) - \hat{\lambda}| &= \mathcal{O}\left((t_g)^2\right), \quad \text{if } T(\cdot)^H = T(\cdot), \\ |p(u) - \hat{\lambda}| &= \mathcal{O}\left((t_g)\right), \quad \text{otherwise} \end{aligned}$$

where t_g denotes the generalized tangent between u and $\hat{v}$.

Proof. From (2.14) we have

$$|p(u) - \hat{\lambda}| = \begin{cases} \mathcal{O}\left(\tan^2\left(\measuredangle(u, \hat{v})\right)\right) & \text{if } T(\cdot) \text{ is Hermitian} \\ \mathcal{O}\left(\tan\left(\measuredangle(u, \hat{v})\right)\right) & \text{otherwise.} \end{cases}$$

Combining this result with the property

$$\tan\left(\measuredangle(u, \hat{v})\right) \leqslant C t_g$$

from (2.21) completes the proof. □

2.4.3 The Residual

In this subsection we consider some properties of the residual

$$r^{(k)} := T(\lambda^{(k)}) x^{(k)}. \tag{2.22}$$

The residual can be used to measure the quality of the current iterates for the eigenpair. But a residual with a small norm does not guarantee that this iterate is close to the solution of the eigenvalue problem.

Some results about the residual of a nonlinear eigenvalue problem are summarized in the following proposition.

Proposition 2.27. *Let $T : D \subset \mathbb{R} \to \mathbb{C}^{n \times n}$ be Lipschitz continuous with respect to λ in a neighborhood $\mathcal{U}(\hat{\lambda})$ of a simple eigenvalue $\hat{\lambda}$. Moreover, let $\lambda^{(k)} \in \mathcal{U}(\hat{\lambda})$, the left eigenvector, $\hat{w}$, not orthogonal to $C\hat{v}$ and $x^{(k)} \in \mathbb{C}^n$ be sufficiently close to the corresponding eigenvector $\hat{v}$, and let the residual be defined as in* (2.22)*. Then,*

(i) $\left\| r^{(k)} \right\|_2 = \mathcal{O}(s_k) + \mathcal{O}(|\lambda^{(k)} - \hat{\lambda}|)$, *where s_k denotes the generalized sine of $x^{(k)}$ compared with the eigenvector $\hat{v}$.*

(ii) If $\lambda^{(k)} = p(x^{(k)})$ is chosen, then $r^{(k)} \perp x^{(k)}$.

Proof. The orthogonality in (ii) is obvious, since the second condition in Definition 2.15, $\left(x^{(k)}\right)^H T\left(p(x^{(k)})\right) x^{(k)} = 0$, is satisfied.

We consider

$$\begin{aligned} r^{(k)} = T(\lambda^{(k)}) x^{(k)} &= \left(T(\lambda^{(k)}) - T(\hat{\lambda}) + T(\hat{\lambda})\right) x^{(k)} \\ &= \left(T(\lambda^{(k)}) - T(\hat{\lambda})\right) x^{(k)} + T(\hat{\lambda}) x^{(k)}. \end{aligned}$$

For the latter term we apply the decomposition from (2.18) to $x^{(k)}$.

$$\begin{aligned} T(\lambda^{(k)}) x^{(k)} &= \left(T(\lambda^{(k)}) - T(\hat{\lambda})\right) x^{(k)} + T(\hat{\lambda}) \gamma^{(k)} \left(c_k \hat{v} + s_k g^{(k)}\right) \\ &= \left(T(\lambda^{(k)}) - T(\hat{\lambda})\right) x^{(k)} + \gamma^{(k)} c_k \underbrace{T(\hat{\lambda}) \hat{v}}_{=0} + \gamma^{(k)} s_k T(\hat{\lambda}) g^{(k)}. \end{aligned}$$

By exploiting the Lipschitz continuity of T, we finally obtain

$$\begin{aligned}
\left\|r^{(k)}\right\|_2 &= \left\|\left(T(\lambda^{(k)}) - T(\hat{\lambda})\right) x^{(k)} + \gamma^{(k)} s_k T(\hat{\lambda}) g^{(k)}\right\|_2 \\
&\leqslant \left\|T(\lambda^{(k)}) - T(\hat{\lambda})\right\|_2 \left\|x^{(k)}\right\|_2 + \gamma^{(k)} s_k \left\|T(\hat{\lambda}) g^{(k)}\right\|_2 \\
&\leqslant L|\lambda^{(k)} - \hat{\lambda}| \left\|x^{(k)}\right\|_2 + \gamma^{(k)} s_k \left\|T(\hat{\lambda}) g^{(k)}\right\|_2 \qquad (2.23)
\end{aligned}$$

In (2.20) it was shown that the $\left\|g^{(k)}\right\|$ can be bounded independently of s_k. Therefore, the second part of (2.23) behaves as follows,

$$\gamma^{(k)} s_k \left\|T(\hat{\lambda}) g^{(k)}\right\|_2 = \mathcal{O}(s_k). \qquad (2.24)$$

Finally, inserting (2.24) into (2.23) yields

$$\left\|r^{(k)}\right\|_2 = \mathcal{O}\left(|\lambda^{(k)} - \hat{\lambda}|\right) + \mathcal{O}\left(s_k\right).$$

□

The following lemma considers the dependence of the residual on the sine of the angle between the current vector iterate and an eigenvector.

Lemma 2.28. *Let $T : D \subset \mathbb{R} \to \mathbb{C}^{n\times n}$ be Lipschitz continuous with respect to λ in a neighborhood $\mathcal{U}(\hat{\lambda})$ of an eigenvalue $\hat{\lambda}$. Moreover, let $\lambda^{(k)} \in \mathcal{U}(\hat{\lambda})$ and $x^{(k)} \in \mathbb{C}^n$ be sufficiently close to the corresponding eigenvector $\hat{v}$ and the residual defined as in (2.22). Then,*

(i) $\left\|r^{(k)}\right\|_2 = \mathcal{O}(\sin(\phi_k)) + \mathcal{O}(|\lambda^{(k)} - \hat{\lambda}|)$, where ϕ_k denotes the the angle $\measuredangle(x^{(k)}, \hat{v})$.

(ii) If $\lambda^{(k)} = p(x^{(k)})$ is chosen, then $r^{(k)} \perp x^{(k)}$.

Proof. This lemma is proven in the same way as Proposition 2.27. We only use the decomposition from (2.17) for $x^{(k)}$, which means here that $\hat{v} \perp g^{(k)}$ and $\gamma^{(k)} = 1$. □

Corollary 2.29. *Let all prerequisites of Proposition 2.27 be fulfilled.*

1. *If $|\lambda^{(k)} - \hat{\lambda}| = \mathcal{O}\left(s_k\right)$ is satisfied[4], then*

$$\left\|r^{(k)}\right\|_2 = \mathcal{O}(s_k).$$

2. *If $|\lambda^{(k)} - \hat{\lambda}| = \mathcal{O}\left(\sin(\phi_k)\right)$ is satisfied, then*

$$\left\|r^{(k)}\right\|_2 = \mathcal{O}(\sin(\phi_k)).$$

Proof. Insert $|\lambda^{(k)} - \hat{\lambda}| = \mathcal{O}\left(s_k\right)$ and $|\lambda^{(k)} - \hat{\lambda}| = \mathcal{O}\left(\sin(\phi_k)\right)$ into Proposition 2.27 or Lemma 2.28, respectively. This completes the proof. □

[4] This can be assumed, cf. (2.14), Corollary 2.26.

2.5 Numerical Methods for Nonlinear Eigenvalue Problems

Now we will give an overview of numerical methods for eigenvalue problems. The following aspects should be considered for the selection of a suitable method.

- Is the eigenvalue problem linear, generalized linear or nonlinear (polynomial)?
- Do the matrices have a special structure (e.g. Hermitian)?
- Are the matrices full or sparse?
- Are we interested in one, $k \ll n$ or all eigenpairs?

2.5.1 Classical Methods

We will now consider numerical methods for nonlinear eigenvalue problems. To begin with, methods based on Newton's method are presented, followed by special methods for nonlinear eigenvalue problems. Finally, an overview of methods for large and sparse nonlinear eigenvalue problems is given.

Newton-like methods

One possibility to solve equation (2.1) is to append a scaling equation and solve the nonlinear system

$$G(z) = \begin{pmatrix} T(\lambda)x \\ l^H x - 1 \end{pmatrix} \overset{!}{=} 0 \quad \text{with } z = \begin{pmatrix} x \\ \lambda \end{pmatrix} \tag{2.25}$$

using Newton's method. Here, $l \in \mathbb{C}^n$ denotes a scaling vector which has to fulfill $l^H \hat{v} \neq 0$[5]. This approach is similar to inverse iteration for linear eigenvalue problems. Hence, the resulting method is called *Inverse Iteration for Nonlinear Eigenvalue Problems*.

[5] Even though the eigenvector is usually unknown we can assume that this condition is fulfilled, since in finite precision on a computer it is nearly impossible to obtain a vector l which is orthogonal to the sought eigenvector.

Algorithm 2.1: Inverse Iteration for Nonlinear Eigenvalue Problems

```
 input : initial guess for the eigenvector and eigenvalue $y^{(0)} \in \mathbb{C}^n$, $\lambda^{(0)} \in \mathbb{C}$, a
         scaling vector $l \in \mathbb{C}^n$ with $l^H \hat{v} \neq 0$
 output: approximation $\left(\lambda^{(k)}, x^{(k)}\right)$ to a solution $\hat{\lambda}$ and $\hat{v}$ of (2.1)
1 $\kappa^{(0)} = l^H y^{(0)}$;
2 $x^{(0)} = \frac{y^{(0)}}{\kappa^{(0)}}$;
3 for $k=0$ to convergence do
4     Solve $T(\lambda^{(k)})y^{(k+1)} = T'(\lambda^{(k)})x^{(k)}$ for $y^{(k+1)}$;
5     $\kappa^{(k+1)} = l^H y^{(k+1)}$;
6     $\lambda^{(k+1)} = \lambda^{(k)} + \frac{1}{\kappa^{(k+1)}}$;
7     $x^{(k+1)} = \frac{y^{(k+1)}}{\kappa^{(k+1)}}$;
8 end for
```

Theorem 2.30.
Inverse Iteration for Nonlinear Eigenvalue Problems, cf. Algorithm 2.1, converges locally quadratically to simple eigenvalues if Assumption 2.14 is satisfied.

Proof. See [54]. □

If a minmax characterization is available[6], Algorithm 2.1 can be accelerated by using the Rayleigh functional to derive an approximation for the eigenvalue. Therefore, the lines 5, 6 and 7 are replaced by

$$\begin{aligned} x^{(k+1)} &= \frac{y^{(k+1)}}{\|y^{(k+1)}\|_2} \\ \lambda^{(k+1)} &= p(x^{(k+1)}). \end{aligned}$$

The resulting method is called *Rayleigh functional iteration* because of its similarity to the Rayleigh quotient iteration, which is known for Hermitian linear eigenvalue problems.

[6] This replacement can also be applied if a functional with similar properties is available. But the improvement of the method can only be guaranteed for a Rayleigh functional with minmax characterization.

Algorithm 2.2: Rayleigh functional Iteration

```
input  : initial guess for the eigenvector $y^{(0)} \in D(p)$
output: approximations to a solution $\hat{\lambda}$ and $\hat{v}$ of (2.1)
$x^{(0)} = \frac{y^{(0)}}{\|y^{(0)}\|_2}$;
$\lambda^{(0)} = p(x^{(0)})$;
for $k$=0 to convergence do
    Solve $T(\lambda^{(k)})y^{(k+1)} = T'(\lambda^{(k)})x^{(k)}$ for $y^{(k+1)}$;
    $x^{(k+1)} = \frac{y^{(k+1)}}{\|y^{(k+1)}\|_2}$;
    $\lambda^{(k+1)} = p(x^{(k+1)})$;
end for
```

Theorem 2.31 ([59])**.**
Let Assumptions 2.7 and 2.8 be satisfied and $\hat{\lambda}$ a simple eigenvalue. Then the Rayleigh Functional Iteration from Algorithm 2.2 converges locally cubically.

Because solving a linear system in every iteration step with a varying matrix is expensive, Neumaier [49] introduced a method based on the simplified Newton method, avoiding a varying matrix. This method, called *Residual Inverse Iteration*, is shown in the following algorithm.

Algorithm 2.3: Residual Inverse Iteration

```
input  : initial guess for the eigenpair $(\lambda^{(0)}, x^{(0)})$ and a normalization vector
         $l$ with $l^H\hat{v} \neq 0$ and $l^H x^{(0)} = 1$
output: approximation to a solution $\hat{\lambda}$ and $\hat{v}$ of (2.1)
1 for $k = 0$ to convergence do
2     solve $l^H T(\lambda^{(0)})^{-1} T(\lambda^{(k+1)}) x^{(k)} = 0$ for $\lambda^{(k+1)}$;
3     solve $T(\lambda^{(0)}) u^{(k)} = T(\lambda^{(k+1)}) x^{(k)}$ for $u^{(k)}$;
4     $v^{(k+1)} = x^{(k)} - u^{(k)}$;
5     $x^{(k+1)} = \frac{1}{l^H v^{(k+1)}} v^{(k+1)}$;
6 end for
```

If $T(\cdot)$ is Hermitian, line 2 in Algorithm 2.3 can be replaced by evaluating the Rayleigh functional p at $x^{(k)}$, thus $\lambda^{(k+1)} = p(x^{(k)})$. Although the quadratic convergence had to be abandoned, the method is effective since a decomposition, like Cholesky or LU, of $T(\lambda^{(0)})$ can be made at once with the beginning of the iteration process.

The convergence behavior of Residual Inverse Iteration is, for instance described, in [49] and [81]. If $T(\cdot)$ is twice continuously differentiable and $\hat{\lambda}$ is simple, the Residual Inverse Iteration converges for every $(\lambda^{(0)}, x^{(0)})$ sufficiently close to the eigenpair with

$$\frac{\left\|x^{(k+1)} - \hat{v}\right\|_2}{\left\|x^{(k)} - \hat{v}\right\|_2} = \mathcal{O}(|\lambda^{(0)} - \hat{\lambda}|) \tag{2.26}$$

and

$$|\lambda^{(k)} - \hat{\lambda}| = \begin{cases} \mathcal{O}\left(\left\|x^{(k)} - \hat{v}\right\|_2^2\right) & \text{if } T(\cdot) \text{ is Hermitian,} \\ \mathcal{O}\left(\left\|x^{(k)} - \hat{v}\right\|_2\right) & \text{else.} \end{cases} \tag{2.27}$$

Safeguarded Iteration

Let Assumptions 2.7 and 2.8 be satisfied and consider the parameter dependent linear eigenvalue problem

$$T(\lambda)u = \mu(\lambda)u.$$

Then for $\lambda = \lambda_j$, there exists at least one eigenvalue $\mu_j(\lambda_j) = 0$. This property is used by Voss and Werner to determine a minmax characterization for non-overdamped nonlinear eigenvalue problem in [82]. Furthermore, they present a method, where this property is exploited, since $T(\lambda^{(k)})$, where $\lambda^{(k)}$ is close to an eigenvalue λ_j, has an eigenvalue $\mu(\lambda^{(k)})$ close to zero. This behavior is exploited in the following algorithm:

Algorithm 2.4: Safeguarded Iteration

```
input  : initial guess for an eigenvector $x^{(0)} \in D(p)$
output: the $m$-th eigenvalue and a corresponding eigenvector of (2.1)
for $k = 0$ to convergence do
    $\lambda^{(k)} = p(x^{(k)})$;
    determine the $m$-th largest eigenvalue and an corresponding eigenvector
      $x^{(k)}$ of $T(\lambda^{(k)})x = \mu x$;
end for
```

The properties and convergence behavior are summarized in the following lemma:

Lemma 2.32 ([50, 81])**.** *If Assumptions 2.7 and 2.8 are satisfied, the Safeguarded Iteration from Algorithm 2.4 has the following properties:*

(i) If $\lambda_1 := \inf_{x \in D(p)} p(x)$ and $x^{(0)} \in D(p)$, then the safeguarded iteration with $m = 1$ converges globally to λ_1.

(ii) If $T(\cdot)$ is continuously differentiable and λ_m is a simple eigenvalue, then the safeguarded iteration converges locally quadratically to λ_m.

(iii) Let $T(\cdot)$ be twice continuously differentiable, λ_m be simple and $T'(\lambda_m)$ be positive definite. If $x^{(k)}$ in step 2 is determined as the m-th largest eigenvalue of the generalized linear eigenvalue problem $T(\lambda^{(k)})x = \mu T'(\lambda^{(k)})x$, then the Safeguarded Iteration converges cubically.

The name of this method originates from the fact that we can determine, with certainty, the k-th eigenvalue of the nonlinear problem. This guarantee is not given in any other method.

The Safeguarded Iteration is used if one is seeking the m-th smallest eigenvalue of a given problem. But for large n, the method is expensive because every iteration step requires the solution of a linear eigenvalue problem. Solving a linear eigenvalue problem for large n is a process where several iteration steps are necessary.

Rothe presented a simplification in [59] where the solution of the linear eigenvalue problem in step 3 of Algorithm 2.4 is replaced by one step of Inverse Iteration[7] to the linear eigenvalue problem $T(\lambda^{(k)})x = \mu x$.

$$y^{(k+1)} := \left(T(\lambda^{(k)})\right)^{-1} x^{(k)}, \qquad x^{(k+1)} = \frac{y^{(k+1)}}{\left\|y^{(k+1)}\right\|_2}.$$

This leads to the following algorithm:

Algorithm 2.5: Derivative-free Rayleigh functional Iteration

```
input  : initial guess for an eigenvector $x^{(0)}$
output: an eigenvalue and a corresponding eigenvector of (2.1)
for $k = 0$ to convergence do
    $\lambda^{(k)} = p(x^{(k)})$;
    solve $T(\lambda^{(k)})y^{(k+1)} = x^{(k)}$ for $y^{(k+1)}$;
    $x^{(k+1)} = \frac{y^{(k+1)}}{\|y^{(k+1)}\|_2}$;
end for
```

Theorem 2.33 ([59]).
Algorithm 2.5 converges locally quadratically if $T(\cdot)$ is real symmetric and twice differentiable with respect to λ, and, additionally, $\hat{\lambda}$ is a simple eigenvalue.

Theorem 2.33 shows that the convergence order of Algorithm 2.4 can be maintained for local convergence of Algorithm 2.5. But we had to abandon the global convergence properties which are provided by the Safeguarded Iteration method.

If the iteration step of the Inverse Iteration is applied to the generalized eigenvalue problem $T(\lambda^{(k)})x = \mu T'(\lambda^{(k)})x$, the algorithm reduces to Rayleigh functional iteration as in Algorithm 2.2, which converges cubically as well.

Iterative Projection Methods

For sparse nonlinear eigenvalue problems of large size there exist iterative projection methods, which are partially adapted from those for the linear eigenvalue problem.

The following algorithm describes generally an iterative projection method for nonlinear eigenvalue problems:

[7] Here, the algorithm Inverse Iteration for linear eigenvalues is considered.

Algorithm 2.6: Iterative Projection Method

```
input  : a matrix $V$ containing a basis of an initial subspace in which the
         eigenvectors are assumed, with $V^H V = I$
output: eigenvalues and corresponding eigenvectors of (2.1)

1  $m = 0$;
2  while $m <$ number of wanted eigenvalues do
3      compute the wanted eigenvalue $\theta$ and corresponding eigenvector $y$ of the
         projected problem $V^H T(\theta) V y \overset{!}{=} 0$;
4      determine Ritz vector $u = Vy$ and residual $r = T(\theta)u$;
5      if $\|r\| \,/\, \|u\| < \epsilon$ then
6          accept approximate eigenpair $\lambda_m = \mu$, $x^m = u$;
7          $m = m + 1$ ;
8          if $m ==$ number of wanted eigenvalues then
9              STOP;
10         end if
11         if number of columns of $V$ is too large then
12             restart;
13         end if
14         choose approximations $\mu$ and $u$ to the next eigenpair;
15         determine residual $r = T(\theta)u$;
16     end if
17     determine a subspace expansion $v$;
18     $v = v - VV^H v$;
19     $\tilde{v} = v / \|v\|$;
20     $V = [V, \tilde{v}]$;
21     reorthogonalize $V$ if necessary;
22 end while
```

In line 3, the *wanted* eigenvalue has to be determined. If the dimension of the subspace is greater than one, there might be more than one eigenvalue available. Possible choices for the wanted eigenvalue are for instance

- the smallest/largest algebraic eigenvalue (if the eigenvalue is real),
- the smallest eigenvalue in magnitude,
- the closest eigenvalue to a given shift.

In line 12, a restart is executed. For this, the algorithm is started from the beginning again with a new initial basis of V. The new initial basis can be obtained, for instance, by dropping all Ritz vectors whose Ritz values are not close to the wanted eigenvalue. Alternatively, an invariant subspace of the linear eigenvalue problem $T(\sigma)$ can be used as the new subspace, where σ denotes a shift.

There exist different possibilities to determine a subspace expansion for line 13. We choose a Ritz pair and apply one step of a vector iteration (cf. Algorithms 2.1, 2.2, 2.3, 2.5).

Voss suggests in [74] to expand the subspace by the correction obtained by Residual Inverse Iteration (Algorithm 2.3):

$$u^{(k)} = T(\sigma)^{-1}T(\lambda^{(k+1)})x^{(k)},$$

where σ denotes a shift for the eigenvalue. The method is called *Nonlinear Arnoldi method*. Since solving the linear system

$$T(\sigma)u^{(k)} = T(\lambda^{(k+1)})x^{(k)}$$

for $u^{(k)}$ in every iteration step is too expensive for large matrices, a preconditioner $M \approx T(\sigma)^{-1}$ is used instead of $T(\sigma)^{-1}$. Thus, the search space expansion is computed by

$$v = MT(\lambda^{(k+1)})x^{(k)}. \tag{2.28}$$

Further details for the Nonlinear Arnoldi method can be found in [46, 74].

Alternatively, the subspace can be expanded by the next iterate of a Newton-like method. Unfortunately, these linear systems, like

$$T(\lambda^{(k)})v = Cx^{(k)}, \qquad \text{for } C \in \{T'(\lambda^{(k)}), I\},$$

might result in a subspace expansion v which is very sensitive. Small perturbations in the subspace expansion might yield to greater changes in the subspace. This is discussed in detail in Theorem 3.1.

Therefore, a simplification, as in (2.28), cannot be used. In Chapter 3, the Jacobi-Davidson method is presented. This method is based on Newton-like methods and provides the computation of a search space expansion in a more robust way.

2.5.2 Inexact Methods

We have seen that many methods (e.g. Inverse Iteration or Rayleigh functional iteration) require the solution of the linear system

$$T(\lambda^{(k)})y^{(k+1)} = Cx^{(k)}, \qquad \text{for } C \in \{T'(\lambda^{(k)}), I, ...\}, \tag{2.29}$$

in every iteration step. For methods like the Inverse Iteration or the Rayleigh functional iteration, the matrix $T(\lambda^{(k)})$ varies in every k. So information, e.g. a matrix decomposition, cannot be recycled.

For large matrices a decomposition is too expensive, especially if the matrix is not available explicitly. Therefore, those methods have been modified such that in every iteration step only an approximative solution of the linear system is used for the next iterate.

In this context, we will call the iterations of the method to compute the eigenpairs *outer iterations* and, consequently, the iterations to solve the linear system approximately *inner iterations*.

Firstly, we present inexact methods to solve linear eigenvalue problems. Then, inexact methods for nonlinear eigenvalue problems are introduced.

Linear Eigenvalue Problems

```
Algorithm 2.7: Inexact Inverse Iteration
input : initial guess for the eigenvector and eigenvalue $y^{(0)}$, $\lambda^{(0)}$, a scaling
        vector $l \in \mathbb{C}^n$ with $l^H \hat{v} \neq 0$
output: approximations for an eigenvalue $\hat{\lambda}$ closest to $\lambda^{(0)}$ and the
        corresponding eigenvector $\hat{v}$
$\kappa^{(0)} = l^H y^{(0)}$;
$x^{(0)} = \frac{y^{(0)}}{\kappa^{(0)}}$;
for $k$=0 to convergence do
    Solve $(A - \lambda^{(k)} B) y^{(k+1)} = x^{(k)}$ for $y^{(k+1)}$ approximately such that
    $\left\| (A - \lambda^{(k)} B) y^{(k+1)} - x^{(k)} \right\| \leqslant \tau_k$;
    $\kappa^{(k+1)} = l^H y^{(k+1)}$;
    $\lambda^{(k+1)} = \lambda^{(k)} + \frac{1}{\kappa^{(k+1)}}$;
    $x^{(k+1)} = \frac{y^{(k+1)}}{\kappa^{(k+1)}}$;
end for
```

The approximation for the eigenvalue can be improved by using the Rayleigh quotient, which leads to the well known Rayleigh quotient iteration. For Hermitian eigenvalue problems, it converges locally cubically. The following algorithm describes its inexact variant:

```
Algorithm 2.8: Inexact Rayleigh quotient Iteration
input : initial guess for the eigenvector $y^{(0)}$, matrix $A$, initial residual
        tolerance $\tau_0$
output: approximations for $\hat{\lambda}$ and $\hat{v}$
for $k$=0 to convergence do
    $x^{(k)} = \frac{y^{(k)}}{\left\| y^{(k)} \right\|_B}$;
    $\theta^{(k)} = \left( x^{(k)} \right)^H A x^{(k)}$;
    if $k > 0$ then
        Determine $\tau_k$;
    end if
    Solve $(A - \theta^{(k)} B) y^{(k+1)} = x^{(k)}$ for $y^{(k+1)}$ approximately such that
    $\left\| (A - \theta^{(k)} B) y^{(k+1)} - x^{(k)} \right\| \leqslant \tau_k$;
end for
```

The choice of the error bound τ_k is an important part of the discussion about the method described above, which can be found for different cases in [6, 7, 15, 16, 18, 21, 39, 52, 68].

The convergence of the Inexact Inverse Iteration with a fixed shift is discussed by Smit and Paardekooper in [68], Freitag and Spence in [18] and by Lai, Lin and

Lin in [39] for the standard linear eigenvalue problem[8]. The convergence from the "exact" case can only be maintained if τ_k is chosen such that

$$\tau_0 > \tau_1 > \ldots > \tau_k > \tau_{k+1} > .... > 0$$

and

$$\tau_k \xrightarrow{k \to \infty} 0.$$

Lai, Lin and Lin suggest choosing $\tau_k \sim \frac{1}{k}$ and Smit and Paardekooper present a formula in [68], proving that if

$$\tau_k \leqslant C \cos(\phi_k) \sin(\phi_k) \quad \text{with} \quad \phi_k = \measuredangle(x^{(k)}, \hat{v})$$

is satisfied, then the convergence is maintained.

The Inexact Inverse Iteration (cf. Algorithm 2.7) is discussed in [5, 6, 7, 15, 16] for the generalized eigenvalue problem. The results are outlined in the following theorem.

Theorem 2.34.
Let $(\lambda^{(0)}, x^{(0)})$ be a sufficiently good initial approximation of the eigenpair $(\hat{\lambda}, \hat{v})$, then the Inexact Inverse Iteration (Algorithm 2.7) converges

- *quadratically if $\tau_k \leqslant C \left\| r^{(k)} \right\|$.*
- *linearly if τ_k is constant.*

Inexact Rayleigh quotient iteration is analyzed in [5, 68, 52]. As it has already been shown for the Inverse Iteration, the convergence[9] can be maintained, if the tolerance τ_k for the solution of the linear system is chosen decreasing proportionally to $\left\| r^{(k)} \right\|$. If $\tau^{(k)} = \tau_{\max}$ is fixed, then the order of convergence is reduced by one. Therefore, the Inexact Rayleigh quotient iteration still converges quadratically.

In [15, 18], Freitag and Spence present special preconditioning strategies to accelerate the algorithm to solve (2.29).

They suggest to solve

$$\left(A - \lambda^{(k)} M\right) y^{(k+1)} = \frac{1}{\lambda^{(k)}} \mathbb{P}_k x^{(k)} \tag{2.30}$$

instead of (2.29). Here, $\mathbb{P}_k$ denotes a modified preconditioner varying in every iteration step. This preconditioner has to satisfy

$$\mathbb{P}_k x^{(k)} = A x^{(k)}. \tag{2.31}$$

[8] $Ax = \lambda x$

[9] Here the convergence is cubic.

If the classical preconditioner[10] is denoted by P, a tuned one can easily be determined with a rank one update.

$$\mathbb{P}_k := P + f^{(k)} \left(c_k\right)^H, \tag{2.32}$$

where $f^{(k)} := Ax^{(k)} - Px^{(k)}$ and c_k is chosen such that $\left(c_k\right)^H x^{(k)} = 1$.

The inverse of this preconditioner can easily be applied by using the Sherman-Morrison-Woodbury formula, if the inverse of P can be applied easily.

In [18] a different approach is presented. Here the preconditioner has to fulfill

$$\tilde{\mathbb{P}}_k x^{(k)} = x^{(k)}. \tag{2.33}$$

A comparison of this preconditioner and another one for the Jacobi-Davidson correction equation are provided by Freitag and Spence in [17]. They show that the generated Krylov subspaces are very similar. Therefore, the modified preconditioner from (2.33) can be considered as executing a preconditioned Jacobi-Davidson method. A positive definite preconditioner for Hermitian eigenvalue problems is introduced in [5] by Berns-Müller, Graham and Spence.

As further improvements Freitag, Kürschner and Pestana present choices of polynomial preconditioners,

$$\mathbb{P}_p^{-1} := q\left(\left(A - \sigma I\right)\right), \quad q \in \Pi_d,$$

in [14]. Here Π_d denotes the set of all polynomials of maximum degree d. Combined with stopping criteria for the GMRES method they can achieve small numbers of inner iterations. Such a polynomial q is chosen such that

$$\left(A - \sigma I\right) q\left(\left(A - \sigma I\right)\right) \approx I.$$

Freitag, Kürschner and Pestana use a common construction process based on reciprocal Chebychev nodes. That process requires information about the location[11] of the spectrum of A. Depending on the matrix structure this might lead to further computational costs.

Nonlinear Eigenvalue Problems

For nonlinear eigenvalue problems, the inexact methods can be adapted from the methods presented in Subsection 2.5.1. This leads to the following algorithm for the Inverse Iteration (Algorithm 2.1):

[10] e.g. incomplete LU factorization

[11] i.e. lower and upper bound of the real and imaginary part (if complex)

Algorithm 2.9: Inexact Inverse Iteration for Nonlinear Eigenvalue Problems

```
input  : initial guess for the eigenvector and eigenvalue $y^{(0)}$, $\lambda^{(0)}$, a scaling
         vector $l \in \mathbb{C}^n$ with $l^H \hat{v} \neq 0$
output: approximation to a solution $\hat{\lambda}$ and $\hat{v}$ of (2.1)
$\kappa^{(0)} = l^H y^{(0)}$;
$x^{(0)} = \frac{y^{(0)}}{\kappa^{(0)}}$;
for $k=0$ to convergence do
    Solve $T(\lambda^{(k)})y^{(k+1)} = T'(\lambda^{(k)})x^{(k)}$ for $y^{(k+1)}$ approximately such that
      $\left\| T(\lambda^{(k)})y^{(k+1)} - T'(\lambda^{(k)})x^{(k)} \right\| \leqslant \tau_k \left\| T'(\lambda^{(k)})x^{(k)} \right\|$;
    $\kappa^{(k+1)} = l^H y^{(k+1)}$;
    $\lambda^{(k+1)} = \lambda^{(k)} - \frac{1}{\kappa^{(k+1)}}$;
    $x^{(k+1)} = \frac{y^{(k+1)}}{\kappa^{(k+1)}}$;
end for
```

The same is done for the Rayleigh functional iteration (Algorithm 2.2).

Algorithm 2.10: Inexact Rayleigh functional Iteration

```
input  : initial guess for the eigenvector $y^{(0)}$
output: approximation to a solution $\hat{\lambda}$ and $\hat{v}$ of (2.1)
$x^{(0)} = \frac{y^{(0)}}{\left\| y^{(0)} \right\|}$;
$\lambda^{(0)} = p(x^{(0)})$;
for $k=0$ to convergence do
    Solve $T(\lambda^{(k)})y^{(k+1)} = T'(\lambda^{(k)})x^{(k)}$ for $y^{(k+1)}$ approximately such that
      $\left\| T(\lambda^{(k)})y^{(k+1)} - T'(\lambda^{(k)})x^{(k)} \right\| \leqslant \tau_k \left\| T'(\lambda^{(k)})x^{(k)} \right\|$;
    $x^{(k+1)} = \frac{y^{(k+1)}}{\left\| y^{(k+1)} \right\|}$;
    $\lambda^{(k+1)} = p(x^{(k+1)})$;
end for
```

A good overview of inexact Newton-type methods and their convergence is given in [70]. Furthermore, the nonlinear Arnoldi method can also be seen as an inexact method. In this case, the linear system is not even solved by an iterative solver, but the system matrix is replaced by a preconditioner.

Szyld and Xue provide the following two theorems for the convergence of the Inexact Inverse Iteration (Algorithm 2.9) and the Inexact Rayleigh functional iteration (Algorithm 2.10):

Theorem 2.35 ([70]).
Let $(\hat{\lambda}, \hat{v})$ be a simple eigenpair of the nonlinear eigenvalue problem (2.1)*. Let $\eta > 0$ such that $\|J_G(z_*)^{-1}\| \leqslant \eta$, where J_G denotes the Jacobian of the function G from* (2.25)*. Then, for some sufficiently small r and τ_{max}, $z^{(0)} \in B(z_*, r)$ and $\tau^{(i)} \leqslant \tau_{max}$, Algorithm 2.9 converges to $z_* = \begin{pmatrix} \hat{v} \\ \hat{\lambda} \end{pmatrix}$ at least linearly. If $\tau^{(i)} \leqslant C \left\| e^{(i)} \right\|_2 \leqslant \tau_{max}$,*[12] *where C is a constant independent of i, then Algorithm 2.9 converges at least quadratically.*

Theorem 2.36 ([70]).
Let $(\hat{\lambda}, \hat{v})$ be a simple eigenpair of (2.1)*, and w^H the corresponding left eigenvector. Assume that there exists a small $r > 0$ and $\zeta > 0$ such that $\|T'(\mu)x\| \leqslant \zeta$ for all $(\mu, x) \in \overline{B(z_*, r)}$ where $z_* = \begin{pmatrix} \hat{v} \\ \hat{\lambda} \end{pmatrix}$. Let $x^{(0)} = \gamma^{(0)} \left(c^{(0)} \hat{v} + s^{(0)} g^{(0)} \right)$ (see* (2.18)*) be a vector such that $\left(p(x^{(0)}), x^{(0)} \right) \in \overline{B(z_*, r)}$. For a given $d \in (0, |c^{(0)}|)$, let $\tau_{max} < \frac{|c^{(0)}| - d}{\zeta \|w\|}$ be an upper bound for the tolerance for the inner solve of Algorithm 2.10.*

Then, if $x^{(0)}$ is close to $\hat{v}$ in direction, and if $\tau^{(i)} = \tau < \tau_{max}$ is an appropriately small fixed tolerance , Algorithm 2.10 with $\mu^{(i)} = p(x^{(i)})$ converges at least linearly to $(\hat{\lambda}, \hat{v})$, and it converges at least quadratically if the local symmetry[13] *of $T(\lambda)$ is present and the Rayleigh functional defined in Definition 2.15 is applied. In addition, if $\tau^{(i)} \leqslant Ct^{(i)} \leqslant \tau_{max}$* [14] *for some C independent of i, this algorithm converges at least quadratically and at least cubically, respectively, if the local symmetry of $T(\lambda)$ is absent, or if it is present and the Rayleigh functional defined in Definition 2.15 is applied.*

These two theorems show that the same results as in the linear case can be adapted to nonlinear eigenvalue problems. Hence, the order of convergence can be maintained if $\tau^{(k)}$ is decreased from one step to the next. Otherwise the convergence order is reduced by one if $\tau^{(k)} = \tau^{(0)}$.

2.6 Numerical Example

The following example shows the performance of Inexact Rayleigh functional iteration from Algorithm 2.10:

Example 2.37. We consider the 3D Quantum-Dot problem. This problem was discussed by Voss in [76]. This nonlinear eigenvalue problem originates by the Schrödinger equation

$$-\nabla \left(\frac{\hbar^2}{2m(\lambda, x)} \nabla \psi \right) + V(x)\psi = \lambda \psi, \quad x \in \Omega_q \cup \Omega_m, \tag{2.34}$$

where the wave functions ψ and the energy levels λ are sought. Here $\Omega_q \subset \mathbb{R}^3$ denotes the domain occupied by the quantum dot and $\Omega_m \subset \mathbb{R}^3$ a bounded matrix

[12] Here $s^{(i)}$ denotes the generalized sine of $\measuredangle(x^{(i)}, \hat{v})$, cf. Subsection 2.4.2

[13] i.e. $T(\lambda) = T^T(\lambda)$, for all λ in a neighborhood of $\hat{\lambda}$

[14] $t^{(i)}$ denotes the generalized tangent of $\measuredangle(x^{(i)}, \hat{v})$, cf. Subsection 2.4.2

quantum dot		matrix	
Parameter	Value	Parameter	Value
P_q	0.8503	P_m	0.8878
$E_{g,q}$	0.42	$E_{g,m}$	1.52
Δ_q	0.48	Δ_m	0.34
V_q	0	V_m	0.7

Table 2.1: Parameters for the quantum dot problem example

of different material. Furthermore, $\hbar$ denotes the reduced Planck constant. The effective mass of the electron is constant on Ω_q and Ω_m for a fixed energy level λ. Therefore, $m_j(\lambda)$ was defined by

$$\frac{1}{m_j(\lambda)} := \left.\frac{1}{m(\lambda,x)}\right|_{x\in\Omega_j} = \frac{P_j^2}{\hbar^2}\left(\frac{2}{\lambda+E_{g,j}-V_j}+\frac{1}{\lambda+E_{g,j}-V_j+\Delta_j}\right), \quad j\in\{q,m\}.$$

The confinement potential $V_j := V|_{x\in\Omega_j}$ is piecewise constant and P_j denotes the momentum element, $E_{g,j}$ the band gap and Δ_j the spin-orbit splitting in the valence band for the quantum dot material and the matrix, respectively. The following boundary conditions are chosen:

- homogenous Dirichlet boundary conditions $\psi = 0$ on the horizontal part of the outer boundary of Ω_m,
- Neumann boundary conditions $\frac{\partial\psi}{\partial n} = 0$ on the vertical part of the outer boundary of Ω_m,
- Ben Daniel-Duke condition on the interface between quantum dot material and the matrix:

$$\left.\frac{1}{m_q}\frac{\partial\psi}{\partial n_q}\right|_{\partial\Omega_q} = \left.\frac{1}{m_m}\frac{\partial\psi}{\partial n_m}\right|_{\partial\Omega_m}, \quad x\in\partial\Omega_q\cap\partial\Omega_m.$$

For this example a pyramidal quantum dot width 12.4 nm and height 6.2 nm embedded in a cuboid matrix of size 24.8 nm × 24.8 nm × 18.6 nm is considered. The parameters are chosen as given in Table 2.1 (cf [30]).

Discretizing (2.34) with FEMLAB by the finite element method with quadratic Lagrangian elements on a tetrahedral grid yields the following nonlinear eigenvalue problem (cf [43]).

$$T(\lambda)x := \lambda M x - \frac{1}{m_q(\lambda)} A_q x - \frac{1}{m_m(\lambda)} A_m x - Bx = 0, \tag{2.35}$$

where

$$\begin{aligned} A_j &= \left(\int_{\Omega_j} \nabla\phi_k \cdot \nabla\phi_l \, dx \right)_{k,l}, \quad j \in \{q, m\} \\ M &= \left(\int_{\Omega} \phi_k \phi_l \, dx \right)_{k,l}, \\ B &= \left(V_q \int_{\Omega_q} \phi_k \phi_l \, dx + V_m \int_{\Omega_m} \phi_k \phi_l \, dx \right)_{k,l}, \end{aligned}$$

where $\Omega := \bar{\Omega}_q \cup \Omega_m$ and $\phi_i \in H := \{\psi \in H^1(\Omega) \mid \psi = 0 \text{ on } \partial\Omega_h\}$ are basis functions of the ansatz space.

The sparsity pattern of $A_q + A_m$ is shown in Figure 2.1 and the size of these matrices is $n = 96640$.

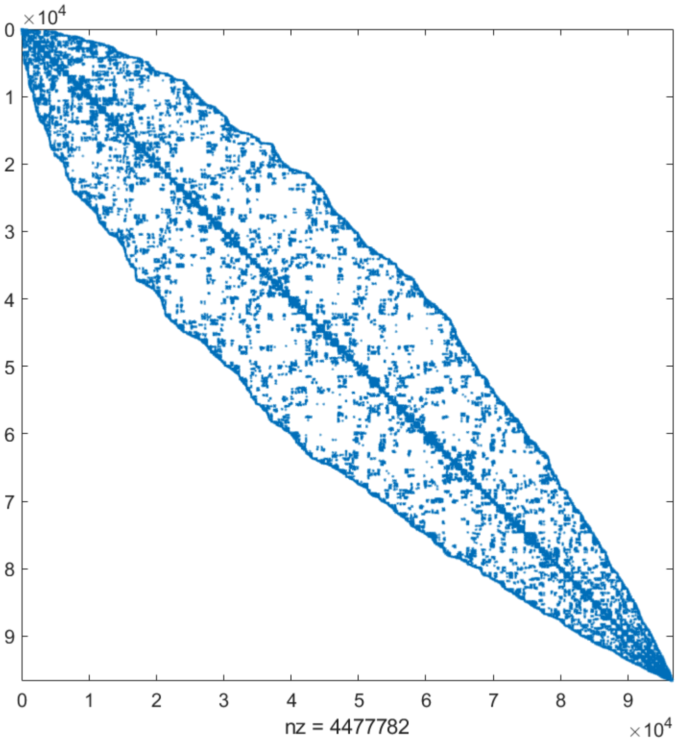

Figure 2.1: The sparsity pattern of the matrix $A_q + A_m$

We applied Inexact Rayleigh functional iteration (Algorithm 2.10) for computing the smallest eigenpair. Since the smallest eigenvalue, which has been computed before, has only a very small convergence radius, the starting vector $x^{(0)}$ was chosen with a very small angle to the eigenspace belonging to the smallest eigenvalue. The linear system in each iteration step was solved with the MINRes method. As preconditioner an incomplete Cholesky factorization $T(0) \approx RR^H$ was used for every

k	$\sin(\phi_k)$	$\lvert\lambda^{(k)}-\lambda_1\rvert$	$\left\Vert r^{(k)}\right\Vert$	$\left\Vert res^{(k)}\right\Vert$	**inner iterations**
0	0.012	0.0467	0.0488	0.094	61
1	0.017	1.183e-4	9.221e-5	0.092	168
2	2.366e-5	3.190e-10	2.203-8	0.098	358
3	3.877e-11	4.419e-14	6.041e-13	0.61*	603*
4	2.810-10	4.430e-14	2.534e-11	n.a.	n.a.

