Nonlinear Eigenvalue Problems: Newton-type Methods and Nonlinear Rayleigh Functionals

vorgelegt von

Dipl.-Math. Kathrin Schreiber

aus Weimar

Von der Fakultät II - Mathematik und Naturwissenschaften
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktorin der Naturwissenschaften

- Dr. rer. nat. -

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr. John M. Sullivan
Berichter:	Prof. Dr. Hubert Schwetlick
Berichter:	Prof. Dr. Volker Mehrmann
Gutachter:	Prof. Dr. Heinrich Voß

Tag der wissenschaftlichen Aussprache: 09.05.2008

Berlin 2008

D 83

Danksagung

Die vorliegende Dissertation wäre nicht ohne Prof. Dr. Hubert Schwetlick entstanden, dem ich auf das Herzlichste danken möchte. Ich danke ihm für das Vertrauen, das er mir entgegenbrachte, für die Freiheit, bei der Wahl der Schwerpunkte meinen Interessen nachzugehen, und für die Geduld, die er mir und meinen Fragen entgegenbrachte, sowie für die vielen kurzweiligen Stunden, in denen ich von ihm lernen konnte.

Ich danke Prof. Dr. Volker Mehrmann ebenso herzlich für seine fachliche Unterstützung und die vielen Anmerkungen, Hinweise und kurzfristiges Korrekturlesen, wie dafür, dass er mir die Möglichkeit gab, nach Berlin zu kommen in eine wundervolle, inspirierende Arbeitsumgebung.

Beiden muss ich dafür danken, dass sie Mittel und Wege fanden, mich über die letzten drei Jahre zu finanzieren.

Herrn Prof. Dr. Voß danke ich für seine Hinweise zur Geschichte von Rayleigh Funktionalen, und für die freundliche Bereiterklärung, als Gutachter zu fungieren.

Ich danke Elias Jarlebring für die Bereitstellung der Matrizen eines Beispiels.

Ohne meine Freunde hätte ich kaum die Kraft, Ausdauer und Geduld gehabt, die diese Arbeit abverlangten. Deswegen möchte ich allen für die schöne Zeit in Dresden und Berlin – für die Gestaltung der Zeit jenseits vom scharfen Nachdenken – danken, sowohl Freunden als auch Kollegen. Insbesondere Dr. Sonja Schmelter, die so oft meine erste Ansprechpartnerin war bei allen allgemeinen und LaTeX 2_ε-Fragen, versüßte mir den Dissertationsalltag und die zahlreichen Tassen Kaffee.

Ganz besonders möchte ich meiner Familie danken, auf die ich mich immer verlassen konnte. Meine Eltern (die besten, die es gibt) gaben mir Mut, Rückhalt und Kraft, seit ich denken kann. Alexander Hoyer danke ich aus tiefstem Herzen für die Unterstützung, dass er für mich da ist und mir so viel Liebe schenkt.

To my parents

Abstract

Nonlinear eigenvalue problems arise in many fields of natural and engineering sciences. Theoretical and practical results are scattered in the literature and in most cases they have been developed for a certain type of problem. In this thesis we consider the most general nonlinear eigenvalue problem without assumptions on the structure or spectrum. We start by providing basic facts on the conditioning of a simple eigenvalue and an inverse operator representation in terms of the singular value decomposition. The main part of the thesis connects Newton-type methods for nonlinear eigenvalue problems and nonlinear Rayleigh functionals.

The one-sided and the two-sided/generalized Rayleigh functional are introduced with complex range, in contrast to the one-sided functional of the literature. Such functionals are the generalizations of Rayleigh quotients for matrices. Local uniqueness and bounds for the distance to the exact eigenvalue in terms of the angles of eigenvectors and approximations to eigenvectors are derived. We obtain a first order perturbation bound which is used to show stationarity, and which implies the Lipschitz continuity of the functionals.

With the so gained knowledge on Rayleigh functionals, we design new basic methods for the computation of eigenvalues and -vectors: The two–sided Rayleigh functional iteration is the analogon to the two-sided Rayleigh quotient iteration for nonnormal matrices, and is shown to converge locally cubically as well. These methods are important for subspace extension methods like Jacobi–Davidson and nonlinear Arnoldi, that need to solve small dimensional projected nonlinear problems. We compare the methods regarding computational costs and initial assumptions, and show numerical results.

A technique to accelerate convergence for methods computing left and right eigenvector is introduced and the convergence improvement is shown in terms of the R-order. The new methods are called half-step methods. The half-step two-sided Rayleigh functional iteration is shown to converge with R-order 4.

Using the previous results, we discuss various nonlinear Jacobi–Davidson methods. The proposed methods are compared theoretically regarding asymptotic condition numbers,

and in practice with examples.