Table 2.2: Results of Inexact Rayleigh functional iteration with fixed τ

k	$\sin(\phi_k)$	$\lvert\lambda^{(k)}-\lambda_1\rvert$	$\left\Vert r^{(k)}\right\Vert$	$\left\Vert res^{(k)}\right\Vert$	**inner iterations**
0	0.012	0.0467	0.0488	0.094	61
1	0.017	1.184e-4	9.221e-5	2e-4	279
2	8.130e-6	3.285e-11	5.039e-8	0.095*	868*
3	1.099e-13	4.585e-14	5.908e-14	n.a.	n.a.

Table 2.3: Results of Inexact Rayleigh functional iteration with decreasing τ

linear system. Firstly, we chose a fixed tolerance for solving the linear system, such that

$$\left\Vert T\left(\lambda^{(k)}\right)y - T'\left(\lambda^{(k)}\right)x^{(k)}\right\Vert_2 \leqslant 0.1\left\Vert T'\left(\lambda^{(k)}\right)x^{(k)}\right\Vert_2 .$$

The results are shown in Table 2.2. The entries marked with stars denote results where the MinRes method terminated without convergence. We computed the convergence rate based on this data and obtained

$$\kappa_{fixed} \approx 2.5432.$$

This shows that the Algorithm 2.10 converges faster than quadratically. Theorem 2.36, which forecasts at least quadratic convergence for a fixed τ if local symmetry is provided, can be verified by it.

The decreasing tolerance for solving the corresponding linear system in each iteration step is chosen by $\tau_1 = 0.1$ and $\tau_{k+1} = \tau_k \frac{\left\Vert r^{(k+1)}\right\Vert_2}{\left\Vert r^{(k)}\right\Vert_2}$. Hence, the linear system was solved such that

$$\left\Vert T\left(\lambda^{(k)}\right)y - T'\left(\lambda^{(k)}\right)x^{(k)}\right\Vert_2 \leqslant \tau_k\left\Vert T'\left(\lambda^{(k)}\right)x^{(k)}\right\Vert_2$$

was satisfied.

The results are shown in Table 2.3[15]. Based on these results the convergence rate was determined to be

$$\kappa_{decr} \approx 2.8932.$$

[15] The entries marked with stars denote again results where the MinRes method terminated without convergence.

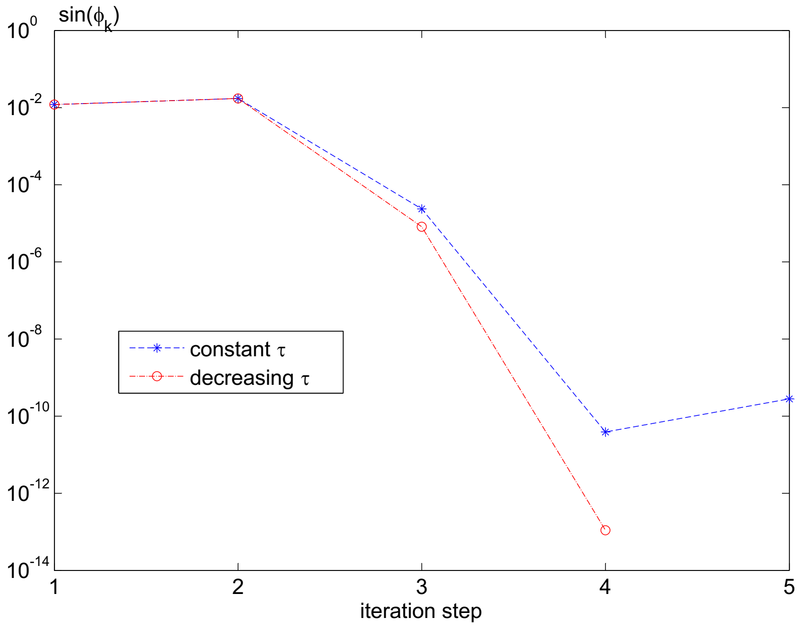

Figure 2.2: Convergence behavior of the two different choices for τ

Theorem 2.36 forecasts cubic convergence, which is nearly reached according to these results. Figure 2.2 illustrates the convergence behavior of Rayleigh functional iteration for the two different properties for the tolerance τ.

Figure 2.3 illustrates the results from Table 2.2 and Table 2.3.

Example 2.37 shows that the forecasted convergence rates from Theorem 2.36 are achievable. But it also shows that these good convergence properties require significantly high numbers of inner iterations for solving the linear system in each iteration step. This costs additional computation time. Furthermore, if the angle between the eigenspace and the current $x^{(k)}$ becomes to small it is no longer possible to solve the linear system to the desired accuracy. Therefore, no further (outer) iteration step can be executed. Finally, the iteration breaks down at this point.

The following two chapters will present an alternative method to achieve a smaller number of inner iterations. In addition, it will still be possible to execute further iteration steps, even though the angle between $x^{(k)}$ and the corresponding eigenspace is significantly small.

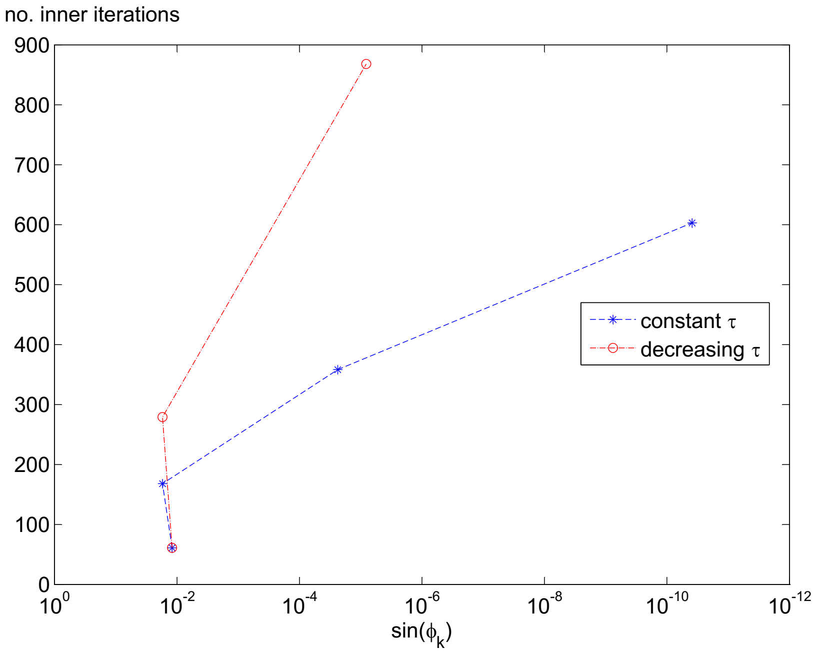

Figure 2.3: Number of inner iterations for the Rayleigh functional iteration depending on the angle between $x^{(k)}$ and the eigenspace

Chapter 3

The Jacobi-Davidson Method

The Jacobi-Davidson method was introduced for linear eigenvalue problems by Sleijpen and van der Vorst in [66]. They presented an alternative way to expand the current subspace based on ideas of Jacobi and Davidson.

The method was then adapted for generalized eigenvalue problems, polynomial eigenvalue problems [67, 65] and general nonlinear eigenvalue problems [8, 78]. In [45] Meerbergen, Schröder and Voss presented a variant for two parameter eigenvalue problems.

In the first section of this chapter the Jacobi-Davidson method is introduced. In the second section the perturbation of search space expansions is discussed. Furthermore, the Jacobi-Davidson method is considered as a robust variant of an existing iterative method. Finally, the (inexact) solution of the correction equation is discussed.

3.1 The Jacobi-Davidson Method

The idea of the Jacobi-Davidson method originates in combining and improving Jacobi's method ([31]) and Davidson's method ([10]). The new method is used for solving the special linear eigenvalue problem

$$Ax = \lambda x, \quad x \neq 0,$$

where A is a real symmetric matrix.

To improve the current iterate $x^{(k)}$ an orthogonal correction is sought. This correction vector is determined by solving the following linear system:

$$\left(I - x^{(k)} \left(x^{(k)}\right)^H\right) (A - \lambda^{(k)} I) \left(I - x^{(k)} \left(x^{(k)}\right)^H\right) t = -r^{(k)}, \tag{3.1}$$

$t \perp x^{(k)}$, $\left\|x^{(k)}\right\|_2 = 1$.

Here, $r^{(k)} = Ax^{(k)} - \lambda^{(k)} x^{(k)}$ denotes the residual after k steps.

For iterative projection methods, this correction vector is also a suitable subspace expansion. This is described in Algorithm 3.1.

Algorithm 3.1: Jacobi-Davidson method [66]

```
   input  : an initial nonzero vector $x$
   output: an approximation of an eigenvalue $\theta$ and a corresponding
           eigenvector
1  Compute $v^{(1)} = x/\|x\|_2$, $w^{(1)} = Av^{(1)}$, $h_{11} = \left(v^{(1)}\right)^H w^{(1)}$ ;
2  Set $V_1 = [v^{(1)}]$, $W_1 = [w^{(1)}]$, $H_1 = [h_{11}]$, $x^{(1)} = v^{(1)}$, $\theta^{(1)} = h_{11}$ ;
3  Compute $r^{(1)} = w^{(1)} - \theta^{(1)}x^{(1)}$ ;
4  for $k = 1$ to $m - 1$ do
5      Solve (approximately) (3.1) for $t$;
6      Orthogonalize $t$ against $V_k$ and expand $V_k$ with this vector $v^{(k+1)}$ to $V_{k+1}$;
7      Compute $w^{(k+1)} = Av^{(k+1)}$ and expand $W_k$ with this vector to $W_{k+1}$;
8      Compute $V_{k+1}^H w^{(k+1)}$ the last column of $H_{k+1}$ ;
9      if $A \neq A^H$ then
10         Compute $\left(v^{(k+1)}\right)^H W_k$ as the last row of $H_{k+1}$;
11     else
12         Take $\left(V_{k+1}^H w^{(k+1)}\right)^H$ as the last row of $H_{k+1}$;
13     end if
14     Compute all eigenpairs of $H_{k+1}$ and choose the desired one $(\theta^{(k+1)}, s)$
         with $\|s\|_2 = 1$ ;
15     Compute the Ritz vector $x^{(k+1)} = V_{k+1}s$, $u = Ax^{(k+1)} = W_{k+1}s$ and
         $r^{(k+1)} = u - \theta^{(k+1)}x^{(k+1)}$;
16     if $\left\|r^{(k)}\right\| \leqslant \varepsilon$ then
17         STOP;
18     end if
19 end for
20 Restart: Set $V_1 = [x^{(k+1)}]$, $W_1 = [u]$, $H_1 = \theta$ and go to 4;
```

The name of this method originates from the fact that the ideas of Jacobi and Davidson are merged to a new method.

The idea can be adapted to the generalized Hermitian eigenvalue problem

$$Av = \lambda Bv, \quad A = A^H,\ B = B^H \text{ and } B \text{ h.p.d.},$$

since with a Cholesky decomposition, $B = CC^H$, the problem can be transformed into the Hermitian special linear eigenvalue problem

$$C^{-1}AC^{-H}y = \lambda y.$$

Consider the nonlinear eigenvalue problem (2.1) and assume furthermore that $T(\cdot)$ allows a minmax characterization for its eigenvalues as well as $\inf_{x\in D} p(x) \in J$ (cf. Section 2.2). The Jacobi-Davidson method to nonlinear eigenvalue problems is presented by Voss in [75]. The following correction equation is presented:

$$\left(I - \frac{q\left(x^{(k)}\right)^H}{\left(x^{(k)}\right)^H q}\right) T(\lambda^{(k)}) \left(I - x^{(k)}\left(x^{(k)}\right)^H\right) t = -T(\lambda^{(k)})x^{(k)},\ t \perp x^{(k)}. \tag{3.2}$$

This is a general approach which allows different choices for the vector q.

Since $t \perp x^{(k)}$, the right projector can be neglected and (3.2) reduces to

$$\left(I - \frac{q\left(x^{(k)}\right)^H}{\left(x^{(k)}\right)^H q}\right) T(\lambda^{(k)})t = -T(\lambda^{(k)})x^{(k)}, \quad t \perp x^{(k)}. \tag{3.3}$$

If $\lambda^{(k)}$ is not an eigenvalue of (2.1), $T(\lambda^{(k)})$ is invertible and we determine the solution of the homogenous equation

$$\left(I - \frac{q\left(x^{(k)}\right)^H}{\left(x^{(k)}\right)^H q}\right) T(\lambda^{(k)})t = 0,$$

which is $t_h = T(\lambda^{(k)})^{-1}q$. A particular solution of (3.3) is given by $t = -x^{(k)}$. Finally, we have an explicit expression for all possible solutions t:

$$t = \beta T(\lambda^{(k)})^{-1}q - x^{(k)}. \tag{3.4}$$

To ensure that the next iterate of one step of inverse iteration or Rayleigh functional[1] iteration is included in the next subspace, the vector q is chosen by

$$q = T'(\lambda^{(k)})x^{(k)}. \tag{3.5}$$

Furthermore, $\beta \neq 0$ has to be satisfied. The value of β is then determined by inserting (3.4) into the orthogonality condition.

3.2 Global Convergence Behavior

Global convergence for the Jacobi-Davidson method for the real symmetric eigenproblem

$$Av = \lambda v, \quad A \in \mathbb{R}^{n\times n}$$

was proven by Aishima in [1]. Aishima proved for Algorithm 3.1 that if the largest Ritz Value $\theta^{(k)}$ is always chosen, this Ritz value converges to an eigenvalue λ_i of A. This proof assumes that (3.1) is solved exactly. Empirical tests show that the largest Ritz value usually converges to the largest eigenvalue.

The global convergence property remains true for the restarted Jacobi-Davidson method. But it does not provide a convergence rate.

The proof was generalized for iterative projection methods in [2].

[1] for linear eigenvalue problems: Rayleigh quotient

3.3 Jacobi-Davidson Method Without Subspace Acceleration

3.3.1 Basic Principle

Equations (3.4) and (3.5) imply that

$$x^{(k)} + t = \beta T(\lambda^{(k)})^{-1} T'(\lambda^{(k)}) x^{(k)}.$$

This shows that the solution of the Jacobi-Davidson correction equation can also be used to obtain the direction of the next iterate of inverse iteration (or Rayleigh quotient/functional iteration respectively). This suggests the method *Jacobi-Davidson method without Subspace Acceleration.* In the literature this method is also named by *Simplified Jacobi-Davidson method* or *Single Vector Jacobi-Davidson method* ([17, 36, 52, 64, 70]).

The convergence behavior of the Jacobi-Davidson method without subspace acceleration is analyzed by Jia and Wang in [36] for linear eigenproblems. Szyld and Xue present in [70] convergence proofs for nonlinear eigenvalue problems. Since large linear system cannot be solved exactly within an appropriate computing time, the influence of inexact solves of (3.2) is discussed.

Algorithm 3.2: Jacobi-Davidson method without Subspace Acceleration

```
input  : an initial nonzero vector $x^{(0)}$ with unit norm
output: an approximation for an eigenpair $(\hat{\lambda}, \hat{v})$
$\lambda^{(0)} = p(x^{(0)})$;
$r^{(0)} = T(\lambda^{(0)}) x^{(0)}$;
for $k = 0$ to convergence do
    Solve (approximately) (3.2) for $t$;
    $y^{(k+1)} = x^{(k)} + t$;
    $x^{(k+1)} = y^{(k+1)} / \left\| y^{(k+1)} \right\|$;
    $\lambda^{(k+1)} = p(x^{(k+1)})$;
    $r^{(k+1)} = T(\lambda^{(k+1)}) x^{(k+1)}$;
    Check $r^{(k+1)}$ for convergence;
end for
```

Algorithm 3.2 describes the Jacobi-Davidson method without subspace acceleration for the Hermitian nonlinear eigenvalue problem with a Rayleigh functional p.

3.3.2 Inexact Jacobi-Davidson without Subspace Acceleration

For the inexact variant of the Jacobi-Davidson method without subspace acceleration the approximate solution $\tilde{t}$ is considered. This vector still satisfies the orthogonality condition $\tilde{t} \perp x^{(k)}$ with a bounded relative residual

$$\frac{\left\| res^{(k)} \right\|_2}{\left\| r^{(k)} \right\|_2} \leqslant \tau^{(k)} < 1, \tag{3.6}$$

where

$$res^{(k)} := \left(I - \frac{T'(\lambda^{(k)})x^{(k)} \left(x^{(k)}\right)^H}{\left(x^{(k)}\right)^H T'(\lambda^{(k)})x^{(k)}} \right) T(\lambda^{(k)}) \left(I - x^{(k)} \left(x^{(k)}\right)^H \right) \tilde{t} + T(\lambda^{(k)})x^{(k)}. \tag{3.7}$$

In [70] Szyld and Xue prove the convergence behavior for different settings and properties. They assume that the initial vector $x^{(0)}$ is sufficiently close to the sought eigenvector and that the conditions of Theorem 2.36 are fulfilled. Those results are summarized in Table 3.1. Here t_k denotes the generalized tangent of $\measuredangle(x^{(k)}, \hat{v})$ and $\tau^{(k)}$ denotes the stop criterion for the inner iteration while solving the Jacobi-Davidson correction equation (3.2) inexactly (cf. (3.6)).

	Nonnsymm.		Symm.		
	$\tau^{(k)} = \tau^{(0)}$	$\tau_k = Ct^{(k)}$	$\tau^{(k)} = \tau^{(0)}$	$\tau^{(k)} = Ct_k$	$\tau^{(k)} = C{t_k}^2$
$\left\Vert res^{(k)} \right\Vert$	$\mathcal{O}(t_k)$	$\mathcal{O}(t_k^2)$	$\mathcal{O}(t_k)$	$\mathcal{O}(t_k^2)$	$\mathcal{O}(t_k^3)$
Convergence	Linear	Quadratic	Linear	Quadratic	Cubic

Table 3.1: Convergence behavior of the inexact Jacobi-Davidson method without subspace acceleration

3.4 Robustness of the Jacobi-Davidson Process Regarding the Reduced Influence of Perturbations

In this section we discuss how a perturbation in the subspace expansion influences the new subspace. These perturbations originate from inexact solves of the corresponding linear system in every iteration step.

In [77] Voss considers the influence of a perturbed subspace expansion. Here the current iterate x is expanded by a vector v to a two dimensional plane E. A perturbed expansion $\tilde{v} := v + e$ results in a perturbed plane $\tilde{E}$. The following theorem yields the maximal possible angle ϑ between the planes E and $\tilde{E}$ depending on the angle $\phi_0 := \angle(x, v)$ and the length of the perturbation e.

Theorem 3.1 ([77]).
The maximal possible acute angle between the planes E and $\tilde{E}$ is

$$\vartheta(\varepsilon, \phi_0) = \begin{cases} \arccos \sqrt{1 - \frac{\varepsilon^2}{\sin^2(\phi_0)}} & \text{if } \varepsilon \leqslant |\sin \phi_0| \\ \frac{\pi}{2} & \text{if } \varepsilon > |\sin \phi_0| \end{cases}$$

where $\varepsilon = \|e\|_2$.

Regarding the case $\varepsilon \leqslant |\sin \phi_0|$, we conclude that the angle ϑ is minimal for a fixed ε if $\phi_0 = \pm\frac{\pi}{2}$. This orthogonality condition is met if the search space expansion is determined by solving the Jacobi-Davidson correction equation.

As a result of Theorem 3.1, it is obvious that a subspace expansion by the next iterate of a vector iteration like Inverse Iteration or Rayleigh quotient iteration[2] is much more sensitive to perturbations than the Jacobi-Davidson method, since ϕ_0 tends to zero when the method converges.

Alternatively, methods exist, which do not provide $\left(x^{(k)}\right)^H v = 0$, but a convergence to zero, i.e.

$$\frac{\left(x^{(k)}\right)^H v^{(k)}}{\|v^{(k)}\|} \longrightarrow 0, \tag{3.8}$$

if k tends to infinity.

In [41] and [44] the Cayley transform is used to modify the subspace expansion for generalized linear eigenvalue problems such that

$$t_C = (A - \sigma B)^{-1}(A - \lambda^{(k)} B)x^{(k)},$$

which satisfies (3.8) for $v = t_C$.

For nonlinear eigenproblems, Voss shows in [78] that (3.8) is satisfied for the nonlinear Arnoldi method if

$$t_{NA} = T(\sigma)^{-1} T(\lambda^{(k)}) x^{(k)}$$

is used to expand the subspace instead of

$$t_{RIV} = x^{(k)} - T(\sigma)^{-1} T(\lambda^{(k)}) x^{(k)}.$$

Figure 3.1 illustrates that the perturbation of t is much stronger than the one in $x^{(k+1)}$ to obtain the same perturbation angle between the planes E and $\tilde{E}$.

In the following section we will present, how methods can be transformed into a robust method, satisfying $\left(x^{(k)}\right)^H v = 0$. We will see, that this is possible for every vector iteration.

3.5 The Jacobi-Davidson Method Based on any Iterative Method

In the previous section it was shown that the most robust way to expand a search space is to have an expansion t, which is orthogonal to the current iterate $x^{(k)}$. So far the Jacobi-Davidson method has been presented as a procedure to stabilize the Newton method in order to solve the linear or nonlinear eigenproblem. In this section we present another approach by Voss [80]. Each vector iteration to determine an eigenpair can be stabilized by determining a subspace expansion in a more robust way. The scheme originates from the same approach as the Jacobi-Davidson method, which is why we will continue calling this group of methods “Jacobi-Davidson method”.

[2] for nonlinear problems: Rayleigh functional iteration

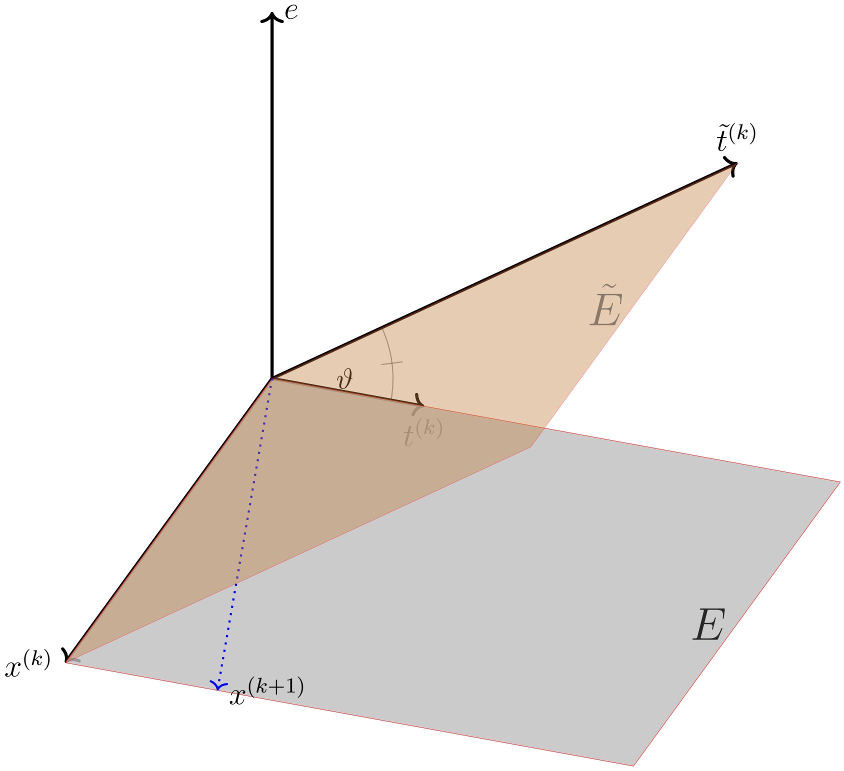

Figure 3.1: The desired subspaces vs the perturbed subspace

In Chapter 4 we will analyze the local convergence behavior when a stabilized method is applied on a perturbed subspace. This will be based on the convergence rates of the original method.

The considered iteration method to find an eigenvector can be described by

$$x^{(k+1)} = \alpha M_k x^{(k)}, \quad \alpha \neq 0, \tag{3.9}$$

where $M_k \in \mathbb{C}^{n \times n}$ is a matrix which performs a vector iteration from $x^{(k)}$ to a nonzero multiple of $x^{(k+1)}$. Additionally, a Rayleigh functional $p : \mathbb{C}^n \supseteq U \to \mathbb{C}$ is available. We consider the subspace spanned by the two iterates $x^{(k)}$ and $x^{(k+1)}$.

Possible methods are:

- Rayleigh quotient iteration for linear (generalized) eigenvalue problems

$$T(\lambda) = A - \lambda B, \quad M_k = \left(A - p\left(x^{(k)}\right) B\right)^{-1} B, \quad p(x) = \frac{x^H A x}{x^H B x}$$

- Rayleigh functional iteration for nonlinear eigenvalue problems

$$M_k = T\left(\lambda^{(k)}\right)^{-1} T'\left(\lambda^{(k)}\right)$$

- Derivative free variant of Rayleigh functional iteration (cf. [59])

$$M_k = T\left(\lambda^{(k)}\right)^{-1}$$

- Residual Inverse Iteration for nonlinear problems

$$M_k = I - T(\sigma)^{-1} T\left(\lambda^{(k)}\right)$$

The idea is to obtain a vector $t \in \mathbb{C}^n$ such that

$$\operatorname{span}\left\{x^{(k)}, x^{(k+1)}\right\} = \operatorname{span}\left\{x^{(k)}, t\right\} \quad \text{and} \quad t \perp x^{(k)},$$

as a solution of a linear system.

The ansatz for the linear system is motivated by the Jacobi-Davidson correction equation,

$$\left(I - \frac{C x^{(k)} {x^{(k)}}^H}{{x^{(k)}}^H C x^{(k)}}\right) S \left(I - \frac{x^{(k)} {x^{(k)}}^H B}{{x^{(k)}}^H B x^{(k)}}\right) t = -S x^{(k)}, \quad t \perp_B x^{(k)}, \tag{3.10}$$

where the matrix S is the matrix function $T(\cdot)$, evaluated at a certain value $\mu^{(k)} \neq \hat{\lambda}$, which may vary from step to step. If the eigenvectors can be chosen B-orthogonal, this property can be exploited by expanding the search space by a B-orthogonal correction instead of an orthogonal correction with respect to the standard scalar product. Therefore, the matrix B is inserted into the right projector. If the eigenvectors cannot be chosen B-orthogonal, we set $B = I$.

We will discuss different choices for the matrix C. We will see later that for $C = T'(\lambda^{(k)})$ the standard Newton based methods as Inverse Iteration or Rayleigh quotient / functional iteration are described.

Although the matrices S and C change after each iteration step, we neglect the indices for a better readability.

The B-orthogonality between t and $x^{(k)}$ simplifies (3.10) to

$$\left(I - \frac{Cx^{(k)}{x^{(k)}}^H}{{x^{(k)}}^H Cx^{(k)}}\right) St = -Sx^{(k)}, \quad t \perp_B x^{(k)}, \tag{3.11}$$

which is equivalent to

$$St - Cx^{(k)} \underbrace{\frac{{x^{(k)}}^H St}{{x^{(k)}}^H Cx^{(k)}}}_{=:\beta} = -Sx^{(k)}, \quad t \perp_B x^{(k)}. \tag{3.12}$$

Since S is invertible, t can be obtained by

$$t = \beta S^{-1} Cx^{(k)} - x^{(k)}.$$

Then β has to be chosen, such that $t \perp_B x^{(k)}$ holds.

Thus,

$$t = \frac{{x^{(k)}}^H Bx^{(k)}}{{x^{(k)}}^H BS^{-1}Cx^{(k)}} S^{-1}Cx^{(k)} - x^{(k)}. \tag{3.13}$$

The newly obtained vector t has to lie in $\operatorname{span}\{x^{(k)}, x^{(k+1)}\}$ so that

$$x^{(k+1)} = S^{-1}Cx^{(k)} = M_k x^{(k)},$$

and, therefore,

$$M_k = S^{-1}C.$$

This leads to the following different kinds of the correction equations

- Rayleigh quotient iteration in the linear case:

$$\left(I-\frac{Bx^{(k)}\left(x^{(k)}\right)^{H}}{\left(x^{(k)}\right)^{H}Bx^{(k)}}\right)\left(A-\lambda^{(k)}B\right)\left(I-\frac{x^{(k)}{x^{(k)}}^{H}B}{{x^{(k)}}^{H}Bx^{(k)}}\right)t$$
$$=-\left(A-\lambda^{(k)}B\right)x^{(k)},\quad t\quad \perp_{B} x^{(k)}.$$

- Rayleigh functional iteration in the nonlinear case:

$$\left(I-\frac{T'\left(\lambda^{(k)}\right)x^{(k)}{x^{(k)}}^{H}}{{x^{(k)}}^{H}T'\left(\lambda^{(k)}\right)x^{(k)}}\right)T\left(\lambda^{(k)}\right)\left(I-\frac{x^{(k)}{x^{(k)}}^{H}}{{x^{(k)}}^{H}x^{(k)}}\right)t=-T\left(\lambda^{(k)}\right)x^{(k)},$$
$$t\perp x^{(k)}$$

- Derivative free Rayleigh functional iteration in the nonlinear case:

$$\left(I-\frac{x^{(k)}{x^{(k)}}^{H}}{{x^{(k)}}^{H}x^{(k)}}\right)T\left(\lambda^{(k)}\right)\left(I-\frac{x^{(k)}{x^{(k)}}^{H}}{{x^{(k)}}^{H}x^{(k)}}\right)t=-T\left(\lambda^{(k)}\right)x^{(k)},\quad t\perp x^{(k)}$$

- Residual Inverse Iteration:

$$\left(I-\frac{(T(\sigma)-T(\lambda^{(k)}))x^{(k)}{x^{(k)}}^{H}}{{x^{(k)}}^{H}(T(\sigma)-T(\lambda^{(k)}))x^{(k)}}\right)T(\sigma)\left(I-\frac{x^{(k)}{x^{(k)}}^{H}}{{x^{(k)}}^{H}x^{(k)}}\right)t=-T(\sigma)x^{(k)},$$
$$t\perp x^{(k)}$$

For Residual Inverse Iteration, one can also think about using the Nonlinear Arnoldi subspace expansion

$$t_{NA}=T(\sigma)^{-1}T(\lambda^{(k)})x^{(k)}$$

as basic iterative method. This would yield $S=T(\sigma)$ and $C=T(\lambda^{(k)})$. But then we would divide by zero in the left projector, since $x^{H}Cx=0$. Therefore, we had to choose the matrices as above.

However, this approach is still quite theoretical, since for a small $\left\|T(\sigma)-T(\lambda^{(k)})\right\|$ the left projector might cause some computational problems. Then the numerator and the denominator become very small. It is still recommended to exploit the stable properties as mentioned above of the Nonlinear Arnoldi method.

3.6 The Correction Equation

We consider the general correction equation:

$$\overbrace{\left(I - \frac{Cxx^H}{x^HCx}\right) S \left(I - xx^HB\right)}^{=:S_{JD}} t = -r, \qquad t \perp_B x, \quad \text{and} \quad \|x\|_B = 1. \tag{3.15}$$

We assume, that S is invertible, i.e. the matrix function T is evaluated at any value $\lambda \neq \hat{\lambda}$.

Solving this equation efficiently is one of the largest challenges in the Jacobi-Davidson method. In this section some properties of the correction equation are presented. We will discuss the solution, preconditioning and the condition number of the corresponding matrix.

3.6.1 Solution of the Correction Equation

Considering the left projector in (3.15)

$$P_l := \left(I - \frac{Cxx^H}{x^HCx}\right),$$

we see that its range – and hence the range of the complete product of matrices – is the subspace $x^\perp$. Furthermore, the right projector

$$P_r := \left(I - xx^HB\right)$$

projects onto the subspace $x^{\perp_B} := \left\{z \in \mathbb{C}^n \mid x^HBz = 0\right\}$.

Therefore, this map can be considered to be restricted to these two subspaces as domain and range, and, thus,

$$\mathcal{S}_{JD} : \left\{ \begin{array}{rcl} x^{\perp_B} & \to & x^\perp \\ v & \mapsto & P_lSv. \end{array} \right.$$

The right projector P_r can be neglected because of

$$P_rv = v, \qquad \forall v \in x^{\perp_B}.$$

The matrix S_{JD} is singular, since $S_{JD}x = 0$. But if we consider the map $\mathcal{S}_{JD}$, then the inverse map can be determined. We solve

$$\left(I - \frac{Cxx^H}{x^HCx}\right) S \left(I - xx^HB\right) v = w, \quad t \perp_B x, \tag{3.16}$$

for $v \in x^{\perp_B}$ with any $w \in x^\perp$. Following (3.11) to (3.13) yields

$$\begin{aligned} v &= -S^{-1}Cx\frac{x^HBS^{-1}w}{x^HBS^{-1}Cx} + S^{-1}w \\ &= \left(I - \frac{S^{-1}Cxx^HB}{x^HBS^{-1}Cx}\right) S^{-1}w \\ &= S^{-1}\left(I - \frac{Cxx^HBS^{-1}}{x^HBS^{-1}Cx}\right) w. \end{aligned} \tag{3.17}$$

Hence, the matrix

$$S^{-1}\left(I - \frac{Cxx^HBS^{-1}}{x^HBS^{-1}Cx}\right) = \left(I - \frac{S^{-1}Cxx^HB}{x^HBS^{-1}Cx}\right)S^{-1}$$

can be seen as the inverse of S_{JD} restricted to the subspaces $x^{\perp}$ and $x^{\perp_B}$. This is summarized in the following proposition:

Proposition 3.2. *The inverse map of $\mathcal{S}_{JD}$ is denoted by $\mathcal{S}_{JD}^{-1} : x^{\perp} \to x^{\perp_B}$ and applied*

$$\mathcal{S}_{JD}^{-1}(v) = S_{JD}^{+}v,$$

with

$$S_{JD}^{+} = S^{-1}\left(I - \frac{Cxx^HBS^{-1}}{x^HBS^{-1}Cx}\right).$$

Proof. We obtain

$$S_{JD}^{+}S_{JD} = I - xx^HB,$$

and

$$S_{JD}S_{JD}^{+} = I - \frac{Cxx^H}{x^HCx}.$$

Since

$$\left(I - xx^HB\right)u = u, \qquad \forall u \in x^{\perp_B}$$

and

$$\left(I - \frac{Cxx^H}{x^HCx}\right)w = w, \qquad \forall w \in x^{\perp}$$

the proof is completed. □

Note that we would never compute S_{JD}^{+} explicitly.

In Chapter 4, convergence analysis for this stabilization procedure will be presented.

3.6.2 Preconditioning

If the Jacobi-Davidson correction equation, (3.15), is solved with an iterative solver, the method can be accelerated by using a preconditioner. We refer to [4, 23, 61] for further details. For the solution of the Jacobi-Davidson correction equation, the preconditioner has to be adapted, which we will explain in this section.

Let $K \approx S$ be a preconditioner for the matrix S, then it can be adapted to a preconditioner for S_{JD},

$$K_{JD} := \left(I - \frac{Cxx^H}{x^HCx}\right)K\left(I - xx^HB\right).$$

The inverse restricted to $x^{\perp}$ is obtained using Proposition 3.2 and replacing S by K, thus,

$$K_{JD}^{+} = \left(I - \frac{K^{-1}Cxx^{H}B}{x^{H}BK^{-1}Cx} \right) K^{-1}.$$

During the iteration process to solve the correction equation, this preconditioner has to be applied to any vector $u \in \mathbb{C}^n$ in every iteration step.

We now present an efficient application of this preconditioner algorithmically.

In Algorithm 3.3 variables are computed which do not vary during the iteration process. Algorithm 3.4 is performed before the first iteration starts.

Algorithm 3.3: Preprocessing to apply the Jacobi-Davidson preconditioner

```
input  : $K$ a preconditioner for $S$, the vector $x$ and the matrices $B$ and $C$
output: The vector $z$ and the scalar $\eta$ to be used in Algorithm 3.4
Solve $Kz = Cx$ for $z$;
Compute $\eta = x^H Bz$;
```

The second algorithm is executed in every iteration step and uses the output from Algorithm 3.3.

Algorithm 3.4: Application of the Jacobi-Davidson preconditioner to any right hand side $b \in x^{\perp}$

```
input  : $K$ a preconditioner for $S$, the vector $x$, the right hand side $b$, $\eta$, $z$
         and the matrix $B$
output: the result vector $y = K_{JD}^{+}b$
Solve $Ku = b$ for $u$ ;
Compute $y = u - \frac{x^H Bu}{\eta} z$;
```

Chapter 4

Perturbed Subspaces

In this chapter we consider the nonlinear eigenvalue problem

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

We generally assume that T is twice differentiable with respect λ.

Different setups for this problem are considered: In Section 4.2 and Subsection 4.3.2 we require $T(\lambda)$ to be real symmetric or Hermitian, respectively, for all λ in $J \subset \mathbb{R}$ for which we can use minmax-characterization. In Subsection 4.3.3 we assume $T(\cdot)$ to be a family of real matrices but do not require symmetry. We restrict ourselves in analyzing the behavior of the convergence to simple eigenvalues.

In Section 3.4 it was explained that the angle between the two subspaces E and $\tilde{E}$ is minimized, if the subspace expansion is chosen orthogonal[1] to the current eigenvector approximation $x^{(k)}$.

We consider the case that an existing search space is expanded orthogonally by a further direction. For simplicity, we consider the k-th search space containing $x^{(k)}$ as the one-dimensional subspace $\mathcal{V}_k = \text{span}\{x^{(k)}\}$. This subspace is then expanded by a direction $\hat{t}^{(k)} \neq 0$ with $\hat{t}^{(k)} \perp_B x^{(k)}$, thus, $\mathcal{V}_{k+1} = \text{span}\{x^{(k)}, \hat{t}^{(k)}\}$. Then we consider a perturbed search space expansion $\tilde{t}^{(k)}$ still satisfying $\tilde{t}^{(k)} \perp x^{(k)}$. The perturbed search space, $\tilde{\mathcal{V}}_{k+1}$, is given by $\tilde{\mathcal{V}}_{k+1} = \text{span}\{x^{(k)}, \tilde{t}^{(k)}\}$.

We chose this simplification to analyze the search spaces with a geometrical consideration.

This simplification is summarized generically in the following algorithm.

[1] B-orthogonal, if the eigenvectors can be chosen B-orthogonal

Algorithm 4.1: The Jacobi-Davidson with two dimensional subspaces

```
   input : an initial nonzero vector $x^{(0)}$ with unit $B$-norm
   output: approximations for an eigenpair $(\hat{\lambda}, \hat{v})$
1  Set $\lambda^{(0)} := p(x^{(0)})$;
2  Set $r^{(0)} := T(\lambda^{(0)})x^{(0)}$;
3  for $k = 0$ to convergence do
4      Solve approximately (3.2) for $t$ and set $\tilde{t}^{(k)} \perp_B x^{(k)}$ be the approximative
         solution;
5      Set $\tilde{V} := \left[ x^{(k)} \ \frac{\tilde{t}}{\|\tilde{t}\|_B} \right]$;
6      Choose either a left projector $\tilde{W} \in \mathbb{C}^{n\times 2}$ satisfying $\tilde{W}^H B\tilde{V} = I$ or set
         $\tilde{W} = \tilde{V}$;
7      Compute a solution $(\hat{\theta}, \hat{z})$ of the projected eigenvalue problem
         $\tilde{W}^H T(\theta)\tilde{V}z = 0$;
8      Set $x^{(k+1)} := \tilde{V}\hat{z}$;
9      Set $\lambda^{(k+1)} := p(x^{(k+1)})$;
10     Set $r^{(k+1)} := T(\lambda^{(k+1)})x^{(k+1)}$;
11 end for
```

The matrix B is used if the eigenvectors for the problem can be chosen to be B-orthogonal to each other, otherwise $B = I$.

For the functional p in step 9 we choose the Rayleigh functional if $T(\cdot)$ is Hermitian, otherwise p denotes a functional to extract an eigenvalue approximation if an eigenvector approximation is given. Requirements for those kinds of functionals are given in Assumption 4.23.

We begin in Section 4.1 with the condition numbers of eigenvalue problems. The projected problem changes, if the corresponding subspace is perturbed. Hence, the sensitivity of the eigenvalues towards these perturbations is analyzed. In Section 4.2 we will prove local convergence for the Ritz values on perturbed subspaces towards the extremal eigenvalues using minmax-characterization. Furthermore, in Section 4.3 we turn our consideration to interior eigenvalues. Finally, we will apply the Jacobi-Davidson process to solve the problems from the numerical examples in Chapter 2. The results will be compared with those determined by classical methods.

We usually consider the step from the k-th subspace expanded by $\hat{t}^{(k)}$ or $\tilde{t}^{(k)}$, respectively, to the $k+1$-th subspace. Therefore, we skip the indices for t for a better readability.

4.1 Condition of Eigenvalues and Eigenvectors

Condition numbers are a very interesting tool in the field of numerical mathematics. They measure the sensitivity of the result to perturbations of the input data.

For eigenvalue problems, the condition is also of interest. Here the sensitivity of eigenvalue and eigenvector toward perturbations of the matrix is considered.

Firstly, the condition of linear eigenvalue problems is introduced, then it is expanded to nonlinear problems. Thirdly, the condition number is described if the perturbation in the matrix traces back to a perturbation of the search space.

4.1.1 Condition of Linear Eigenvalue Problems

Considering the real linear eigenvalue problem with a perturbed matrix A with n distinct eigenvalues and a simple eigenvalue

$$(A + \varepsilon F)x(\varepsilon) = \lambda(\varepsilon)x(\varepsilon),$$

where $\|F\|_2 = 1$ and $\|x(\varepsilon)\|_2 = 1, \quad \forall \varepsilon \in \mathbb{R}$, the derivatives evaluated at 0 of λ and x are given by

$$\left|\dot{\lambda}(0)\right| = \left|\frac{\left(w^k\right)^H Fx(0)}{\left(w^k\right)^H x(0)}\right| \leqslant \frac{1}{|\left(w^k\right)^H x(0)|}$$

$$\dot{x}(0) = -\sum_{\substack{l=1\\ l\neq k}}^{n} \frac{\left(w^l\right)^H Fx(0)}{(\lambda_l - \lambda(0))} v^l$$

Here w^k with $\left\|w^k\right\|_2 = 1$ denotes a left eigenvector belonging to the eigenvalue $\lambda_k = \lambda(0)$ and $(\lambda_l, v^l)_{l=1,\ldots n}$ denote the eigenpairs of A.

The sensitivity of eigenvalues to perturbations in the matrix is presented for linear problems in [20], in detail.

4.1.2 Condition of Nonlinear Eigenvalue Problems

The ideas from the linear eigenvalue problem can be transferred to nonlinear eigenvalue problems. This is, for instance, done in [71] for polynomial eigenvalue problems. A more general approach is presented in [62]. Here a perturbation vector $\varepsilon \in \mathbb{C}^d$ is introduced by $\varepsilon = (\varepsilon_l)_{l=1}^d$. For linear eigenvalue problems the perturbed matrix is expressed as a function of the parameters ε. Here, in the nonlinear case, the matrix depends additionally on the eigenparameter λ such that $T(\lambda, \varepsilon)$, where $T(\lambda, 0) = T(\lambda), \quad \forall \lambda \in D$.

The condition number is then explained in Lemma 4.1 (cf. [62]).