The special case of nonlinear complex symmetric eigenvalue problems is examined. We show the appropriate definition of a complex symmetric Rayleigh functional, which is used to derive a complex symmetric Rayleigh functional iteration which converges locally cubically, and the complex symmetric residual inverse iteration method. A complex symmetric Jacobi–Davidson algorithm is formulated and tested numerically.

Zusammenfassung

Nichtlineare Eigenwertprobleme entstehen in vielen Bereichen der Natur- und Ingenieurwissenschaften. Obwohl sie seit mehreren Jahrzehnten Gegenstand der mathematischen Forschung sind, ist die Lösung oftmals schwierig und benötigt besonderen Aufwand, insbesondere durch die Vielfalt und unterschiedlichen Anforderungen der einzelnen Problemstellungen. In der Literatur finden sich verstreute Beiträge zur Analysis und verschiedene Algorithmen zur Lösung, oft angepasst an die Struktur einer speziellen Klasse von Problemen oder eines einzelnen Problems.

Diese Arbeit beschäftigt sich mit dem nichtlinearen Eigenwertproblem in seiner allgemeinsten Form unter möglichst wenigen Annahmen. Es werden zuerst Darstellungen für die Konditionszahl eines einfachen Eigenwertes und den inversen Operator entwickelt. Der Hauptteil der Arbeit befasst sich mit Newton-Verfahren in Zusammenhang mit nichtlinearen Rayleigh-Funktionalen, das bedeutet, die Eigenvektoren werden durch Newtonschritte verbessert und die Eigenwertnäherungen als Rayleigh-Funktionale neu bestimmt.

Das einseitige und das zweiseitige, auch verallgemeinerte, Rayleigh-Funktional werden eingeführt auf einer komplexen Menge – im Gegensatz zum in der Literatur gebräuchlichen einseitigen reellen Funktional. Die so definierten Funktionale sind die natürliche Erweiterung der Rayleigh-Quotienten für Matrizen auf nichtlineare Eigenwertprobleme. Lokale Eindeutigkeit und Schranken für den Abstand von Funktional und exaktem Eigenwert in Termen der Winkel zwischen Eigenvektoren und Eigenvektorapproximationen werden bewiesen. Eine Störungsschranke erster Ordnung wird hergeleitet, mit deren Hilfe Stationarität unter bestimmten Voraussetzungen sowie die Lipschitz-Stetigkeit des Funktionals gezeigt werden.

Mit dem Wissen über Rayleigh-Funktionale werden neue Verfahren für die Berechnung von Eigenwerten und -vektoren entwickelt und deren Konvergenz bewiesen. Diese Verfahren sind essentiell für Unterraum-erweiternde Verfahren wie das Jacobi–Davidson-Verfahren oder das nichtlineare Arnoldi Verfahren, da diese niedrig dimensionierte projizierte nichtlineare Eigenwertprobleme im Inneren der Algorithmen lösen müssen. Es wird bewiesen,

dass die zweiseitige Rayleigh-Funktional-Iteration lokal kubisch konvergiert. Die neuen Verfahren werden bezüglich Aufwand und Voraussetzungen wie auch anhand numerischer Beispiele mit bekannten verglichen.

Eine Technik, die die R-Ordnung der Konvergenz von den Verfahren, die Links- und Rechtseigenvektoren berechnen, erhöht, wird aufgezeigt und analysiert. Die so entstandenen neuen Verfahren erhalten das zusätzliche Attribut *halb-Schritt*. Die halb-Schritt zweiseitige Rayleigh-Funktional-Iteration konvergiert mit R-Ordnung 4.

Aufbauend auf den Aussagen über Rayleigh-Funktionale und ein- und zwei-Vektor Methoden werden verschiedene Varianten des nichtlinearen Jacobi–Davidson Verfahrens vorgeschlagen und analysiert, und zwar unter anderem von theoretischer Seite bezüglich asymptotischer Konditionszahlen, als auch anhand von diversen numerischen Beispielen.

Der Spezialfall des nichtlinearen komplex symmetrischen Eigenwertproblems wird diskutiert und ein entsprechendes komplex symmetrisches Rayleigh-Funktional definiert. Störungsabschätzungen erster Ordnung und die Stationarität des Funktionals werden gezeigt, lokal kubische Konvergenz für die komplex symmetrische Rayleigh-Funktional-Iteration bewiesen. Diese bildet die Grundlage für das komplex symmetrische Jacobi–Davidson-Verfahren, das formuliert und getestet wird.

Notation

λ_*	exact eigenvalue
x_*	exact right eigenvector
y_*	exact left eigenvector
(λ_*, x_*)	eigenpair
(λ_*, x_*, y_*)	eigentriple
u	right eigenvector approximation
v	left eigenvector approximation
$\bar{\lambda}$	complex conjugate of λ
M^H	$\bar{M}^T$
α	$y_*^H \dot{T}(\lambda_*) x_*$
$\tilde{\alpha}$	$x_*^H \dot{T}(\lambda_*) x_*$
$I,\ I_n$	identity matrix of order n
e_n	n-th unit vector
(u, v)	$v^H u$, scalar product
$(u, v)_T$	$v^T u$
$S(\lambda, \tau)$	$\{\lambda_0 \in \mathbb{C} :\ \lvert\lambda_0 - \lambda\rvert < \tau\}$
$\bar{S}(\lambda, \tau)$	$\{\lambda_0 \in \mathbb{C} :\ \lvert\lambda_0 - \lambda\rvert \leq \tau\}$
$\mathcal{K}_\varepsilon(x_*)$	$\{u \in \mathbb{C}^n :\ \measuredangle(\mathrm{span}\{u\}, \mathrm{span}\{x_*\}) \leq \varepsilon\}$
$p(u)$	nonlinear (one-sided) Rayleigh functional
$p(u, v)$	nonlinear generalized/two-sided Rayleigh functional
$p_T(u)$	nonlinear complex symmetric Rayleigh functional
$p_L \equiv p_L(\lambda, u, v)$	generalized Rayleigh quotient
$p_N \equiv p_L(\lambda, u, u)$	generalized one-sided Rayleigh quotient
$p_{LT} \equiv p_{LT}(\lambda, u)$	generalized complex symmetric Rayleigh quotient
$q \equiv q(\lambda, u)$	$\dot{T}(\lambda) u$
$\lVert\,.\,\rVert$	Euclidean norm or spectral norm
∇f	grad$f$

$\partial_x f$	$\frac{\partial f}{\partial x}$
$\mathrm{Re}(x)$	real part of x
$\mathrm{Im}(x)$	imaginary part of x

Contents

Chapter 1

Introduction

1.1 Motivation I

One of the active and still open fields in numerical analysis is the topic of nonlinear eigenvalue problems. Nonlinear eigenvalue problems arise in many applications in physics, chemistry and engineering sciences as is shown in §1.2. One of the most intriguing problems in the field of solving nonlinear eigenproblems is the fact that *"there are essentially no analogous packages that reach the standard of those for linear problems"* [58]. Indeed, it is hard to develop such packages, respectively methods, since nonlinear problems are versatile in structure and size, and it is always advisable to use a method that is particularly suited and adjusted to the structure. Even identifying the structure may be complicated as the example of getting to know whether a quadratic problem is hyperbolic, see [28], shows, cf. §1.2. However, before such advanced methods/packages can be developed, we have to understand the basics completely. We need to understand why some methods work when others do not. This work aims in this direction by assuming as little structure as possible, such that the results are valid for a large class of problems. Applying a certain structure immediately gives adapted results.

A motivation describing also the contents of this thesis is given by the suggestion: *"More research is needed on the development of methods and the appropriate perturbation and error analysis, using the original data of the problem and not the matrix pencil arising from the linearization"* [58]. In this sense, we consider the general nonlinear eigenvalue problem, which, in general, cannot be linearized, and analyze approximations for eigenvalues given by different Rayleigh functionals. This helps to understand the so far unclear parts of (Newton-type) algorithms incorporated in nonlinear Rayleigh–Ritz steps and implies strategies to adapt algorithms for problems of a certain structure.

1.2 The Nonlinear Eigenvalue Problem

The *nonlinear eigenvalue problem* for a nonlinear operator $T(\,.\,) : \mathbb{C} \to \mathbb{C}^{n\times n}$ is to find a pair $(\lambda_*,\, x_*) \in \mathbb{C} \times \mathbb{C}^n \setminus \{0\}$, namely an *eigenpair*, such that

$$T(\lambda_*)x_* = 0. \tag{1.1}$$

If the adjoint operator equation $T(\lambda_*)^H y_* = 0$ is satisfied for $y_* \in \mathbb{C}^n \setminus \{0\}$, then (λ_*, x_*, y_*) is called *eigentriple* for T with *right eigenvector* x_*, *left eigenvector* y_*, respectively, corresponding to the *eigenvalue* λ_*. The set $\lambda(T(\,.\,))$ containing all eigenvalues of $T(\,.\,)$ is called *spectrum*

$$\lambda(T(\,.\,)) := \{\lambda \in \mathbb{C} :\ \lambda \text{ is an eigenvalue of } T(\,.\,)\}.$$