Lemma 4.1. *Let $D \subset \mathbb{C}$ and $E \subset \mathbb{C}^d$ be open sets. Let $T(\cdot) : D \times E \to \mathbb{C}^{n\times n}$ be continuously differentiable, and let $\hat{\lambda} \in D$ be a simple eigenvalue of $T(\cdot, 0)$ and $\hat{v}, \hat{w}$ be the corresponding right and left eigenvectors with unit norm. Let $\hat{\tau} > 0, \hat{\varepsilon} > 0$ be such that $\bar{S}(\hat{\lambda}, \hat{\tau}) \subset D$ and $\bar{S}(0, \hat{\varepsilon}) \subset E$. Then, the first order perturbation expansion of the eigenvalue is given by*

$$\begin{aligned}\lambda(\varepsilon) - \hat{\lambda} &= \frac{\hat{w}^H \frac{\partial T}{\partial \varepsilon}(\hat{\lambda}, 0)[\varepsilon]\hat{v}}{\hat{w}^H T'(\hat{\lambda}, 0)\hat{v}} + \boldsymbol{o}(\|\varepsilon\|_2)\\ &= -\sum_{l=1}^{d} \varepsilon_l \frac{\hat{w}^H \frac{\partial T}{\partial \varepsilon_l}(\hat{\lambda}, 0)\hat{v}}{\hat{w}^H T'(\hat{\lambda}, 0)\hat{v}} + \boldsymbol{o}(\|\varepsilon\|_2).\end{aligned}$$

The normwise condition number for $\hat{\lambda}$ is given by

$$\kappa(\hat{\lambda}) := \limsup_{\|\varepsilon\|_2 \to 0} \frac{|\lambda(\varepsilon) - \hat{\lambda}|}{\|\varepsilon\|_2^2} = \left\| \frac{\partial \lambda}{\partial \varepsilon}(0) \right\|_2 = \sqrt{\sum_{j=1}^{d} \left| \frac{\hat{w}^H \frac{\partial T}{\partial \varepsilon_l}(\hat{\lambda}, 0)\hat{v}}{\hat{w}^H T'(\hat{\lambda}, 0)\hat{v}} \right|^2}.$$

Proof. See [62]. □

The perturbation of the eigenvector can be considered similarly as for the eigenvalue, which is shown in Lemma 4.2.

Lemma 4.2. *Let $D \subset \mathbb{C}$ and $E \subset \mathbb{C}^d$ be open sets. Let $T(\cdot,\cdot) : D \times E \to \mathbb{C}^{n\times n}$ be continuously differentiable, and let $\hat{\lambda} \in D$ a simple eigenvalue and $\hat{v}$ the corresponding (right) eigenvector with unit norm of $T(\cdot, 0)$. Furthermore, let $T(\hat{\lambda}, 0)$ be diagonalizable, $\mu_1, \dots, \mu_n$ its eigenvalues and $u^1, \dots, u^n$ the corresponding right eigenvectors with unit norm. Let l be chosen such that $\mu_l = 0$[2]. Let $w^1, \dots, w^n$ denote the left eigenvectors of $T(\hat{\lambda}, 0)$ fulfilling $(w^i)^H u^i = 1, \quad i = 1, \dots, n$.*

Moreover, we assume

$$\left(w^l\right)^H x(\varepsilon) = 1, \quad \forall \varepsilon \in E.$$

Then, the condition number of the eigenvector $\hat{v}$ can be bounded by

$$\left\| \frac{\partial x}{\partial \varepsilon} \right\|_{\varepsilon=0} \leqslant \sqrt{\sum_{i=1}^{d} \left(\sum_{\substack{j=1 \\ j\neq l}}^{n} \frac{1}{|\mu_j|^2} \left| (w^j)^H \frac{\partial T}{\partial \varepsilon_i} \hat{v} + (w^j)^H \frac{\partial T}{\partial \lambda} \hat{v} \frac{\partial \lambda}{\partial \varepsilon_i} \right| \right)^2} \Bigg|_{\substack{\varepsilon=0 \\ \lambda=\hat{\lambda}}}.$$

Proof. Because $x : \mathbb{C}^d \to \mathbb{C}^n$ the Jacobian $\frac{\partial x}{\partial \varepsilon}$ is a $n \times d$-matrix. The spectral norm can be bounded by the Frobenius norm.

$$\left\| \frac{\partial x}{\partial \varepsilon} \right\| \leqslant \left\| \frac{\partial x}{\partial \varepsilon} \right\|_F$$

The eigenvectors of the diagonalizable matrix $T(\hat{\lambda}, 0)$, $u^1, \dots, u^n$ can be taken as a basis of $\mathbb{C}^n$. So we make the following ansatz for $\frac{\partial x}{\partial \varepsilon}\Big|_{\substack{\varepsilon=0 \\ \lambda=\hat{\lambda}}}$.

$$\frac{\partial x}{\partial \varepsilon_i}\bigg|_{\substack{\varepsilon=0 \\ \lambda=\hat{\lambda}}} = \sum_{j=1}^{n} \alpha_j u^j \tag{4.1}$$

Building the derivative with respect to ε_i of

$$T(\lambda(\varepsilon), \varepsilon)x(\varepsilon) = 0$$
$$\left(w^l\right)^H x(\varepsilon) - 1 = 0$$

[2] We are introducing an enumeration of the eigenvalues of $T(\hat{\lambda}, 0)$, even though it might not be possible to order them (e.g. if there also complex eigenvalues). However, this enumeration is only used to distinguish between one eigenvalue $\mu_l = 0$ and $n - 1$ nonzero eigenvalues.

yields

$$\frac{\partial T}{\partial \varepsilon_i} x + \frac{\partial T}{\partial \lambda}\frac{\partial \lambda}{\partial \varepsilon_i} x + T(\lambda, \varepsilon)\frac{\partial x}{\partial \varepsilon_i} = 0, \tag{4.2a}$$

$$\left(w^l\right)^H \frac{\partial x}{\partial \varepsilon_i} = 0 \tag{4.2b}$$

Evaluating (4.2a) at $\varepsilon = 0$, $\lambda = \hat{\lambda}$ and $x = \hat{v}$ and inserting (4.1) yields

$$0 = \left.\frac{\partial T}{\partial \varepsilon_i}\right|_{\substack{\varepsilon=0\\ \lambda=\hat{\lambda}}} \hat{v} + \left.\frac{\partial T}{\partial \lambda}\frac{\partial \lambda}{\partial \varepsilon_i}\right|_{\substack{\varepsilon=0\\ \lambda=\hat{\lambda}}} \hat{v} + \sum_{j=1}^{n} \underbrace{T(\hat{\lambda}, 0)\alpha_j u^j}_{=\alpha_j \mu_j u^j}. \tag{4.3}$$

Because of the singularity of $T(\hat{\lambda}, 0)$ and the fact that $\hat{\lambda}$ is a simple eigenvalue, there exists exactly one eigenvalue $\mu_l = 0$. The parameters α_k for $k = 1, \ldots, l-1, l+1, \ldots, n$ can be extracted by multiplying (4.3) by $\left(w^k\right)^H$ from the left. Hence,

$$\left(w^k\right)^H \left.\frac{\partial T}{\partial \varepsilon_i}\right|_{\substack{\varepsilon=0\\ \lambda=\hat{\lambda}}} \hat{v} + \left(w^k\right)^H \left.\frac{\partial T}{\partial \lambda}\frac{\partial \lambda}{\partial \varepsilon_i}\right|_{\substack{\varepsilon=0\\ \lambda=\hat{\lambda}}} \hat{v} + \sum_{\substack{j=1\\ j\neq l}}^{n} \alpha_j \mu_j \underbrace{\left(w^k\right)^H u^j}_{=\delta_{jk}} = 0.$$

So α_k for $k \neq l$ can be determined by

$$\alpha_k = -\frac{1}{\mu_k}\left(\left(w^k\right)^H \left.\frac{\partial T}{\partial \varepsilon_i}\right|_{\substack{\varepsilon=0\\ \lambda=\hat{\lambda}}} \hat{v} + \left(w^k\right)^H \left.\frac{\partial T}{\partial \lambda}\frac{\partial \lambda}{\partial \varepsilon_i}\right|_{\substack{\varepsilon=0\\ \lambda=\hat{\lambda}}} \hat{v}\right).$$

To determine the parameter α_l, we consider (4.2b) which yields

$$\sum_{j=1}^{n} \alpha_j \left(w^l\right)^H u^j = 0. \quad \Leftrightarrow \alpha_l = -\frac{1}{\left(w^l\right)^H u^l} \sum_{\substack{j=1\\ j\neq l}}^{n} \underbrace{\left(w^l\right)^H u^j}_{=0} = 0.$$

Finally, we conclude

$$\begin{aligned} \left.\frac{\partial x}{\partial \varepsilon_i}\right|_{\substack{\varepsilon=0\\ \lambda=\hat{\lambda}}} &= \sum_{\substack{j=1\\ j\neq l}}^{n} \alpha_j u^j \\ &= \sum_{\substack{j=1\\ j\neq l}}^{n} -\frac{1}{\mu_j}\left(\left(w^j\right)^H \left.\frac{\partial T}{\partial \varepsilon_i}\right|_{\substack{\varepsilon=0\\ \lambda=\hat{\lambda}}} \hat{v} + \left(w^j\right)^H \left.\frac{\partial T}{\partial \lambda}\frac{\partial \lambda}{\partial \varepsilon_i}\right|_{\substack{\varepsilon=0\\ \lambda=\hat{\lambda}}} \hat{v}\right) u^j. \end{aligned}$$

Therefore, the norm of $\left.\frac{\partial x}{\partial \varepsilon_i}\right|_{\substack{\varepsilon=0\\ \lambda=\hat{\lambda}}}$ can be bounded by

$$\left\| \left.\frac{\partial x}{\partial \varepsilon_i}\right|_{\substack{\varepsilon=0\\ \lambda=\hat{\lambda}}} \right\|_2 \leqslant \sum_{\substack{j=1\\ j\neq l}}^{n} \frac{1}{|\mu_j|} \left| \left(w^j\right)^H \left.\frac{\partial T}{\partial \varepsilon_i}\right|_{\substack{\varepsilon=0\\ \lambda=\hat{\lambda}}} \hat{v} + \left(w^j\right)^H \left.\frac{\partial T}{\partial \lambda}\frac{\partial \lambda}{\partial \varepsilon_i}\right|_{\substack{\varepsilon=0\\ \lambda=\hat{\lambda}}} \hat{v} \right|. \tag{4.4}$$

The Frobenius norm of

$$\left.\frac{\partial x}{\partial \varepsilon}\right|_{\substack{\varepsilon=0\\ \lambda=\hat{\lambda}}} = \left.\begin{pmatrix} \frac{\partial x}{\partial \varepsilon_1} & \cdots & \frac{\partial x}{\partial \varepsilon_d} \end{pmatrix}\right|_{\substack{\varepsilon=0\\ \lambda=\hat{\lambda}}},$$

is given by

$$\left\| \frac{\partial x}{\partial \varepsilon}\bigg|_{\substack{\varepsilon=0\\ \lambda=\hat{\lambda}}} \right\| = \sqrt{\sum_{i=1}^{d} \left\| \frac{\partial x}{\partial \varepsilon_i}\bigg|_{\substack{\varepsilon=0\\ \lambda=\hat{\lambda}}} \right\|^2}. \tag{4.5}$$

Combining (4.4) and (4.5) completes the proof. □

With the results of Lemma 4.1 and Lemma 4.2, the sensitivity of eigenvalues and eigenvectors of a projected eigenvalue problem towards perturbations in the subspace can be described.

In the following section, these condition numbers are determined for a perturbed subspace by an angle ϑ.

4.1.3 Condition Regarding Perturbed Subspaces

Now the concrete case is studied, where a search space is built by the current iterate $x^{(k)}$ and the orthogonal search space expansion obtained by the Jacobi-Davidson method (cf. Chapter 3). So, a new better iterate $x^{(k+1)}$ is sought in the two dimensional subspace spanned by $x^{(k)}$ and $M_k x^{(k)}$. We concentrate on problems with real matrices and regard only real eigenvalues[3], where Assumption 2.14 is satisfied and on Rayleigh-Ritz projections (i.e. left and right search space are equal), which leads to the following projected 2×2-eigenvalue problem:

$$V^T T(\lambda) V s = 0 \tag{4.6}$$

Let $\hat{t}$ be the exact solution of the Jacobi-Davidson correction equation (3.10). Then a perturbation caused by an inexact solve of this equation can be described by

$$\tilde{t}(\vartheta) = \cos(\vartheta)\frac{\hat{t}}{\left\|\hat{t}\right\|_B} + \sin(\vartheta)e, \tag{4.7}$$

where $e \perp_B \operatorname{span}\{x^{(k)}, \hat{t}\}$ and $\|e\|_B = 1$. Then the projected problem can be described depending on ϑ by

$$\underbrace{\tilde{V}(\vartheta)^T T(\lambda) \tilde{V}(\vartheta)}_{=:T_p(\lambda,\vartheta)} s = 0, \quad \text{with} \quad \tilde{V}(\vartheta) = \hat{V} + \begin{pmatrix} 0 & (\cos(\vartheta) - 1)\frac{\hat{t}}{\|\hat{t}\|_B} + \sin(\vartheta)e \end{pmatrix}.$$

Here $\hat{V}$ is given by

$$\hat{V} := \begin{pmatrix} x^{(k)} & \frac{\hat{t}}{\|\hat{t}\|_B} \end{pmatrix},$$

such that its range is the unperturbed search space.

With the derivative of $V(\vartheta)$,

$$\dot{\tilde{V}}(\vartheta) \quad = \quad \begin{pmatrix} 0 & -\sin(\vartheta)\frac{\hat{t}}{\|\hat{t}\|_B} + \cos(\vartheta)e \end{pmatrix},$$

[3] The implicit function theorem cannot be used for complex equations, since the projected problem $T_p(\lambda, \vartheta)s = \tilde{V}(\vartheta)^H T(\lambda) \tilde{V}(\vartheta)s$ is not analytic with respect to ϑ.

the partial derivative of T_p with respect to ϑ can be determined, thus,

$$\frac{\partial T_p}{\partial \vartheta} = \dot{\tilde{V}}(\vartheta)^T T(\lambda)\tilde{V}(\vartheta) + \tilde{V}(\vartheta)^T T(\lambda)\dot{\tilde{V}}(\vartheta). \tag{4.9}$$

Evaluating F at $\vartheta = 0$ yields

$$\begin{aligned} \left.\frac{\partial T_p}{\partial \vartheta}\right|_{\substack{\lambda=\hat{\lambda}\\ \vartheta=0}} &= \dot{\tilde{V}}(0)^T T(\hat{\lambda})\tilde{V}(0) + \tilde{V}(0)^T T(\hat{\lambda})\dot{\tilde{V}}(0) \\ &= \begin{pmatrix} 0 \\ 1 \end{pmatrix} e^T T(\hat{\lambda})\hat{V} + \hat{V}^T T(\hat{\lambda}) e \begin{pmatrix} 0 & 1 \end{pmatrix}, \end{aligned}$$

since

$$\begin{aligned} \dot{\tilde{V}}(0) &= \begin{pmatrix} 0 & e \end{pmatrix} \\ \tilde{V}(0) &= \hat{V}. \end{aligned}$$

With the help of Lemma 4.1

$$\dot{\hat{\lambda}}(0) = \left.\frac{w^T \frac{\partial T_p}{\partial \vartheta} s}{w^T \frac{\partial T_p}{\partial \lambda}(\hat{\lambda}, 0) s}\right|_{\substack{\lambda=\hat{\lambda}\\ \vartheta=0}} \tag{4.10}$$

is obtained where w denotes the corresponding left eigenvector of $\hat{\lambda}$. We assume that the resulting Ritz vector is not orthogonal to the current iterate $x^{(k)}$, since $x^{(k)}$ is already sufficiently close an eigenvector. Therefore, we scale the eigenvector s such that $s_1 = 1$, thus,

$$s = \begin{pmatrix} 1 \\ \alpha \end{pmatrix}$$

(cf. (4.19) in Subsection 4.2.1).

Then we obtain for the numerator of (4.10)

$$\begin{aligned} w^T \left.\frac{\partial T_p}{\partial \vartheta}\right|_{\substack{\lambda=\hat{\lambda}\\ \vartheta=0}} s &= \begin{pmatrix} w_1 & w_2 \end{pmatrix} \left(\begin{pmatrix} 0 \\ 1 \end{pmatrix} e^T T(\hat{\lambda})\hat{V} + \hat{V}^T T(\hat{\lambda}) e \begin{pmatrix} 0 & 1 \end{pmatrix} \right) \begin{pmatrix} 1 \\ \alpha \end{pmatrix} \\ &= w_2 e^T T(\hat{\lambda}) \left(x^{(k)} + \alpha \frac{\hat{t}}{\left\|\hat{t}\right\|} \right) + \alpha w^T \hat{V}^T T(\hat{\lambda}) e \end{aligned} \tag{4.11}$$

According to Assumption 2.14, the eigenvalue $\hat{\lambda}$ is simple, and, therefore, $w^T \left.\frac{\partial T_p}{\partial \lambda}\right|_{\substack{\lambda=\hat{\lambda}\\ \vartheta=0}} s \neq 0$. Inserting (4.11) into the absolute value of (4.10) yields

$$|\dot{\hat{\lambda}}(0)| \leqslant C_1 \left\| r^{(k+1)} \right\| + C_2 |\alpha|, \qquad C_1, C_2 \geqslant 0 \tag{4.12}$$

In Lemma 4.5 in Subsection 4.2.1, we will see that

$$|\alpha| = \mathcal{O}\left(\sin\left(\phi^{(k)}\right)\right), \tag{4.13}$$

where $\phi^{(k)} := \measuredangle\left(x^{(k)}, \hat{v}\right)$.

For the condition number of the eigenvector we obtain, with the help of Lemma 4.2,

$$\|\dot{s}(0)\| \leqslant \frac{1}{|\mu_2|}\left|\left(w^2\right)^T \frac{\partial T_p}{\partial \vartheta}\bigg|_{\substack{\lambda=\hat{\lambda}\\ \vartheta=0}} s + \left(w^2\right)^T \frac{\partial T_p}{\partial \lambda}\bigg|_{\substack{\lambda=\hat{\lambda}\\ \vartheta=0}} s\, \dot{\lambda}(0)\right|. \tag{4.14}$$

Here we consider only a 2×2-nonlinear eigenvalue problem, so there exist only two (linear) eigenvalues μ_1, μ_2 of the matrix $T(\hat{\lambda}, 0)$. We choose the numeration of these eigenvalue, such that

$$\mu_1 = 0 \qquad \text{and} \qquad \mu_2 \neq 0.$$

The second summand of (4.14) can be bounded using (4.12). Considering the first summand yields

$$\left(w^2\right)^T T_\vartheta(\hat{\lambda}, 0)s = w_2^2 e^T r^{(k+1)} + \alpha \left(w^2\right)^T V^T T(\lambda^{(k+1)})e,$$

thus,

$$\left|\left(w^2\right)^T \frac{\partial T_p}{\partial \vartheta}\bigg|_{\substack{\lambda=\hat{\lambda}\\ \vartheta=0}} s\right| \leqslant |w_2^2| \left\|r^{(k+1)}\right\| + \mathcal{O}(|\alpha|). \tag{4.15}$$

Finally, these results can be summarized in Theorem 4.3.

Theorem 4.3.
Let $T : \mathbb{R} \supset J \to \mathbb{R}^{n\times n}$ be a family of real matrices, $\hat{\lambda} \in J$ a simple eigenvalue of $T(\cdot)$ and $\hat{v} \in \mathbb{R}^n$ a corresponding eigenvector. We consider an iteration process $\{x^{(k)}\}_{k=1}^{\infty}$ converging to a nonzero multiple of $\hat{v}$. Let $\delta > 0$ such that $T(\lambda)$ is diagonalizable for all $\lambda \in \bar{S}(\hat{\lambda}, \delta)$. Let $\lambda^{(l)} \in \bar{S}(\hat{\lambda}, \delta)\ \forall l \geqslant k$.

We assume that $x^{(k)}$ is already sufficiently close to $\hat{v}$ and consider a two dimensional subspace spanned by $x^{(k)}$ and $x^{(k+1)}$ and the resulting projected 2×2-nonlinear Rayleigh-Ritz eigenvalue problem (4.6). *For perturbations of the search space expansion, cf.* (4.7), *the condition number behaves as follows.*

$$\dot{\lambda}^{(k+1)}(0) = \mathcal{O}\left(\sin\left(\phi^{(k)}\right)\right), \tag{4.16a}$$

and

$$\|\dot{s}(0)\| = \mathcal{O}\left(\sin\left(\phi^{(k)}\right)\right), \tag{4.16b}$$

where $\phi^{(k)} = \measuredangle\left(x^{(k)}, \hat{v}\right)$.

Proof. Consider Lemma 4.1 and apply it to

$$T_p(\lambda, \vartheta) = \tilde{V}(\vartheta)^T T(\lambda)\tilde{V}(\vartheta)s \overset{!}{=} 0. \tag{4.17}$$

Then (4.12) and (4.13) complete the proof of (4.16a).

For the second part of the proof, Lemma 4.2 is applied to (4.17). We need to provide a scaling of left and right eigenvector such that

$$s_1 = 1$$
$$\text{and}$$
$$\hat{w}^H s = 1$$

is fulfilled.

Since $x^{(k)}$ is sufficiently close to the sought eigenvector we can expect $s_1 \neq 0$ which allows us to scale the vector s such that $s_1 = 1$. The left eigenvector $\hat{w}$ is not orthogonal to the right eigenvector s and therefore it can be scaled as well such that $\hat{w}^H s = 1$. Lemma 4.2 yields (4.14). Finally, with (4.15), the proof is completed. □

Theorem 4.3 shows that the condition number of the eigenvalue and eigenvector tends linearly to 0 with the angle $\measuredangle\left(x^{(k)}, \hat{v}\right)$. Hence, the closer the Ritz pair is to an eigenpair[4], the smaller the influence of perturbations is in the search space. Therefore, the perturbations do not need to be reduced, when the Ritz vectors become closer to the eigenvector.

Unfortunately, this result cannot be used to prove the convergence of iterative projection methods. It only confirms the guess that the Ritz values become more robust against perturbations in the subspace expansion, when $x^{(k)}$ is already close to the eigenvector.

4.2 Error Estimation for Extremal Eigenvalues

In this section we analyze the influence of perturbed subspaces on the convergence to the smallest eigenvalues. We begin with a general geometric consideration, which can be used for linear as well as for nonlinear eigenvalue problems. We continue by examining convergence properties for the generalized linear eigenvalue problem

$$Ax = \lambda Bx,\ A, B \in \mathbb{C}^{n\times n} \tag{4.18}$$

where A is Hermitian and B Hermitian positive definite. The results will, finally, be generalized for nonlinear eigenvalue problems which satisfy Assumptions 2.7 and 2.8.

Through the entire chapter we use the numbering which is defined by the minmax characterization for non-overdamped nonlinear eigenvalue problems[5]. Moreover, we assume that the smallest eigenvalue is simple, i.e.

$$\lambda_1 < \lambda_2 \leqslant \ldots \leqslant \lambda_n.$$

[4] The distance between two eigenpairs can be measured, for instance, by the angle between the Ritz vector and the eigenvector.

[5] cf. Definition 2.10

4.2.1 Geometric Consideration of perturbed Subspaces

Throughout this subsection we use the notation for nonlinear eigenvalue problems, although the results are used for linear and nonlinear eigenvalue problems.

Definition 4.4. We define the two planes E and $\tilde{E}$:

$$\begin{aligned} E &:= \operatorname{span}\{x^{(k)}, \hat{t}\} \\ \tilde{E} &:= \operatorname{span}\{x^{(k)}, \tilde{t}\}, \end{aligned}$$

where $\hat{t}$ denotes the solution of the Jacobi-Davidson correction equation (3.10), which satisfies $x^{(k)} \perp_B \hat{t}$. For the perturbed subspace $\tilde{E}$ we define

$$\tilde{t} := \cos(\vartheta)\frac{\hat{t}}{\left\|\hat{t}\right\|_B} + \sin(\vartheta)e, \quad \text{with } e \perp_B \{x^{(k)}, \hat{t}\} \text{ and } \|e\|_B = 1.$$

Thus, ϑ denotes the angle between the planes E and $\tilde{E}$. The greater ϑ is in magnitude, the greater the perturbation of the subspace will be.

Note that $\tilde{t}$ is always scaled such that $\left\|\tilde{t}\right\|_B = 1$, while $\left\|\hat{t}\right\|_B$ varies from step to step[6].

We scale $x^{(k)}$ to

$$\left\|x^{(k)}\right\|_B = 1,$$

with B as defined in (4.18) for linear eigenvalue problems and $B = I$ for nonlinear eigenvalue problems.

The new iterate $x^{(k+1)} := \gamma M_k x^{(k))}$ (cf. (3.9)) is described by

$$x^{(k+1)} = x^{(k)} + \hat{t} \tag{4.19}$$

with a suitable scaling factor $\gamma \neq 0$.

Let $\tilde{x}^{(k+1)} \in \tilde{E}$ be the Ritz vector corresponding to the smallest Ritz value of the perturbed projected problem. Then we can estimate the error of the perturbed Ritz values by

$$\left|p\left(\tilde{x}^{(k+1)}\right) - p\left(v^1\right)\right| \leqslant \left|p\left(\tilde{y}\right) - p\left(v^1\right)\right|, \quad \forall \tilde{y} \in \tilde{E}. \tag{4.20}$$

We define

$$\tilde{y} := x^{(k+1)} + \alpha e \tag{4.21}$$

where α is chosen such that $\tilde{y} \in \tilde{E}$. We can determine α geometrically. Considering Figure 4.1 where we start at $x^{(k+1)}$ and go orthogonally to E along e until the plane $\tilde{E}$ is reached yields the result.

$$\alpha = \left\|\hat{t}\right\|_B \tan(\vartheta).^7 \tag{4.22}$$

We determine $\left\|\hat{t}\right\|_B$ by considering Figure 4.1 to

$$\left\|\hat{t}\right\|_B = \tan(\delta). \tag{4.23}$$

The following Lemma describes the behavior of $\tan(\delta)$.

[6] When $x^{(k)}$ converges to $\hat{v}$, $\|\hat{t}\|_B \to 0$.

[7] Calculating the intersecting point of the line $x^{(k+1)} + \alpha e$ and $\tilde{E}$ would yield the same result.

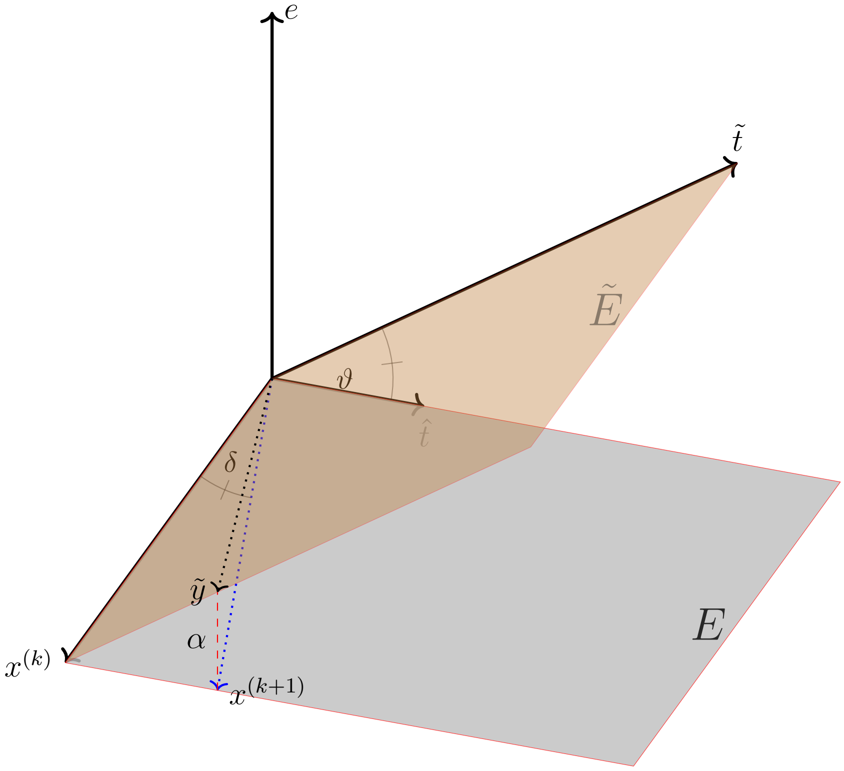

Figure 4.1: The desired subspace vs the perturbed subspace

Lemma 4.5. *Let*

$$x^{(k+1)} = \gamma_k M_k x^{(k)}, \quad \mathbb{C} \ni \gamma_k \neq 0, x^{(k)} \in \mathbb{C}^n, M_k \in \mathbb{C}^{n\times n}$$

describe an iterative method to determine an eigenpair of a linear eigenvalue problem or a nonlinear eigenvalue problem (cf. (3.9)*). Let the angles ϕ_k and ϕ_{k+1} be sufficiently small, i.e.* $|\tan(\phi_k)\tan(\phi_{k+1})| \ll 1$*, then*

$$|\tan(\delta)| = \mathcal{O}\left(|\sin\left(\phi_k\right)|\right) + \mathcal{O}\left(|\sin\left(\phi_{k+1}\right)|\right)$$

holds, while $\phi_k := \measuredangle(x^{(k)}, v^1)$[8]*.*

Proof. From (4.23) we have

$$|\tan(\delta)| = \left\|\hat{t}\right\|_B. \tag{4.24}$$

Therefore, an estimation of the norm of $\hat{t}$ is sought. The subspace expansion $\hat{t}$ is determined by solving the Jacobi-Davidson correction equation, (3.10). In chapter 3 an explicit expression for $\hat{t}$ was developed resulting in (3.13),

$$\hat{t} = \frac{\overbrace{\left(x^{(k)}\right)^H Bx^{(k)}}^{=1}}{\left(x^{(k)}\right)^H B \underbrace{S^{-1}Cx^{(k)}}_{=:y}} \underbrace{S^{-1}Cx^{(k)}}_{=y} - x^{(k)} = \frac{1}{\left(x^{(k)}\right)^H By} y - x^{(k)}. \tag{4.25}$$

For this proof we will decompose each vector, $x^{(k)}$ and y, into a component into the direction of v^1 and another direction orthogonal to v^1. This leads to an expression of $\hat{t}$, where it is decomposed into one direction parallel to v^1 and one orthogonal to it. This result can be used to estimate the norm of $\hat{t}$.

We use the following decompositions[9]

$$x^{(k)} = \cos\left(\phi^{(k)}\right) v^1 + \sin\left(\phi^{(k)}\right) w^{(k)} =: c_k v^1 + s_k w^{(k)} \tag{4.26a}$$

$$\begin{aligned} y &= \eta\left(\cos\left(\phi^{(k+1)}\right) v^1 + \sin\left(\phi^{(k+1)}\right) w^{(k+1)}\right) \\ &=: \eta(c_{k+1}v^1 + s_{k+1}w^{(k+1)}), \end{aligned} \tag{4.26b}$$

with $\|v^1\|_B = \left\|w^{(k)}\right\|_B = \left\|w^{(k+1)}\right\|_B = 1$ and $\{w^{(k)}, w^{(k+1)}\} \perp_B v^1$. Without loss of generality we assume

$$\begin{aligned} c_k, c_{k+1} &\in \mathbb{C} \\ s_k, s_{k+1} &\in \mathbb{R}_{\geqslant 0}. \end{aligned}$$

The latter restriction can be made, since we could easily transform $w^{(k)}$ such that $s_k \in \mathbb{R}_{\geqslant 0}$, if s_k was complex.

$$s_k w^{(k)} = |s_k| \exp(i\psi) w^{(k)} := |s_k| \tilde{w}^{(k)}$$

[8] The angle is defined according to the B-inner product. B is either chosen such that each pair of eigenvectors belonging to two different eigenvalues can be chosen B-orthogonal or $B = I$

[9] For nonlinear eigenvalue problems we set $B = I$

With

$$\left(x^{(k)}\right)^H By = \eta\left(\overline{c_k}c_{k+1} + s_k s_{k+1}\underbrace{\left(w^{(k)}\right)^H Bw^{(k+1)}}_{=\nu}\right)$$

and inserting (4.26a) and (4.26b) into (4.25) we end up with

$$\begin{aligned}\hat{t} &= \frac{\eta(c_{k+1}v^1 + s_{k+1}w^{(k+1)})}{\eta(\overline{c_k}c_{k+1} + s_k s_{k+1}\nu)} - c_k v^1 - s_k w^{(k)}\\ &= \left(\frac{c_{k+1}}{\overline{c_k}c_{k+1} + s_k s_{k+1}\nu} - c_k\right)v^1 + \frac{s_{k+1}}{\overline{c_k}c_{k+1} + s_k s_{k+1}\nu}w^{(k+1)} - s_k w^{(k)}\\ &= \left(\frac{1}{\overline{c_k} + s_k\frac{s_{k+1}}{c_{k+1}}\nu} - c_k\right)v^1 + \frac{s_{k+1}}{\overline{c_k}c_{k+1} + s_k s_{k+1}\nu}w^{(k+1)} - s_k w^{(k)}. \qquad (4.27)\end{aligned}$$

The component into the direction of $\hat{v}$ can be simplified to

$$\begin{aligned}\frac{1}{\overline{c_k} + s_k\frac{s_{k+1}}{c_{k+1}}\nu} - c_k &= \frac{\overbrace{1 - |c_k|^2}^{=|s_k|^2} - s_k c_k\frac{s_{k+1}}{c_{k+1}}\nu}{\overline{c_k} + s_k\frac{s_{k+1}}{c_{k+1}}\nu}\\ &= \frac{|s_k|^2 - s_k c_k\frac{s_{k+1}}{c_{k+1}}\nu}{\overline{c_k} + s_k\frac{s_{k+1}}{c_{k+1}}\nu}\\ &= \frac{c_{k+1}|s_k|^2 - s_{k+1}s_k c_k\nu}{\overline{c_k}c_{k+1} + s_k s_{k+1}\nu} \qquad (4.28)\end{aligned}$$

Inserting (4.28) into (4.27) yields

$$\hat{t} = \frac{c_{k+1}|s_k|^2 - s_{k+1}s_k c_k\nu}{\overline{c_k}c_{k+1} + s_k s_{k+1}\nu}v^1 + \frac{s_{k+1}}{\overline{c_k}c_{k+1} + s_k s_{k+1}\nu}w^{(k+1)} - s_k w^{(k)}. \qquad (4.29)$$

For having a nonzero denominator we need to show that

$$\nu \neq -\frac{\overline{c_k}c_{k+1}}{s_k s_{k+1}}. \qquad (4.30)$$

With the requirement

$$|\tan(\phi_k)\tan(\phi_{k+1})| = \left|\frac{s_k s_{k+1}}{c_k c_{k+1}}\right| \ll 1,$$

the product of the cotangents satisfies

$$|\cot(\phi_k)\cot(\phi_{k+1})| = \frac{1}{|\tan(\phi_k)\tan(\phi_{k+1})|} = \left|\frac{c_k c_{k+1}}{s_k s_{k+1}}\right| = \left|\frac{\overline{c_k}c_{k+1}}{s_k s_{k+1}}\right| \gg 1.$$

On the other hand, $\nu = \left(w^{(k)}\right)^H Bw^{(k+1)} \leqslant \underbrace{\left\|w^{(k)}\right\|_B}_{=1}\underbrace{\left\|w^{(k+1)}\right\|_B}_{=1} = 1$. Thus, the denominator cannot get close to zero.

By taking the B-norm and applying the triangle inequality to (4.29) we obtain

$$\begin{aligned}
\left\|\hat{t}\right\|_B &= \left\| \frac{c_{k+1}s_k^2 - s_{k+1}s_kc_k\nu}{\overline{c_k}c_{k+1} + s_ks_{k+1}\nu}\, v^1 + \frac{s_{k+1}}{\overline{c_k}c_{k+1} + s_ks_{k+1}\nu} w^{(k+1)} - s_k w^{(k)} \right\|_B \\
&\leqslant \left\| \frac{c_{k+1}s_k^2 - s_{k+1}s_kc_k\nu}{\overline{c_k}c_{k+1} + s_ks_{k+1}\nu}\, v^1 \right\|_B + \left\| \frac{s_{k+1}}{\overline{c_k}c_{k+1} + s_ks_{k+1}\nu} w^{(k+1)} \right\|_B + \left\| s_k w^{(k)} \right\|_B \\
&= \left| \frac{c_{k+1}s_k^2 - s_{k+1}s_kc_k\nu}{\overline{c_k}c_{k+1} + s_ks_{k+1}\nu} \right| \underbrace{\left\| v^1 \right\|_B}_{=1} + \left| \frac{s_{k+1}}{\overline{c_k}c_{k+1} + s_ks_{k+1}\nu} \right| \underbrace{\left\| w^{(k+1)} \right\|_B}_{=1} + |s_k| \underbrace{\left\| w^{(k)} \right\|_B}_{=1} \\
&= \underbrace{\left| \frac{c_{k+1}s_k^2 - s_{k+1}s_kc_k\nu}{\overline{c_k}c_{k+1} + s_ks_{k+1}\nu} \right|}_{\mathcal{O}(|s_k^2|)+\mathcal{O}(|s_ks_{k+1}|)} + \underbrace{\left| \frac{s_{k+1}}{\overline{c_k}c_{k+1} + s_ks_{k+1}\nu} \right|}_{=\mathcal{O}(|s_{k+1}|)} + |s_k| \\
&= \mathcal{O}(|s_{k+1}|) + \mathcal{O}(|s_k|) \qquad (4.31)
\end{aligned}$$

which completes the proof using (4.24). □

This result can be used to analyze the convergence behavior for linear and nonlinear eigenvalue problems.

4.2.2 Generalized Linear Eigenvalue Problems

In this subsection we consider the Rayleigh quotient instead of the Rayleigh functional. Therefore, we denote it by $R : \mathbb{C}^n \setminus \{0\} \to \mathbb{R}$ with

$$R(x) = \frac{x^H A x}{x^H B x} \tag{4.32}$$

where A is Hermitian and B Hermitian positive definite.

For the next step of the error estimation the following relation is needed:

$$\tan^2(\phi_k) = \frac{R(x^{(k)}) - \lambda_1}{\eta - R(x^{(k)})} = \mathcal{O}(R(x^{(k)}) - \lambda_1) \tag{4.33}$$

where $\eta = w^H A w$ and $w \in \mathbb{C}^n$ is defined in (2.17).

This relation has been proven by Notay in [52]. It directly implies

$$\sin^2(\phi_k) = \mathcal{O}(R(x^{(k)}) - \lambda_1). \tag{4.34}$$

Now we consider (4.20) for the generalized linear eigenvalue problem, thus,

$$R\left(\tilde{x}^{(k+1)}\right) - R\left(v^1\right) \leqslant R\left(\tilde{y}\right) - R\left(v^1\right), \tag{4.35}$$

with $\tilde{y} := x^{(k+1)} + \alpha e$ where α is determined from (4.22). We assume again that $\left\|x^{(k)}\right\|_B = \|e\|_B = 1$. We can estimate the error to

$$R\left(\tilde{y}\right) - R\left(v^1\right) = R\left(x^{(k+1)} + \alpha e\right) - \lambda_1. \tag{4.36}$$

We exploit the invariance towards scalings of the Rayleigh quotient, thus,

$$\begin{aligned} R\left(x^{(k+1)} + \alpha e\right) &= R\left(\frac{1}{\|x^{(k+1)}\|_B}\left(x^{(k+1)} + \alpha e\right)\right) \\ &= R\left(\frac{x^{(k+1)}}{\|x^{(k+1)}\|_B} + \frac{\alpha}{\|x^{(k+1)}\|_B}\, e\right) \\ &= R\left(\check{x}^{(k+1)} + \check{\alpha} e\right) \qquad (4.37) \end{aligned}$$

where

$$\check{x}^{(k+1)} := \frac{x^{(k+1)}}{\|x^{(k+1)}\|_B} \qquad (4.38a)$$

$$\check{\alpha} := \frac{\alpha}{\|x^{(k+1)}\|_B}. \qquad (4.38b)$$

Combining (4.36) and (4.37) we obtain

$$\begin{aligned} R\left(\tilde{y}\right) - R\left(v^1\right) &= R\left(\check{x}^{(k+1)} + \check{\alpha} e\right) - \lambda_1 \\ &= \frac{\left(\check{x}^{(k+1)} + \check{\alpha} e\right)^H A \left(\check{x}^{(k+1)} + \check{\alpha} e\right)}{\left(\check{x}^{(k+1)} + \check{\alpha} e\right)^H B \left(\check{x}^{(k+1)} + \check{\alpha} e\right)} - \lambda_1 \\ &= \frac{\left(\check{x}^{(k+1)}\right)^H A \check{x}^{(k+1)} + 2\check{\alpha} e^H A \check{x}^{(k+1)} + \check{\alpha}^2 e^H A e}{\|\check{x}^{(k+1)}\|_B^2 + \check{\alpha}^2} - \lambda_1. \qquad (4.39) \end{aligned}$$

We use again the following decompositions

$$x^{(k)} = \cos\left(\phi^{(k)}\right) v^1 + \sin\left(\phi^{(k)}\right) w^{(k)} =: c_k v^1 + s_k w^{(k)} \qquad (4.40a)$$

$$\check{x}^{(k+1)} = \cos\left(\phi^{(k+1)}\right) v^1 + \sin\left(\phi^{(k+1)}\right) w^{(k+1)} =: c_{k+1} v^1 + s_{k+1} w^{(k+1)} \qquad (4.40b)$$

With (4.40b) we have

$$A\check{x}^{(k+1)} = c_{k+1} A v^1 + s_{k+1} A w^{(k+1)} = c_{k+1} \lambda_1 B v^1 + s_{k+1} A w^{(k+1)}. \qquad (4.41)$$

Furthermore, we can exploit that $e \perp_B E$ and therefore $e \perp_B \check{x}^{(k+1)}$. Hence,

$$0 = e^H B \check{x}^{(k+1)} = e^H B \left(c_{k+1} v^1 + s_{k+1} w^{(k+1)}\right).$$

This yields

$$e^H B v^1 = -\frac{s_{k+1}}{c_{k+1}} e^H B w^{(k+1)}. \qquad (4.42)$$

With (4.41) and (4.42) the expression $e^H A\check{x}^{(k+1)}$ can be simplified to

$$\begin{aligned} e^H A\check{x}^{(k+1)} &= c_{k+1}\lambda_1 e^H Bv^1 + s_{k+1}e^H Aw^{(k+1)} \\ &= -s_{k+1}\lambda_1 e^H Bw^{(k+1)} + s_{k+1}e^H Aw^{(k+1)} \\ &= s_{k+1}e^H \left(A - \lambda_1 B\right) w^{(k+1)} \\ &\leqslant |s_{k+1}| \underbrace{\left\|e\right\|_2 \left\|w^{(k+1)}\right\|_2 \left(\lambda_n - \lambda_1\right)}_{=:\check{C}_1}. \end{aligned} \tag{4.43}$$

The following proposition completes the estimation for the error $R\left(\tilde{x}^{(k+1)}\right) - R\left(v^1\right)$.

Proposition 4.6. *Let $|\alpha| < 1$, and $\tilde{x}^{(k+1)}$, $\hat{x}^{(k+1)}$, α, s_{k+1}, A and B as defined above, then*

$$R\left(\tilde{x}^{(k+1)}\right) - \lambda_1 \leqslant R\left(\hat{x}^{(k+1)}\right) - \lambda_1 + 2C_1|s_{k+1}|\alpha + \mathcal{O}(\alpha^2).$$

Proof. We consider (4.35) and (4.39) to determine

$$R\left(\tilde{x}^{(k+1)}\right) - \lambda_1 \leqslant \frac{\left(\check{x}^{(k+1)}\right)^H A\check{x}^{(k+1)} + 2\check{\alpha}e^H A\check{x}^{(k+1)} + \check{\alpha}^2 e^H Ae}{1+\check{\alpha}^2} - \lambda_1 \tag{4.44}$$

With,

$$\check{\alpha} = \frac{\alpha}{\|x^{(k+1)}\|} = \frac{\alpha}{1 + \left\|\hat{t}\right\|_B^2}, \tag{4.45}$$

from (4.38b) and (4.19) the requirement $|\alpha| < 1$ implies $|\check{\alpha}| < 1$.

Hence, we have

$$\begin{aligned} \frac{1}{1+\check{\alpha}^2} &= \frac{1}{1-(-\check{\alpha}^2)} \\ &= \sum_{k=0}^{\infty}\left(-\check{\alpha}^2\right)^k \\ &= 1 + \mathcal{O}(\check{\alpha}^2). \end{aligned}$$

This simplifies (4.44) to

$$\begin{aligned} &R\left(\tilde{x}^{(k+1)}\right) - \lambda_1 \\ &\leqslant \frac{1}{1+\check{\alpha}^2}\left(\left(\check{x}^{(k+1)}\right)^H A\check{x}^{(k+1)} + 2\check{\alpha}e^H A\hat{x}^{(k+1)} + \check{\alpha}^2 e^H Ae\right) - \lambda_1 \\ &= R\left(\check{x}^{(k+1)}\right) + 2\check{\alpha}e^H A\check{x}^{(k+1)} - \lambda_1 + \mathcal{O}(\check{\alpha}^2). \end{aligned} \tag{4.46}$$

Inserting (4.43) into (4.46) completes yields

$$R\left(\tilde{x}^{(k+1)}\right) - \lambda_1 \leqslant R\left(\check{x}^{(k+1)}\right) - \lambda_1 + 2\tilde{C}_1\check{\alpha}|s_{k+1}| + \mathcal{O}(\check{\alpha}^2).$$

Exploiting

$$\begin{array}{rcl} R(\check{x}^{(k+1)}) & = & R(x^{(k+1)}), \\ C_1 & := & \dfrac{\tilde{C}_1}{\|x^{(k+1)}\|_B} \\ & \text{and} & \\ |\check{\alpha}| & \leqslant & |\alpha| \quad (\text{cf. } (4.45)) \end{array}$$

we obtain

$$R\left(\tilde{x}^{(k+1)}\right) - \lambda_1 \leqslant R\left(x^{(k+1)}\right) - \lambda_1 + 2C_1\alpha|s_{k+1}| + \mathcal{O}(\alpha^2).$$

This completes the proof. □

This proposition results in the following theorem:

Theorem 4.7.
We consider the generalized linear eigenvalue problem where $A \in \mathbb{C}^{n\times n}$ is Hermitian and $B \in \mathbb{C}^{n\times n}$ is Hermitian and positive definite. We assume the smallest eigenvalue λ_1 to be simple. Let $x^{(k)}$ be a sufficiently good approximation for a corresponding eigenvector, v^1, to the smallest eigenvalue λ_1 and $\hat{x}^{(k+1)}$ the next iterate of an iterative method with convergence order $\kappa \geqslant 1$.

Then the iterative projection method, where the subspace $E := span\{x^{(k)}, \hat{x}^{(k+1)}\}$ is perturbed by an angle ϑ (cf Figure 4.1) converges linearly for sufficiently small perturbations ϑ, thus,

$$\tilde{\lambda}^{(k+1)} - \lambda_1 \leqslant \hat{\lambda}^{(k+1)} - \lambda_1 + \mathcal{O}(\tan(\vartheta)(\lambda^{(k)} - \lambda_1))$$

where $\hat{\lambda}^{(k+1)} = p(\hat{x}^{(k+1)})$ and $\tilde{\lambda}^{(k+1)}$ is the smallest Ritz value on $\tilde{E}$.

Proof. We use (4.22) and (4.23) to determine

$$\alpha = \tan(\delta)\tan(\vartheta). \tag{4.47}$$

For a sufficiently small angle ϑ we can assume that the requirement $|\alpha| < 1$ for Proposition 4.6 is satisfied. Applying Proposition 4.6 yields

$$\tilde{\lambda}^{(k+1)} - \lambda_1 \leqslant \hat{\lambda}^{(k+1)} - \lambda_1 + 2C_1 s_{k+1}\alpha + \mathcal{O}(\alpha^2).$$

After inserting (4.47) we determine

$$\tilde{\lambda}^{(k+1)} - \lambda_1 \leqslant \hat{\lambda}^{(k+1)} - \lambda_1 + 2C_1 s_{k+1}\tan(\delta)\tan(\vartheta) + \mathcal{O}(\tan^2(\delta)\tan^2(\vartheta)). \tag{4.48}$$

Exploiting the convergence behavior of the iterative method yields

$$\sin(\phi_{k+1}) = \mathcal{O}\left((\sin(\phi_k))^\kappa\right). \tag{4.49}$$

With (4.49) and Lemma 4.5, the behavior of $\tan(\delta)$ can be determined to

$$\tan(\delta) = \mathcal{O}\left(\sin(\phi_k)\right). \tag{4.50}$$

Inserting (4.49) and (4.50) into (4.48) we determine

$$\tilde{\lambda}^{(k+1)} - \lambda_1 \leqslant \hat{\lambda}^{(k+1)} - \lambda_1 + \tan(\vartheta)\mathcal{O}(\sin^2(\phi^{(k)})). \tag{4.51}$$

Lemma (4.34) completes the proof by replacing the last term in (4.51).

□

Theorem 4.7 points out that an iterative projection method on a perturbed subspace converges locally linearly for the extremal eigenvalues. This behavior is similar to the result of Theorem 2.35 and Theorem 2.36. But in this case the local symmetry does not lead to a higher convergence rate. On the other hand, a search space expansion can be determined with less computational cost than the next iterate of Inverse Iteration or Rayleigh functional Iteration.

4.2.3 Nonlinear Eigenvalue Problems

Now we will transform the results we have obtained in Subsection 4.2.2, for nonlinear eigenvalue problems

$$T(\lambda)x = 0.$$

In addition to Assumption 2.7 and Assumption 2.8 from Chapter 2, we make the following assumption

Assumption 4.8. *We assume that*

$$\lambda_1 = \inf_{v \in D(p)} p(v)$$

exists.

We assume that Assumption 2.7, Assumption 2.8 and Assumption 4.8 are satisfied for the whole subsection. Moreover, we set $B = I$ for the whole subsection. Thus, the B-norm $\|\cdot\|_B$ is replaced by the Euclidean norm $\|\cdot\|_2$.

We begin with a general estimation of the error of the Ritz value on the perturbed subspace and complete this section with its interpretation.

Proposition 4.9. *Let $T : \mathbb{R} \supset D \to \mathbb{C}^{n\times n}$ be a family of Hermitian matrices and $p : D(p) \to I \subset \mathbb{R}$ be a Rayleigh functional fulfilling Assumption 2.7 and Assumption 2.8. Let*

$$x^{(k+1)} = \alpha M_k x^{(k)}, \quad \mathbb{C} \ni \alpha \neq 0,\ x^{(k)} \in \mathbb{C}^n,\ M_k \in \mathbb{C}^{n\times n}$$

be an iterative method to determine the first eigenvalue, λ_1, and a corresponding eigenvector v^1 of

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

We consider a perturbation of the search space as described in Figure 4.1. Let $x^{(k)}$ be a sufficiently good approximation for v^1 with $s_k := \sin\left(\measuredangle(x^{(k)}, v^1)\right)$ and $\tilde{x}^{(k+1)}$

the Ritz vector on the perturbed subspace $\tilde{E}$ (cf. Definition 4.4). Then the following error estimation holds,

$$\begin{aligned}\left|p(\tilde{x}^{(k+1)}) - \lambda_1\right| \leqslant |s_{k+1}|^2 + \mathcal{O}\left(|\sin(\delta)|\right)\, |\tan(\vartheta)|\, s_{k+1} \\ + \mathcal{O}\left(|\tan^2(\delta)|\right) \tan^2(\vartheta)\,. \end{aligned} \tag{4.52}$$

Proof. Similar to the linear case (4.20) yields

$$p(\tilde{x}^{(k+1)}) - \lambda_1 \leqslant p(\tilde{y}) - \lambda_1,$$

with $\tilde{y} = x^{(k+1)} + \alpha e$ (cf (4.21)). Exploiting the approximation properties of the Rayleigh functional according to (2.14) and the results from Schwetlick and Schreiber in [63, Cor. 18 and Th. 21] yields

$$\left|p(\tilde{y}) - p(v^1)\right| = \mathcal{O}\left(\left|\tan^2 \measuredangle\left(\tilde{y}, v^1\right)\right|\right).$$

For sufficiently small angles $\measuredangle\left(\tilde{y}, v^1\right)$, the tangent converges like the sine, thus,

$$\left|p(\tilde{y}) - p(v^1)\right| = \mathcal{O}\left(\left|\sin^2 \measuredangle\left(\tilde{y}, v^1\right)\right|\right). \tag{4.53}$$

We consider

$$\left|\measuredangle\left(\tilde{y}, v^1\right)\right| \leqslant \left|\measuredangle\left(\tilde{y}, x^{(k+1)}\right)\right| + \left|\measuredangle\left(x^{(k+1)}, v^1\right)\right|.$$

The sine function is strictly monotonicly increasing around 0, therefore, the following inequality is fulfilled if $\measuredangle\left(\tilde{y}, v^1\right)$, $\measuredangle\left(\tilde{y}, x^{(k+1)}\right)$ and $\measuredangle\left(\tilde{x}^{(k+1)}, v^1\right)$ are sufficiently small in magnitude.

$$\begin{aligned}\sin\left(\left|\measuredangle\left(\tilde{y}, v^1\right)\right|\right) &\leqslant \sin\left(\left|\measuredangle\left(\tilde{y}, x^{(k+1)}\right)\right| + \left|\measuredangle\left(x^{(k+1)}, v^1\right)\right|\right) \\ &\leqslant \sin\left(\left|\measuredangle\left(\tilde{y}, x^{(k+1)}\right)\right|\right) + \sin\left(\left|\measuredangle\left(x^{(k+1)}, v^1\right)\right|\right) \\ &= \sin\left(\left|\measuredangle\left(\tilde{y}, x^{(k+1)}\right)\right|\right) + \sin\left(\phi_{k+1}\right) \\ &= \sin\left(\left|\measuredangle\left(\tilde{y}, x^{(k+1)}\right)\right|\right) + s_{k+1}\end{aligned}$$

According to (4.21), $e \perp x^{(k+1)}$ and $\|e\|_2 = 1$ the sine of the angle between $x^{(k+1)}$ and $\tilde{y}$ can be expressed by

$$\sin\left(\measuredangle\left(\tilde{y}, x^{(k+1)}\right)\right) = \frac{\alpha}{\|x^{(k+1)} + \alpha e\|_2} = \mathcal{O}(\alpha).$$

And for $\left|\measuredangle\left(\tilde{y}, x^{(k+1)}\right)\right| \leqslant \pi$ we obtain

$$\sin\left(\left|\measuredangle\left(\tilde{y}, x^{(k+1)}\right)\right|\right) = \left|\sin\left(\measuredangle\left(\tilde{y}, x^{(k+1)}\right)\right)\right| = \mathcal{O}(|\alpha|).$$

Taking the square of $\left|\sin\left(\measuredangle\left(\tilde{y}, v^1\right)\right)\right|$ we end up in

$$\begin{aligned}\left|\sin^2\left(\measuredangle\left(\tilde{y}, v^1\right)\right)\right| &\leqslant \left(\left|\sin\left(\measuredangle\left(\tilde{y}, x^{(k+1)}\right)\right)\right| + |\sin(\phi_{k+1})|\right)^2 \\ &= \left|\sin\left(\measuredangle\left(\tilde{y}, x^{(k+1)}\right)\right)\right|^2 \\ &\quad + 2\left|\sin\left(\measuredangle\left(\tilde{y}, x^{(k+1)}\right)\right)\right| |s_{k+1}| + |s_{k+1}|^2 \\ &= \mathcal{O}\left(|\alpha|^2\right) + \mathcal{O}\left(|\alpha\ \sin(\phi_{k+1})|\right) + |s_{k+1}|^2 \end{aligned} \tag{4.54}$$

Inserting (4.54) into (4.53) and taking $\alpha = \tan(\vartheta)\tan(\delta)$ from (4.22) completes the proof. □

Lemma 4.5 is used to estimate $\tan(\delta)$.

Furthermore, the angles ϕ_k and ϕ_{k+1} have to be estimated depending on the error in the Rayleigh functional. For Hermitian linear eigenvalue problems in Subsection 4.2.2, this was done using (4.34). The following Lemma will show that a similar relation is also applicable for Hermitian Nonlinear eigenvalue problems.

Lemma 4.10. *Let $T : \mathbb{R} \supset D \to \mathbb{C}^{n\times n}$ be a family of Hermitian matrices fulfilling Assumption 2.7 and Assumption 2.8. Let $p : \mathbb{C}^n \to I \subset \mathbb{R}$ be a Rayleigh functional according to Definition 2.15. Then for $x \in \mathbb{C}^n$ with $\phi := \measuredangle(x, v^1)$ sufficiently small,*

$$\sin(\phi)^2 = \mathcal{O}\left(|p(x) - \lambda_1|\right)$$

holds, where (λ_1, v^1) denotes the smallest eigenvalue and a corresponding eigenvector.

Proof. For $p(x)$ with $\|x\|_2 \neq 0$ the following equation is satisfied:

$$x^H T\left(p(x)\right) x = 0 \tag{4.55}$$

Furthermore, we can use a Taylor expansion for T and obtain

$$T(\lambda) = T(\lambda_1) + T'(\lambda_1)(\lambda - \lambda_1) + \mathcal{O}(|\lambda - \lambda_1|^2) \tag{4.56}$$

As in (2.17) x is decomposed into one component into the direction of v^1 and one component $w \in \{v^1\}^\perp$,

$$x = \cos(\phi)v^1 + \sin(\phi)w = cv^1 + sw \tag{4.57}$$

Inserting (4.56) and $\lambda = p(x)$ into (4.55) yields

$$\begin{aligned} x^H\left(T(\lambda_1) + T'(\lambda_1)(\lambda - \lambda_1)\right)x &= \mathcal{O}(|\lambda - \lambda_1|^2) \\ x^H T(\lambda_1)x + (\lambda - \lambda_1)\, x^H T'(\lambda_1)x &= \mathcal{O}(|\lambda - \lambda_1|^2) \end{aligned} \tag{4.58}$$

With (4.57) we obtain

$$\begin{aligned} x^H T(\lambda_1)x &= \left(cv^1 + sw\right)^H T(\lambda_1)\left(cv^1 + sw\right) \\ &= \left(cv^1 + sw\right)^H \left(c\underbrace{T(\lambda_1)v^1}_{=0} + sT(\lambda_1)w\right) \\ &= s\left(c\underbrace{\left(v^1\right)^H T(\lambda_1)}_{=0} + sw^H T(\lambda_1)\right) w \\ &= s^2 w^H T(\lambda_1)w. \end{aligned} \tag{4.59}$$

Hence, (4.58) can be simplified by inserting (4.59).

$$\begin{aligned} x^H T(\lambda_1)x + (\lambda - \lambda_1)\, x^H T'(\lambda_1)x &= \mathcal{O}(|\lambda - \lambda_1|^2) \\ s^2 w^H T(\lambda_1)w &= -x^H T'(\lambda_1)x(\lambda - \lambda_1) + \mathcal{O}(|\lambda - \lambda_1|^2) \end{aligned} \tag{4.60}$$

For the smallest eigenvalue

$$w^H T(\lambda_1)w \neq 0 \tag{4.61}$$

holds for all $w \perp v^1$. Thus, we end up in

$$s^2 = -\frac{x^H T'(\lambda_1)x}{w^H T(\lambda_1)w}(\lambda - \lambda_1) + \mathcal{O}(|\lambda - \lambda_1|^2).$$

Replacing $\lambda = p(x)$ back and $s = \sin(\phi)$ completes the proof.

□

Remark 4.11. The condition (4.61) is mandatory for proving Lemma 4.10. It is equivalent to

$$p\,(w) \neq \lambda_1.$$

If this condition is satisfied for interior eigenvalues, this relation can be proven for interior eigenvalues as well.

Lemma 4.10 and former considerations yield the following theorem:

Theorem 4.12.
Let $T : \mathbb{R} \supset D \to \mathbb{C}^{n\times n}$ a family of Hermitian matrices. Let $p : D(p) \to I \subset \mathbb{R}$ a Rayleigh functional fulfilling Assumption 2.7 and Assumption 2.8. We consider an iterative method

$$x^{(k+1)} = \alpha M_k x^{(k)}, \quad \mathbb{C} \ni \alpha \neq 0,\ x^{(k)} \in \mathbb{C}^n,\ M_k \in \mathbb{C}^{n\times n}$$

converging to the smallest eigenvalue, λ_1, and a corresponding eigenvector v^1 of

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

Let this method converge at least linearly. We consider a perturbation of the search space as described in Figure 4.1. Let $x^{(k)}$ be a sufficiently good approximation for v^1 with $s_k := \sin\left(\measuredangle(x^{(k)}, v^1)\right)$ and $\tilde{x}^{(k+1)}$ the Ritz vector on the perturbed subspace $\tilde{E}$ (cf. Definition 4.4). Then, the perturbed subspace reduces the convergence to linear convergence, i.e.

$$\left|p(\tilde{x}^{(k+1)}) - \lambda_1\right| = \mathcal{O}\left(\left|p(\hat{x}^{(k+1)}) - \lambda_1\right|\right) + \mathcal{O}(|\tan(\vartheta)|\,|p(x^{(k)}) - \lambda_1|). \tag{4.62}$$

Proof. Proposition 4.9 yields

$$\begin{aligned}\left|p(\tilde{x}^{(k+1)}) - \lambda_1\right| \leqslant{}& |s_{k+1}|^2 + \mathcal{O}\left(|\tan(\vartheta)\ \tan(\delta)|\right)\ s_{k+1}\\ &+ \mathcal{O}\left(|\tan^2(\vartheta)\ \tan^2(\delta)|\right)\end{aligned} \tag{4.63}$$

The tangent of δ can be estimated using Lemma 4.5. Then (4.63) changes to

$$\begin{aligned}\left|p(\tilde{x}^{(k+1)}) - \lambda_1\right| \leqslant{}& |s_{k+1}|^2\\ &+ |s_{k+1}|\,\mathcal{O}\left(|\tan(\vartheta)|\right)\left[\mathcal{O}\left(|s_k|\right) + \mathcal{O}\left(|s_{k+1}|\right) + \mathcal{O}\left(\sqrt{s_k s_{k+1}}\right)\right]\\ &+ \mathcal{O}\left(|\tan^2(\vartheta)|\right)\left[\mathcal{O}(|s_k|) + \mathcal{O}(|s_{k+1}|) + \mathcal{O}\left(\sqrt{s_k s_{k+1}}\right)\right]^2\end{aligned} \tag{4.64}$$

Assuming that at least $s_{k+1} = \mathcal{O}(s_k)$ is fulfilled, (4.64) simplifies to

$$\begin{aligned}\left|p(\tilde{x}^{(k+1)}) - \lambda_1\right| \leqslant{}& |s_{k+1}|^2 + \mathcal{O}\left(|\tan(\vartheta)|\right)\mathcal{O}\left(|s_k|^2\right)\\ &+ \mathcal{O}\left(\tan^2(\vartheta)\right)\mathcal{O}\left(|s_k|^2\right)\\ ={}& |s_{k+1}|^2 + \mathcal{O}\left(|\tan(\vartheta)|\right)\mathcal{O}\left(|s_k|^2\right)\end{aligned} \tag{4.65}$$

Finally Lemma 4.10 is exploited, thus,

$$|s_k|^2 = \mathcal{O}\left(\left|p(x^{(k)}) - \lambda_1\right|\right), \tag{4.66a}$$

and

$$|s_{k+1}|^2 = \mathcal{O}\left(\left|p(x^{(k+1)}) - \lambda_1\right|\right). \tag{4.66b}$$

Inserting (4.66a) and (4.66b) into (4.65) yields

$$\left|p(\tilde{x}^{(k+1)}) - \lambda_1\right| = \mathcal{O}(|p(\hat{x}^{(k+1)}) - \lambda_1|) + \mathcal{O}\left(|\tan(\vartheta)|\right) \mathcal{O}\left(|p(x^{(k)}) - \lambda_1|\right).$$

This completes the proof. □

Theorem 4.12 is proven using the assumption that the underlying iterative method converges at least linearly to the smallest eigenvalue and a corresponding eigenvector. However, it is not exploited if the convergence is faster than linear. The following corollary will discuss the convergence on the perturbed subspace, if the underlying iterative method converges quadratically or cubically.

Corollary 4.13. *Let the assumptions of Theorem 4.12 be satisfied. Then, the following convergence properties hold:*

(i) If the underlying iterative method converges at least locally quadratically to the eigenpair, then

$$\begin{aligned}\left|p(\tilde{x}^{(k+1)}) - \lambda_1\right| = &\ \mathcal{O}(|p(\hat{x}^{(k+1)}) - \lambda_1|) + \mathcal{O}\left(|\tan(\vartheta)|\right)\ o\left(|p(x^{(k)}) - \lambda_1|\right)\\ &+ \mathcal{O}\left(|\tan(\vartheta)|^2\right)\ \mathcal{O}\left(|p(x^{(k)}) - \lambda_1|\right)\end{aligned}$$

holds, if the subspace is perturbed by the angle ϑ.

(ii) If the underlying iterative method converges at least locally cubically to the eigenpair, then

$$\begin{aligned}\left|p(\tilde{x}^{(k+1)}) - \lambda_1\right| = &\ \mathcal{O}(|p(\hat{x}^{(k+1)}) - \lambda_1|) + \mathcal{O}\left(|\tan(\vartheta)|\right) \mathcal{O}\left(|p(x^{(k)}) - \lambda_1|^2\right)\\ &+ \mathcal{O}\left(|\tan^2(\vartheta)|\right) \mathcal{O}\left(|p(x^{(k)}) - \lambda_1|\right)\end{aligned}$$

holds, if the subspace is perturbed by the angle ϑ.

Proof. We have to enhance the proof of Theorem 4.12. Therefore, we consider (4.64). For proving part (i) we now assume that $s_{k+1} = \mathcal{O}\left({s_k}^2\right)$. This yields

$$\begin{aligned}\left|p(\tilde{x}^{(k+1)}) - \lambda_1\right| \leqslant &\ \left|s_{k+1}\right|^2\\ &+ \left|s_{k+1}\right| \mathcal{O}\left(|\tan(\vartheta)|\right) \left(\mathcal{O}\left(|s_k|\right) + \mathcal{O}\left(|s_{k+1}|\right) + \mathcal{O}\left(\sqrt{s_k s_{k+1}}\right)\right)\\ &+ \mathcal{O}\left(|\tan^2(\vartheta)|\right)\ \left(\mathcal{O}(|s_k|) + \mathcal{O}(|s_{k+1}|) + \mathcal{O}\left(\sqrt{s_k s_{k+1}}\right)\right)^2\\ = &\ \left|s_{k+1}\right|^2\\ &+ \left|s_k\right|^2 \mathcal{O}\left(|\tan(\vartheta)|\right) \left(\mathcal{O}\left(|s_k|\right) + \mathcal{O}\left(|s_k|^2\right) + \mathcal{O}\left(\sqrt{|s_k|^3}\right)\right)\\ &+ \mathcal{O}\left(|\tan^2(\vartheta)|\right)\ \left(\mathcal{O}(|s_k|) + \mathcal{O}(|s_k^2|) + \mathcal{O}\left(\sqrt{|s_k|^3}\right)\right)^2\\ = &\ \left|s_{k+1}\right|^2 + \mathcal{O}\left(|\tan(\vartheta)|\right)\ \mathcal{O}\left(|s_k|^3\right)\\ &+ \mathcal{O}\left(\left|\tan^2(\vartheta)\right|\right)\ \mathcal{O}\left(|s_k|^2\right)\end{aligned}$$

Applying Lemma 4.10 gives

$$\begin{aligned}\left|p(\tilde{x}^{(k+1)})-\lambda_1\right| &= \mathcal{O}\left(|p(\hat{x}^{(k+1)})-\lambda_1|\right)\\ &+\mathcal{O}\left(|\tan(\vartheta)|\right)\,\mathcal{O}\left(|s_k|\right)\,\mathcal{O}\left(|p(\hat{x}^{(k)})-\lambda_1|\right)\\ &+\mathcal{O}\left(\left|\tan^2(\vartheta)\right|\right)\,\mathcal{O}\left(|p(\hat{x}^{(k)})-\lambda_1|\right)\end{aligned}\tag{4.68}$$

Since s_k is also decreasing towards zero we simplify

$$\mathcal{O}\left(|s_k|\right)\,\mathcal{O}\left(|p(\hat{x}^{(k)})-\lambda_1|\right)=o\left(|p(\hat{x}^{(k)})-\lambda_1|\right).\tag{4.69}$$

Inserting (4.68) into (4.69) completes the proof of (i).

The second part is proven similarly, but $s_{k+1}=\mathcal{O}\left({s_k}^3\right)$ is inserted into (4.64). Then we end up in

$$\begin{aligned}\left|p(\tilde{x}^{(k+1)})-\lambda_1\right| &\leqslant |s_{k+1}|^2\\ &+|s_k|^3\,\mathcal{O}\left(|\tan(\vartheta)|\right)\left(\mathcal{O}\left(|s_k|\right)+\mathcal{O}\left(|s_k|^3\right)+\mathcal{O}\left(\sqrt{s_k^4}\right)\right)\\ &+\mathcal{O}\left(|\tan^2(\vartheta)|\right)\,\left(\mathcal{O}(|s_k|)+\mathcal{O}(|s_k|^3)+\mathcal{O}\left(\sqrt{s_k^4}\right)\right)^2\\ &=|s_{k+1}|^2+\mathcal{O}\left(|\tan(\vartheta)|\right)\,\mathcal{O}\left(|s_k|^4\right)\\ &+\mathcal{O}\left(\left|\tan^2(\vartheta)\right|\right)\,\mathcal{O}\left(|s_k|^2\right)\end{aligned}\tag{4.70}$$

Applying Lemma 4.10 completes the proof for this second part. □

Theorem 4.12 and Corollary 4.13 show that the perturbation of the search space expansion in general reduces the convergence rate to linear convergence. This appears independently of the iterative method $x^{(k+1)}=\alpha M_k x^{(k)}$ where the subspace expansion has been adapted from.

Corollary 4.14. *Let the requirements of Theorem 4.12 be satisfied, then the following convergence properties are fulfilled:*

$$\begin{aligned}|\sin(\measuredangle(\tilde{x}^{(k+1)},v^1))| = \mathcal{O}(|\sin(\hat{x}^{(k+1)},v^1)|)+\mathcal{O}(\sqrt{|\tan(\vartheta)|})\mathcal{O}(|\sin^2(x^{(k)},v^1)|)\\ +\mathcal{O}(\sqrt{|\tan(\vartheta)|})\mathcal{O}(|\sin(x^{(k)},v^1)|).\end{aligned}$$

Proof. Follow the proof of Theorem 4.12 and consider (4.65)

$$\left|p(\tilde{x}^{(k+1)})-\lambda_1\right|\leqslant|s_{k+1}|^2+\mathcal{O}\left(|\tan(\vartheta)|\right)\mathcal{O}\left(|s_k|^2\right),\tag{4.71}$$

where $s_{k+1}=\sin(\hat{x}^{(k+1)},v^1)$ and $s_k=\sin(x^{(k)},v^1)$.

Applying Lemma 4.10 to the angle $\measuredangle(\tilde{x}^{(k+1)},v^1)$ leads to

$$\sin^2(\measuredangle(\tilde{x}^{(k+1)},v^1))=\mathcal{O}\left(\left|p(\tilde{x}^{(k+1)})-\lambda_1\right|\right).\tag{4.72}$$

The equations (4.71) and (4.72) are combined to

$$\sin^2(\measuredangle(\tilde{x}^{(k+1)},v^1))\leqslant|s_{k+1}|^2+\mathcal{O}\left(|\tan(\vartheta)|\right)\mathcal{O}\left(|s_k|^2\right).$$

Taking the square roots on both sides completes the proof. □

We have shown the linear convergence to the smallest eigenvalue of a nonlinear eigenvalue problem so far. For an error estimation of the largest eigenvalue λ_n we consider

$$p(\tilde{x}^{(k+1)}) = \max_{\substack{z \in \tilde{E} \\ \|z\|_B = 1}} p(z).$$

Hence, each vector $\tilde{y} \in \tilde{E}$ satisfies

$$p(\tilde{y}) \leqslant p(\tilde{x}^{(k+1)}).$$

Multiplying by -1 and adding $\lambda_n = p(v^n)$ on both sides yields

$$p(v^n) - p(\tilde{x}^{(k+1)}) \leqslant p(v^n) - p(\tilde{y}).$$

This is similar to (4.20). Replacing v^1 by v^n and λ_1 by λ_n leads to the same convergence results as for the smallest eigenvalue.

4.3 Error Estimation For Interior Eigenvalues

In the previous section a convergence proof for real extremal eigenvalues is shown that can be characterized by a minmax characterization, i.e. Assumptions 2.7 and 2.8 are satisfied. But in real applications eigenvalues of interest often are close to a predefined value.[10] Therefore, a different approach to estimate the perturbation of the corresponding Ritz vector on a perturbed subspace is presented here.

For linear eigenvalue problems, the Rayleigh-Ritz projection method works well for the extraction of well-separated exterior eigenvalues of Hermitian matrices. To improve the extraction of interior eigenvalues, Morgan presented the approach of "harmonic Ritz values" and "harmonic Ritz vectors" in [48]. Here the extremal Ritz values of the matrix $(A - \tau I)^{-1}$ are investigated where τ is a user defined target for the interior eigenvalue. A refined method is given in [35] and convergence analysis about harmonic Ritz values is presented in [84].

Hochstenbach and Sleijpen generalized the harmonic and refined Rayleigh-Ritz method in [28]. They presented the "linearized harmonic Rayleigh–Ritz method for polynomial eigenproblems" where the linearized harmonic Ritz vectors can finally be extracted by solving a projected generalized linear eigenvalue problem, instead of solving a projected polynomial eigenvalue problem. In this context "linearized" means the linear part of a Taylor series of the nonlinear eigenvalue eigenvalue problem, i.e.

$$T(\lambda^{(k)})u = \theta T'(\lambda^{(k)})u,$$

is solved.

In this section we use the notation defined in Definition 4.4 for the planes E and $\tilde{E}$, the vectors $x^{(k)}$, $\hat{t}$, e, $\tilde{t}$ and the angle ϑ. The convergence progress is usually measured by the angle between the sought eigenvector and the current approximation. If the

[10] For example: In mechanical engineering eigenvalues close to external excitation frequencies are sought.

"distance" between a vector and a subspace is mentioned, we determine the angle between the vector and its projection onto the subspace.

In Section 4.2 we did not have to determine the exact location of the new Ritz vector in $\tilde{E}$. The fact that the sought minimal eigenvalue is the global minimum of the Rayleigh functional allowed us to estimate an upper bound of the error by choosing any vector in the plane $\tilde{E}$. For eigenvectors belonging to interior eigenvalues, we cannot exploit this property anymore, therefore, we will determine a triangle $\mathcal{T}$ in the plane $\tilde{E}$, in which the Ritz vector is located. $\mathcal{T}$ can be described by

$$\operatorname{span}\{x^{(k)}, \hat{t}\} \supset \mathcal{T} := \{u \in \tilde{E} \mid \tau_1 \leqslant |\tan(\measuredangle(x^{(k)}, u))| \leqslant \tau_2\},$$

with $\tau_2 \geqslant \tau_1 \geqslant 0$.

For an illustration see Figure 4.2.

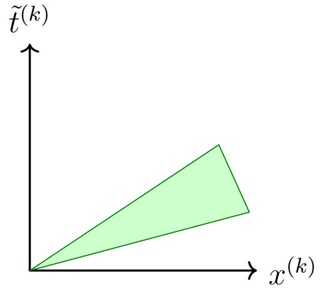

Figure 4.2: Position of $\tilde{x}^{(k+1)}$ in the plane $\tilde{E}$

In [1] and [2] Aishima proves the global convergence for the Jacobi-Davidson method for the real symmetric linear eigenvalue problem. Szyld, Vecharynski and Xue present a convergence proof for a special variant of the nonlinear Jacobi-Davidson method in [69]. We keep our consideration limited to a very general approach where the Jacobi-Davidson method is regarded as a procedure to stabilize a vector iteration in order to determine an eigenpair of a nonlinear eigenvalue problem (cf. Section 3.5).

The assumptions for this section are summarized as follows.

Assumption 4.15. *In this section we restrict our considerations to real*[11] *nonlinear eigenvalue problems.*

Moreover, we assume that all eigenvalues under consideration are real and simple, i.e. if $\hat{\lambda}$ is a real eigenvalue with corresponding eigenvector $\hat{v}$, then $\hat{v}^T T'(\hat{\lambda})\hat{v} \neq 0$.

After a general geometric consideration, we will present convergence proofs for different prerequisites. First, we consider eigenvalue problems where Assumption 2.7 and Assumption 2.8 are satisfied. Finally, the ideas are adapted for problems where the eigenvalues cannot be minmax characterized anymore.

[11] This means, that $T : \mathbb{R} \supseteq J \to \mathbb{R}^{n\times n}$, and we consider only real eigenvalues.

4.3.1 A Geometric Approach

We consider the three dimensional subspace $\mathcal{S} \subset \mathbb{R}^n$, which is spanned by $x^{(k)}$, $x^{(k+1)}$ and the orthogonal perturbation direction[12] e. Moreover, Definition 4.4 yields that the vectors

$$\left\{ x^{(k)}, \frac{\hat{t}}{\left\|\hat{t}\right\|_B}, e \right\}$$

provide an orthonormal basis of $\mathcal{S}$ and that E and $\tilde{E}$ are subspaces of $\mathcal{S}$. To determine the angles between $x^{(k+1)}$ and vectors in $\tilde{E}$ we can apply geometric methods to the coefficient vectors, as used in three dimensional vector calculus. We denote these coefficient vectors by the subscript $\mathcal{S}$ and obtain the following vectors:

$$x_{\mathcal{S}}^{(k)} = (1, 0, 0)^T \tag{4.73a}$$

$$\hat{t}_{\mathcal{S}} = \left(0, \left\|\hat{t}\right\|_B, 0\right)^T \tag{4.73b}$$

$$e_{\mathcal{S}} = (0, 0, 1)^T \tag{4.73c}$$

$$x_{\mathcal{S}}^{(k+1)} = \left(1, \left\|\hat{t}\right\|_B, 0\right)^T \tag{4.73d}$$

$$\tilde{t}_{\mathcal{S}} = (0, \cos(\vartheta), \sin(\vartheta))^T \tag{4.73e}$$

$$\tilde{y}_{\mathcal{S}} = \left(1, \left\|\hat{t}\right\|_B, \alpha\right)^T \tag{4.73f}$$

The vector $\tilde{y}$ is defined in (4.21) as

$$\tilde{y} = x^{(k)} + \hat{t} + \alpha e,$$

and ϑ denotes the perturbation angle between the planes E and $\tilde{E}$ (cf Definition 4.4). For estimating the error for extremal eigenvalues in Section 4.2, the vector $\tilde{y}$ and the inequality

$$|p(\tilde{x}^{(k+1)}) - p(\hat{v})| \leqslant |p(\tilde{y}) - p(\hat{v})|$$

were used. Here, for interior eigenvalues, we cannot expect that $\tilde{y}$ is a Ritz vector, which solves the projected Rayleigh Ritz eigenvalue problem[13] on the perturbed subspace $\tilde{E}$. Hence, we consider all vectors in the affine space,

$$\tilde{E}_M := \left\{ y \in \tilde{E} \subset \mathbb{R}^n \mid \left(x^{(k)}\right)^T By = 1 \right\}, \tag{4.74}$$

as candidates for the sought Ritz vector. Here $B \in \mathbb{R}^{n \times n}$ denotes a positive definite matrix, as defined in (3.10) in Section 3.5[14].

Assuming that the sought Ritz vector is not orthogonal[15] to $x^{(k)}$, the manifold $\tilde{E}_M$ contains a multiple of the sought Ritz vector. Thus, the Ritz vector u can then be described as

$$u = \tilde{y} + \zeta \tilde{t}, \tag{4.75}$$

[12] The orthogonal perturbation vector is defined in Definition 4.4.

[13] This can be linear or nonlinear.

[14] Either B is chosen such that the eigenvectors can be chosen B-orthogonal or, if this is not possible, we set $B = I$.

[15] This property is satisfied if $x^{(k)}$ is sufficiently close an eigenvector.

for a suitable $\zeta \in \mathbb{R}$. We will discuss later how ζ can be determined. Using (4.73e) and (4.73f) we obtain

$$u_{\mathcal{S}} = \tilde{y}_{\mathcal{S}} + \zeta \tilde{t}_{\mathcal{S}} = \begin{pmatrix} 1 \\ \left\|\hat{t}\right\|_B \\ \alpha \end{pmatrix} + \zeta \begin{pmatrix} 0 \\ \cos(\vartheta) \\ \sin(\vartheta) \end{pmatrix} = \begin{pmatrix} 1 \\ \left\|\hat{t}\right\|_B + \zeta\cos(\vartheta) \\ \alpha + \zeta\sin(\vartheta) \end{pmatrix}$$

for the coefficient vector of u.

Proposition 4.16. *We assume that Assumption 4.15 is satisfied. Let $\{x^{(k)}\}_{k=0}^{\infty}$ be iteration vectors converging to an eigenvector $\hat{v}$ of the nonlinear eigenvalue problem*

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

and E and $\tilde{E}$ are defined according to Definition 4.4. If k is sufficiently large, then for a Ritz vector u as defined in (4.75) *it holds*

$$\left|\sin\left(\measuredangle(x^{(k+1)}, u)\right)\right| = \mathcal{O}(|\alpha|) + \mathcal{O}(|\zeta|).$$

Here α denotes the distance from $x^{(k+1)}$ along e until e crosses the plane $\tilde{E}$, cf (4.22) *and Figure 4.1.*

Proof. We use the fact that

$$\measuredangle\left(x^{(k+1)}, u\right) = \measuredangle\left(x_{\mathcal{S}}^{(k+1)}, u_{\mathcal{S}}\right). \tag{4.76}$$

In $\mathbb{R}^3$ we can determine the sine of the angle between two vectors by the following equation:

$$\left|\sin\left(\measuredangle\left(x_{\mathcal{S}}^{(k+1)}, u_{\mathcal{S}}\right)\right)\right| = \frac{\left\|x_{\mathcal{S}}^{(k+1)} \times u_{\mathcal{S}}\right\|_2}{\left\|x_{\mathcal{S}}^{(k+1)}\right\|_2 \left\|u_{\mathcal{S}}\right\|_2} \tag{4.77}$$

We calculate

$$x_{\mathcal{S}}^{(k+1)} \times u_{\mathcal{S}} = \begin{pmatrix} 1 \\ \left\|\hat{t}\right\|_B \\ 0 \end{pmatrix} \times \begin{pmatrix} 1 \\ \left\|\hat{t}\right\|_B + \zeta\cos(\vartheta) \\ \alpha + \zeta\sin(\vartheta) \end{pmatrix} = \begin{pmatrix} \left\|\hat{t}\right\|_B (\alpha + \zeta\sin(\vartheta)) \\ -\alpha - \zeta\sin(\vartheta) \\ \zeta\cos(\vartheta). \end{pmatrix}$$

The euclidean norm can be bounded by the 1-norm such that

$$\begin{aligned} \left\|x_{\mathcal{S}}^{(k+1)} \times u_{\mathcal{S}}\right\|_2 &\leqslant \left\|x_{\mathcal{S}}^{(k+1)} \times u_{\mathcal{S}}\right\|_1 \\ &= \left\|\hat{t}\right\|_B |\alpha + \zeta\sin(\vartheta)| + |-\alpha - \zeta\sin(\vartheta)| + |\zeta\cos(\vartheta)| \\ &= \mathcal{O}(|\alpha|) + \mathcal{O}(|\zeta|). \end{aligned}$$

The B-norm of $\hat{t}$ also converges to zero (cf. (4.23) and Lemma 4.5 in Chapter 4.2).

For the denominator of the right hand side of (4.77), we have

$$\left\|x_{\mathcal{S}}^{(k+1)}\right\|_2 = \sqrt{1 + \left\|\hat{t}\right\|_B^2} \geqslant 1,$$

and

$$\|u_{\mathcal{S}}\|_2 = \sqrt{1 + (\left\|\hat{t}\right\|_B + \zeta \cos(\vartheta))^2 + (\alpha + \zeta \sin(\vartheta))^2} \geqslant 1.$$

Inserting these results into (4.77) yields

$$\left|\sin\left(\measuredangle\left(x_{\mathcal{S}}^{(k+1)}, u_{\mathcal{S}}\right)\right)\right| \leqslant \mathcal{O}(|\alpha|) + \mathcal{O}(|\zeta|). \tag{4.78}$$

By exploiting (4.76) we can convert (4.78) into

$$\left|\sin\left(\measuredangle\left(x^{(k+1)}, u\right)\right)\right| \leqslant \mathcal{O}(|\alpha|) + \mathcal{O}(|\zeta|).$$

This completes the proof. □

Corollary 4.17. *Proposition 4.16 provides two useful properties:*

1. *Using* (4.22)*,* (4.23) *and Lemma 4.5 we obtain*

$$\left|\sin\left(\measuredangle\left(x^{(k+1)}, u\right)\right)\right| \leqslant \mathcal{O}(|\tan(\vartheta)|)\mathcal{O}(|\sin(\phi_k)|) + \mathcal{O}(|\zeta|),$$

while at least $\sin(\phi_{k+1}) = \mathcal{O}(\sin(\phi_k))$ *is assumed.*

2. *Using* $\zeta = 0$ *yields*

$$\left|\sin\left(\measuredangle\left(x^{(k+1)}, \tilde{y}\right)\right)\right| \leqslant \mathcal{O}(|\tan(\vartheta)|)\mathcal{O}(|\sin(\phi_k)|).$$

Corollary 4.17 points out that no higher convergence rate than linear convergence can be expected if the search space is perturbed by a non decreasing (for $k \to \infty$) angle ϑ.