Definition (1.1) includes the *linear eigenvalue problem*

$$(A - \lambda I)x = 0, \quad A \in \mathbb{C}^{n\times n},$$

where $T(\lambda) = A - \lambda I$, the *generalized eigenvalue problem* $T(\lambda)x = (A - \lambda B)x = 0$, and the *quadratic eigenvalue problem*

$$T(\lambda)x = (\lambda^2 M + \lambda C + K)x = 0, \tag{1.2}$$

where, in general, M is referred to as *mass matrix*, K as *stiffness matrix* and C as *damping matrix*, if the problem governs a vibrating system, but it could also be originated in a *gyroscopic* system, cf. [87]. These examples are special cases of the *polynomial eigenvalue problem* of degree m and order n given by

$$T(\lambda)x = \sum_{l=0}^{m} \lambda^l A_l x = 0 \quad \text{with} \quad A_l \in \mathbb{C}^{n\times n}, \tag{1.3}$$

where $T(\lambda)$ is sometimes called *matrix polynomial* or λ-*matrix*. Polynomial eigenvalue problems often arise in the analysis of vibrating systems like machines, buildings and vehicles, where solutions of the eigenproblem are called *eigenfrequencies* and *eigenmodes*, cf. [25, 49].

Non-polynomial eigenproblems can be split into *rational eigenvalue problems*, for instance

$$T(\lambda)x = -Kx + \lambda Mx + \sum_{l=1}^{m} \frac{\lambda}{\sigma_l - \lambda} C_l x = 0$$

[96], and *irrational eigenvalue problems*, which are genuinely nonlinear, for instance

$$T(\lambda)x = \left(K - \lambda M + \mathrm{i} \sum_{l=1}^{k} \sqrt{\lambda - \sigma_l^2}\, W_l \right) x = 0, \tag{1.4}$$

where K, M, W_l are $n \times n$ real symmetric matrices, K is positive semi-definite, M is positive definite and σ_l are given nonnegative scalars [52]. This problem arises, e.g., from a Nedelec-type finite element discretization of the frequency domain Maxwell equation with waveguide boundary conditions in waveguide loaded cavity design.

The energy level calculation of semiconductor nanostructures yields symmetric rational eigenvalue problems based on the Schrödinger equation, which admits variational principles in order to compute the stationary states, see [11, 12].

Truly nonlinear eigenvalue problems arise in the stability analysis of vibrating systems under state feedback control [40, 84]. Delay-differential equations emerge when phenomena like transmission, transportation and inertia are modelled [62]. When determining properties of such time-delay systems, it is useful to consider the associated nonlinear eigenvalue problem

$$T(\lambda)x = \left(-\lambda I + A_0 + \sum_{l=1}^{m} A_k e^{-h_l \lambda}\right) x = 0,$$

given here with m delays, where h_l is a positive real number and $A_l \in \mathbb{R}^{n\times n}$ [43]. See [15] for a nonlinear eigenvalue problem that occurs in the investigation of time-dependent vortex-structures in a Taylor–Couette apparatus.

In most cases, the underlying issue that produces the eigenproblem, is to solve the homogeneous equation corresponding to an eigenvalue problem of a (partial) differential equation, where matrices arise when operators on spaces of infinite dimension are discretized. Therefore, the finer the grid is chosen, the larger the dimension $n \gg 1$ becomes. In general, the discretization will be large and sparse, and, mostly, just some special eigenvalues $\lambda \in \mathbb{C}$ are of interest, for example those of smallest absolute value, of smallest real part or those which are closest to the imaginary axis. However, as we will see in Chapter 6, the key to the solution of the large and sparse eigenproblems is the solution of small dimensional dense eigenproblems, which occur when the problem is projected by promising eigenvector approximations. Hence, when we want to develop and analyze methods for large sparse nonlinear eigenvalue problems, it is important to study methods for dense nonlinear eigenvalue problems as well. We will do this in Chapter 4.

The survey on nonlinear eigenvalue problems [58] provides a wide range of further examples.

Let us give some technical definitions. We call $T(\lambda)$ *regular* if $\det T(\lambda)$ is not identically zero for all λ, and *singular* otherwise. The most common definition of a *simple eigenvalue*

λ_* states that λ_* is *algebraically simple*, i.e.,

$$\left.\frac{\partial(\det(T(\lambda)))}{\partial\lambda}\right|_{\lambda_*} \neq 0. \tag{1.5}$$

An algebraically simple eigenvalue λ_* is *geometrically simple* automatically, i.e., there exist nonzero vectors x_*, y_* such that

$$\ker(T(\lambda_*)) = \operatorname{span}\{x_*\}, \quad \ker(T(\lambda_*)^H) = \operatorname{span}\{y_*\}, \tag{1.6}$$

which is not obvious in case of nonlinear eigenvalue problems and will therefore be shown in Chapter 2.

However, in [49, p. 42f] another definition of a simple eigenvalue for matrix polynomials is given. There, a regular λ-matrix is said to be simple, if for an eigenvalue with algebraic multiplicity $m_* \geq 1$, the geometric multiplicity is m_*, too. Moreover, $T(\lambda)$ is called a simple λ-matrix polynomial if $T(\lambda_*)$ is simple for all eigenvalues λ_*. In this thesis, Definition (1.5) will be used.

Common practice and the classical approach for solving a regular polynomial eigenvalue problem of degree m and order n is to reduce the problem to a generalized linear eigenvalue problem $(A - \lambda B)\tilde{x} = 0$ by enlarging the system matrix via *linearization*, see [58] for a definition. The eigenvalues are equal for both problems. A well-known linearization is given by the *first companion form*

$$\left(\lambda\begin{bmatrix} A_m & 0 & \cdots & 0 \\ 0 & I_n & 0 & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & I_n \end{bmatrix} + \begin{bmatrix} A_{m-1} & A_{m-2} & \cdots & A_0 \\ -I_n & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & -I_n & 0 \end{bmatrix}\right)\tilde{x} = 0 \quad \text{with} \quad \tilde{x} = \begin{bmatrix} \lambda^{m-1}x \\ \lambda^{m-2}x \\ \vdots \\ x \end{bmatrix}. \tag{1.7}$$

Besides the first companion form various linearizations, as the second companion form, exist and have been analyzed in the past few years in order to find representations matching the structure of the original problem or to treat particular parts of the spectrum, which are of interest, especially well, cf. [34, 36, 54]. A comprehensive theory for matrix polynomials including linearization is given in [25].

Rational eigenvalue problems can be turned into polynomial eigenvalue problems by multiplying with an appropriate scalar polynomial in λ. However, as in the case of linearizing polynomial problems, this may worsen the performance of numerical solvers since spurious eigenvalues are generated by the roots of the denominators and the degree of the new polynomial problem may become large.

Another way of distinguishing eigenvalue problems is to consider and separate them by the underlying structure of the involved matrices. This may reveal valuable information on the spectrum. Even more important is the fact that one should always take into account this structure when treating a special problem and use this knowledge in perturbation analysis and numerical issues.

We identify nonlinear eigenvalue problems as follows

- $T(\cdot)$ is *real symmetric* if $T(\cdot) : \mathbb{R} \to \mathbb{R}^{n\times n}$, $T(\lambda) = T(\lambda)^T,\ \forall\lambda \in \mathbb{R}$
- $T(\cdot)$ is *complex symmetric* if $T(\cdot) : \mathbb{C} \to \mathbb{C}^{n\times n}$, $T(\lambda) = T(\lambda)^T,\ \forall\lambda \in \mathbb{C}$
- $T(\cdot)$ is *Hermitian* if $T(\cdot) : \mathbb{C} \to \mathbb{C}^{n\times n}$, $T(\lambda)^H = T(\bar{\lambda}),\ \forall\lambda \in \mathbb{C}$.

These identifiers imply the properties given in Table 1.1 indicating the type of eigentriples corresponding to the structure.

The complex symmetric structure is observed easily, since one will know whether the problem is symmetric, and eigenvalues are complex in most cases, which leads to a complex symmetric problem. However, one cannot tell at first sight whether the problem is Hermitian with real eigenvalues only. For a Hermitian matrix A the linear problem $(A-\lambda I)x = 0$ has only real eigenvalues, and right and left eigenvectors are the same. This is not the case for the generalized linear problem $Ax = \lambda Bx$ with Hermitian A and B, in general, and all higher order polynomial eigenproblems with Hermitian coefficients, where with λ also $\bar{\lambda}$ is an eigenvalue, but left and right eigenvectors corresponding to λ are different.

A small example for a generalized problem with Hermitian matrices is given by

$$T(\lambda)x = (A - \lambda B)x = 0 \qquad \text{with} \quad A = \begin{bmatrix} 0 & 1 \\ 1 & 2 \end{bmatrix},\ B = \begin{bmatrix} 0 & \mathrm{i} \\ -\mathrm{i} & 0 \end{bmatrix}, \tag{1.8}$$

the eigenvalues of which are $\lambda_1 = \mathrm{i}$ and $\lambda_2 = -\mathrm{i}$. Here we have $T(\lambda) \neq T(\lambda)^H$ and $T(\lambda) \neq T(\lambda)^T$, but $T(\lambda)^H = T(\bar{\lambda})$, i.e., $T(\lambda)$ is Hermitian.