In [69] Szyld, Vecharynski and Xue assume for the perturbed Ritz vector $\tilde{x}^{(k+1)}$ that

$$\sin\left(\measuredangle(\tilde{x}^{(k+1)}, \hat{v})\right) \leqslant C \sin\left(\measuredangle(\tilde{E}, \hat{v})\right) \tag{4.79}$$

is satisfied for a small constant C. With this assumption the results from the convergence proof for the Jacobi-Davidson method without subspace acceleration can be used to prove the convergence of the Jacobi-Davidson projection method based on nonlinear inverse iteration.

We use this assumption in our more general context. This yields the following theorem:

Theorem 4.18.
Let $\{x^{(k)}\}_{k=0}^{\infty}$ be a vector iteration which converges at least linearly to an eigenvector $\hat{v}$ of

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

We assume that Assumption 4.15 is satisfied. Let $x^{(k)}$ be sufficiently close to an eigenvector, i.e. there exists a small $\delta > 0$ with $\left\| \begin{pmatrix} \measuredangle(x^{(k)}, \hat{v}) \\ p(x^{(k)}) - \hat{\lambda} \end{pmatrix} \right\| \leqslant \delta$. Suppose that the Ritz vector u on the perturbed subspace $\tilde{E}$ fulfills

$$\sin\left(\measuredangle(\tilde{x}^{(k+1)}, \hat{v})\right) \leqslant C \sin\left(\measuredangle(\tilde{E}, \hat{v})\right)$$

for a small constant C.

Then Algorithm 4.1 converges locally linearly and

$$\sin\left(\measuredangle(\tilde{x}^{(k+1)}, \hat{v})\right) = \mathcal{O}\left(|\sin\left(\measuredangle(\hat{x}^{(k+1)}, \hat{v})|\right)|\right) + \mathcal{O}(|\tan(\vartheta)|)\mathcal{O}(|\sin(\phi_k)|)$$

is satisfied.

Proof. Since $\tilde{y} \in \tilde{E}$ we conclude

$$\begin{aligned} |\measuredangle(\tilde{E}, \hat{v})| &\leqslant |\measuredangle(\tilde{y}, \hat{v})| \\ &\leqslant |\measuredangle(\tilde{y}, \hat{x}^{(k+1)})| + |\measuredangle(\hat{x}^{(k+1)}, \hat{v})|. \end{aligned} \tag{4.80}$$

For sufficiently small angles, i.e. $|\measuredangle(\tilde{y}, \hat{x}^{(k+1)})| + |\measuredangle(\hat{x}^{(k+1)}, \hat{v})| < \frac{\pi}{2}$, the relation in (4.80) remains valid, if the sine of the angles is taken on both sides. Thus,

$$\sin\left(|\measuredangle(\tilde{E}, \hat{v})|\right) \leqslant \sin\left(|\measuredangle(\tilde{y}, \hat{x}^{(k+1)})| + |\measuredangle(\hat{x}^{(k+1)}, \hat{v})|\right). \tag{4.81}$$

Using the sine's addition theorem, the right hand side of (4.81) can be estimated to

$$\begin{aligned} &\sin\left(|\measuredangle(\tilde{y}, \hat{x}^{(k+1)})| + |\measuredangle(\hat{x}^{(k+1)}, \hat{v})|\right) \\ &\qquad \leqslant \sin\left(|\measuredangle(\tilde{y}, \hat{x}^{(k+1)})|\right) + \sin\left(|\measuredangle(\hat{x}^{(k+1)}, \hat{v})|\right) \end{aligned} \tag{4.82}$$

For an angle ϕ with $|\phi| < \frac{\pi}{2}$, $\sin(|\phi|) = |\sin(\phi)|$ holds. Therefore, we can transform (4.82) to

$$\sin\left(|\measuredangle(\tilde{E}, \hat{v})|\right) \leqslant \left|\sin\left(\measuredangle(\tilde{y}, \hat{x}^{(k+1)})\right)\right| + \left|\sin\left(\measuredangle(\hat{x}^{(k+1)}, \hat{v})|\right)\right|. \tag{4.83}$$

After applying Corollary 4.17 and inserting its result,

$$\left|\sin\left(\measuredangle\left(\hat{x}^{(k+1)}, \tilde{y}\right)\right)\right| = \mathcal{O}(|\tan(\vartheta)|)\mathcal{O}(|\sin(\phi_k)|),$$

into the right hand side of (4.83), we obtain

$$\sin\left(|\measuredangle(\tilde{E}, \hat{v})|\right) \leqslant \left|\sin\left(\measuredangle(\hat{x}^{(k+1)}, \hat{v})|\right)\right| + \mathcal{O}(|\tan(\vartheta)|)\mathcal{O}(|\sin(\phi_k)|). \tag{4.84}$$

Finally, we exploit the assumption for this theorem that

$$\sin\left(\measuredangle(\tilde{x}^{(k+1)}, \hat{v})\right) \leqslant C \sin\left(\measuredangle(\tilde{E}, \hat{v})\right). \tag{4.85}$$

Inserting (4.84) into the right hand side of (4.85), we end up with

$$\begin{aligned}\sin\left(\measuredangle(\tilde{x}^{(k+1)},\hat{v})\right) &= \mathcal{O}\left(|\sin\left(\measuredangle(\hat{x}^{(k+1)},\hat{v})|\right)|\right) + \mathcal{O}(|\tan(\vartheta)|)\mathcal{O}(|\sin(\phi_k)|)\\ &= \mathcal{O}(|\sin(\phi_k)|).\end{aligned}$$

The latter step can be executed since the vector iteration $\left\{x^{(k)}\right\}_{k=0}^{\infty}$ converges at least linearly to an eigenvector.

This completes the proof. □

In the following sections we discuss under which conditions the assumption (4.79) is satisfied.

4.3.2 Error Estimation for Real Eigenproblems with Minmax Characterization

We assume that Assumption 2.7 and Assumption 2.8 are satisfied, and, moreover, that T is at least once continuously differentiable with respect to λ for this entire subsection.

In the previous subsection we have shown that Algorithm 4.1 converges locally linearly, if the assumption

$$\sin\left(\measuredangle(\tilde{x}^{(k+1)},\hat{v})\right) \leqslant C\sin\left(\measuredangle(\tilde{E},\hat{v})\right) \tag{4.86}$$

is satisfied.

We will now show that this assumption is satisfied on perturbed two-dimensional subspaces.

We define y_p as a scaled projection of the desired eigenvector $\hat{v}$[16] onto the subspace $\tilde{E}$. We consider

$$\begin{aligned}P_{\tilde{E}}\hat{v} &= \tilde{V}\tilde{V}^T\hat{v}\\ &= x^{(k)}\left({x^{(k)}}^T\hat{v}\right) + \tilde{t}\;(\overbrace{\tilde{t}^T\hat{v}}^{=:\nu})\\ &= \cos(\phi_k)x^{(k)} + \nu\tilde{t}.\end{aligned}$$

where $\tilde{V} = (x^{(k)}, \tilde{t})$ and the angles are taken from the orthogonal decomposition of $x^{(k)}$ into components along $\hat{v}$ and orthogonal to it in (2.17).

Determining the direction of the Ritz vector of interest on this plane is challenging. However, we will present an approach to determine an area in which a Ritz vector is located. The area's size decreases with the angle of the plane $\tilde{E}$ and the eigenvector $\hat{v}$.

Hence, we define the vector

$$y_p := \frac{1}{\cos(\phi_k)}P_{\tilde{E}}\hat{v} = \frac{1}{\cos(\phi_k)}\tilde{V}\tilde{V}^T\hat{v} = x^{(k)} + \underbrace{\frac{\nu}{\cos(\phi_k)}}_{=:\tilde{\nu}}\tilde{t}. \tag{4.87}$$

[16] with $\|\hat{v}\|_B = 1$

Since we restrict the Rayleigh functional to the manifold $\tilde{E}_M$, as defined in (4.74), we define the following real scalar function as

$$\rho : \begin{cases} \mathbb{R} \supset I & \to \quad J \subset \mathbb{R} \\ \qquad \zeta & \mapsto \quad p(y_p + \zeta\tilde{t}). \end{cases} \tag{4.88}$$

Lemma 4.19. *Let Assumptions 2.7, 2.8 and 4.15 be satisfied and T be at least once continuously differentiable with respect to λ. Let ρ be defined as in* (4.88)*. Let $\tilde{E}$ be sufficiently close to the desired eigenvector $\hat{v}$. Then*

$\rho'(\hat{\zeta}) = 0 \Leftrightarrow$ *$y_p + \hat{\zeta}\tilde{t}$ is a Ritz vector of the projected eigenvalue problem onto $\tilde{E}$.*

Proof. We build the first derivative of ρ with respect to ζ.

$$\rho'(\zeta) = \tilde{t}^T \nabla p(y_p + \zeta\tilde{t}) \tag{4.89}$$

According to Lemma 2.17 the gradient of p is given by

$$\nabla p(x) = -\frac{2}{x^T T'(p(x))x} T(p(x))x.$$

Evaluation of ∇p at $y_p + \zeta\tilde{t}$ yields

$$\nabla p\left(y_p + \zeta\tilde{t}\right) = -\frac{2\,T\left(p\left(y_p + \zeta\tilde{t}\right)\right)\left(y_p + \zeta\tilde{t}\right)}{\left(y_p + \zeta\tilde{t}\right)^T T'\left(p\left(y_p + \zeta\tilde{t}\right)\right)\left(y_p + \zeta\tilde{t}\right)}. \tag{4.90}$$

After inserting (4.90) into (4.89) we obtain

$$\rho'(\zeta) = -\frac{2\,\tilde{t}^T T\left(p\left(y_p + \zeta\tilde{t}\right)\right)\left(y_p + \zeta\tilde{t}\right)}{\left(y_p + \zeta\tilde{t}\right)^T T'\left(p\left(y_p + \zeta\tilde{t}\right)\right)\left(y_p + \zeta\tilde{t}\right)}.$$

According to Assumption 4.15 the denominator is not zero, since the Ritz vector is close to an eigenvector of $T(\cdot)$. We conclude that $\rho'(\hat{\zeta}) = 0$ implies

$$\tilde{t}^T T\left(p\left(y_p + \hat{\zeta}\tilde{t}\right)\right)\left(y_p + \hat{\zeta}\tilde{t}\right) = 0. \tag{4.91}$$

We consider three different cases which are equivalent to the condition (4.91):

(i) $\tilde{t} = 0$: In this case the correction is zero and $x^{(k)}$ has already converged to a nonzero multiple of $\hat{v}$.

(ii) $T\left(p\left(y_p + \hat{\zeta}\tilde{t}\right)\right)\left(y_p + \hat{\zeta}\tilde{t}\right) = 0$: In this case $y_p + \hat{\zeta}\tilde{t} = \hat{v}$, and, therefore, $y_p + \hat{\zeta}\tilde{t}$ is also a Ritz vector.

(iii) $\tilde{t} \perp T\left(p\left(y_p + \hat{\zeta}\tilde{t}\right)\right)\left(y_p + \hat{\zeta}\tilde{t}\right)$.

For the third case, we also consider the standard condition of the Rayleigh functional evaluated at $y_p + \hat{\zeta}\tilde{t}$,

$$\left(y_p + \hat{\zeta}\tilde{t}\right)^T T\left(p\left(y_p + \hat{\zeta}\tilde{t}\right)\right)\left(y_p + \hat{\zeta}\tilde{t}\right) = 0.$$

This and the condition of case (iii) is equivalent to

$$T\left(p\left(y_p + \hat{\zeta}\tilde{t}\right)\right)\left(y_p + \hat{\zeta}\tilde{t}\right) \perp \operatorname{span}\left\{\tilde{t},\, y_p + \hat{\zeta}\tilde{t}\right\}.$$

With the fact that $\operatorname{span}\left\{\tilde{t},\, y_p + \hat{\zeta}\tilde{t}\right\} = \operatorname{span}\left\{y_p,\, \tilde{t}\right\} = \tilde{E}$, we conclude that the Rayleigh Ritz condition

$$\tilde{V}^T T\left(p\left(\tilde{y} + \hat{\zeta}\tilde{t}\right)\right)\tilde{V}z = 0, \qquad \text{with } \tilde{V} = \left(x^{(k)},\, \tilde{t}\right),$$

where z is chosen such that $\tilde{V}z = y_p + \hat{\zeta}\tilde{t}$, is satisfied.

This completes the proof. □

We cannot derive an explicit expression for the $\hat{\zeta}$ solving $\rho'(\hat{\zeta}) = 0$. But we can enclose the stationary point of ρ.

The following Lemma helps to enclose the stationary point of the real function ρ on $\tilde{E}$.

Lemma 4.20. *Let Assumptions 2.7, 2.8 and 4.15 be satisfied. Let $x^{(k)}$ be an sufficiently good approximation for an eigenvector $\hat{v}$ of the eigenvalue $\hat{\lambda}$. Furthermore, assume that $|\tilde{t}^T T(p(\tilde{y}))\tilde{t}| \geqslant C > 0$ is satisfied. Then there exists a stationary point of ρ (defined in* (4.88)*) at $\hat{\zeta}$ with*

$$|\hat{\zeta}| = \mathcal{O}(\|T(p(y_p))y_p\|_2).$$

Proof. For the proof, we first determine the value of η such that $\rho(0) = \rho(\eta)$. We will then use Rolle's theorem to show the existence of a stationary point of ρ between 0 and η.

The definition of ρ in (4.88) yields

$$\rho(0) = p(y_p + 0\tilde{t}).$$

In general, we cannot expect a stationary point of ρ at $\zeta = 0$. We will show that there is a stationary point of ρ in the neighborhood of $\zeta = 0$. Therefore, we are seeking a value $0 \neq \eta \in \mathbb{R}$ fulfilling

$$\rho(0) = \rho(\eta). \tag{4.92}$$

We can transform (4.92) into

$$p(y_p) = p(y_p + \eta\tilde{t}).$$

The unknown parameter η can be determined by solving

$$\left(y_p + \eta\tilde{t}\right)^T T(p(y_p)) \left(y_p + \eta\tilde{t}\right) = 0 \tag{4.93}$$

for η. Expanding (4.93)[17] yields

$$y_p^T T(p(y_p))y_p + 2\eta\tilde{t}^T T(p(y_p))y_p + \eta^2\tilde{t}^T T(p(y_p))\tilde{t} = 0. \tag{4.94}$$

We exploit $y_p^T T(p(y_p))y_p = 0$, since this is the condition for the Rayleigh functional and exclude η from the remaining part of (4.94). This reduces (4.94) to

$$\eta\left(2\tilde{t}^T T(p(y_p))y_p + \eta\tilde{t}^T T(p(y_p))\tilde{t}\right) = 0.$$

Hence, we determine one non trivial solution for η:

$$\eta = -\frac{2\tilde{t}^T T(p(y_p))y_p}{\tilde{t}^T T(p(y_p))\tilde{t}}. \tag{4.95}$$

We consider three different cases:

(i) $\eta > 0$

(ii) $\eta < 0$

(iii) $\eta = 0$

For the unlikely case of $\eta = 0$, we see that (4.91) is fulfilled, and, therefore, p has a stationary point at y_p[18]. For cases (i) and (ii), we apply Rolle's theorem[19] to the function ρ on the interval

$$I_{stat} := \begin{cases} [0,\eta], & \text{if } \eta > 0 \\ [\eta,0], & \text{if } \eta < 0. \end{cases}$$

Thus, there exists a $\hat{\zeta} \in I_{stat}$, satisfying $\rho'(\hat{\zeta}) = 0$. The diameter of I_{stat} is given by $|\eta|$. We conclude

$$|\hat{\zeta}| \leqslant |\eta|. \tag{4.96}$$

For an estimation of $|\eta|$ we consider (4.95). The denominator is bounded and does not tend to zero. The numerator shows the inner product of the perturbed subspace extension $\tilde{t}$ and the residual $T(p(y_p))y_p$. It can be bounded with

$$|\tilde{t}^T T(p(y_p))y_p| \leqslant \left\|\tilde{t}\right\|_2 \|T(p(y_p))y_p\|_2 .$$

After inserting this expression into (4.96), we end up with

$$|\hat{\zeta}| = \mathcal{O}(\|T(p(y_p))y_p\|_2).$$

This completes the proof. □

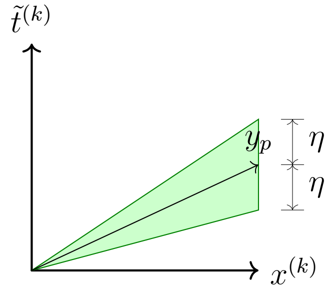

Figure 4.3: Position of $\tilde{x}^{(k+1)}$ in $x^{(k)}$-$\tilde{t}^{(k)}$ plane

Figure 4.3 illustrates the area where the sought Ritz vector can be located.

Lemma 4.20 leads us to the following theorem. This will finally show that the assumption in (4.79) is satisfied for our two dimensional approach:

Theorem 4.21.
Let Assumptions 2.7, 2.8 and 4.15 be satisfied. Let $x^{(k)}$ be a sufficiently good approximation for an eigenvector $\hat{v}$ of the eigenvalue $\hat{\lambda}$. Consider the closest Ritz pair to the eigenpair $(\hat{\lambda}, \hat{v})$ and let $\tilde{x}^{(k+1)} \in \tilde{E}$ be a corresponding Ritz vector.

Then,

$$\sin(\measuredangle(\tilde{x}^{(k+1)}, \hat{v})) \leqslant C \sin(\measuredangle(\tilde{E}, \hat{v})).$$

Proof. Lemma 4.20 provides

$$|\hat{\zeta}| = \mathcal{O}(\|T(p(y_p))y_p\|_2).$$

We now apply Lemma 2.28 and Corollary 2.29 to the residual $T(p(y_p))y_p$:

$$\|T(p(y_p))y_p\|_2 = \mathcal{O}(\sin(\phi_p)), \qquad \text{with} \quad \phi_p := \measuredangle(y_p, \hat{v}).$$

Combining the last two equations leads to

$$|\hat{\zeta}| = \mathcal{O}(\sin(\phi_p)). \tag{4.97}$$

However, the vector y_p does not have unit length, but its length tends to 1 when $x^{(k)}$ converges to $\hat{v}$. Therefore, the behavior of the residual is not affected much by the scaling. Thus, we have shown that the distance between Ritz vector and the projection of y_p onto $\tilde{E}$ tends to zero with $\sin(\phi_p)$. For estimating the sine of the angle $\measuredangle(\tilde{x}^{(k+1)}, \hat{v})$, we split the vector y_p into two orthogonal directions, similarly to (2.17) and (4.57).

[17] The symmetry of T is exploited as well.

[18] We consider here the Rayleigh functional p restricted to the subspace $\tilde{E}$.

[19] Rolle's theorem can be found in every book about one dimensional real analysis. A further requirement of it, that ρ is continuously differentiable, is fulfilled since $T(\cdot)$ is continuously differentiable and, therefore, also p.

$$y_p = \gamma_p \left(\cos(\phi_p)\hat{v} + \sin(\phi_p)w^p\right)$$

with $\gamma_p := \|y_p\|_2$, $w^p \perp \hat{v}$ and $\|w^p\|_2 = 1$. Thus, we can describe the Ritz vector $\tilde{x}^{(k+1)}$ by

$$\begin{aligned}\tilde{x}^{(k+1)} &= y_p + \hat{\zeta}\tilde{t}\\ &= \gamma_p \left(\cos(\phi_p)\hat{v} + \sin(\phi_p)w^p\right) + \hat{\zeta}\tilde{t}\\ &= \left(\gamma_p \cos(\phi_p) + \hat{\zeta}\hat{v}^T\tilde{t}\right)\hat{v} + \gamma_p \sin(\phi_p)w^p + \hat{\zeta}(I - \hat{v}\hat{v}^T)\tilde{t} \qquad (4.98)\end{aligned}$$

In (4.98) we have a decomposition of $\tilde{x}^{(k+1)}$ into the direction of $\hat{v}$ and a direction orthogonal to it. This implies

$$|\tan(\measuredangle(\tilde{x}^{(k+1)}, \hat{v}))| = \frac{\left\|\gamma_p \sin(\phi_p)w^p + \hat{\zeta}(I - \hat{v}\hat{v}^T)\tilde{t}\right\|_2}{|\gamma_p \cos(\phi_p) + \hat{\zeta}\hat{v}^T\tilde{t}|}. \qquad (4.99)$$

Using (4.97), we see that the numerator in (4.99) tends to zero with $\mathcal{O}(|\sin(\phi_p)|)$, since $|\hat{v}^T\tilde{t}| \leqslant 1$. For the denominator we consider[20]

$$\gamma_p = \|y_p\|_2 = \left\|x^{(k)} + \tilde{\nu}\tilde{t}\right\|_2 \geqslant 1$$

and the fact that ϕ_k is close to zero. Thus, $|\gamma_p \cos(\phi_k)| \gg |\hat{\zeta}\hat{v}^T\tilde{t}| \geqslant 0$ which assures that the denominator does not tend to zero.

With $|\sin(\measuredangle(\tilde{x}^{(k+1)}, \hat{v})))| \leqslant |\tan(\measuredangle(\tilde{x}^{(k+1)}, \hat{v}))|$ we end up with

$$\sin(\measuredangle(\tilde{x}^{(k+1)}, \hat{v})) \leqslant C \sin(\measuredangle(y_p, \hat{v})) = C \sin(\measuredangle(\tilde{E}, \hat{v})). \qquad (4.100)$$

This completes the proof. □

Remark 4.22. Combining Theorem 4.21 with Theorem 4.18 yields the linear convergence of Algorithm 4.1. Here linear convergence factor depends linearly on the angle $\vartheta = \measuredangle(E, \tilde{E})$.

4.3.3 General Error Estimation if Minmax Characterization is not Available

If no minmax characterization is given for the eigenvalue problem we cannot exploit the fact that eigenvectors are stationary points of the Rayleigh functional as done so far. We consider iterative methods to solve a real nonlinear eigenvalue problem[21] as described in (3.9). The assumptions for this method are summarized in the following assumption:

Assumption 4.23. *We consider the nonlinear eigenvalue problem*

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

[20] cf. (4.87)

[21] $T(\cdot)$ does not need to be symmetric.

where $T(\cdot)$ is a family of real matrices.

The iterative method

$$x^{(k+1)} = \alpha M_k x^{(k)} \tag{4.101}$$

is used determine an eigenpair $(\hat{\lambda}, \hat{v}) \in \mathbb{R} \times \mathbb{R}^n$ where the matrix M_k describes the mapping from the k-th vector iterate to a nonzero multiple of the $k+1$-th iterate (cf (3.9)*). Furthermore, the method converges at least linearly, e.g.*

$$\left|\frac{\sin(\measuredangle(x^{(k+1)}, \hat{v}))}{\sin(\measuredangle(x^{(k)}, \hat{v}))}\right| \leqslant C < 1.$$

Moreover, the method provides a functional

$$p : \left\{ \begin{array}{ccc} \mathbb{R}^n & \to & \mathbb{R} \\ x^{(k+1)} & \mapsto & \lambda^{(k+1)} = p(x^{(k+1)}) \end{array} \right.$$

satisfying

$$p(\hat{v}) = \hat{\lambda}$$

and

$$|p(x^{(k)}) - \hat{\lambda}| = \mathcal{O}(|\sin(\measuredangle(x^{(k)}, \hat{v}))|).$$

Remark 4.24. There exist several approaches for the functional p in Assumption 4.23. Possible examples for p are:

- **Standard Rayleigh functional:** Solve $\left(x^{(k)}\right)^T T(\lambda^{(k)})x^{(k)} = 0$ for $\lambda^{(k)}$ and set $p(x^{(k)}) = \lambda^{(k)}$.
- **One sided Rayleigh functional:** Solve $l^T T(\lambda^{(k+1)})x^{(k+1)} = 0$ for $\lambda^{(k+1)}$ and set $p(x^{(k+1)}) = \lambda^{(k+1)}$, where $l \in \mathbb{R}^n \setminus \{0\}$ denotes a left scaling vector.
- **Two sided Rayleigh functional:** Solve $\left(u^{(k+1)}\right)^T T(\lambda^{(k+1)})x^{(k+1)} = 0$ for $\lambda^{(k+1)}$ and set $p(x^{(k+1)}) = \lambda^{(k+1)}$, where $u^{(k+1)}$ denotes an approximation for a left eigenvector belonging to the sought eigenvalue.
- **Residual Inverse iteration:** Solve $l^T T(\lambda^{(0)})^{-1} T(\lambda^{(k+1)})x^{(k+1)} = 0$ for $\lambda^{(k+1)}$ and set $p(x^{(k+1)}) = \lambda^{(k+1)}$.[22]

If the eigenvalues can not be characterized as stationary points of the Rayleigh functional, a Ritz vector cannot be determined by the solution of a scalar equation like (4.91). Here the nonlinear system of equations resulting from the projected nonlinear eigenvalue problems has to be considered.

For non-symmetric or non-Hermitian eigenvalue problems, the Petrov-Galerkin approach can be used as an alternative to the Rayleigh-Ritz approach. In addition, the Petrov-Galerkin approach is used for harmonic Ritz values and vectors (cf [35, 28,

[22] cf. Algorithm 2.3.

48, 84]). Here the left and right search space do not need to be identical. Therefore, we consider the following projected problem

$$W^T T(\lambda)\tilde{V}s = 0, \qquad \text{with } \tilde{V} = \left(x^{(k)}, \tilde{t}\right), \quad W = (w^1, w^2) \tag{4.102}$$

with $\operatorname{rank}(W)^{23} = 2$. This general representation covers the special case $W = V$, which describes the Rayleigh-Ritz approach.

To determine a unique solution (λ_*, s_*) of (4.102), a scaling condition is required, as already introduced in Section 2.5. We choose

$$\left(e^1\right)^T s = s_1 \overset{!}{=} 1.$$

This leads to the following nonlinear system:

$$G(z) := \begin{pmatrix} W^T T(\lambda)\tilde{V}s \\ \left(e^1\right)^T s - 1 \end{pmatrix} \overset{!}{=} \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \text{with} \quad z := \begin{pmatrix} s \\ \lambda \end{pmatrix}. \tag{4.103}$$

We consider the length of the component s_2, which denotes the coordinate along the direction of the search space expansion $\tilde{t}$. A possible way to obtain a bound for this length is to consider

$$z^{(0)} := \begin{pmatrix} s^{(0)} \\ \lambda^{(k+1,0)} \end{pmatrix} \tag{4.104}$$

and estimate the error $\left\|z^{(0)} - z_*\right\|$ with the help of the convergence theory of systems of nonlinear equations where

$$s^{(0)} = \begin{pmatrix} 1 \\ \tilde{\nu} \end{pmatrix}, \qquad \tilde{\nu} = \frac{\hat{v}^T \tilde{t}}{\cos(\phi_k)} \qquad \text{and} \qquad \lambda^{(k+1,0)} := p(Vs^{(0)}). \tag{4.105}$$

To distinguish between the inner (to solve this small projected problem) and outer iteration we choose double indices.

The error can be bounded with Banach's fixed-point Theorem ([53]). This is summarized in Lemma 4.25.

Lemma 4.25. *Let z_* be a solution of*

$$G(z) \overset{!}{=} 0$$

and the Jacobian $\frac{\partial G}{\partial z}\big|_{z=z_*}$ be non-singular. Let $\mathcal{U}(z_*)$ be a neighborhood of z_* such that, $\frac{\partial G}{\partial z}$ is non-singular for each $z \in \mathcal{U}(z_*)$ and

$$\Phi(z) := z - \left(\frac{\partial G}{\partial z}\right)^{-1} G(z) \tag{4.106}$$

fulfills the assumptions for Banach's fixed-point Theorem:

1. $\Phi : \mathcal{U}(z_*) \to \mathcal{U}(z_*)$

[23] For practical applications, the numerical rank should be considered.

2. $\|\Phi(z^2) - \Phi(z^1)\| \leqslant q \|z^2 - z^1\|, \quad q < 1, \quad \forall z^1, z^2 \in \mathcal{U}(z_*).$

Furthermore, let the norm of the inverse of the Jacobian of G have an upper bound such that

$$\left\| \left(\frac{\partial G}{\partial z} \right)^{-1} \right\| \leqslant \mu < \infty, \quad \forall z \in \mathcal{U}(z_*).$$

Then the initial error can be bounded using the constant C by

$$\left\| z_* - z^{(0)} \right\| \leqslant C\mu \left\| G(z^{(0)}) \right\|,$$

for every starting value $z^{(0)} \in \mathcal{U}(z_*)$.

Proof. The proof follows immediately from Banach's fixed point theorem. □

If the following assumption is fulfilled Lemma 4.25 can be applied for convergence analysis of the projected problem.

Assumption 4.26. *Let G be at least three times continuously differentiable with respect to z. Moreover, we assume that $z^{(0)}$, as defined in* (4.104)*, fulfills $z^{(0)} \in \mathcal{U}(z_*)$ such that Φ (defined in* (4.106)*) on $\mathcal{U}(z_*)$ fulfills the requirements of Banach's fixed point theorem:*

1. $\Phi : \mathcal{U}(z_*) \to \mathcal{U}(z_*)$

2. $\|\Phi(z^2) - \Phi(z^1)\| \leqslant q \|z^2 - z^1\|, \quad q < 1, \quad \forall z^1, z^2 \in \mathcal{U}(z_*).$

Remark 4.27. Ostrowski's Theorem ([53]) ensures the existence of a neighborhood $\mathcal{U}(z_*)$ such that each $z^{(0)} \in \mathcal{U}(z_*)$ converges to z_*. Since the Jacobian $\Phi'(z_*) = O$, Ostrowski's convergence theorem yields quadratic convergence to z_*.

With this result, we can describe a Ritz vector as the solution of (4.102). Therefore, a Ritz vector's distance to the previous iterate can be estimated. This is shown in Lemma 4.28.

Lemma 4.28. *Let Assumptions 2.14, 4.23 and 4.26 be satisfied and $\left(x^{(k)}, \lambda^{(k)}\right)$ be a sufficiently good approximation to the eigenpair $(\hat{v}, \hat{\lambda})$. Additionally, let $x^{(k)}$ be decomposed to*

$$x^{(k)} = \gamma^{(k)} \left(c_k \hat{v} + s_k g^{(k)} \right),$$

as described in Section 2.4. Then the Ritz vector $\tilde{x}^{(k+1)}$ is a multiple of

$$u := y_p + \hat{\zeta} \tilde{t}$$

where $\tilde{t}$ denotes an inexact solution of the Jacobi-Davidson correction equation, y_p is defined in (4.87) *and the real scalar $\hat{\zeta}$ can be bounded by*

$$|\hat{\zeta}| = \mathcal{O}\left(\|T(p(y_p))y_p\| \right).$$

Proof. Let $\hat{\zeta}$ and $\lambda^{(k+1)}$ be chosen such that $\left(1, \tilde{\nu}+\hat{\zeta}, \lambda^{(k+1)}\right)^T$ solves (4.103). Taking the starting values from (4.105) and applying Lemma 4.25 yields

$$\begin{aligned}
|\hat{\zeta}| &= \left\| \begin{pmatrix} 1 \\ \tilde{\nu} \end{pmatrix} - \begin{pmatrix} 1 \\ \tilde{\nu}+\hat{\zeta} \end{pmatrix} \right\| \\
&\leqslant \left\| \underbrace{\begin{pmatrix} 1 \\ \tilde{\nu} \\ p(Vs^{(0)}) \end{pmatrix}}_{=z^{(0)}} - \underbrace{\begin{pmatrix} 1 \\ \tilde{\nu}+\hat{\zeta} \\ \lambda^{(k+1)} \end{pmatrix}}_{=z_*} \right\| \\
&\leqslant C\mu \left\| G\left(z^{(0)}\right) \right\|.
\end{aligned}$$

For $G\left(z^{(0)}\right)$ we obtain

$$G\left(z^{(0)}\right) = \begin{pmatrix} W^T T(p(Vs^{(0)}))(x^{(k)} + \tilde{\nu}\tilde{t}) \\ 0 \end{pmatrix}. \tag{4.107}$$

When using (4.87):

$$x^{(k)} + \tilde{\nu}\tilde{t} = Vs^{(0)} = y_p$$

we are able to simplify (4.107) to

$$G(z^{(0)}) = \begin{pmatrix} W^T T(p(y_p))y_p \\ 0 \end{pmatrix}.$$

The norm of the residual $G(z^{(0)})$ can be bounded by

$$\left\| G\left(z^{(0)}\right) \right\| \leqslant \left\| T(p(y_p))y_p \right\|.$$

This completes the proof. □

Lemma 4.28 gives the same information on the Ritz vector's direction as Lemma 4.20 does for problems with minmax characterization. If, additionally, Assumption 4.15 is satisfied, the convergence for this case can be proven in a similar way to Theorem 4.21. This leads to Theorem 4.29.

Theorem 4.29.
Consider the nonlinear eigenvalue problem

$$T(\lambda)v \overset{!}{=} 0, \qquad \textit{with} \quad T : \mathbb{R} \supseteq J \to \mathbb{R}^{n \times n}.$$

Let Assumptions 2.14, 4.15 and 4.23 be satisfied. Let the pair $(x^{(k)}, p(x^{(k)}))$ be a sufficiently good approximation of the real eigenpair $(\hat{v}, \hat{\lambda})$. Apply one step of Algorithm 4.1, including solving the projected Petrov-Galerkin problem

$$W^T T(\lambda)\tilde{V}s = 0$$

for s_* and λ_* and set $\tilde{x}^{(k+1)} = \tilde{V}s_*$.

Then the sine of $\measuredangle(\tilde{x}^{(k+1)}, \hat{v})$[24] is bounded by

$$|\sin(\measuredangle(\tilde{x}^{(k+1)}, \hat{v}))| \leqslant C \sin(\tilde{E}, \hat{v}) \tag{4.108}$$

Proof. Follow the proof of Theorem 4.21, but use Lemma 4.28 instead of Lemma 4.20. □

4.4 Numerical Examples

In this section we consider different examples where the Jacobi-Davidson method is applied.

Example 4.30. The example is taken from the Harwell-Boeing Collection, available on the matrix market ([47]). It consists of a real symmetric sparse matrix $A \in \mathbb{R}^{n \times n}$ with $n = 4410$ and the sparsity pattern of A can be seen in Figure 4.4.

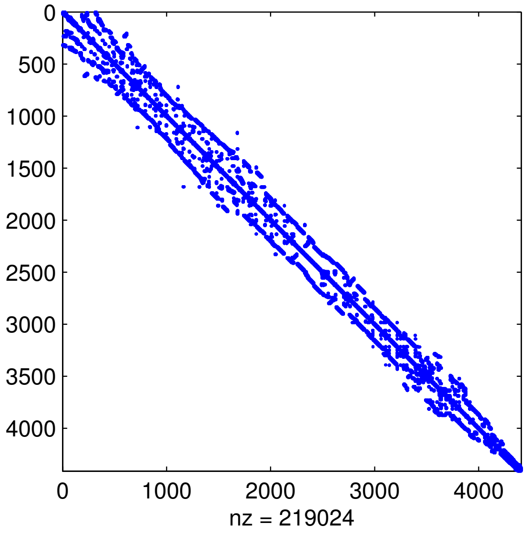

Figure 4.4: Sparsity Pattern of the System Matrix A

We are seeking the smallest eigenvalue $\lambda_1 = 0.8142\ldots$. The initial guess is chosen by perturbing the computed eigenvector randomly, such that $\sin(\phi_0) = 0.08$, where ϕ_k denotes the angle between the current iterate $x^{(k)}$ and the eigenvector v^1.

The problem is solved with two different families of numerical methods: the Rayleigh quotient iteration (Algorithm 2.8) and the Jacobi-Davidson method.

[24] the angle between the Ritz vector and the direction of the sought eigenvector.

k	$\sin(\phi_k)$	$\|\lambda^{(k)} - \lambda_1\|$	$\left\|r^{(k)}\right\|$	$\frac{\left\|res^{(k)}\right\|}{\left\|r^{(k)}\right\|}$	**inner iterations**
0	0.080	1.2e4	7.7e5	0.0897	1
1	0.033	225.5	6.9e4	0.0947	2
2	0.023	24.17	6.5e3	0.0627	8
3	0.014	1.076	408.8	0.0937	18
4	0.012	0.074	38.32	0.0999	70
5	0.011	0.008	3.826	0.0992	277
6	0.005	3.7e-4	0.380	0.0996	538
7	0.001	9.7e-6	0.038	0.0998	487
8	1.9e-4	8.7e-8	0.004	0.0997	549
9	2.7e-5	1.5e-9	3.8e-4	0.0995	659
10	1.1e-6	3.0e-11	3.7e-5	0.0987	641
11	2.9e-7	3.7e-11	3.7e-6	0.0998	670
12	1.1e-6	3.0e-11	3.7e-5	0.0987	641
13	1.0e-8	3.3e-11	3.7e-7	conv.	conv

Table 4.1: Results of the Inexact Jacobi-Davidson method without subspace acceleration with fixed τ

For the inexact Rayleigh quotient iteration, we choose the threshold $\tau^{(k)}$ to stop the inner iterations in two different ways:

1. $\tau^{(k)} = 0.1$, for $k = 0, 1, \ldots$,
2. $\tau_0 = 0.25, \qquad \tau_k = \frac{\left\|r^{(k)}\right\|_2}{\left\|r^{(k-1)}\right\|_2}\tau_{k-1}, \quad$ for $k = 1, 2, \ldots$

We solve the linear system in every outer iteration step using the MinRes method until the residual $\left\|(A - \lambda^{(k)}I)\tilde{y}^{(k+1)} - x^{(k)}\right\|_2$ is less than the threshold τ_k.

For the Jacobi-Davidson method, we use the zero vector as initial guess for solving the Jacobi-Davidson correction equation. The matrix A is preconditioned by an incomplete Cholesky decomposition. We converted the preconditioner to a Jacobi-Davidson preconditioner, cf. Subsection 3.6.2.

We apply two different variants of the Jacobi-Davidson method. Firstly, the problem is solved using the inexact Jacobi-Davidson method without subspace acceleration[25]. The Jacobi-Davidson correction equation was always solved with the MinRes method until the relative residual, $\frac{\left\|res^{(k)}\right\|}{\left\|res^{(0)}\right\|}$, is smaller than 0.1. The number of iterations is not bounded[26]. The results are displayed in Table 4.1.

Additionally, we solve the problem using Jacobi-Davidson method, where the problem is projected with a Rayleigh-Ritz projection onto a subspace, then the smallest Ritz value and the corresponding Ritz vector are used as the next iterates.

[25] cf. Section 3.3.

[26] This was only done for demonstration. In general, an upper bound of iterations is defined.

k	$\sin(\phi_k)$	$\lvert\lambda^{(k)} - \lambda_1\rvert$	$\lVert r^{(k)}\rVert$	$\frac{\lVert res^{(k)}\rVert}{\lVert r^{(k)}\rVert}$	**inner iterations**
0	0.080	1.2e4	7.7e5	0.0897	1
1	0.033	224.5	6.5e4	0.0934	2
2	0.022	22.95	1.3e4	0.0854	5
3	0.016	2.140	1.0e3	0.0977	12
4	0.012	0.212	129.1	0.0947	27
5	0.013	0.026	11.05	0.0993	219
6	0.008	0.002	1.062	0.0997	281
7	0.003	6.6e-6	0.097	0.0995	586
8	2.9e-4	2.5e-7	0.009	0.0988	646
9	3.9e-5	4.5e-9	4.5e-4	0.0983	648
10	5.1e-7	5.4e-11	6.9e-5	0.0991	226
11	1.2e-7	5.5e-11	1.0e-5	0.0997	375
12	3.1e-8	5.5e-11	1.1e-6	0.0992	515
13	8.8e-10	5.4e-11	1.0e-7	conv.	conv

Table 4.2: Results of the Inexact Jacobi-Davidson method with fixed τ

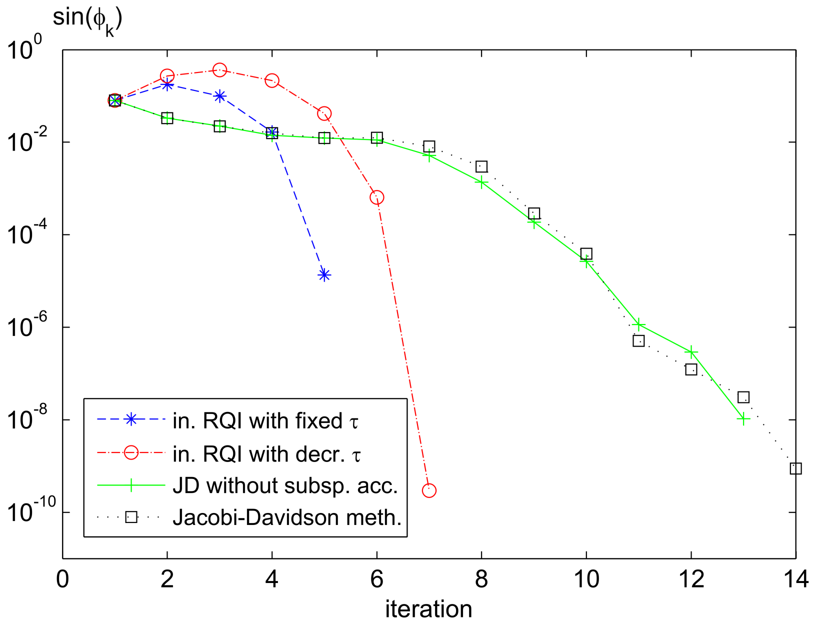

Figure 4.5: Convergence behavior of inexact Rayleigh quotient iteration compared with the Jacobi-Davidson method

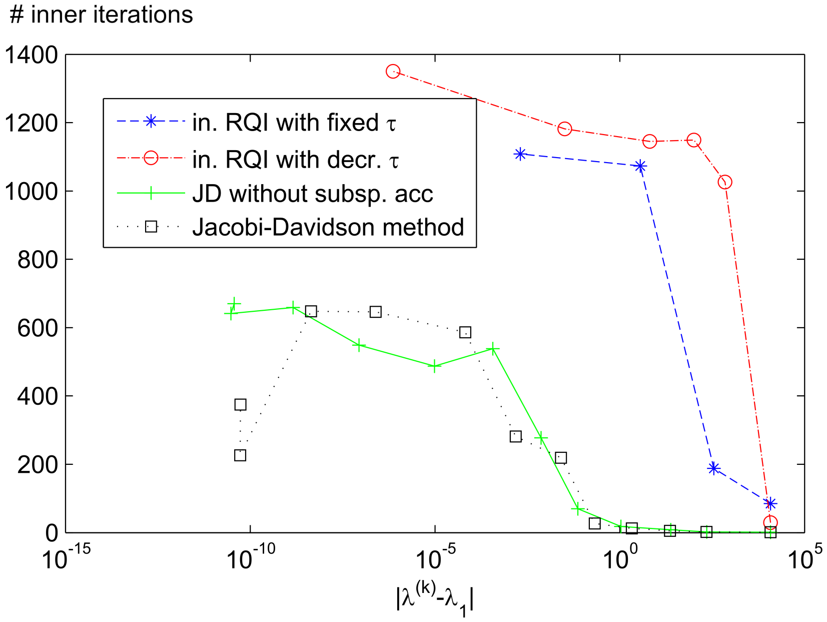

Figure 4.6: The error in the eigenvalues vs. the number of inner iterations for all four cases

The dimension of the subspaces is limited to 5. If this limit is reached, then the subspace is reduced to span$\{u^1, u^2\}$, where u^1 and u^2 are the Ritz vectors belonging to the two smallest Ritz values. The results are shown in Table 4.2.