Type of T	Eigentriple
real symmetric	(λ, x, x) if $\lambda \in \mathbb{R}$
complex symmetric	$(\lambda, x, \bar{x})$
Hermitian	(λ, x, y) and $(\bar{\lambda}, y, x)$, if $\lambda \in \mathbb{R}$: (λ, x, x)

Table 1.1: *Type of the nonlinear eigenproblem and corresponding form of eigentriples containing eigenvalue, right eigenvector and left eigenvector.*

An example for a quadratic problem is given by the nonoverdamped mass-spring system, see [87], where $T(\lambda) = \lambda^2 M + \lambda C + K$ with real symmetric M, C, K, which has complex eigenvalues. Obviously, the operator is Hermitian, but $T(\lambda) \neq T(\bar{\lambda})$, since there are complex eigenvalues. If it had only real eigenvalues, it would be real symmetric, but the complex eigenvalues imply that $T(\lambda)$ is also complex symmetric.

If $T(\lambda)$ is complex symmetric and Hermitian, then the properties shown in Table 1.1 add up, i.e., if λ is an eigenvalue, then $\bar{\lambda}$ is also an eigenvalue and the corresponding eigentriples are given by $(\lambda, x, \bar{x})$ and $(\bar{\lambda}, \bar{x}, x)$.

The problem given in (1.4) is complex symmetric but not Hermitian.

Hermitian matrix polynomials (1.3) with positive definite matrices A_l $(l = 0, \ldots, m)$ that have only real eigenvalues are called *hyperbolic*, [55, p. 169], see also [35]. Hyperbolicity is hard to show in most practical applications, cf. [28], where an iteration to check whether a quadratic eigenvalue problem is hyperbolic is developed.

A generalized Hermitian linear problem $(A - \lambda B)x = 0$ has only real eigenvalues if it is a definite pencil (A, B) [85, Chapter 6], i.e.,

$$\min \left\{ \left((z^H A z)^2 + (z^H B z)^2 \right)^{1/2} : z \in \mathbb{C}^n, \|z\| = 1 \right\} > 0.$$

A sufficient condition to have only real eigenvalues for a Hermitian quadratic eigenproblem is the *overdamping condition* (1.10), which can be obtained by extracting λ_* from the quadratic form

$$x_*^H T(\lambda_*) x_* = \lambda_*^2 x_*^H M x_* + \lambda_* x_*^H C x_* + x_*^H K x_* = 0,$$

which holds for arbitrary matrices M, C, K, and, in case $x_*^H M x_* \neq 0$, yields the eigenvalues of (1.2) as one of the two possibilities

$$\lambda_*^{\pm} \equiv \lambda^{\pm}(x_*) = \frac{-x_*^H C x_* \pm \sqrt{(x_*^H C x_*)^2 - 4(x_*^H K x_*)(x_*^H M x_*)}}{2 x_*^H M x_*}. \tag{1.9}$$

In general, only one of the two is an eigenvalue, the other one is spurious. These roots are real for a Hermitian problem if M and C are positive definite, K is positive semidefinite, and

$$(x_*^H C x_*)^2 - 4(x_*^H K x_*)(x_*^H M x_*) > 0, \tag{1.10}$$

see also [87].

The noteworthiness of complex symmetric problems—though they have no underlying special spectral properties, since any matrix is similar to a complex symmetric matrix [39, p. 209]—is that one can apply methods for Hermitian eigenvalue problems with real

eigenvalues, by changing the inner product $(x, y) = y^H x$ to the bilinear form $(x, y)_T := y^T x$. This is due to the fact, that if (λ_*, x_*) is an eigenpair for T, then $x_*^T T(\lambda_*) = 0$ for complex symmetric problems, whereas the Hermitian problem with real eigenvalues satisfies $x_*^H T(\lambda_*) = 0$ if $T(\lambda_*)x_* = 0$. We discuss complex symmetric nonlinear eigenvalue problems and methods in Chapter 7.

After the outer structure of the problem has been recognized, a second look may reveal inner structure, which is hidden in blocks of matrices. Particularly structured polynomial eigenproblems have received a lot of attention lately. Generalizations of Hamiltonian matrices are represented by *even/odd* polynomials, studied in [59,60], whereas generalizations of symplectic matrices are called *palindromic* polynomials [53]. Several substructures exist and all of them imply certain symmetries of the spectra.