The convergence behavior of all four computations is visualized in Figure 4.5. For tables with detailed results, see Appendix B.

Considering Figure 4.5, we recognize that the inexact Rayleigh quotient iteration converges after fewer outer iteration steps than the inexact Jacobi-Davidson method. The difference of the convergence velocity between the Jacobi-Davidson method without subspace acceleration and the iterative projection Jacobi-Davidson method is neglectable.

Additionally, the number of inner iterations is illustrated depending on the convergence progress with respect to the eigenvalue in Figure 4.6. We can see that solving the Jacobi-Davidson correction equation requires significantly fewer inner iteration steps than solving the linear system in the inexact Rayleigh quotient iteration.

Moreover, the total number of inner iterations is shown vs. the reached error regarding the eigenvalues in Figure 4.7. We notice that much more inner iteration steps are required in total to solve the problem by inexact Rayleigh quotient iteration, although Rayleigh quotient iteration with a decreasing tolerance τ achieves the fastest convergence regarding the number of outer iteration steps.

We finally conclude, that in terms of total inner iteration steps and, therefore, with respect to of computational costs, both Jacobi-Davidson iterations are better than inexact Rayleigh quotient iteration.

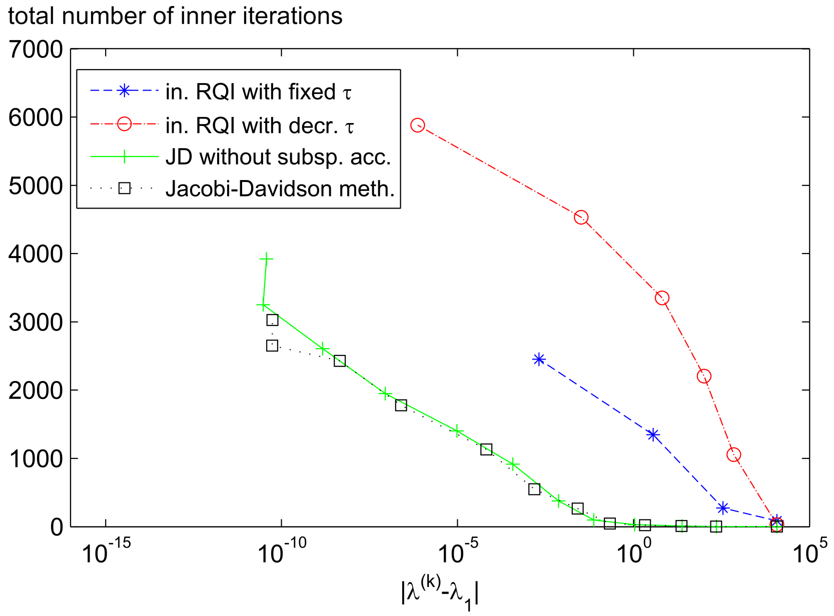

Figure 4.7: The error in the eigenvalues vs. the number of total inner iterations

Example 4.31. In this example, we consider the nonlinear eigenvalue problem resulting from the 3D Quantum-Dot problem again (cf. Example 2.37):

$$T(\lambda)x := \lambda Mx - \frac{1}{m_q(\lambda)}A_q x - \frac{1}{m_m(\lambda)}A_m x - Bx = 0,$$

where

$$\begin{aligned} A_j &= \left(\int_{\Omega_j} \nabla\phi_k \cdot \nabla\phi_l \, dx\right)_{k,l}, \qquad j \in \{q, m\} \\ M &= \left(\int_{\Omega} \phi_k\phi_l \, dx\right)_{k,l}, \\ B &= \left(V_q \int_{\Omega_q} \phi_k\phi_l \, dx + V_m \int_{\Omega_m} \phi_k\phi_l \, dx\right)_{k,l}. \end{aligned}$$

We compute the smallest eigenvalue λ_1 with the inexact Jacobi-Davidson method based on the standard Rayleigh functional iteration, i.e. $C = T'(\lambda^{(k)})$ and $B = I$. Since for $C \neq B$, the Jacobi-Davidson correction equation is not symmetric. The correction equation is solved with the GMRes method in every iteration step. The number of inner iterations is limited to 100. The tolerance for the relative residual is set to $\tau = 0.1$. This means that for the subspace expansion $\tilde{t}$ fulfills

$$\left\| \left(I - \frac{T'(\lambda^{(k)})x^{(k)}\left(x^{(k)}\right)^H}{\left(x^{(k)}\right)^H T'(\lambda^{(k)})x^{(k)}} \right) T(\lambda^{(k)}) \left(I - x^{(k)}\left(x^{(k)}\right)^H \right) + r^{(k)} \right\| \leqslant \tau \left\| r^{(k)} \right\|.$$

The preconditioner is determined with the help of an incomplete LU-decomposition of $T(\sigma)$ where $\sigma = 0$ denotes the initial shift. The size of the subspaces is limited to

5 and, after a restart, the current iterate for the eigenvector and one further vector were resumed for the new subspace. Each projected nonlinear eigenvalue problem is solved by Safeguarded Iteration (cf. Algorithm 2.4).

In Table 4.3 the results of the computation are illustrated.

We see that more than 20 seconds are required until the first iteration step could be performed. This behavior is caused by the fact that the time for the incomplete LU decomposition is included. Furthermore, we see that the number of inner iterations decreases during the Jacobi-Davidson algorithm.

k	$\lvert\lambda^{(k)} - \lambda_1\rvert$	$\left\lVert r^{(k)}\right\rVert$	$\frac{\left\lVert res^{(k)}\right\rVert}{\left\lVert r^{(k)}\right\rVert}$	**inner iterations**	computation time [s]
1	0.648	0.24013	0.094	79	23.610
2	0.556	0.10625	0.094	58	29.554
3	0.483	0.083210	0.095	37	33.443
4	0.440	0.026620	0.097	21	35.631
5	0.404	0.049776	0.085	28	36.773
6	0.358	0.014584	0.074	23	38.337
7	0.356	0.0019264	0.080	18	39.578
8	0.345	0.034228	0.098	16	40.532
9	0.286	0.0074241	0.087	16	41.396
10	0.248	0.042660	0.077	12	42.264
11	0.130	0.015478	0.085	13	42.906
12	0.009	0.011658	0.098	5	43.563
13	$7.90e-5$	$7.33e-4$	0.084	7	43.870
14	$1.28e-6$	$4.12e-5$	0.076	10	44.262
15	$4.19e-9$	$4.58e-6$	0.082	8	44.786
16	$1.36e-11$	$3.63e-7$	0.086	7	45.223
17	$1.27e-13$	$3.48e-8$	0.080	7	45.617

Table 4.3: Results of (classical) Inexact Jacobi-Davidson method to compute the smallest eigenvalue

Secondly, the problem is solved with the same initial guess but with the Jacobi-Davidson procedure based on Derivative-free Rayleigh functional iteration (cf. Algorithm 2.5). Here $C = I$ and since $B = I$, the Jacobi-Davidson correction equation (3.10) is symmetric. Thus, we do not have to use a non-symmetric solver for the linear systems. We perform both possibilities: we solved the correction equation by GMRes and by MinRes as well. Since MinRes requires a symmetric positive definite preconditioner, an incomplete Cholesky factorization of $T(0)$ is computed and then adapted to a Jacobi-Davidson preconditioner. The results are shown in Table 4.4 (with GMRes) and Table 4.5 (with MinRes).

Finally, we compare the computation results from the Jacobi Davidson method and the nonlinear Arnoldi method (cf. [74], page 24). Since the nonlinear Arnoldi method does not have any inner iterations[27], the number of (outer) iterations is larger than for the Jacobi-Davidson method. But each iteration step requires less time. Therefore, the velocity of these methods is compared by considering the error $|\lambda^{(k)} - \lambda_1|$ depending on the computation time in Figure 4.10.

k	$\|\lambda^{(k)} - \lambda_1\|$	$\left\|r^{(k)}\right\|$	$\frac{\left\|res^{(k)}\right\|}{\left\|r^{(k)}\right\|}$	**inner iterations**	computation time [s]
1	0.648	0.240	0.093	17	23.980
2	0.414	0.039	0.095	73	24.879
3	0.348	0.029	0.092	30	30.209
4	0.335	0.029	0.074	27	31.848
5	0.241	0.012	0.093	10	33.301
6	0.067	0.016	0.096	5	33.844
7	$4.25e-4$	0.003	0.090	5	34.152
8	$7.93e-6$	$1.34e-4$	0.087	9	34.462
9	$1.05e-7$	$1.16e-5$	0.081	10	34.958
10	$1.89e-10$	$1.44e-6$	0.082	7	35.502
11	$9.20e-13$	$5.97e-8$	0.084	9	35.906
12	$8.19e-14$	$5.21e-9$	0.084	9	36.372

Table 4.4: Results modified Inexact Jacobi-Davidson method to compute the smallest eigenvalue using GMRes

We used an incomplete LU-decomposition of $T(\sigma)$ with $\sigma = 0$, the same as for the Jacobi-Davidson method.

We see that the nonlinear Arnoldi method converges faster than all variants of the Jacobi-Davidson method, although more iteration steps are required.

But the results also show that the derivative free variant of the Jacobi-Davidson method requires less computation time than the standard variant. The same behavior can be seen if we compare the (outer) iteration steps.

Regarding the different variants of the Jacobi-Davidson method, we see clearly, that using the derivative free variant of Rayleigh functional iteration as base method results here in faster convergence. Exploiting its symmetry of the Jacobi-Davidson correction equation, by using the MinRes method instead of the more expensive GMRes method, accelerates the convergence even further.

[27] Here only a preconditioner is applied instead of solving a linear system.

k	$\lvert\lambda^{(k)} - \lambda_1\rvert$	$\left\lVert r^{(k)}\right\rVert$	$\frac{\lVert res^{(k)}\rVert}{\lVert r^{(k)}\rVert}$	**inner iterations**	computation time [s]
1	0.648	0.240	0.087	9	24.576
2	0.367	0.037	0.093	8	25.198
3	0.150	0.031	0.087	5	25.763
4	0.010	0.004	0.079	8	26.135
5	$1.09e-4$	$3.09e-4$	0.075	9	26.704
6	$5.20e-7$	$2.81e-5$	0.099	9	27.324
7	$8.49e-9$	$4.01e-6$	0.098	9	27.951
8	$7.62e-11$	$3.47e-7$	0.092	7	28.597
9	$5.75e-13$	$3.05e-8$	0.086	8	29.115
10	$6.76e-14$	$2.49e-9$	0.080	10	29.700

Table 4.5: Results modified Inexact Jacobi-Davidson method to compute the smallest eigenvalue using MinRes

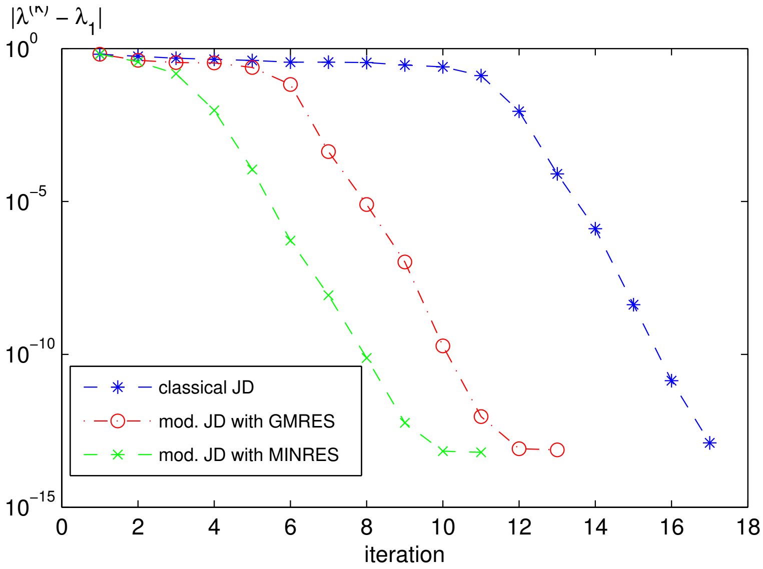

Figure 4.8: The error in the eigenvalues depending on the iteration step

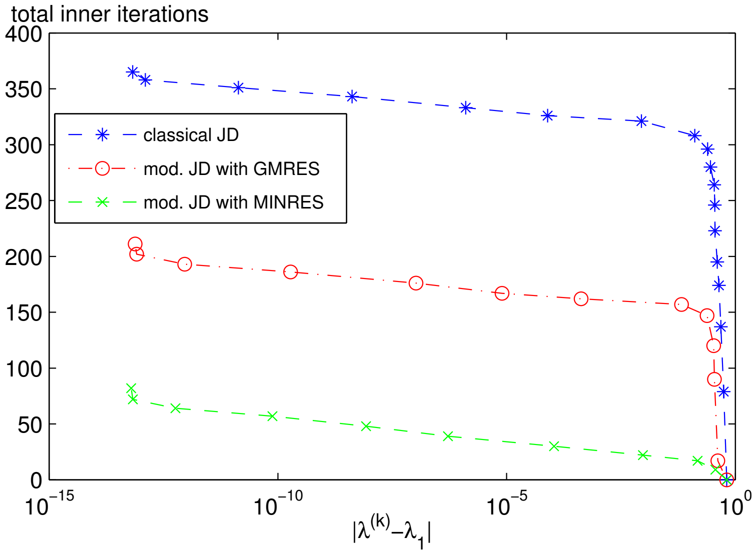

Figure 4.9: The total number of inner iterations depending on the error in the eigenvalues

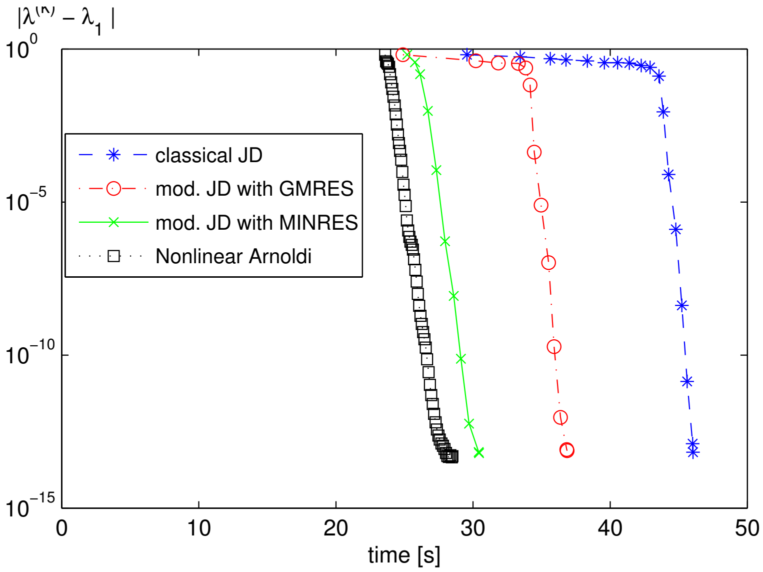

Figure 4.10: The error in the eigenvalues depending on the computation time

Chapter 5

Two-Parameter Eigenvalue Problems

In this chapter, we consider a special type of nonlinear eigenproblems:

$$T(\omega,\tau) := i\omega M + A + e^{-i\omega\tau}B \overset{!}{=} 0, \quad M, A, B \in \mathbb{C}^{n\times n},\ \omega, \tau \in \mathbb{R},\ \tau > 0.$$

Here the family of matrices depends on two real parameters instead of one parameter[1]. After a short introduction into the class of problems and numerical methods to solve these problems, we discuss different approaches for functionals having similar properties as the Rayleigh functional[2].

This problem is discussed in detail in [22] where its solutions lead to Hopf bifurcations in systems of linear differential equations. Meerbergen, Schröder and Voss discuss in [45] a variant of the Jacobi-Davidson method for a special class of these problems. Using this method they obtained fast convergence which they assumed to be quadratic. However we will show that it is only linear.

5.1 The Problem Description

Linear dynamic systems with a time delay τ can be described by the following system of *delay-differential equations*:

$$M\dot{x}(t) + Ax(t) + Bx(t-\tau) = 0,$$

where $M, A, B \in \mathbb{C}^{n\times n}$ are given matrices describing the system and $\mathbb{R} \ni \tau \geqslant 0$ is a given time delay. Using the standard ansatz function for systems of linear differential equations, $x(t) = \exp(\lambda t)\, u$, yields the following nonlinear eigenvalue problem

$$\lambda M u + A u + e^{-\lambda\tau} B u = 0.$$

For such nonlinear eigenvalue problems, each eigenpair (λ_i, u^i) depends on the delay parameter τ. We are interested in those values of τ for which an eigenvalue crosses the imaginary axis, since then the system's stability can change from stable into unstable or vice versa (see for example [24]). Therefore, we set $\lambda = i\omega$, $\omega \in \mathbb{R}$ and consider

$$i\omega M u + A u + e^{-i\omega\tau} B u = 0. \tag{5.1}$$

[1] which might be complex.

[2] cf. Section 2.3

We denote this two parameter eigenvalue problem by

$$T(\omega,\tau)u \overset{!}{=} 0, \tag{5.2}$$

with

$$T(\omega,\tau) := i\omega M + A + e^{-i\omega\tau}B. \tag{5.3}$$

The problem (5.2) is one example of a *two (real) parameter eigenvalue problem.* Here, a pair of parameters $(\hat{\omega},\hat{\tau})$ is sought such that $T(\hat{\omega},\hat{\tau})u = 0$ has a nontrivial complex solution $\hat{u}$; whereas $T : \mathbb{R}\times\mathbb{R} \to \mathbb{C}^{n\times n}$ maps a pair of real parameters to a family of complex matrices. Further details can be found in [22]. Moreover, applications in other disciplines (e.g. hydrodynamics [11] and chemical engineering [26]) exist where the second parameter, τ, describes other physical parameters.

We will keep our consideration on the concrete problem in (5.3).

Since we only consider simple eigentriples the simplicity of eigentriples for this kind of eigenvalue problems is defined in the following definition.

Definition 5.1. A pair of eigenparameters $(\hat{\omega},\hat{\tau})$ is called simple if

$$\operatorname{rank}(T(\hat{\omega},\hat{\tau})) = n-1,$$

and $U^H K$ has full row rank (which is two) with

$$K = \begin{pmatrix} \hat{T}_\omega\hat{u} & \hat{T}_\tau\hat{u} \\ \overline{\hat{T}_\omega\hat{u}} & \overline{\hat{T}_\tau\hat{u}} \end{pmatrix}, \quad U = \begin{pmatrix} \hat{u} & o \\ o & \overline{\hat{u}} \end{pmatrix}$$

where $\hat{T}_\omega$ and $\hat{T}_\tau$ denote the partial derivatives of T, with respect to ω and τ and are evaluated at the eigenparameters $\hat{\omega}$ and $\hat{\tau}$.

5.2 Numerical Methods for Delay Eigenvalue Problems

We consider two different numerical methods to solve (5.2). Firstly, an approach is presented where this problem is transformed into a quadratic eigenvalue problem. Additionally, an approach for large problems is introduced where this problem can be solved by an adapted Jacobi-Davidson method.

5.2.1 Transforming the Nonlinear Delay Eigenproblem into a Quadratic Eigenproblem

The idea to transform the delay eigenvalue problem (5.2) into a polynomial eigenvalue problem is discussed in [9, 13, 32, 33, 42].

Equation (5.1) is combined with its conjugate complex counterpart,

$$-i\omega\overline{Mu} + \overline{Au} + e^{i\omega\tau}\overline{Bu} = 0. \tag{5.4}$$

Additionally,

$$\mu := e^{-i\omega\tau} \tag{5.5}$$

is introduced. Since μ is located on the unit circle in the complex plane,

$$\overline{\mu} = \frac{1}{\mu} \tag{5.6}$$

holds. Using (5.2), (5.4), (5.5) and (5.6) yields

$$\begin{aligned} 0 &= -(i\omega Mu) \otimes \overline{Mu} - Mu \otimes (-i\omega \overline{Mu}) \\ &= (Au + \mu Bu) \otimes \overline{Mu} + Mu \otimes \left(\overline{Au} + \frac{1}{\mu}\overline{Bu} \right) \\ &= \underbrace{\left((A + \mu B) \otimes \overline{M} + M \otimes \left(\overline{A} + \frac{1}{\mu}\overline{B} \right) \right)}_{=:R(\mu)} (u \otimes \overline{u}), \end{aligned} \tag{5.7}$$

where $\otimes$ denotes the Kronecker product. For further details see [29]. Multiplying (5.7) by μ yields the following quadratic eigenvalue problem:

$$P(\mu)z := \mu^2(B \otimes \overline{M})z + \mu(A \otimes \overline{M} + M \otimes \overline{A})z + (M \otimes \overline{B})z = 0, \quad z = u \otimes \overline{u}. \tag{5.8}$$

To determine eigentriples[3] $(\hat{\omega}, \hat{\tau}, \hat{u})$ of (5.1), eigenpairs $(\hat{\mu}, \hat{z})$ of (5.8) have to be determined. If the matrices B and M are regular, then $B \otimes \overline{M}$ is also regular, and, hence, (5.8) has $2n^2$ finite eigenvalues. We are only interested in those satisfying

$$|\hat{\mu}| = 1.$$

After μ and z have been determined, ω and τ can be calculated directly by

$$\omega = -\frac{\operatorname{Im}\left((Mu)^H (Au + \mu Bu) \right)}{\|Mu\|_2^2}, \tag{5.9}$$

$$\tau = -\frac{\operatorname{Im}(\ln(\mu))}{\omega}. \tag{5.10}$$

To distinguish between the eigenvectors of (5.1) and those from (5.8), we denote the eigenvectors of (5.1) by $u \in \mathbb{C}^n$. The eigenvectors of (5.8) and (5.12) are denoted by $z \in \mathbb{C}^{n^2}$.

Fassbender, Mackey, Mackey and Schröder present in [13] a structure preserving eigenvalue solver for (5.8). Therefore, they transform the eigenvalues by a Möbius transformation from μ to θ and the other way round.

$$\theta = i\frac{\mu - 1}{\mu + 1} \tag{5.11a}$$

$$\Leftrightarrow$$

$$\mu = \frac{i + \theta}{i - \theta} \tag{5.11b}$$

[3] Since we are seeking the two real parameters ω and τ and a complex vector u, we call it eigentriples instead of eigenpairs.

We see that the unit circle is mapped to the real axis, thus, we can consider the transformed quadratic eigenvalue problem

$$Q(\theta)z := \theta^2 \left(A_2 - A_1 + A_0\right) z + \theta \left(2iA_2 - 2iA_0\right) z + \left(-A_2 - A_1 - A_0\right) z \overset{!}{=} 0, \quad (5.12)$$

with

$$\begin{aligned} \mathbb{C}^{n^2\times n^2} &\ni A_0 = M \otimes \overline{B}, \\ \mathbb{C}^{n^2\times n^2} &\ni A_1 = A \otimes \overline{M} + M \otimes \overline{A}, \\ \mathbb{C}^{n^2\times n^2} &\ni A_2 = B \otimes \overline{M}, \end{aligned}$$

and seek for real eigenvalues θ.

Fassbender et al. exploit the property of the Kronecker product, that there exists a symmetric permutation matrix P satisfying

$$A \otimes B = P(B \otimes A)P.$$

Finally, they end up in the following algorithm to solve (5.1).

Algorithm 5.1: Algorithm to compute the eigentriples of a delay-eigenproblem

```
input  : the matrices $M, A, B \in \mathbb{C}^{n\times n}$
output: approximations for all solution triples $(\omega_k, \tau_k, u^k)_{k=1,\dots}$ of (5.1)
1  $A_0 = M \otimes \overline{B}$, $A_1 = A \otimes \overline{M} + M \otimes \overline{A}$;
2  Construct a permutation matrix $P$;
3  $C = \frac{1}{2}\begin{pmatrix} I & iP \\ P & -iI \end{pmatrix}^H \begin{pmatrix} A_1 - P\overline{A}_0 P & A_0 \\ A_0 & A_0 \end{pmatrix} \overline{\begin{pmatrix} I & iP \\ P & -iI \end{pmatrix}}$;
4  Compute all eigenpairs $(\theta_j, x^j)$ of $\mathrm{Re}(C)x = \theta\mathrm{Im}(C)x$ where $\theta_j$ is real;
5  for $j = 1, \dots$ do
6      $\mu_j = \frac{i+\theta_j}{i-\theta_j}$;
7      $z^j = \begin{pmatrix} I & -iP \end{pmatrix} x^j$;
8      Compute $u^j$ as the dominant singular vector of $\overline{\mathrm{mat}(z^j)}$;
9      $\omega_j = -\frac{\mathrm{Im}\left((Mu^j)^H (Au^j + \mu Bu^j)\right)}{\|Mu^j\|_2^2}$;
10     $\tau_j = -\frac{\mathrm{Im}(\ln(\mu_j))}{\omega_j}$;
11 end for
```

- The permutation matrix can be constructed in MATLAB by using the following commands:

```
p = reshape(reshape(1:n^2,n,n)',1,n^2);
I = eye(n^2);
P = I(:,p);
```

- A possible solver for the linear eigenvalue problem in step 4 is the QZ-Algorithm (see for instance [20]).

- In step 4 a generalized linear eigenvalue problem has to be solved, where the matrices have the size $2n^2 \times 2n^2$. Therefore, this method might become very time consuming for larger n.

- For steps 1 to 4 this algorithm has a computational complexity of $\mathcal{O}(n^6)$.

An explicit representation of P can be found in [29]:

$$P = \sum_{j=1}^{n} e^j \otimes I_n \otimes \left(e^j\right)^T = \sum_{j=1}^{n} \left(e^j\right)^T \otimes I_n \otimes e^j. \tag{5.13}$$

We recognize that this permutation depends only on the size of the matrices, n, but not on the entries of the matrices A, B and M.

Proposition 5.2. *Let $a, b \in \mathbb{C}^n$ and P the permutation matrix as defined in* (5.13)*. Then*

$$P(a \otimes b) = b \otimes a.$$

Proof. We consider

$$\begin{aligned}
P(a \otimes b) &= \sum_{j=1}^{n} \left(e^j \otimes I_n \otimes \left(e^j\right)^T\right)(a \otimes b) \\
&= \sum_{j=1}^{n} \left(\left(e^j \otimes I_n\right) \otimes \left(e^j\right)^T\right)(a \otimes b) \\
&= \sum_{j=1}^{n} \left(e^j \otimes I_n\right) a \otimes \underbrace{\left(e^j\right)^T b}_{=b_j} \\
&= \sum_{j=1}^{n} b_j \left(e^j \otimes a\right) \\
&= \begin{pmatrix} b_1 a \\ b_2 a \\ \vdots \\ b_n a \end{pmatrix} \\
&= b \otimes a.
\end{aligned}$$

□

5.2.2 The Jacobi-Davidson Method for Delay Eigenproblems

We have seen that the costs of Algorithm 5.1 increase very fast. Solving, for instance, a two parameter nonlinear eigenvalue problem $n = 50$, results in solving a linear eigenvalue problem of size 5000. Therefore, more efficient methods are needed for large n. Meerbergen, Schröder and Voss presented a variant of the Jacobi-Davidson method (cf Chapter 3) in [45]. They constructed the following algorithm:

Algorithm 5.2: Jacobi-Davidson method for two parameter eigenvalue problems

```
   input  : $T : \mathbb{R} \times \mathbb{R} \to \mathbb{C}^{n \times n}$
   output: approximations for solution triples $(\omega_k, \tau_k, u^k)_{k=1,\dots} \in \mathbb{R} \times \mathbb{R} \times \mathbb{C}^n$ of
           $T(\omega, \tau)u = 0$
1  Construct a suitable basis $V \in \mathbb{C}^{n \times k_0}$, $V^H V = I$ of the initial search space;
2  for $k = k_0, k_0 + 1, \dots$ do
3      Solve the projected eigenvalue problem $V^H T(\omega, \tau) V z = 0$ using
         Algorithm 5.1 ;
4      Extract the eigentriples $(\omega_i, \tau_i, z^i)$ with $\omega_i, \tau_i \in \mathbb{R}$ ;
5      Compute the associated Ritz vectors $u^i = V z^i$ ;
6      Check for convergence $\|T(\omega_i, \tau_i) u^i\| \leqslant \varepsilon$;
7      if $(\omega_i, \tau_i, u^i)$ has converged then
8          if $\overline{u}_i \notin span(V)$ then
9              expand $V$ by $\overline{u}_i$;
10         end if
11         stop;
12     end if
13     Select an approximate eigentriple to continue the JD method;
14     Reduce space, if necessary;
15     Determine expansion direction $t$;
16     Orthogonalize $t$ against the columns of $V$;
17     Expand $V = \left[V, \frac{t}{\|t\|}\right]$;
18 end for
```

Steps 8 to 10 can be skipped if the matrices M, A and B are not real. For real matrices we can exploit the fact, that $(-\omega_i, \tau_i, \overline{u}^i)$ is also an eigentriple after the eigentriple (ω_i, τ_i, u^i) has been determined.

The reduction of the search space in step 14 is recommended if k becomes too large, since the computational complexity of solving the projected problem in step 3 is $\mathcal{O}(k^6)$, cf Algorithm 5.1.

The search space expansion is explained in detail in [45]. Based on Newton's method, an adapted Jacobi-Davidson correction equation is presented:

$$\left(I - K(U^H K)^{-1} U^H\right) \mathbb{T} \left(I - U U^H\right) \begin{pmatrix} t \\ \overline{t} \end{pmatrix} = -\left(I - K(U^H K)^{-1} U^H\right) \begin{pmatrix} r \\ \overline{r} \end{pmatrix}$$

with

$$
\begin{aligned}
\mathbb{T} &:= \begin{pmatrix} T^{(i)} & O \\ O & \overline{T^{(i)}} \end{pmatrix} \\
K &:= \begin{pmatrix} T^{(i)}_{\omega} u^i & T^{(i)}_{\tau} u^i \\ \overline{T^{(i)}_{\omega} u^i} & \overline{T^{(i)}_{\tau} u^i} \end{pmatrix} \\
U &:= \begin{pmatrix} u^i & o \\ o & \overline{u^i} \end{pmatrix} \\
T^{(i)} &:= T(\omega_i, \tau_i), \qquad T^{(i)}_{\omega} := T_{\omega}(\omega_i, \tau_i) \qquad T^{(i)}_{\tau} := T_{\tau}(\omega_i, \tau_i)
\end{aligned}
$$

Moreover, there are preconditioners presented and possibilities to implement an efficient matrix-vector multiplication in order to solve the correction equation iteratively.

Very often, no Ritz pair can be found in the initial search space.[4] For this problem in the first steps, Meerbergen, Schröder and Voss [45] present alternative search space expansions. They suggest to solve the following minimization problem

$$\min_{\substack{z \in \mathbb{C}^k \\ \|z\|=1}} \|(A + \sigma B) V z\| \,.$$

Then $u = Vz$ is used instead of a Ritz vector and the search space is expanded such that $(A+\sigma B)^{-1} M u$ is included in the new search space. Therefore, the Jacobi-Davidson correction equation can be adapted as we did in Chapter 3.

5.3 Rayleigh Functional for Delay Eigenvalue Problems

In this section the question is discussed if projecting the problem onto a one-dimensional subspace and solving the resulting nonlinear scalar equation for ω and τ defines a functional, which is comparable to the Rayleigh functional described in Section 2.3. We will show that the conditions given in Definition 2.15 are fulfilled. Furthermore, this functional is analyzed for its stationarity at eigentriples.

This question is motivated by the numerical results from Meerbergen, Schröder and Voss in [45]. They solved problems with large matrices by projecting them using Rayleigh-Ritz projections onto small subspaces. This method worked well and it seemed that the method converged quadratically.

Since we are dealing with two eigenvalue problems, (5.1) and (5.12), whose solvability and solutions are related to each other, we will also have three different functionals. To point out this difference, we denote the functionals related to (5.1) by p_ω and p_τ and the Rayleigh functional for the large quadratic eigenvalue problem (5.12) by Φ_θ.

[4] If, for instance, no information regarding the location of the eigenvalues is known, random vectors are used to span initial search space.

5.3.1 Stationarity of the Polynomial Eigenproblem

We consider the polynomial eigenvalue problem (5.12). We multiply this equation by $z^H = (u \otimes \overline{u})^H$ from the left and obtain

$$(u \otimes \overline{u})^H \Big(\theta^2 (A_2 - A_1 + A_0) + 2i\theta (A_2 - A_0) + (-A_2 - A_1 - A_0)\Big) (u \otimes \overline{u}) = 0.$$

Applying the rules for the Kronecker product yields

$$\theta^2 \left(\underbrace{b\overline{m} - a\overline{m} - m\overline{a} + m\overline{b}}_{=2\mathrm{Re}(b\overline{m}) - 2\mathrm{Re}(a\overline{m})} \right) + 2i\theta \left(\underbrace{b\overline{m} - m\overline{b}}_{=2i\mathrm{Im}(b\overline{m})} \right) + \left(\underbrace{-b\overline{m} - a\overline{m} - m\overline{a} - m\overline{b}}_{-2\mathrm{Re}(b\overline{m}) - 2\mathrm{Re}(a\overline{m})} \right) = 0, \tag{5.14}$$

with

$$\begin{aligned} A_0 &= M \otimes \overline{B}, && (5.15a)\\ A_1 &= A \otimes \overline{M} + M \otimes \overline{A}, && (5.15b)\\ A_2 &= B \otimes \overline{M}, && (5.15c)\\ a &:= u^H A u, && (5.15d)\\ b &:= u^H B u, && (5.15e)\\ m &:= u^H M u. && (5.15f) \end{aligned}$$

For a better readability we define

$$\begin{aligned} \beta_0 &:= -2\mathrm{Re}(b\overline{m}) - 2\mathrm{Re}(a\overline{m}) = z^H(-A_2 - A_1 - A_0)z, && (5.16a)\\ \beta_1 &:= -4\mathrm{Im}(b\overline{m}) = 2iz^H(A_2 - A_0)z, && (5.16b)\\ \beta_2 &:= 2\mathrm{Re}(b\overline{m}) - 2\mathrm{Re}(a\overline{m}) = z^H(A_2 - A_1 + A_0)z. && (5.16c) \end{aligned}$$

Thus, (5.14) simplifies to

$$\beta_2\theta^2 + \beta_1\theta + \beta_0 = 0 \tag{5.17}$$

The existence of such a functional depends on the question if (5.17) has at least one real root θ. This condition has to be satisfied to be able to transform θ to μ on the complex unit circle and finally evaluate p_ω and p_τ.

Therefore, the following lemma gives conditions when (5.17) has real solutions.

Lemma 5.3. *The roots of the polynomial in* (5.17) θ_1 *and* θ_2 *are real iff*

$$|Re(\overline{m}a)| \leqslant |\overline{m}b|.$$

Furthermore,

$$|Re(\overline{m}a)| = |\overline{m}b| \Leftrightarrow \theta_1 = \theta_2$$

holds.

Proof. We consider

$$\beta_2\theta^2 + \beta_1\theta + \beta_0 = 0 \Leftrightarrow \theta^2 + \frac{\beta_1}{\beta_2}\theta + \frac{\beta_0}{\beta_2} = 0.$$

We apply the pq-formula, thus,

$$\theta_{1,2} = -\frac{\beta_1}{2\beta_2} \pm \sqrt{\frac{\beta_1^2}{4\beta_2^2} - \frac{\beta_0}{\beta_2}}.$$

A real solution is equivalent to a non-negative discriminant, since $\beta_0, \beta_1, \beta_2 \in \mathbb{R}$. Thus,

$$\begin{aligned}
& \frac{\beta_1^2}{4\beta_2^2} - \frac{\beta_0}{\beta_2} \geqslant 0 \\
\Leftrightarrow \quad & \beta_1^2 \geqslant 4\beta_0\beta_2 \\
\Leftrightarrow \quad & 16\ \mathrm{Im}(b\overline{m})^2 \geqslant 4\left(-2\mathrm{Re}((b\overline{m}) - 2\mathrm{Re}((a\overline{m})\right)\left(2\mathrm{Re}((b\overline{m}) - 2\mathrm{Re}((a\overline{m})\right) \\
\Leftrightarrow \quad & \mathrm{Im}(b\overline{m})^2 \geqslant -\left(\mathrm{Re}((b\overline{m}) + \mathrm{Re}((a\overline{m})\right)\left(\mathrm{Re}((b\overline{m}) - \mathrm{Re}((a\overline{m})\right) \\
\Leftrightarrow \quad & \mathrm{Im}(\overline{m}b)^2 \geqslant -\left(\mathrm{Re}(\overline{m}b)^2 - \mathrm{Re}(\overline{m}a)^2\right) \\
\Leftrightarrow \quad & \mathrm{Im}(\overline{m}b)^2 \geqslant \mathrm{Re}(\overline{m}a)^2 - \mathrm{Re}(\overline{m}b)^2 \\
\Leftrightarrow \quad & |\overline{m}b|^2 \geqslant \mathrm{Re}(\overline{m}a)^2 \\
\Leftrightarrow \quad & |\mathrm{Re}(\overline{m}a)| \leqslant |\overline{m}b|.
\end{aligned}$$

In case of equality, the discriminant vanishes, thus, the root has a multiplicity of two.

□

Corollary 5.4. *Let $(\hat{\omega}, \hat{\tau})$ be a simple pair of eigenparameters and $\hat{u}$ a corresponding eigenvector of (5.1), then $\hat{\theta} := \frac{e^{-i\hat{\omega}\hat{\tau}}-1}{e^{-i\hat{\omega}\hat{\tau}}+1}$ is a root of*

$$\left(2Re(\hat{b}\overline{\hat{m}}) - 2Re(\hat{a}\overline{\hat{m}})\right)\theta^2 - 4Im(\hat{b}\overline{\hat{m}})\theta + \left(-2Re(\hat{b}\overline{\hat{m}}) - 2Re(\hat{b}\overline{\hat{m}})\right) = 0$$

with

$$\hat{a} := \hat{u}^H A\hat{u}, \quad \hat{b} := \hat{u}^H B\hat{u} \quad \textit{and} \quad \hat{m} := \hat{u}^H M\hat{u}.$$

Using Lemma 5.3 yields

$$|Re(\overline{\hat{m}}\hat{a})| \leqslant |\overline{\hat{m}}\hat{b}|.$$

Since m, a and b depend on u continuously, there exists a neighborhood around $\hat{u}$ where the roots of the corresponding polynomial are real. Since the roots θ_1 and θ_2 are invariant to any nonzero scaling of u, the neighborhood can be seen as a cone around the vector $\hat{u}$. We will denote this cone by

$$\mathcal{K}_\varepsilon(\hat{u}) := \{u \in \mathbb{C}^n \mid \measuredangle(u, \hat{u}) \leqslant \varepsilon\}.$$

Hence, the domain of the Rayleigh functional is given by $\mathcal{K}_\varepsilon(\hat{u})$.

Similarly, there exists a cone around $\hat{z}$ in $\mathbb{C}^{n^2 \times n^2}$ where we can define the two unique Rayleigh functionals:

$$\Phi_{\theta_+} \quad : \quad \begin{cases} \mathcal{K}_\varepsilon(\hat{z}) & \rightarrow \quad \mathbb{R} \\ \qquad z & \mapsto \quad \dfrac{-\beta_1 - \sqrt{\beta_1^2 - 4\beta_0\beta_2}}{2\beta_2} \end{cases}$$

$$\Phi_{\theta_-} \quad : \quad \begin{cases} \mathcal{K}_\varepsilon(\hat{z}) & \rightarrow \quad \mathbb{R} \\ \qquad z & \mapsto \quad \dfrac{-\beta_1 + \sqrt{\beta_1^2 - 4\beta_0\beta_2}}{2\beta_2} \end{cases}$$

where β_0, β_1 and β_2 are defined in (5.16).

If we further assume A, B and M to be Hermitian, the values a, b and m are real.[5] Under this assumption and $m \neq 0$,[6] (5.14) reduces to

$$\begin{aligned} 2m(b-a)\theta^2 - 2m(b+a) &= 0 \\ \Leftrightarrow (b-a)\theta^2 - (b+a) &= 0 \end{aligned}$$

with the solutions

$$\theta_{1,2} = \pm\sqrt{\frac{b+a}{b-a}}. \tag{5.18}$$

The stationarity in this special case is explained in Proposition 5.6. But first we introduce Assumption 5.5.

Assumption 5.5. *Let $A, B, M \in \mathbb{C}^{n \times n}$ be Hermitian and $\hat{z}$ a left and right eigenvector belonging to a real simple eigenvalue $\hat{\theta} \neq 0$ of $Q(\theta)$ from* (5.12). *Moreover, let $z^H(A_2 - A_1 + A_0)z \neq 0$ be satisfied for all $z \in \mathcal{K}_\varepsilon(\hat{z})$.*

Proposition 5.6. *Let Assumption 5.5 be satisfied, then Φ_{θ_-} and Φ_{θ_+} define Rayleigh functionals which are stationary at the eigenvector $\hat{z} = \hat{u} \otimes \overline{\hat{u}}$ for perturbations in the way such that $\hat{z} + \Delta z =: \tilde{z} = \tilde{u} \otimes \overline{\tilde{u}}$.*

The proof of Proposition 5.6 is very long and technical. We will present here a short summary of the proof. The detailed proof can be found in Appendix A.1

Proof. We consider the scalar equation with $\tilde{z} = \hat{z} + \Delta z$ and $\tilde{\theta} := \hat{\theta} + \Delta\theta$.

$$0 \overset{!}{=} \tilde{z}^H Q(\tilde{\theta})\tilde{z}.$$

Expanding, reordering[7] and exploiting the fact that $\hat{z}$ is a right eigenvector of (5.12) and that $\hat{z}^H$ is a left eigenvector yields the following polynomial in $\Delta\theta$:

$$(\Delta\theta)^2 + 2\hat{\theta}\Delta\theta + \frac{(\Delta z)^H Q(\hat{\theta})\Delta z}{\tilde{z}^H Q_2 \tilde{z}} \overset{!}{=} 0. \tag{5.19}$$

[5] They are real since for $\|u\| = 1$ they are the Rayleigh quotients from the corresponding matrix.

[6] If $m = 0$ (5.14) reduces to the zero polynomial which has roots for each $z \in \mathbb{C}$.

[7] Separating the parts with $\hat{z}$ and $\hat{\theta}$ from the other parts.

Assumption 5.5 assures that $\tilde{z}^H Q_2 \tilde{z} \neq 0$. $\Delta\theta$ can be determined by taking the roots of (5.19), depending on Δz. Applying a Taylor approximation to the function describing $\Delta\theta$ yields

$$|\Delta\theta| = \mathcal{O}\left(\|\Delta z\|_2^2\right).$$

□

5.3.2 From the Polynomial Problem to the Two Parameter Eigenproblem

We want to prove the stationarity of the functionals p_ω and p_τ evaluated at eigenvectors of (5.1). Therefore, we have to convert the result from Proposition 5.6 to the two parameter eigenvalue problem (5.1). Proposition 5.7 describes how the error in the eigenvectors of the large quadratic eigenvalue problem, $\|z - \hat{z}\|$, is related to the error in the eigenvectors of the two parameter nonlinear eigenvalue problem (5.1), $\|u - \hat{u}\|$.