A natural and one of the oldest approaches for solving equation (1.1) is to apply Newton's method to the extended system

$$F_w(x, \lambda) = \begin{pmatrix} T(\lambda)x \\ w^H x - 1 \end{pmatrix} = 0, \tag{1.11}$$

for given starting value and vector (λ_0, x_0), cf. [5, 90] for the linear case. The second equation is a normalization condition where the normalizing vector w, $\|w\| = 1$, has to satisfy $w^H x_* \neq 0$ where x_* is the exact eigenvector. The Newton equation

$$0 = F_w(x_k, \lambda_k) + \partial F_w(x_k, \lambda_k) \begin{bmatrix} x_{k+1} - x_k \\ \lambda_{k+1} - \lambda_k \end{bmatrix}$$

for (1.11) at the current approximation (x_k, λ_k) is equivalent to

$$\begin{bmatrix} T(\lambda_k) & \dot{T}(\lambda_k)x_k \\ w^H & 0 \end{bmatrix} \begin{bmatrix} x_{k+1} - x_k \\ \lambda_{k+1} - \lambda_k \end{bmatrix} = - \begin{bmatrix} T(\lambda_k)x_k \\ w^H x_k - 1 \end{bmatrix} \Leftrightarrow \begin{array}{rcl} T(\lambda_k)x_{k+1} &=& (\lambda_k - \lambda_{k+1})\dot{T}(\lambda_k)x_k \\ w^H x_{k+1} &=& 1. \end{array} \tag{1.12}$$

Multiplying the upper equation by $w^H T(\lambda_k)^{-1}$ gives the actual iteration

$$\lambda_{k+1} = \lambda_k - \frac{1}{w^H T(\lambda_k)^{-1}\dot{T}(\lambda_k)x_k}, \tag{1.13}$$

$$x_{k+1} = \frac{T(\lambda_k)^{-1}\dot{T}(\lambda_k)x_k}{w^H T(\lambda_k)^{-1}\dot{T}(\lambda_k)x_k} = (\lambda_k - \lambda_{k+1})T(\lambda_k)^{-1}\dot{T}(\lambda_k)x_k, \tag{1.14}$$

provided that $\lambda_k \notin \lambda(T(\,.\,))$ and $w^H T(\lambda_k)^{-1}\dot{T}(\lambda_k)x_k \neq 0$. Whenever a new vector x_{k+1} is obtained as

$$x_{k+1} = \omega_k T(\lambda_k)^{-1}\dot{T}(\lambda_k)x_k,$$

i.e., as solution of $T(\lambda_k)x_{k+1} = \omega_k \dot{T}(\lambda_k)x_k$, for some scalar $\omega_k \neq 0$, we also speak of an *inverse iteration* step with shift λ_k applied to x_k.

Variations of Newton's method and the inverse iteration method have been studied further in [4,61,65,72]. Surveys of methods for nonlinear eigenvalue problems are given in [58,96].

We will discuss various methods for computing eigenpairs and -triplets in Chapters 4, 5 and 6, and give a detailed analysis of the Newton step (1.12), in different versions.

1.3 A Summary of Variational Principles for the Characterization of Eigenvalues

Although the assumptions needed in this section are rather strict, we want to review existing variational principles for nonlinear eigenvalue problems. Variational principles are useful for characterizing and locating the eigenvalues of an eigenproblem. For linear Hermitian operators such results are the principle of Rayleigh [69], the minmax characterization of Poincaré [68] and the maxmin principle of Courant [18], Fischer [21] and Weyl [103]. In order to find such principles for nonlinear eigenvalue problems the concept of a *Rayleigh functional* was introduced in [19] for quadratic overdamped systems and further treated in [30, 31, 70] for multi-parametric and nonlinear operators. We use the definition from [30] for real symmetric T: Let p be a continuous real-valued functional on $\mathbb{R}^n \setminus \{0\}$ subject to the following properties

$$p : u \in \mathbb{R}^n \setminus \{0\} \longrightarrow p(u) \in (a, b), \tag{1.15}$$

$$p(cu) = p(u) \quad \text{for all } c \neq 0, \tag{1.16}$$

$$(T(p(u))u, u) = 0, \tag{1.17}$$

$$(\dot{T}(p(u))u, u) > 0. \tag{1.18}$$

The functional p is called *Rayleigh functional* for T, and the set

$$W(T) = \{p(u) : \ u \in \mathbb{R}^n \setminus \{0\}\} \subset \mathbb{R}$$

is called the *range* of T.

All cited authors considered mostly real-valued, continuous p and symmetric operators. If they have considered also complex vectors, then they have assumed existence and differentiability of $p(u)$ beforehand. For real vectors u, the implicit function theorem can be applied to obtain locally unique existence and differentiability of p up to the same order as $T(\lambda)$ is differentiable, provided that u is sufficiently close to the exact eigenvector x_*.