Proposition 5.7. *Let $\hat{u} \in \mathbb{C}^n$ be a solution of* (5.1)*, $u \in \mathcal{K}_\varepsilon(\hat{u})$, $z := u \otimes \overline{u}$ and $\hat{z}$ the solution of* (5.8)*. Then,*

(i)

$$\|z - \hat{z}\|_2 = \mathcal{O}(\|u - \hat{u}\|_2),$$

(ii)

$$|\sin(\measuredangle(z, \hat{z}))| \leqslant \sqrt{2}|\sin(\measuredangle(u, \hat{u}))|$$

holds.

Proof. The first part is proven by

$$\begin{aligned}
\|z - \hat{z}\|_2 &= \left\|(u \otimes \overline{u}) - \hat{u} \otimes \overline{\hat{u}}\right\|_2 \\
&= \left\|(\hat{u} + u - \hat{u}) \otimes \overline{u} - \hat{u} \otimes (\overline{u} + \overline{\hat{u}} - \overline{u})\right\|_2 \\
&= \left\|\hat{u} \otimes \overline{u} + (u - \hat{u}) \otimes \overline{u} - \hat{u} \otimes \overline{u} + \hat{u} \otimes (\overline{u} - \overline{\hat{u}})\right\|_2 \\
&= \left\|(u - \hat{u}) \otimes \overline{u} + \hat{u} \otimes (\overline{u - \hat{u}})\right\|_2 \\
&\leqslant \|(u - \hat{u}) \otimes \overline{u}\|_2 + \left\|\hat{u} \otimes (\overline{u - \hat{u}})\right\|_2 \\
&= \|u - \hat{u}\|_2 \|\overline{u}\|_2 + \|\hat{u}\|_2 \left\|\overline{u - \hat{u}}\right\|_2 \\
&= \mathcal{O}(\|u - \hat{u}\|_2).
\end{aligned}$$

For the second part we assume $\|\hat{u}\|_2 = \|u\|_2 = 1$, which implies $\left\|\overline{\hat{u}}\right\|_2 = \|\overline{u}\|_2 = 1$ and $\|\hat{z}\|_2 = \|z\|_2 = 1$. We consider

$$\begin{aligned}
|\cos(\measuredangle(z, \hat{z}))| &= \left|z^H \hat{z}\right| \\
&= \left|\left(u^H \otimes \overline{u}^H\right)\left(\hat{u} \otimes \overline{\hat{u}}\right)\right| \\
&= \left|\left(u^H \hat{u}\right) \otimes \overline{(u^H \hat{u})}\right| \\
&= |\cos(\measuredangle(u, \hat{u}))|^2.
\end{aligned}$$

With

$$\begin{aligned}\sin\left(\measuredangle(z,\hat{z})\right)^2 &= 1-\cos\left(\measuredangle(z,\hat{z})\right)^2\\ &= 1-\left|\cos\left(\measuredangle(u,\hat{u})\right)\right|^4\\ &= \left(1-\left|\cos\left(\measuredangle(u,\hat{u})\right)\right|^2\right)\left(1+\left|\cos\left(\measuredangle(u,\hat{u})\right)\right|^2\right)\\ &\leqslant 2\left(1-\left|\cos\left(\measuredangle(u,\hat{u})\right)\right|^2\right),\end{aligned}$$

thus,

$$|\sin\left(\measuredangle(z,\hat{z})\right)| \leqslant \sqrt{2}|\sin\left(\measuredangle(u,\hat{u})\right)|.$$

This completes the proof.

□

For the functionals p_ω and p_τ, the parameters ω and τ have to be determined from θ. Therefore, we consider the complex equation

$$u^H T(\omega,\tau)u = 0.$$

It can be transformed similarly into a quadratic polynomial as in Section 5.2.1,

$$u^H T(\omega,\tau)u = i\omega m + a + e^{-i\omega\tau}b = 0. \tag{5.20}$$

Its conjugate complex equation is given by

$$-i\omega\overline{m} + \overline{a} + e^{i\omega\tau}\overline{b} = 0. \tag{5.21}$$

After multiplying (5.20) by $\overline{m}$ and (5.21) by m we obtain

$$\begin{aligned} i\omega|m|^2 + \overline{m}a + e^{-i\omega\tau}\overline{m}b &= 0 && (5.22a)\\ -i\omega|m|^2 + m\overline{a} + e^{i\omega\tau}m\overline{b} &= 0 && (5.22b)\end{aligned}$$

with

$$\mu := e^{-i\omega\tau}, \tag{5.23}$$

and adding (5.22a) with (5.22b), we can eliminate the ω from the two equations and end up in

$$\mu\,\overline{m}b + 2\mathrm{Re}\,(\overline{m}a) + \overline{\mu}\,m\overline{b} = 0. \tag{5.24}$$

Since $|\mu| = 1$,

$$\overline{\mu} = \mu^{-1}$$

is satisfied.

After (5.24) has been multiplied by μ, the value of μ can be determined by computing the roots of the polynomial

$$\mu^2\,\overline{m}b + 2\mu\mathrm{Re}\,(\overline{m}a) + \; m\overline{b} = 0. \tag{5.25}$$

With the assumption, that A, B and M are Hermitian and $m \neq 0$, (5.25) simplifies to

$$b\mu^2 + 2a\mu + b = 0.$$

Hence, the first step evaluating the Rayleigh functionals p_ω and p_τ at u is to evaluate Φ_θ at $u \otimes \overline{u}$ and apply the Möbius transformation from (5.11b).

After $\theta = \Phi_\theta(u \otimes \overline{u})$ has been computed, ω can be derived by transforming (5.20) into

$$\omega = i\left(\frac{a+\mu b}{m}\right),$$

and τ with the help of (5.23) by

$$\tau = i\frac{\ln(\mu)}{\omega}.$$

Applying the Möbius transformation by inserting $\mu = \frac{i+\theta}{i-\theta}$ yields

$$\begin{aligned}
\omega &= i\left(\frac{a+\frac{i+\theta}{i-\theta}b}{m}\right)\\
&= \frac{i}{m(1+\theta^2)}\left(\theta^2(a-b)-2i\theta b+a+b\right)\\
&= \frac{2b\theta}{m(1+\theta^2)} && (5.26)\\
\tau &= i\frac{\ln\left(\frac{i+\theta}{i-\theta}\right)}{\omega}. && (5.27)
\end{aligned}$$

The latter simplification for ω results from exploiting $\theta^2 = \frac{b+a}{b-a}$ from (5.18), which is possible since we have Hermitian matrices.

This yields the two functionals p_ω and p_τ:

$$p_\omega : \mathcal{K}_\varepsilon(\hat{u}) \to \mathbb{R}$$

and

$$p_\tau : \mathcal{K}_\varepsilon(\hat{u}) \to \mathbb{R}$$

with

$$p_\omega(u) = \frac{2u^H Bu\,\Phi_\theta(u\otimes\overline{u})}{u^H Mu\,(1+\Phi_\theta(u\otimes\overline{u})^2)} \tag{5.28}$$

$$p_\tau(u) = i\frac{\ln\left(\frac{i+\Phi_\theta(u\otimes\overline{u})}{i-\Phi_\theta(u\otimes\overline{u})}\right)}{p_\omega(u)} \tag{5.29}$$

Here we have skipped the "+" and the "-" for the functional Φ_θ, because all following ideas can be applied to both functionals, Φ_{θ_+} and Φ_{θ_-}.

Theorem 5.8 and Theorem 5.9 show that an eigenvector of (5.1), $\hat{u}$, is a stationary point of the functional p_ω and p_τ, respectively.

Theorem 5.8.
Let Assumption 5.5 be satisfied. Furthermore, let $\hat{z}$ be a stationary point of Φ_θ and let the domain $\mathcal{K}_\varepsilon(\hat{u})$ of p_ω, be chosen such that

$$\left|\frac{u^H Bu}{u^H Mu\,(1+\Phi_\theta(u\otimes\overline{u})^2)}\right| \leqslant C < \infty, \quad \forall u \in \mathcal{K}_\varepsilon(\hat{u}).$$

Then, $\hat{u}$ is a stationary point of p_ω, i.e.

$$|p_\omega(\hat{u}+\Delta u) - p_\omega(\hat{u})| = \mathcal{O}\left(\|\Delta u\|_2^2\right), \forall \Delta u \textit{ such that } \hat{u}+\Delta u \in \mathcal{K}_\varepsilon(\hat{u}).$$

Proof. We consider

$$\begin{aligned}
&|p_\omega(\hat{u}+\Delta u) - p_\omega(\hat{u})| \\
&= 2\left|\frac{(\hat{u}+\Delta u)^H B\,(\hat{u}+\Delta u)\ \Phi_\theta\left((\hat{u}+\Delta u)\otimes\overline{(\hat{u}+\Delta u)}\right)}{(\hat{u}+\Delta u)^H M\,(\hat{u}+\Delta u)\ \left(1+\Phi_\theta\left((\hat{u}+\Delta u)\otimes\overline{(\hat{u}+\Delta u)}\right)^2\right)}\right. \\
&\left. - \frac{\hat{u}^H B\hat{u}\,\Phi_\theta(\hat{u}\otimes\overline{\hat{u}})}{\hat{u}^H M\hat{u}\left(1+\Phi_\theta(\hat{u}\otimes\overline{\hat{u}})^2\right)}\right| \\
&= 2\left|\Psi(\hat{u}+\Delta u)\Phi_\theta\left((\hat{u}+\Delta u)\otimes\overline{(\hat{u}+\Delta u)}\right) - \Psi(\hat{u})\Phi_\theta(\hat{u}\otimes\overline{\hat{u}})\right|,
\end{aligned}$$

with

$$\Psi(u) := \frac{u^H Bu}{u^H Mu\,(1+\Phi_\theta(u\otimes\overline{u})^2)}.$$

Using

$$\hat{\Psi} := \sup_{u\in\mathcal{K}_\varepsilon(\hat{u})} |\Psi(u)|,$$

we obtain

$$|p_\omega(\hat{u}+\Delta u) - p_\omega(\hat{u})| \leqslant 2\hat{\Psi}|\Phi_\theta\left((\hat{u}+\Delta u)\otimes\overline{(\hat{u}+\Delta u)}\right) - \Phi_\theta(\underbrace{\hat{u}\otimes\overline{\hat{u}}}_{\hat{z}})|.$$

Defining

$$\Delta z =: (\hat{u}+\Delta u)\otimes\overline{(\hat{u}+\Delta u)} - \hat{z},$$

and applying Proposition 5.6 yields

$$|p_\omega(\hat{u}+\Delta u) - p_\omega(\hat{u})| = \mathcal{O}\left(\|\Delta z\|_2^2\right).$$

Finally, Proposition 5.7 completes the proof, since

$$\|\Delta z\|_2 = \mathcal{O}\left(\|\Delta u\|_2\right).$$

□

Theorem 5.9.
Let Assumption 5.5 be satisfied. Furthermore, let $\hat{z}$ be a stationary point of Φ_θ and let the domain of p_τ, $\mathcal{K}_\varepsilon(\hat{u})$, be chosen such that

$$\left|\frac{u^H Bu}{u^H Mu\,(1+\Phi_\theta(u\otimes\overline{u})^2)}\right| \leqslant C < \infty, \quad \forall u \in \mathcal{K}_\varepsilon(\hat{u}).$$

Then, $\hat{u}$ is a stationary point of p_τ, i.e.

$$|p_\tau(\hat{u}+\Delta u) - p_\tau(\hat{u})| = \mathcal{O}\left(\|\Delta u\|_2^2\right), \forall u+\Delta u \in \mathcal{K}_\varepsilon(\hat{u}).$$

Proof. Since p_τ is differentiable with respect to ω and θ we can apply Taylor expansion such that

$$\begin{aligned} p_\tau(\hat{u}+\Delta u) = p_\tau(\hat{u}) &+ \left.\frac{\partial p_\tau}{\partial\theta}\right|_{\substack{\omega=p_\omega(\hat{u})\\ \theta=\Phi_\theta(\hat{z})}} (\Phi_\theta(\hat{z}+\Delta z) - \Phi_\theta(\hat{z}))\\ &+ \left.\frac{\partial p_\tau}{\partial\omega}\right|_{\substack{\omega=p_\omega(\hat{u})\\ \theta=\Phi_\theta(\hat{z})}} (p_\omega(\hat{u}+\Delta u) - p(\hat{u}))\\ &+ o\left(|\Phi_\theta(\hat{z}+\Delta z) - \Phi_\theta(\hat{z})|\right)\\ &+ o\left(|p_\omega(\hat{u}+\Delta u) - p_\omega(\hat{u})|\right). \end{aligned} \tag{5.30}$$

The partial derivatives can be determined using (5.27).

$$\frac{\partial p_\tau}{\partial\theta} = \frac{2}{\omega(1+\theta^2)} \tag{5.31}$$

$$\frac{\partial p_\tau}{\partial\omega} = -\frac{i}{\omega^2}\ln\left(\frac{i+\theta}{i-\theta}\right). \tag{5.32}$$

For further information regarding the partial derivative of p_τ, see Appendix A.2.

Equation (5.30) implies

$$\begin{aligned} |p_\tau(\hat{u}+\Delta u) - p_\tau(\hat{u})| &= \mathcal{O}\left(|\Phi_\theta(\hat{z}+\Delta z) - \Phi_\theta(\hat{z})|\right)\\ &\quad + \mathcal{O}\left(|p_\omega(\hat{u}+\Delta u) - p_\omega(\hat{u})|\right). \end{aligned}$$

Finally, Proposition 5.6, Proposition 5.7 and Theorem 5.8 complete the proof. □

5.3.3 Conditions for Stationarity

The theory about stationarity of the Rayleigh functional that we have discussed so far, depends on the question of whether Assumption 5.5 is satisfied. Specifically, the requirement that $\hat{z}^H$ has to be a left eigenvector of $Q(\hat{\theta})$ yields further conditions to assure that Assumption 5.5 is satisfied. Therefore, we consider

$$Q^H(\hat{\theta})\hat{z} = \hat{\theta}^2\left(A_2 - A_1 + A_0\right)^H \hat{z} - 2i\hat{\theta}\left(A_2 - A_0\right)^H \hat{z} + \left(-A_2 - A_1 - A_0\right)^H \hat{z} = 0$$

where the matrix polynomial Q is defined in (5.12).

With our assumption, that A, B and M are Hermitian, this property is transferred to the large matrices A_0, A_1 and A_2, thus,

$$A_2 = A_2^H, \quad A_1 = A_1^H, \quad \text{and} \quad A_0 = A_0^H$$

and

$$\hat{\theta}^2 \left(A_2 - A_1 + A_0\right) \hat{z} - 2i\hat{\theta} \left(A_2 - A_0\right) \hat{z} + \left(-A_2 - A_1 - A_0\right) \hat{z} = 0.$$

Subtracting (5.12)

$$Q(\hat{\theta})\hat{z} = \hat{\theta}^2 \left(A_2 - A_1 + A_0\right) \hat{z} + 2i\hat{\theta} \left(A_2 - A_0\right) \hat{z} + \left(-A_2 - A_1 - A_0\right) \hat{z} = 0,$$

yields

$$-4i\hat{\theta} \left(A_2 - A_0\right) \hat{z} = 0. \tag{5.33}$$

After applying the rules of the Kronecker product, we confirm that (5.33) is equivalent to

$$4i\hat{\theta} \left((M\hat{u}) \otimes (\overline{B\hat{u}}) - (B\hat{u}) \otimes (\overline{M\hat{u}})\right) = 0.$$

This equation is, in general, not fulfilled, since $\hat{\theta} \neq 0$ for a simple eigenvalue. Even if we assume $M\hat{u}$ and $B\hat{u}$ to be real, the second part is usually also nonzero, due to the rules of the Kronecker product, which is not commutative.

Therefore, our assumption that A, B and M are Hermitian is not sufficient for stationarity of the Rayleigh functional.

We will see in Example 5.16 that there exist cases where the Rayleigh functional seems to be stationary, although z^H is no left eigenvector of (5.12). This observation can be explained by considering the polynomial (5.12). The only matrix destroying the Hermitian structure is the coefficient matrix of the linear part, $Q_1 = 2i(A_2 - A_0)$. Hence, if $\hat{\theta}$ is very small or very large in magnitude, the influence of this non-Hermitian part to the matrix $Q(\hat{\theta})$ is comparably small. Thus,

$$\frac{\left\|Q(\hat{\theta}) - Q^H(\hat{\theta})\right\|_2}{\left\|Q(\hat{\theta})\right\|_2} \ll 1,$$

and the behavior of the Rayleigh functional can be explained. A small or large absolute value of $\hat{\theta}$ is equivalent to a μ close to the real axis such that this behavior can also be regarded with respect to the polynomial eigenvalue problem from (5.8). Additionally, this implies a small value of $\hat{\omega}$, such that the matrix $T(\hat{\omega}, \hat{\tau})$ is close to being Hermitian. With this property the false assumption of quadratic convergence behavior by Meerbergen, Schröder and Voss in [45] is explained.

5.3.4 Two Sided and Complex Symmetric Approach

Similar to the standard nonlinear eigenvalue problem, we consider the idea of a two-sided Rayleigh functional. Therefore, the equation

$$y^H T(\omega, \tau) u = 0$$

is solved for the two real parameters ω and τ for given $y, u \in \mathbb{C}^n$. Here u denotes an approximation for the sought (right) eigenvector and y can be chosen differently, which will be discussed later.

The following proposition shows the relationship between the Rayleigh functional of the two-parameter eigenvalue problem and the Rayleigh functional of the polynomial eigenvalue problem (5.8).

Proposition 5.10. *The equation*

$$w^H P(\mu) z = 0,$$

with P and z from (5.8) *and $w := y \otimes \overline{y}$, is equivalent to*

$$(b\overline{m})\mu^2 + (a\overline{m} + m\overline{a})\mu + \overline{m}b = 0$$

with

$$m := y^H M u, \quad a := y^H A u \quad \textit{and} \quad b := y^H B u.$$

Proof. Multiplying (5.8) by $w^H = (y \otimes \overline{y})^H$ from the left leads to

$$\mu^2 (y\otimes\overline{y})^H (B\otimes\overline{M})(u\otimes\overline{u}) + \mu (y\otimes\overline{y})^H (A\otimes\overline{M} + M\otimes\overline{A})(u\otimes\overline{u}) + (y\otimes\overline{y})^H (M\otimes\overline{B})(u\otimes\overline{u}) = 0$$

Applying the calculation rules for the Kronecker product (see [29]) leads to

$$\mu^2 \left(\underbrace{y^H B u}_{=b} \; \underbrace{\overline{y^H M u}}_{=\overline{m}} \right) + \mu \left(\underbrace{y^H A u}_{=a} \; \underbrace{\overline{y^H M u}}_{=\overline{m}} + \underbrace{y^H M u}_{=m} \; \underbrace{\overline{y^H A u}}_{=\overline{a}} \right) + \left(\underbrace{y^H M u}_{=m} \; \underbrace{\overline{y^H B u}}_{=\overline{b}} \right) = 0$$

This completes the proof. □

Complex symmetric Rayleigh functional

If the matrices $A, B, M \in \mathbb{C}^{n\times n}$ are symmetric, the (complex) symmetry is preserved for the family of matrices $T(\omega,\tau)$[8], i.e.,

$$T(\omega,\tau) = T(\omega,\tau)^T, \quad \forall \omega,\tau \in \mathbb{R}. \tag{5.34}$$

For symmetric matrices, the left eigenvector is always the conjugate complex of the right eigenvector. Therefore, an approximation for the left eigenvector is available without any further computation if an approximation for the right eigenvector has been determined. Then, the following equation is considered to determine ω and τ:

$$u^T T(\omega,\tau) u = 0.$$

Using the rules for the Kronecker product, we easily see, that for the polynomial from (5.8),

$$P(\mu) = P(\mu)^T,$$

[8] In the Hermitian case, this property is not preserved.

is satisfied if (5.34) is fulfilled.

For those problems, the left eigenvector of the quadratic eigenvalue problem (5.8) is given by

$$\hat{w} = \overline{\hat{z}}.$$

We denote the symmetric Rayleigh functional for the problem (5.8) by Φ_T.

Schreiber discussed this special Rayleigh functional in [62]. It is shown that the Rayleigh functional satisfies the following properties:

$$\begin{array}{rcl} \Phi_T & : & \left\{ \begin{array}{rcl} \mathcal{K}_\varepsilon(\hat{z}) & \to & R \subset \mathcal{S}^1 \\ z & \mapsto & \Phi_T(z) \end{array} \right. \\ \Phi_T(\hat{z}) & = & \hat{\mu}, \\ \Phi_T(\beta z) & = & \Phi_T(z), \quad \forall \beta \in \mathbb{C} \setminus \{0\} \\ z^T P(\Phi_T(z)) z & = & 0 \\ z^T P'(\Phi_T(z)) z & \neq & 0, \end{array}$$

where $\hat{\mu}$ denotes an eigenvalue and $\hat{z}$ a corresponding eigenvector. The set R denotes an arc on the complex unit circle $\mathcal{S}^1$. Hence, two disjoint arcs on the unit circle have to be chosen, then the uniqueness of this functional is satisfied.

In contrast to the standard Rayleigh functional, this Rayleigh functional Φ_T is holomorphic in z. Thus, the stationarity of Φ_T in $\hat{z}$ can be shown by building the derivative Φ_T', with respect to z. Then the derivative of the Rayleigh functional is derived as in Lemma 2.17 in Chapter 2. Thus, we derive

$$\Phi_T'(z) = -\frac{2}{z^T P'\left(\Phi_T(z)\right) z} P\left(\Phi_T(z)\right) z.$$

This implies for an eigenvector $\hat{z}$,

$$\Phi_T'(\hat{z}) = 0.$$

Now, we define the functional

$$p_{\mu,T} : \left\{ \begin{array}{rcl} \mathcal{K}_\varepsilon(\hat{u}) & \to & R \subset \mathcal{S}^1 \\ u & \mapsto & \Phi_T(u \otimes \overline{u}). \end{array} \right. \tag{5.35}$$

However, this functional is not holomorphic with respect to u anymore, because the transformation from u to $u \otimes \overline{u}$ includes the complex conjugate $\overline{u}$.

Therefore, Definition 2.19 is used here.

Lemma 5.11. *Let the matrices $A, B, M \in \mathbb{C}^{n \times n}$ be symmetric and let ($\hat{\mu} = p_{\mu,T}(\hat{u}), \hat{z} = \hat{u} \otimes \overline{\hat{u}}$) be a simple eigenpair of* (5.8)*. Then, for the functional $p_{\mu,T}$, defined in* (5.35)*, there exists an $\varepsilon > 0$ such that for every $u \in \mathcal{K}_\varepsilon(\hat{u})$ the following error estimation holds*

$$|p_{\mu,T}(u) - p_{\mu,T}(\hat{u})| = \mathcal{O}(\|u - \hat{u}\|_2^2).$$

Proof. We consider

$$\begin{aligned}|p_{\mu,T}(u) - p_{\mu,T}(\hat{u})| &= |\Phi_T(u \otimes \overline{u}) - \Phi_T(\hat{u} \otimes \overline{\hat{u}})| \\ &= |\Phi_T(z) - \Phi_T(\hat{z})| \\ &= |\underbrace{\Phi'_T(\hat{z})}_{=0}(z-\hat{z})| + \mathcal{O}(\|z-\hat{z}\|_2^2) \\ &= \mathcal{O}(\|z-\hat{z}\|_2^2).\end{aligned}$$

Using Proposition 5.7, the error in z can be transferred to an error in u. Thus,

$$|p_{\mu,T}(u) - p_{\mu,T}(\hat{u})| = \mathcal{O}(\|u-\hat{u}\|_2^2).$$

This completes the proof. □

Remark 5.12. With (5.28) and (5.29), we define the functionals to determine the parameters ω and τ:

$$p_{\omega,T} : \begin{cases} \mathcal{K}_\varepsilon & \to & J_\omega \subset \mathbb{R} \\ u & \mapsto & i\frac{u^T Au + p_{\mu,T}(u) u^T Bu}{u^T Mu}, \end{cases} \tag{5.36}$$

$$p_{\tau,T} : \begin{cases} \mathcal{K}_\varepsilon & \to & J_\tau \subset \mathbb{R} \\ u & \mapsto & i\frac{\ln(p_{\mu,T}(u))}{p_{\omega,T}(u)}. \end{cases} \tag{5.37}$$

Finally, the stationarity of these functionals is summarized in the following theorem:

Theorem 5.13.
Let $A, B, M \in \mathbb{C}^{n\times n}$ be complex symmetric. Furthermore, let $(\hat{\omega}, \hat{\tau}, \hat{u})$ be an eigentriple satisfying $\hat{u}^T M\hat{u} \neq 0$ and $\hat{\omega}\hat{\tau} \neq \pm\pi$. Then there exists $\varepsilon > 0$ such that for all $u \in \mathcal{K}_\varepsilon(\hat{u}) \subset \mathbb{C}^n$ there exists the two functionals:

$$\begin{aligned} p_{\omega,T} &: \mathcal{K}_\varepsilon(\hat{u}) \to J_\omega \subset \mathbb{R} \\ p_{\tau,T} &: \mathcal{K}_\varepsilon(\hat{u}) \to J_\tau \subset \mathbb{R} \end{aligned}$$

satisfying

$$u^T T(p_{\omega,T}(u), p_{\tau,T}(u))u = 0,$$

and the additional properties of a Rayleigh functional, where T is given in (5.1).

The two functionals, $p_{\omega,T}$ and $p_{\tau,T}$, are stationary evaluated at the eigenvector $\hat{u}$.

The proof is very long and technical. Therefore, we give a summary of the proof. For further details see Appendix A.3.

Proof. Without loss of generality we consider only perturbations, Δu, which fulfill

$$\hat{u}^T M\Delta u = 0.^9 \tag{5.38}$$

[9] If this condition is not satisfied, we split Δu into $\Delta u = \frac{\hat{u}^T M\Delta u}{\hat{u}^T M\hat{u}}\hat{u} + \left(I - \frac{\hat{u}\hat{u}^T M}{\hat{u}^T M\hat{u}}\right)\Delta u$. The first part is a multiple of the eigenvector and the latter part satisfies the above condition.

We exploit the fact that $p_{\omega,T}$ is holomorphic with respect to u. Applying Taylor expansion to $p_{\omega,T}(\hat{u} + \Delta u)$ yields

$$p_{\omega,T}(\hat{u} + \Delta u) = p_{\omega_T}(\hat{u}) + 2\underbrace{\frac{\omega \hat{u}^T M \Delta u}{\hat{u}^T M \hat{u}}}_{=0} + \mathcal{O}(|p_\mu(\hat{u} + \Delta u) - p_{\mu_T}(\hat{u})|) + \mathcal{O}(\|\Delta u\|^2).$$

By using Lemma 5.11 we end up in

$$|p_{\omega,T}(\hat{u} + \Delta u) - p_{\omega,T}(\hat{u})| = \mathcal{O}(\|\Delta u\|^2). \tag{5.39}$$

For the functional $p_{\tau,T}$, we use the same approach: We exploit that the complex logarithm is holomorphic[10] and apply a Taylor expansion to $p_{\tau,T}$. Finally, we have

$$|p_{\tau,T}(\hat{u}+\Delta u) - p_{\tau,T}(\hat{u})| = \mathcal{O}(|p_{\omega,T}(\hat{u}+\Delta u) - p_{\omega,T}(\hat{u})|) + \mathcal{O}(|p_{\mu,T}(\hat{u}+\Delta u) - p_{\mu,T}(\hat{u})|).$$

Lemma 5.11 and (5.39) complete the proof. □

Two sided Rayleigh functional

If the matrices A, B and M are not symmetric the approximation property of the Rayleigh functional can be improved by choosing an approximation for the left eigenvector, y. According to two-sided Rayleigh quotient iteration ([55, 56]) and two-sided Rayleigh functional iteration (cf [62]), a two sided Rayleigh functional fulfills the stationarity property at eigenvectors here, too.

To obtain approximations for the left eigenvector, a further linear equation containing the varying matrix $T(\omega^{(k)}, \tau^{(k)})^H$ has to be solved in every iteration step.[11] Furthermore, we then have two different residuals. The right residual, $T(\omega^{(k)}, \tau^{(k)})u^{(k)}$, and the left residual, $\left(y^{(k)}\right)^H T(\omega^{(k)}, \tau^{(k)})$.

5.4 Numerical Examples

Example 5.14. This example is presented in [34]. We consider the partial differential equation

$$\frac{\partial x(\xi, t)}{\partial t} = \frac{\partial^2 x(\xi, t)}{\partial \xi^2} + a_0(\xi)v(\xi, t) + a_1(\xi)v(\pi - \xi, t - \tau), \tag{5.40}$$

with

$$\frac{\partial}{\partial \xi}x(0, t) = \frac{\partial}{\partial \xi}x(\pi, t) = 0. \tag{5.41}$$

Discretizing (5.40) with central differences[12] with respect to ξ yields the following delay ordinary differential equation

$$M\dot{x}(t) + Ax(t) - Bx(t - \tau) = 0,$$

[10] We consider the complex logarithm only on the unit circle, here it is holomorphic except at -1. With the assumption $\hat{\omega}\hat{\tau} \neq \pm\pi$, we excluded this case.

[11] Alternatively, the adjoint Jacobi-Davidson correction equation has to be solved.

[12] i.e. $x(t) := (x_1(t), \dots, x_n(t))^T$ with $x_j(t) = x((j-1)h, t)$

with

$$
\begin{aligned}
M &= I_n,\\
A &= \frac{1}{h^2}\begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix} + \begin{pmatrix} a_0(0) & & & & \\ & a_0(h) & & & \\ & & \ddots & & \\ & & & a_0(\pi-h) & \\ & & & & a_0(\pi) \end{pmatrix},\\
B &= -\begin{pmatrix} & & & & a_1(0) \\ & & & a_1(h) & \\ & & \cdot^{\cdot^{\cdot}} & & \\ & a_1(\pi-h) & & & \\ a_1(\pi) & & & & \end{pmatrix},\\
h &= \frac{\pi}{n-1}.
\end{aligned}
$$

With the well known ansatz, $x(t) = e^{\lambda t}u$, we finally end up in the following nonlinear eigenvalue problem, which depends additionally on the parameter τ.

$$\left(\lambda M + A + e^{-\lambda\tau}B\right)u = 0.$$

In order to determine those values for τ where an eigenvalue changes from the stable to unstable state, or vice versa, we set $\lambda = i\omega$ and end up in

$$\left(i\omega M + A + e^{-i\omega\tau}B\right)u = 0,$$

which has the same structure as (5.1). We choose

$$
\begin{aligned}
a_0(\xi) &= -2\sin(\xi) \quad \text{and}\\
a_1(\xi) &= -1, \forall \xi \in [0,\pi]
\end{aligned}
$$

to assure that $\hat{z}^H$ is a left eigenvector of (5.12). This is an assumption for Theorem 5.8 and Theorem 5.9. Hence, a_1 is chosen differently from [34] and [45]. They choose $a_1(\xi) = 2\sin(\xi)$ or $a_1(\xi) = 2\sin(\xi)+1$. This causes a non Hermitian matrix B. Since our goal is confirming the results of Theorem 5.8 and Theorem 5.9, we choose a very rough discretization grid, i.e. $n = 5$. We compute four eigentriples with Algorithm 5.1 and choose the eigentriple

$$(\hat{\omega}, \hat{\tau}, \hat{u})$$

with $\hat{\omega} = 0.726...$ and $\hat{\tau} = 3.207....$

We perturb the eigenvector by a random complex vector e with $e^H\hat{u} = 0$ and $\|e\| \approx 0.1$. The Rayleigh functionals are evaluated on vectors along $\{w \in \mathbb{C}^n \mid w = \hat{u} + \alpha e,\ \alpha \in (0,1] \subset \mathbb{R}\}$.

The results are illustrated in Figure 5.1. These computations confirm Theorem 5.8 and Theorem 5.9, since $\left|p_\omega\left(u^{(k)}\right) - \hat{\omega}\right|$ and $\left|p_\tau\left(u^{(k)}\right) - \hat{\tau}\right|$ have the same slope as $\left\|u^{(k)} - \hat{u}\right\|^2$ in logarithmic scale.

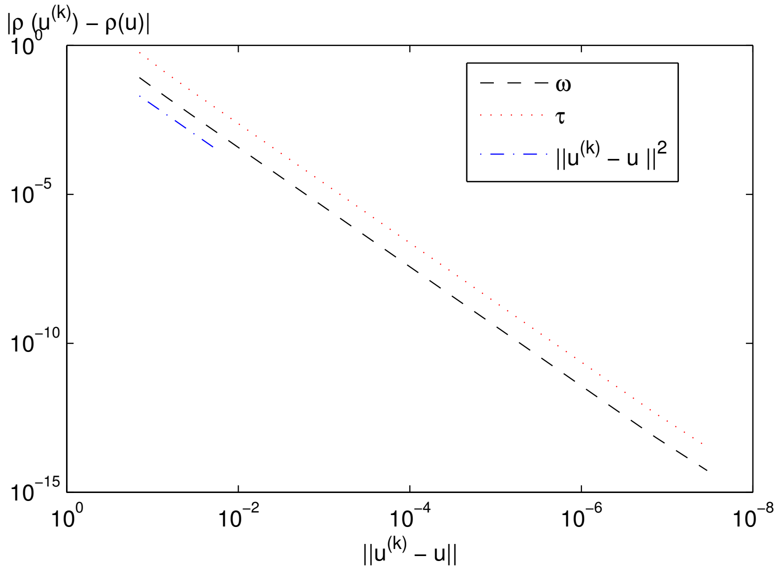

Figure 5.1: The error of the Rayleigh functionals vs. the error in the eigenvector

Example 5.15. Secondly, we consider the same problem as in Example 5.14 but we choose

$$a_1(\xi) = 2\sin(\xi) + 1,$$

according to [45]. We choose $n = 15$ and select the eigentriple

$$(\hat{\omega}, \hat{\tau}, \hat{u})$$

with $\hat{\omega} = 1.8140...$ and $\hat{\tau} = 0.9151...$ after all eigentriples have been computed using Algorithm 5.1. Similar to Example 5.14, we perturb the eigenvector in a random direction e which is orthogonal to $\hat{u}$. Then we evaluate the standard Rayleigh functional p_ω and p_τ along the perturbed vector. Furthermore, the symmetric Rayleigh functionals, $p_{\omega,T}$ and $p_{\tau,T}$, are evaluated.

In this example, $\hat{z}^H$ is no longer a left eigenvector of (5.12). Consequently, Theorem 5.8 and Theorem 5.9 are not applicable. Therefore, we cannot expect p_ω and p_τ to be stationary at the eigenvector. But since A, B, M and all their complex linear combinations are complex symmetric, we can expect the functionals $p_{\omega,T}$ and $p_{\tau,T}$ to be stationary at the eigenvector, according to Theorem 5.13.

Figure 5.2 confirms these expectations. The slopes belonging to the symmetric Rayleigh functionals $p_{\omega,T}$ and $p_{\tau,T}$ are steeper and about to be parallel to the one of $\|u - \hat{u}\|_2^2$. The ones of the standard Rayleigh functional are less steep, as expected.

The observations from the logarithmic plots in Figure 5.2 are highlighted by the convergence order. It can be determined by the slope in the logarithmic scale. Since we expect that the error in the Rayleigh functionals behaves like

$$|p(u) - p(\hat{u})| \approx C \, \|u - \hat{u}\|_2^\kappa ,$$

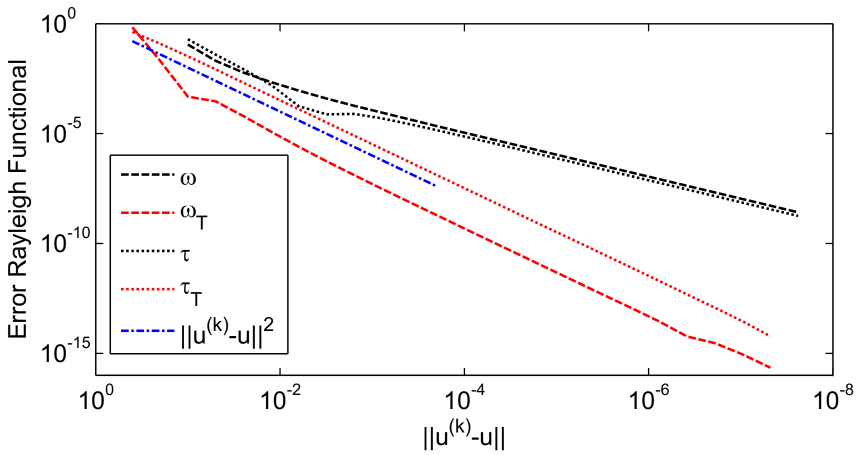

Figure 5.2: The error of the Rayleigh functionals vs. the error in the eigenvector

Rayleigh functional	κ
p_ω	1.1127
p_τ	1.2814
$p_{\omega,T}$	2.0412
$p_{\tau,T}$	2.0037

Table 5.1: Convergence orders of the Rayleigh functionals

the values of C and κ can be determined by solving the least squares problem,

$$\sum_{j=1}^{m}\left(\ln(C)+\kappa\ln\left(\left\|u^{(j)}-\hat{u}\right\|_2\right)-\ln\left(\left|p(u^{(j)})-p(\hat{u})\right|\right)\right)^2\overset{!}{=}\min,$$

for different $p\in\{p_\omega,p_\tau,p_{\omega,T},p_{\tau,T}\}$.

The results for the four different Rayleigh functionals are noted in Table 5.1.

We recognize, obviously, that the error of the standard Rayleigh functionals tends to zero almost linearly with the error in the arguments. This confirms our previous results in this chapter that the stationarity is not provided if $\hat{z}^H$ is not a left eigenvector of (5.12).

But Table 5.1 also confirms that the complex symmetric Rayleigh functionals, $p_{\omega,T}$ and $p_{\tau,T}$, are stationary at the eigenvector $\hat{u}$ as proven in Theorem 5.13.

Example 5.16. The last example is also presented in [45]. It originates in the discretization of the following partial differential equation:

$$u_t-\nabla\left(\left(1+x^2+y^2\right)\nabla u\right)-\alpha(1+xy)u(t-\tau)=0, \tag{5.42}$$

where u depends on the coordinates x and y as well as on time t. We consider a discretization on $\Omega=(0,1)\times(0,1)$ with Dirichlet boundary conditions, $u=0$ on $\partial\Omega$, with respect to the positions x and y. We obtain a system of $n=104257$ ordinary

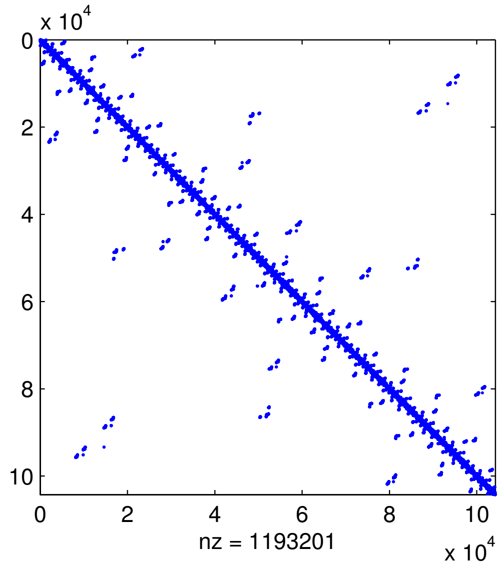

Figure 5.3: The nonzero pattern of the discretized version of (5.42)

delay differential equations. This is transformed into a two parameter eigenvalue problem as described at the beginning of this chapter. The nonzero pattern of $A + B + M$ is shown in Figure 5.3.

We choose $\alpha = 89$, since the system becomes unstable for $\alpha \geqslant 22.6$ and this is what we were interested in. The problem is solved using Algorithm 5.2 and we select the eigenparameters

$$\hat{\omega} = 13.3004 \qquad \text{and} \qquad \hat{\tau} = 0.2274.$$

We perturb the eigenvector orthogonally as it has been done in Example 5.14 and Example 5.15 and evaluate the Rayleigh functionals at the perturbed eigenvector. Then the errors of the functionals are compared to the error in the vector.

The results are plotted in Figure 5.4.

We observe, that the slope of the error in the Rayleigh functionals is nearly equal to the slope of the error in u squared. Table 5.2 confirms this observation.

Although this problem does not satisfy Assumption 5.5 the Rayleigh functionals seem to be stationary at the eigenvector $\hat{u}$. Even the calculated convergence orders in Table 5.2 pretend this. This behavior can be explained by considering Section 5.3.3. We have $\hat{\mu} = -0.9931 - 0.1174i$ which yields $\hat{\theta} = 16.9697$. Thus, the quadratic part of $Q(\hat{\theta})$ is dominant and therefore

$$\left\| Q(\hat{\theta}) - Q^H(\hat{\theta}) \right\|_2 \ll \left\| Q(\hat{\theta}) \right\|_2 .$$

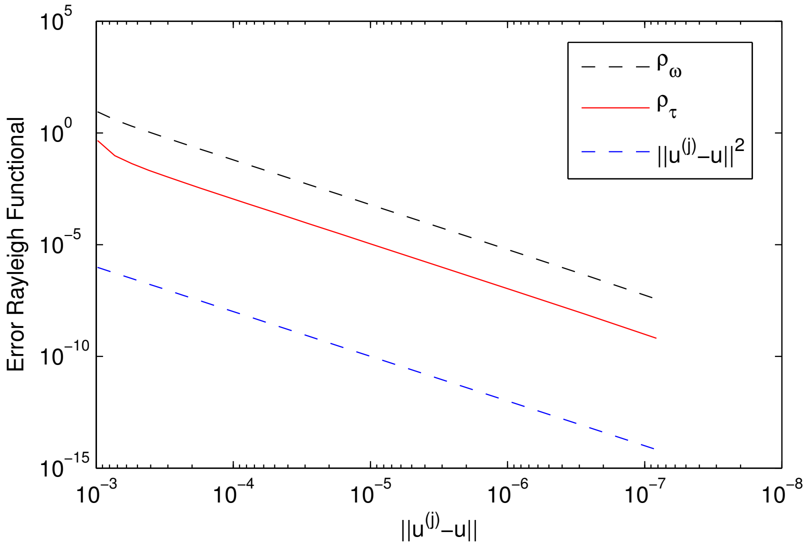

Figure 5.4: The error of the Rayleigh functionals vs. the error in the eigenvector

Rayleigh functional	κ
p_ω	2.0252
p_τ	2.0671

Table 5.2: Convergence orders of the Rayleigh functionals

Chapter 6

Conclusions and Outlook

The Jacobi-Davidson Algorithm can be regarded as a procedure to make an existing iterative method more robust against perturbations. These perturbations are usually caused by an inexact solve of linear systems occurring in each iteration step. Pointing out this fact is the most important aim of this thesis. In Chapter 3 we introduced a very general approach of the Jacobi-Davidson method based on an iterative method.

For variants of this approach, we developed convergence proofs for the Jacobi-Davidson method on a two-dimensional subspace. We were able to confine the Ritz vectors' location on a two dimensional plane. The influence of the orthogonal search space expansion decreases the closer the Ritz vector is to an eigenvector. Thus, the impact of a perturbed subspace expansion decreases in the same manner. As a result, we proved that the perturbation of a Ritz vector decreases with the sine of the angle between $x^{(k)}$ and the target eigenvector. The convergence constant depends linearly on the sine of the perturbation angle ϑ (cf. Theorems 4.7, 4.12, 4.18, 4.29).

This perturbation occurs independently to the original iterative method to solve the (nonlinear) eigenvalue problem. Hence, less expensive methods like Derivative-free Rayleigh functional iteration (cf. Algorithm 2.5) reduce the computational costs, but provide nearly the same convergence behavior as more expensive methods like classical Rayleigh functional iteration.

We restricted ourselves to two dimensional principles to demonstrate the convergence behavior. Usually, the subspace dimension is limited by a greater number than two. We can surely suppose that search spaces with a dimension greater than two provide at least the same convergence rate as our two dimensional approach. Further analysis can answer the question whether better convergence rates can be proven for larger subspaces or not.

We pointed out that the orthogonal subspace expansion is much more robust against perturbation than inexact variants of Rayleigh quotient iteration, Rayleigh functional iteration or inverse iteration. Solving the linear system requires the more inner iteration steps the closer we are at an eigenvector and eigenvalue for all these methods. Usually the iterative solver for the linear system breaks down due to the bad condition of the linear system at the end. The inexact Jacobi-Davidson correction equation can be solved iteratively with an acceptable number of inner iterations. And the outcome improves the current approximation of the eigenpair.

In Chapter 5 we considered a special class of two parameter eigenvalue problems. We could present different approaches for a Rayleigh functional for this kind of problems. The greatest challenge was to show the stationarity of this functional at eigenvectors (cf. Theorems 5.8 and 5.9). This property was proven under certain assumptions for the matrices M, A and B. These assumptions (cf. Assumption 5.5) are much more restrictive than for other problems, since, for instance, the matrix M is multiplied by $i\omega$. Additionally, we introduced a symmetric Rayleigh functional for the case of complex symmetric matrices (cf. Theorem 5.13). Here we exploit that the left eigenvector is given by the complex conjugate of the right eigenvector, and that the Rayleigh functional is holomorphic.

These observations and theory imply a possible question for future work: How fast does the Jacobi-Davidson Algorithm for the two-parameter eigenvalue problem converge if the correction equation is solved only inexactly?

Furthermore, the existence of a Rayleigh functional allows further methods to solve this special eigenvalue problem. For instance, it can be analyzed if a special version of Safeguarded Iteration (Algorithm 2.4) or Rayleigh functional iteration (with derivative and without derivative, see Algorithm 2.2 and Algorithm 2.5) leads to similar results as for conventional nonlinear eigenvalue problems.

There also exist quadratic eigenvalue problems with a time delay, like

$$\left(\lambda^2 A_2 + \lambda A_1 + A_0 + e^{-\lambda\tau} B\right) v = 0.$$

By analyzing for critical delays and setting $\lambda = i\omega$, we obtain the two-parameter eigenvalue problem

$$\left(-\omega^2 A_2 + i\omega A_1 + A_0 + e^{-i\omega\tau} B\right) v = 0.$$

Numerical methods for problems of this type can be studied in the future.

Moreover, the robustness of a block-variant of the Jacobi-Davidson method can be considered. Currently, the Jacobi-Davidson correction equation

$$\left(I - \frac{Cx^{(k)}\left(x^{(k)}\right)^H}{\left(x^{(k)}\right)^H Cx^{(k)}}\right) S \left(I - x^{(k)}\left(x^{(k)}\right)^H B\right) t = -Sx^{(k)}$$

is solved such that $\left(x^{(k)}\right)^H Bt = 0$. But we can also find a search space expansion which is B-orthogonal to all vectors of the current subspace. Let $V \in \mathbb{C}^{n\times k}$ contain an orthogonal basis of the current k- dimensional subspace with $k \ll n$, then the Jacobi-Davidson correction equation is adapted to

$$\left(I - CV(V^H CV)^{-1}V^H\right) S \left(I - VV^H B\right) t = -Sx^{(k)}, \quad V^H t = 0. \tag{6.1}$$

However, for $C \neq I$, the matrix vector multiplication with this matrix requires the solution of a $k \times k$ linear system. The same problem occurs if the preconditioner for the Jacobi-Davidson correction equation is applied. Here, even for $C = I$ a $k \times k$ linear system has to be solved. The matrix $V^H CV$ is constant over all inner iterations, thus, a decomposition can be made at the beginning of the iteration process.

Appendix A

Proofs

A.1 Proof of Proposition 5.6

Proposition 5.6. *Let Assumption 5.5 be satisfied, then Φ_{θ_-} and Φ_{θ_+} define Rayleigh functionals which are stationary at the eigenvector $\hat{z} = \hat{u} \otimes \overline{\hat{u}}$ for perturbations in the way such that $\hat{z} + \Delta z =: \tilde{z} = \tilde{u} \otimes \overline{\tilde{u}}$.*

Proof. We consider the following scalar equation with $\tilde{z} = \hat{z} + \Delta z$ and $\tilde{\theta} := \hat{\theta} + \Delta\theta$.

$$\begin{aligned} 0 &= \tilde{z}^H Q(\tilde{\theta})\tilde{z} \\ &= (\hat{z}^H + \Delta z^H)\left(Q(\hat{\theta}) + Q(\tilde{\theta}) - Q(\hat{\theta})\right)(\hat{z} + \Delta z) \\ &= (\hat{z}^H + \Delta z^H)Q(\hat{\theta})(\hat{z} + \Delta z) + \underbrace{(\hat{z}^H + \Delta z^H)}_{\tilde{z}^H}\left(Q(\tilde{\theta}) - Q(\hat{\theta})\right)\underbrace{(\hat{z} + \Delta z)}_{\tilde{z}} \end{aligned} \tag{A.1}$$

We exploit that $\hat{z}$ is the left and right eigenvector belonging to the eigenvalue $\hat{\theta}$.

$$\begin{aligned} 0 &= (\hat{z}^H + \Delta z^H)\left(\underbrace{Q(\hat{\theta})\hat{z}}_{=0} + Q(\hat{\theta})\Delta z\right) + (\tilde{z}^H)\left(Q(\tilde{\theta}) - Q(\hat{\theta})\right)\tilde{z} \\ &= \underbrace{\hat{z}^H Q(\hat{\theta})\Delta z}_{=0} + \Delta z^H Q(\hat{\theta})\Delta z + \tilde{z}^H\left(Q(\tilde{\theta}) - Q(\hat{\theta})\right)\tilde{z} \\ &= \Delta z^H Q(\hat{\theta})\Delta z + \tilde{z}^H\left(Q(\tilde{\theta}) - Q(\hat{\theta})\right)\tilde{z} \end{aligned} \tag{A.2}$$

The polynomial Q is defined in (5.12), by

$$Q(\theta) = \theta^2 \underbrace{(A_2 - A_1 + A_0)}_{=:Q_2} + \theta\underbrace{(2iA_2 - 2iA_0)}_{=:Q_1} + \underbrace{(-A_2 - A_1 - A_0)}_{=:Q_0}.$$

Moreover, we consider

$$\begin{aligned} Q(\tilde{\theta}) - Q(\hat{\theta}) &= \tilde{\theta}^2 Q_2 + \tilde{\theta}Q_1 + Q_0 - \hat{\theta}^2 Q_2 - \hat{\theta}Q_1 - Q_0 \\ &= (\tilde{\theta}^2 - \hat{\theta}^2)Q_2 + (\tilde{\theta} - \hat{\theta})Q_1. \end{aligned}$$

Thus, (A.2) simplifies to

$$\begin{aligned} 0 &= \Delta z^H Q(\hat{\theta})\Delta z + \tilde{z}^H \left(Q(\tilde{\theta}) - Q(\hat{\theta})\right) \tilde{z} \\ &= \Delta z^H Q(\hat{\theta})\Delta z + \tilde{z}^H \left((\tilde{\theta}^2 - \hat{\theta}^2)Q_2 + (\tilde{\theta} - \hat{\theta})Q_1\right) \tilde{z} \\ &= \Delta z^H Q(\hat{\theta})\Delta z + (\tilde{\theta}^2 - \hat{\theta}^2)\tilde{z}^H Q_2 \tilde{z} + (\tilde{\theta} - \hat{\theta})\tilde{z}^H Q_1 \tilde{z}. \end{aligned} \tag{A.3}$$

The matrices M, A and B are assumed to be Hermitian, which implies that $\tilde{u}^H M \tilde{u}$ and $\tilde{u}^H B \tilde{u}$ are real. With (5.14) we have

$$\begin{aligned} \tilde{z}^H Q_1 \tilde{z} &= 2i\tilde{z}^H \left(A_2 - A_0\right) \tilde{z} \\ &= 2i(\tilde{u} \otimes \overline{\tilde{u}})^H (B \otimes \overline{M} - M \otimes \overline{B})(\tilde{u} \otimes \overline{\tilde{u}}) \\ &= 2i\left(\tilde{u}^H B\tilde{u}\, \overline{\tilde{u}^H M\tilde{u}} - \tilde{u}^H M\tilde{u}\, \overline{\tilde{u}^H B\tilde{u}}\right) \\ &= 2i\left(\tilde{u}^H B\tilde{u}\, \tilde{u}^H M\tilde{u} - \tilde{u}^H M\tilde{u}\, \tilde{u}^H B\tilde{u}\right) \\ &= 0, \end{aligned} \tag{A.4}$$

as long $\tilde{z}$ can be built by $\tilde{z} = \tilde{u} \otimes \overline{\tilde{u}}$. Thus, (A.3) reduces to

$$\Delta z^H Q(\hat{\theta})\Delta z + (\tilde{\theta}^2 - \hat{\theta}^2)\tilde{z}^H Q_2 \tilde{z} = 0. \tag{A.5}$$

Inserting $\tilde{\theta} = \hat{\theta} + \Delta\theta$, (A.5) yields

$$\begin{aligned} 0 &= \Delta z^H Q(\hat{\theta})\Delta z + (\tilde{\theta}^2 - \hat{\theta}^2)\tilde{z}^H Q_2 \tilde{z} \\ &= \Delta z^H Q(\hat{\theta})\Delta z + (\hat{\theta}^2 + 2\hat{\theta}\Delta\theta + \Delta\theta^2 - \hat{\theta}^2)\tilde{z}^H Q_2 \tilde{z} \\ &= \tilde{z}^H Q_2 \tilde{z}\Delta\theta^2 + 2\hat{\theta}\tilde{z}^H Q_2 \tilde{z}\Delta\theta + \Delta z^H Q(\hat{\theta})\Delta z. \end{aligned} \tag{A.6}$$

Hence, we end up in a polynomial in $\Delta\theta$ presented in (A.6). Its roots are the sought perturbations of the Rayleigh functional, $\Delta\theta$.

According to Assumption 5.5, we assume $\tilde{z}^H Q_2 \tilde{z} \neq 0.$ and $\hat{z} \neq 0$. Then (A.6) is transferred to

$$(\Delta\theta)^2 + 2\hat{\theta}\Delta\theta + \frac{\Delta z^H Q(\hat{\theta})\Delta z}{\tilde{z}^H Q_2 \tilde{z}} \overset{!}{=} 0, \tag{A.7}$$

and can be solved by standard methods.

Solving (A.7) for $\Delta\theta$ yields

$$\begin{aligned}
\Delta\theta_{1,2} &= -\hat{\theta} \pm \sqrt{\hat{\theta}^2 - \frac{\Delta z^H Q(\hat{\theta})\Delta z}{\tilde{z}^H Q_2 \tilde{z}}} \\
&= -\hat{\theta} \pm \sqrt{\hat{\theta}^2 - \hat{\theta}^2 \frac{\Delta z^H Q(\hat{\theta})\Delta z}{\hat{\theta}^2\ \tilde{z}^H Q_2 \tilde{z}}} \\
&= -\hat{\theta} \pm \underbrace{|\hat{\theta}|}_{=\mathrm{sgn}\ (\hat{\theta})\hat{\theta}} \sqrt{1 - \frac{(\Delta z)^H Q(\hat{\theta})\Delta z}{\hat{\theta}^2\ \tilde{z}^H Q_2 \tilde{z}}} \\
&= \hat{\theta}\left(-1 \pm \mathrm{sgn}(\hat{\theta})\sqrt{1 - \frac{(\Delta z)^H Q(\hat{\theta})\Delta z}{\hat{\theta}^2\ \tilde{z}^H Q_2 \tilde{z}}}\right). \qquad \text{(A.8)}
\end{aligned}$$

However, we obtain two solutions of (A.7). This originates from the fact that (A.1) has two solutions for $\tilde{\theta}$. We are interested in a solution $\tilde{\theta}$ in the neighborhood of $\hat{\theta}$. Thus, we consider the solution $\Delta\theta_1$, if $\mathrm{sgn}(\hat{\theta}) = 1$. And, if $\mathrm{sgn}(\hat{\theta}) = -1$, we consider $\Delta\theta_2$ instead. The case $\hat{\theta} = 0$ is excluded by assumption.

With the help of the Taylor expansion of

$$f : \left\{ \begin{array}{rcl} (-\infty, 1) & \rightarrow & \mathbb{R}^+ \\ x & \mapsto & -1 + \sqrt{1 - x} \end{array} \right.$$

we easily obtain

$$|\Delta\theta| = \mathcal{O}\left(\|\Delta z\|_2^2\right).$$

□

A.2 Derivative of the Complex Logarithm

We will prove (5.31):

$$\frac{\partial p_\tau}{\partial\theta} = \frac{2}{\omega(1+\theta^2)}$$

In (5.27) we have

$$\tau = \frac{i}{\omega}\ln\left(\frac{i+\theta}{i-\theta}\right).$$

Inserting

$$\mu = \frac{i+\theta}{i-\theta} \qquad \text{(A.9)}$$

from (5.11b) yields

$$\tau = \frac{i}{\omega}\ln(\mu).$$

We recognize in (A.9) that $\mu \neq -1$, $\forall\theta \in \mathbb{R}$. The complex logarithm is holomorphic in the entire complex plane except the real negative axis. Since μ is always located on the unit circle, $\mu = -1$ is the only point we need to be aware of. But this point is never reached. The derivative of the complex logarithm is given by

$$\frac{d}{d\mu}\ln(\mu) = \frac{1}{\mu}.$$

Inserting (A.9) and determining the inner derivative yields

$$\begin{aligned}\frac{d}{d\theta}\ln\left(\frac{i+\theta}{i-\theta}\right) &= \frac{i-\theta}{i+\theta}\quad\frac{d}{d\theta}\left(\frac{i+\theta}{i-\theta}\right)\\ &= \frac{i-\theta}{i+\theta}\,\frac{(i-\theta)-(i+\theta)(-1)}{(i-\theta)^2}\\ &= \frac{i-\theta}{i+\theta}\,\frac{2i}{(i-\theta)^2}\\ &= \frac{2i}{(i+\theta)(i-\theta)}\\ &= \frac{-2i}{1+\theta^2}.\end{aligned}$$

Finally, we end up with

$$\begin{aligned}\frac{\partial\tau}{\partial\theta} &= \frac{i}{\omega}\quad\frac{d}{d\theta}\ln\left(\frac{i+\theta}{i-\theta}\right)\\ &= \frac{i}{\omega}\frac{-2i}{1+\theta^2}\\ &= \frac{2}{\omega(1+\theta^2)}.\end{aligned}$$

A.3 Proof of Theorem 5.13

Theorem 5.13.
Let $A, B, M \in \mathbb{C}^{n\times n}$ be complex symmetric. Furthermore, let $(\hat\omega, \hat\tau, \hat u)$ be an eigentriple satisfying $\hat u^T M \hat u \neq 0$ and $\hat\omega\hat\tau \neq \pm\pi$. Then there exists $\varepsilon > 0$ such that for all $u \in \mathcal{K}_\varepsilon(\hat u) \subset \mathbb{C}^n$ there exists the two functionals:

$$\begin{aligned}p_{\omega,T} &: \mathcal{K}_\varepsilon(\hat u) \to J_\omega \subset \mathbb{R}\\ p_{\tau,T} &: \mathcal{K}_\varepsilon(\hat u) \to J_\tau \subset \mathbb{R}\end{aligned}$$

satisfying

$$u^T T(p_{\omega,T}(u), p_{\tau,T}(u))u = 0,$$

and the additional properties of a Rayleigh functional, where T is given in (5.1).

The two functionals, $p_{\omega,T}$ and $p_{\tau,T}$, are stationary evaluated at the eigenvector $\hat u$.

Proof. Without loss of generality, we consider only perturbations, Δu, which fulfill

$$\hat u^T M \Delta u = 0. \tag{A.10}$$

Lemma 5.11 yields

$$|p_{\mu,T}(\hat{u}+\Delta u) - p_{\mu,T}(\hat{u})| = \mathcal{O}(\|\Delta u\|^2). \tag{A.11}$$

We consider

$$\begin{aligned} p_{\omega,T}(u+\Delta u) = i \Bigg(& \underbrace{\frac{(\hat{u}+\Delta u)^T A(\hat{u}+\Delta u)}{(\hat{u}+\Delta u)^T M(\hat{u}+\Delta u)}}_{=:\Psi_1(\hat{u}+\Delta u)} \\ & + p_{\mu,T}(\hat{u}+\Delta u) \underbrace{\frac{(\hat{u}+\Delta u)^T B(\hat{u}+\Delta u)}{(\hat{u}+\Delta u)^T M(\hat{u}+\Delta u)}}_{=:\Psi_2(\hat{u}+\Delta u)} \Bigg). \end{aligned} \tag{A.12}$$

Since Ψ_1 and Ψ_2 are holomorphic with respect to u, we can use Taylor expansion at $\hat{u}$, thus,

$$\Psi_k(\hat{u}+\Delta u) = \Psi_k(\hat{u}) + \nabla\Psi_k(\hat{u})^T \Delta u, \quad k = 1, 2,$$

with

$$\begin{aligned} \nabla\Psi_1(u)^T &= 2\frac{\left(u^T M u\right) u^T A - \left(u^T A u\right) u^T M}{\left(u^T M u\right)^2}, \\ \nabla\Psi_2(u)^T &= 2\frac{\left(u^T M u\right) u^T B - \left(u^T B u\right) u^T M}{\left(u^T M u\right)^2}. \end{aligned}$$

Using (A.10), we obtain

$$\begin{aligned} \nabla\Psi_1(\hat{u})^T \Delta u &= 2\frac{\hat{u}^T A \Delta u}{\hat{u}^T M \hat{u}}, \\ \nabla\Psi_2(\hat{u})^T \Delta u &= 2\frac{\hat{u}^T B \Delta u}{\hat{u}^T M \hat{u}}. \end{aligned}$$

Hence, we can simplify (A.12) to

$$\begin{aligned}
p_{\omega,T}(\hat{u}+\Delta u) &= i\left(\Psi_1(\hat{u}) + 2\frac{\hat{u}^T A\Delta u}{\hat{u}^T M\hat{u}} + p_{\mu,T}(\hat{u}+\Delta u)\left(\Psi_2(\hat{u}) + 2\frac{\hat{u}^T B\Delta u}{\hat{u}^T M\hat{u}}\right)\right) + \mathcal{O}(\|\Delta u\|^2)\\
&= i\left(\Psi_1(\hat{u}) + p_{\mu,T}(\hat{u})\Psi_2(\hat{u}) + p_{\mu,T}(\hat{u}+\Delta u)\Psi_2(\hat{u}) - p_{\mu,T}(\hat{u})\Psi_2(\hat{u})\right)\\
&+ i\left(2\frac{\hat{u}^T A\Delta u}{\hat{u}^T M\hat{u}} + 2(p_{\mu,T}(\hat{u}) - p_{\mu,T}(\hat{u}) + p_{\mu,T}(\hat{u}+\Delta u))\frac{\hat{u}^T B\Delta u}{\hat{u}^T M\hat{u}}\right)\\
&+ \mathcal{O}(\|\Delta u\|^2)\\
&= i\left(\Psi_1(\hat{u}) + p_{\mu,T}(\hat{u})\Psi_2(\hat{u})\right) + i\Psi_2(\hat{u})\left(p_{\mu,T}(\hat{u}+\Delta u) - p_{\mu,T}(\hat{u})\right)\\
&+ 2i\frac{\hat{u}^T(A + p_{\mu,T}(\hat{u})\, B)\Delta u}{\hat{u}^T M\hat{u}} + 2i\frac{\hat{u}^T B\Delta u}{\hat{u}^T M\hat{u}}\left(p_{\mu,T}(\hat{u}+\Delta u) - p_{\mu,T}(\hat{u})\right)\\
&+ \mathcal{O}(\|\Delta u\|^2)\\
&= i\left(\frac{\hat{u}^T A\hat{u}}{\hat{u}^T M\hat{u}} + p_{\mu,T}(\hat{u})\frac{\hat{u}^T B\hat{u}}{\hat{u}^T M\hat{u}}\right) + 2i\frac{\hat{u}^T(A + p_{\mu,T}(\hat{u})\, B)\Delta u}{\hat{u}^T M\hat{u}}\\
&+ \mathcal{O}(|p_{\mu,T}(\hat{u}+\Delta u) - p_{\mu,T}(\hat{u})|) + \mathcal{O}(\|\Delta u\|^2) \qquad \text{(A.13)}
\end{aligned}$$

We exploit now that $\hat{u}^T$ is a left eigenvector of (5.1), thus,

$$i\hat{\omega}\hat{u}^T M + \hat{u}^T A + \underbrace{e^{-i\hat{\omega}\hat{\tau}}}_{=\hat{\mu}=p_{\mu,T(\hat{u})}} \hat{u}^T B = 0,$$

and, therefore,

$$\hat{u}^T A + p_{\mu,T(\hat{u})}\hat{u}^T B = -i\hat{\omega}\hat{u}^T M.$$

Inserting this result into the penultimate row of (A.13) and setting

$$p_{\omega,T}(\hat{u}) = i\left(\frac{\hat{u}^T A\hat{u}}{\hat{u}^T M\hat{u}} + p_{\mu,T}(\hat{u})\frac{\hat{u}^T B\hat{u}}{\hat{u}^T M\hat{u}}\right)$$

we end up in

$$p_{\omega,T}(\hat{u}+\Delta u) = p_{\omega,T}(\hat{u}) + 2\underbrace{\frac{\hat{\omega}\hat{u}^T M\Delta u}{\hat{u}^T M\hat{u}}}_{=0} + \mathcal{O}(|p_{\mu,T}(\hat{u}+\Delta u) - p_{\mu,T}(\hat{u})|) + \mathcal{O}(\|\Delta u\|^2).$$

With (A.11), we end up in

$$|p_{\omega,T}(\hat{u}+\Delta u) - p_{\omega,T}(\hat{u})| = \mathcal{O}(\|\Delta u\|^2). \qquad \text{(A.14)}$$

For the functional $p_{\tau,T}$ the proof is similar. We have

$$p_{\tau,T}(\hat{u}+\Delta u) = i\frac{\ln\left(p_{\mu,T}(\hat{u}+\Delta u)\right)}{p_{\omega,T}(\hat{u}+\Delta u)}.$$

Since the logarithm is holomorphic on the unit circle, except $\mu = -1$, (which has been excluded), we can linearize with the help of a Taylor expansion.

Therefore,

$$p_{\tau,T}(\hat{u}+\Delta u) = i\frac{\ln\left(p_{\mu,T}(\hat{u})\right) + \overline{p_{\mu,T}(\hat{u})}(p_{\mu,T}(\hat{u}+\Delta u) - p_{\mu,T}(\hat{u}))}{p_{\omega,T}(\hat{u}+\Delta u)} + \mathcal{O}(\|\Delta u\|_2^2)$$

and

$$\begin{aligned}
|p_{\tau,T}(\hat{u}+\Delta u) - p_{\tau,T}(\hat{u})| &= \left|\frac{\ln\left(p_{\mu,T}(\hat{u})\right) + \overline{p_{\mu,T}(\hat{u})}(p_{\mu,T}(\hat{u}+\Delta u) - p_{\mu,T}(\hat{u}))}{p_{\omega,T}(\hat{u}+\Delta u)} - \frac{\ln\left(p_{\mu,T}(\hat{u})\right)}{p_{\omega,T}(\hat{u})}\right| \\
&= \left|\frac{\left[\ln\left(p_{\mu,T}(\hat{u})\right) + \overline{p_{\mu,T}(\hat{u})}(p_{\mu,T}(\hat{u}+\Delta u) - p_{\mu,T}(\hat{u}))\right] p_{\omega,T}(\hat{u})}{p_{\omega,T}(\hat{u}+\Delta u)\, p_{\omega,T}(\hat{u})}\right. \\
&\quad \left. - \frac{\ln\left(p_{\mu,T}(\hat{u})\right)\; p_{\omega,T}(\hat{u}+\Delta u)}{p_{\omega,T}(\hat{u}+\Delta u)\, p_{\omega,T}(\hat{u})}\right| \\
&= \left|\frac{\ln(p_{\mu,T}(\hat{u}))(p_{\omega,T}(\hat{u}) - p_{\omega,T}(\hat{u}+\Delta u))}{p_{\omega,T}(\hat{u}+\Delta u)p_{\omega,T}(\hat{u})}\right. \\
&\quad \left. + \frac{\overline{p_{\mu,T}(\hat{u})}(p_{\mu,T}(\hat{u}+\Delta u) - p_{\mu,T}(\hat{u}))}{p_{\omega,T}(\hat{u}+\Delta u)}\right| \\
&= \mathcal{O}(|p_{\omega,T}(\hat{u}+\Delta u) - p_{\omega,T}(\hat{u})|) \\
&\quad + \mathcal{O}(|p_{\mu,T}(\hat{u}+\Delta u) - p_{\mu,T}(\hat{u})|).
\end{aligned}$$

Finally, (A.11) and (A.14) yield

$$|p_{\tau,T}(\hat{u}+\Delta u) - p_{\tau,T}(\hat{u})| = \mathcal{O}\left(\|\Delta u\|^2\right).$$

This completes the proof.

□

Appendix B

Further Results of Example 4.30

In this section we present further result tables of Example 4.30.

k	$\sin(\phi_k)$	$\lvert\lambda^{(k)} - \lambda_1\rvert$	$\left\lVert r^{(k)} \right\rVert$	$\left\lVert res^{(k)} \right\rVert$	**inner iterations**
0	0.080	1.2e4	7.7e5	0.0987	85
1	0.177	350.2	2.5e3	0.0996	188
2	0.100	3.60	80.28	0.0983	1073
3	0.017	2.0e-3	0.508	0.0945	1108
4	1.34e-5	4.3e-10	1.97e-4	n.a.	n.a.

Table B.1: Results of Inexact Rayleigh Quotient Iteration with fixed τ

In Table B.1 the linear system in the last step could not be solved anymore, because the returned value was equal to the starting value.

k	$\sin(\phi_k)$	$\lvert\lambda^{(k)} - \lambda_1\rvert$	$\left\lVert r^{(k)} \right\rVert$	τ_k	$\left\lVert res^{(k)} \right\rVert$	**inner iterations**
0	0.08	1.2e4	7.7e5	0.25	0.2494	29
1	0.2713	710.00	3.96e3	1.3e-3	1.3e-3	1026
2	0.3658	101.60	318.20	1.03-4	9.71e-5	1149
3	0.2158	6.55	48.25	1.57e-5	1.4564e-5	1145
4	0.0417	0.0327	1.56	5.08e-7	3.2678e-4	1182*
5	6.42e-4	7.44e-7	1.4e-3	4.49e-10	13.74	1350*

Table B.2: Results of Inexact Rayleigh Quotient Iteration with decreasing τ

Bibliography

[1] K. Aishima. Global convergence of the restarted Lanczos and Jacobi–Davidson methods for symmetric eigenvalue problems. *Numerische Mathematik*, 131(3):405–423, 2015.

[2] K. Aishima. On convergence of iterative projection methods for symmetric eigenvalue problems. *Journal of Computational and Applied Mathematics*, 311:513 – 521, 2017.

[3] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst. *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide.* Software, Environments and Tools. Society for Industrial and Applied Mathematics, 2000.

[4] R. Barrett, M. Berry, T.F. Chan, J. Demmel, J.M. Donato, J. Dongarra, V. Eijkhout, R. Pozo, Ch. Romine, and H. van der Vorst. *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods.* Pa. Society for Industrial and Applied Mathematics, 1994.

[5] J. Berns-Müller, I.G. Graham, and A. Spence. Inexact inverse iteration for symmetric matrices. *Linear Algebra and its Applications*, 416:389–413, 2006.

[6] J. Berns-Müller and A. Spence. Inexact inverse iteration and GMRES. Technical report, Department of Mathematics, University of Bath, Bath, UK, 2005.

[7] J. Berns-Müller and A. Spence. Inexact inverse iteration with variable shift for nonsymmetric generalized eigenvalue problems. *SIAM Journal on Matrix Analysis and Applications*, 28(4):1069 – 1082, 2006.

[8] T. Betcke and H. Voss. A Jacobi-Davidson-type projection method for nonlinear eigenvalue problems. *Future Generation Computer Systems*, 20(3):363–372, 2004.

[9] J. Chen, G. Gu, and C.N. Nett. A new method for computing delay margins for stability of linear delay systems. In *Decision and Control, 1994., Proceedings of the 33rd IEEE Conference on*, volume 1, pages 433–437, Dec 1994.

[10] E.R. Davidson. The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real-symmetric matrices. *Journal of Computational Physics*, 17:87–94, 1975.

[11] P.G. Drazin. *Introduction to hydrodynamic stability.* Cambridge University Press, Cambridge, UK, 2002.

[12] R.J. Duffin. *A Minimax Theory for Overdamped Networks.* Pittsburgh. Carnegie Institute of Technology. Department of Mathematics. Technical report. Carnegie Institute of Technology, 1954.

[13] H. Fassbender, D.S. Mackey, N. Mackey, and C. Schrödder. Structured polynomial eigenproblems related to time-delay systems. *Electronic Transactions on Numerical Analysis*, pages 306–330, 2008.

[14] M.A. Freitag, P. Kürschner, and J. Pestana. Gmres convergence bounds for eigenvalue problems. *Computational Methods in Applied Mathematics*, 18(2):203–222, 2017.

[15] M.A. Freitag and A. Spence. Convergence of inexact inverse iteration with application to preconditioned iterative solves. *BIT Numerical Mathematics*, 47:27 – 44, 2007.

[16] M.A. Freitag and A. Spence. Convergence theory for inexact inverse iteration applied to the generalised nonsymmetric eigenproblem. *Electronic Transactions on Numerical Analysis*, 28:40 – 64, 2007.

[17] M.A. Freitag and A. Spence. Rayleigh quotient iteration and simplified Jacobi-Davidson method with preconditioned iterative solves. *Linear Algebra and its Applications*, 428:2049 – 2060, 2008.

[18] M.A. Freitag and A. Spence. A tuned preconditioner for inexact inverse iteration applied to Hermitian eigenvalue problems. *IMA Journal of Numerical Analysis*, 28:522 – 551, 2008.

[19] I. Gohberg, P. Lancaster, and L. Rodman. *Matrix Polynomials.* Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 1982.

[20] G.H. Golub and C.F. Van Loan. *Matrix Computations.* Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 1984.

[21] G.H. Golub and Q. Ye. Inexact inverse iteration for the eigenvalue problems. *BIT Numerical Mathematics*, 40(4):671 – 684, 2000.

[22] W. Govaerts. *Numerical methods for bifurcations of dynamical equilibria.* SIAM, Philadelphia, PA, USA,, 2000.

[23] A. Greenbaum. *Iterative Methods for Solving Linear Systems.* Frontiers in Applied Mathematics. Society for Industrial and Applied Mathematics, 1997.

[24] K. Gu, V.L. Kharitonov, and J. Chen. *Stability of Time-Delay Systems.* Birkhäuser, Boston, 2003.

[25] St. Güttel and Fr. Tisseur. The nonlinear eigenvalue problem. *Acta Numerica*, 26:1–94, 2017.

[26] R. F. Heinemann and A. B. Poore. Multiplicity, stability, and oscillatory dynamics of the tubular reactor. *Chemical Engineering Science*, 36(8):1411 – 1419, 1981.

[27] M. Hochstenbach and Y. Notay. Controlling Inner Iterations in the Jacobi-Davidson Method. *SIAM Journal on Matrix Analysis and Applications*, 31(2):460–477, 2009.

[28] M. Hochstenbach and G. Sleijpen. Harmonic and refined Rayleigh–Ritz for the polynomial eigenvalue problem. *Numerical Linear Algebra with Applications*, 15(1):35–54, 2008.

[29] R.A. Horn and Ch.R. Johnson. *Topics in matrix analysis.* Cambridge Univ. Press, Cambridge [u.a.], 1991.

[30] T.-M. Hwang, W.-W. Lin, W.-Ch. Wang, and W. Wang. Numerical simulation of three dimensional pyramid quantum dot. *J. Comput. Phys.*, 196(1):208–232, 2004.

[31] C.G.J. Jacobi. Über ein leichtes Verfahren die in der Theorie der Säcularstörungen vorkommenden Gleichungen numerisch aufzulösen. *Journal für die reine und angewandte Mathematik*, 30:51–94, 1846.

[32] E. Jarlebring. Critical delays and polynomial eigenvalue problems. *Journal of Computational and Applied Mathematics*, 224(1):296 – 306, 2009.

[33] E. Jarlebring and M.E. Hochstenbach. Polynomial two-parameter eigenvalue problems and matrix pencil methods for stability of delay-differential equations. *Linear Algebra and its Applications*, 431(3–4):369 – 380, 2009. Special Issue in honor of Henk van der Vorst.

[34] E. Jarlebring, K. Meerbergen, and W. Michiels. A Krylov Method for the Delay Eigenvalue Problem. *SIAM Journal on Scientific Computing*, 32(6):3278–3300, 2010.

[35] Z. Jia. The Convergence of Harmonic Ritz Values, Harmonic Ritz Vectors, and Refined Harmonic Ritz Vectors. *Mathematics of Computation*, 74(251):1441–1456, 2004.

[36] Z.X. Jia and Z. Wang. A convergence analysis of the inexact Rayleigh quotient iteration and simplified Jacobi-Davidson method for the large Hermitian matrix eigenproblem. *Science in China Series A: Mathematics*, 51(12):2205–2216, 2008.

[37] K. Königsberger. *Analysis 1.* Springer, Berlin [u.a.], 6., durchges. Aufl. edition, 2004.

[38] K. Königsberger. *Analysis 2.* Springer, Berlin [u.a.], 5., korrigierte Aufl. edition, 2004.

[39] Y.-L. Lai, K.-Y. Lin, and W.-W. Lin. An inexact inverse iteration for large sparse eigenvalue problems. *Numerical Linear Algebra with Applications*, 4:425 – 437, 1997.

[40] P. Lancaster. A generalised Rayleigh quotient iteration for lambda-matrices. *Archive for Rational Mechanics and Analysis*, 8(1):309–322, 1961.

[41] R. Lehoucq and K. Meerbergen. Using generalized Cayley transformations within an inexact rational Krylov sequence method. *SIAM Journal on Matrix Analysis and Applications*, 20(1):131–148, 1998.

[42] J. Louisell. A matrix method for determining the imaginary axis eigenvalues of a delay system. *IEEE Transactions on Automatic Control*, 46(12):2008–2012, Dec 2001.

[43] M. Markiewicz and H. Voss. Electronic states in three dimensional quantum dot/wetting layer structures. In M. Gavrilova, O. Gervasi, V. Kumar, C. J. K. Tan, D. Taniar, A. Laganá, Y. Mun, and H. Choo, editors, *Computational Science and Its Applications - ICCSA 2006*, pages 684–693, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[44] K. Meerbergen and D. Roose. The Restarted Arnoldi Method Applied to Iterative Linear System Solvers for the Computation of Rightmost Eigenvalues. *SIAM Journal on Matrix Analysis and Applications*, 18(1):1–20, 1997.

[45] K. Meerbergen, C. Schröder, and H. Voss. A Jacobi-–Davidson method for two-real-parameter nonlinear eigenvalue problems arising from delay-differential equations. *Numerical Linear Algebra with Applications*, 20(5):852–868, 2013.

[46] V. Mehrmann and H. Voss. Nonlinear eigenvalue problems: A Challenge for Modern Eigenvalue Methods, GAMM-Reports, 2004.

[47] A. Mera, R. Cigel, and J Lewis. Harwell-Boeing collection, Martrix Market. http://math.nist.gov/MatrixMarket/collections/hb.html, 1986.

[48] R.B. Morgan. Computing interior eigenvalues of large matrices. *Linear Algebra and its Applications*, 154/156:289 – 309, 1991.

[49] A. Neumaier. Residual inverse iteration for the nonlinear eigenvalue problem. *SIAM Journal on Numerical Analysis*, 22:914–923, 1985.

[50] V. Niendorf and H. Voss. Detecting hyperbolic and definite matrix polynomials. *Linear Algebra and its Applications*, 432:1017–1035, 2010.

[51] Y. Notay. Combination of Jacobi–Davidson and conjugate gradients for the partial symmetric eigenproblem. *Numerical Linear Algebra with Applications*, 9:21 – 44, 2002.

[52] Y. Notay. Convergence analysis of inexact Rayleigh quotient iteration. *SIAM Journal on Matrix Analysis and Applications*, 24:627 – 644, 2003.

[53] J.M. Ortega and W.C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables.* Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 1970.

[54] M. R. Osborne. A new method for the solution of eigenvalue problems. *The Computer Journal*, 7(3):228–232, 1964.

[55] A.M. Ostrowski. On the convergence of the Rayleigh quotient iteration for the computation of the characteristic roots and vectors. III. *Archive for Rational Mechanics and Analysis*, 3(1):325–340, 1959.

[56] A.M. Ostrowski. On the convergence of the Rayleigh quotient iteration for the computation of the characteristic roots and vectors. IV. *Archive for Rational Mechanics and Analysis*, 3(1):341–347, 1959.

[57] B.N. Parlett. *The symmetric eigenvalue problem.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1998.

[58] E.H. Rogers. A minmax theory for overdamped systems. *Archive for Rational Mechanics and Analysis*, 16(2):89–96, 1964.

[59] K. Rothe. *Lösungsverfahren für nichtlineare Matrixeigenwertaufgaben mit Anwendung auf die Ausgleichselementmethode.* PhD thesis, Universität Hamburg, 1989.

[60] A. Ruhe. Algorithms for the Nonlinear Eigenvalue Problem. *SIAM Journal on Numerical Analysis*, 10(4):674–689, 1973.

[61] Y. Saad. *Iterative Methods for Sparse Linear Systems: Second Edition.* Society for Industrial and Applied Mathematics, 2003.

[62] K. Schreiber. *Nonlinear Eigenvalue Problems: Newton-type Methods and Nonlinear Rayleigh Functionals.* VDM Verlag Dr. Müller, 2008.

[63] H. Schwetlick and K. Schreiber. Nonlinear Rayleigh functionals. *Linear Algebra and its Applications*, 436:3991–4016, 2012.

[64] V. Simoncini and L. Eldén. Inexact Rayleigh quotient-type methods for eigenvalue computations. *BIT*, 42:159 – 182, 2002.

[65] G.L.G. Sleijpen, A.G.L. Booten, D.R. Fokkema, and H.A. van der Vorst. Jacobi–Davidson type methods for generalized eigenproblems and polynomial eigenproblems. *BIT Numerical Mathematics*, 36(3):595–633, 1996.

[66] G.L.G. Sleijpen and H.A. van der Vorst. A Jacobi-Davidson iteration method for linear eigenvalue problems. *SIAM Journal on Matrix Analysis and Applications*, 17:401–425, 1996.

[67] G.L.G. Sleijpen, H.A. van der Vorst, and M. van Gijzen. Quadratic eigenproblems are no problem. *SIAM News*, 8:9–10, 1996.

[68] P. Smit and M.H.C. Paardekooper. The effects of inexact solvers in algorithms for symmetric eigenvalue problems. *Linear Algebra and its Applications*, 287:337–357, 1999. Special issue celebrating the 60th birthday of Ludwig Elsner.

[69] D.B. Szyld, E. Vecharynski, and F. Xue. Preconditioned Eigensolvers for Large-Scale Nonlinear Hermitian Eigenproblems with Variational Characterizations. II. Interior Eigenvalues. *SIAM Journal on Scientific Computing*, 37:A2969–A2997, 2015.

[70] D.B. Szyld and F. Xue. Local convergence analysis of several inexact Newton-type algorithms for general nonlinear eigenvalue problems. *Numerische Mathematik*, 123(2):333–362, 2013.

[71] F. Tisseur. Backward Error and Condition of Polynomial Eigenvalue Problems. *Linear Algebra and its Applications*, 309:339–361, 1999.

[72] F. Tisseur and K. Meerbergen. The Quadratic Eigenvalue Problem. *SIAM Review*, 43(2):235–286, 2001.

[73] H. Voss. A maxmin principle for nonlinear eigenvalue problems with application to a rational spectral problem in fluid-solid vibration. *Applications of Mathematics*, 48(6):607–622, Dec 2003.

[74] H. Voss. An Arnoldi Method for Nonlinear Eigenvalue Problems. *BIT Numerical Mathematics*, 44(2):387–401, 2004.

[75] H. Voss. A Jacobi-–Davidson Method for Nonlinear Eigenproblems. In Marian Bubak, Geert Dick van Albada, Peter M.A. Sloot, and Jack Dongarra, editors, *Computational Science - ICCS 2004*, volume 3037 of *Lecture Notes in Computer Science*, pages 34–41. Springer Berlin Heidelberg, 2004.

[76] H. Voss. Iterative projection methods for computing relevant energy states of a quantum dot. *Journal of Computational Physics*, 217(2):824 – 833, 2006.

[77] H. Voss. A new justification of the Jacobi-–Davidson method for large eigenproblems. *Linear Algebra and its Applications*, 424:448 – 455, 2006.

[78] H. Voss. A Jacobi—Davidson method for nonlinear and nonsymmetric eigenproblems. *Computers and Structures*, 85(17–18):1284 – 1292, 2007. Computational Structures Technology.

[79] H. Voss. A minmax principle for nonlinear eigenproblems depending continuously on the eigenparameter. *Numerical Linear Algebra with Applications*, 16(11-12):899–913, 2009.

[80] H. Voss. private communication, 2012.

[81] H. Voss. Nonlinear eigenvalue problems. In L. Hogben, editor, *Handbook of Linear Algebra, Second Edition*, Discrete Mathematics and Its Applications, chapter 60. Taylor & Francis, 2013.

[82] H. Voss and B. Werner. A Minmax Principle for Nonlinear Eigenvalue Problems with Applications to Nonoverdamped Systems. *Mathematical Methods in the Applied Sciences*, 4(1):415–424, 1982.

[83] J.H. Wilkinson, editor. *The Algebraic Eigenvalue Problem.* Oxford University Press, Inc., New York, NY, USA, 1988.

[84] G. Wu. The Convergence of Harmonic Ritz Vectors and Harmonic Ritz Values, Revisited. *SIAM Journal on Matrix Analysis and Applications*, 38(1):118–133, 2017.