The assumption (1.18) is also referred to as overdamping condition, although it is not equivalent to the overdamping condition (1.10) for quadratic eigenvalue problems. To see this, consider $T(\lambda) = \lambda^2 M + \lambda C + K$, hence $\dot{T}(\lambda) = 2\lambda M + C$. Then, for $u = x_*$, $p(u) = x_*$, condition (1.18) gives $\lambda_* > -x_*^H C x_*/(2x_*^H M x_*)$—a condition that is different from (1.10).

Nevertheless, (1.18) guarantees that the desired eigenvalue is algebraically simple for Hermitian operators with real eigenvalues, provided that it is geometrically simple.

It is not obvious how to find this continuous mapping. Consider for instance the quadratic eigenvalue problem $T(\lambda)x = (\lambda^2 M + \lambda C + K)x = 0$. Then, the Rayleigh functional is given by equation (1.9). Here, we could assign

$$p(u) = \lambda^+(u) = \frac{-u^H C u + \sqrt{(u^H C u)^2 - 4(u^H K u)(u^H M u)}}{2u^H M u}$$

in order to have a continuous mapping, but it is not clear whether this will give the eigenvalue. This question is neither asked nor answered in the cited publications.

Chapter 3 discusses and extends this restricted definition of a Rayleigh functional to arbitrary problems with complex eigenvalues and introduces appropriate functionals, where we focus on a local approach.

However, in order to obtain variational principles it is necessary to have Hermitian operators and subsets of real eigenvalues. For continuous, real-valued p such that (1.15)-(1.18) hold, the following Lemma was shown in [70].

Lemma 1.1 *[70] Suppose that $T(\lambda)$ is Hermitian with real eigenvalues and let* (1.15)-(1.18) *hold. Suppose that $\lambda \in W(T)$ and $u \neq 0$. Then, we have*

$$\begin{aligned} p(u) < \lambda &\iff (T(\lambda)u, u) > 0, \\ p(u) = \lambda &\iff (T(\lambda)u, u) = 0, \\ p(u) > \lambda &\iff (T(\lambda)u, u) < 0. \end{aligned}$$

This lemma states the crucial point on the way to the minmax characterization given in the next theorem, which was formulated by Rogers [70].

Theorem 1.2 *[30] Under the assumptions of Lemma 1.1, let the eigenvalues of T be counted regarding their geometric multiplicity. Then, there exist n eigenvalues in $\lambda(T(\,.\,))$, namely $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. There exists a basis $\{x_j\}$ of eigenvectors in $\mathbb{R}^n$, such that $T(\lambda_j)x_j = 0$, for $j = 1, \ldots n$. Moreover, the eigenvalues are characterized by the minimum-maximum principle*

$$\lambda_j = \min_{\mathbb{R}^j} \max_{u \in \mathbb{R}^j \setminus \{0\}} p(u) \qquad (j = 1, \ldots, n),$$

where $\mathbb{R}^j$ denotes an arbitrary subspace of $\mathbb{R}^n$ of dimension j.

This theorem was also proved for finite dimensional quadratic overdamped eigenproblems in [19].

In [30] the validity of the Rayleigh principle is proved for overdamped systems given in Theorem 1.3 by showing the existence of a complete set of orthogonal eigenvectors with respect to the following generalized inner product

$$[x,y] = \begin{cases} (\Delta(p(x),p(y))x,y) & \text{if} \quad x,y \neq 0 \\ 0 & \text{if} \quad x=0 \text{ or } y=0, \end{cases} \tag{1.19}$$

where, for $\alpha, \beta \in (a,b)$,

$$\Delta(\alpha,\beta) = \begin{cases} \frac{1}{\alpha-\beta}[T(\alpha)-T(\beta)] & \text{if} \quad \alpha \neq \beta \\ \dot{T}(\alpha) & \text{if} \quad \alpha = \beta, \end{cases}$$

which is symmetric, definite and homogeneous, but in general not bilinear.

Theorem 1.3 *[30] Under the assumptions of Lemma 1.1, let the eigenvalues of T be ordered in the following way $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. Let $x_1, \ldots, x_n$ be a system of eigenvectors such that $[x_i, x_j] = \delta_{ij}$ and $T(\lambda_j)x_j = 0$. Then, we have*

$$\lambda_j = \min_{[x_k,x]=0,\, x\neq 0,\, k=1,\ldots,j-1} p(x) \qquad (j=1,\ldots,n).$$

However, the existence of a system of such linearly independent eigenvectors is not guaranteed, as we will see in the example in equation (2.3), which satisfies (1.15)–(1.18).

Generalizations of the minmax principle by Poincaré and the maxmin principle of Courant, Fischer and Weyl for infinite dimensional overdamped systems were addressed by Hadeler [31], Langer [50], Rogers [70], Turner [88, 89] and Werner [102].

Voss and Werner [99] examined nonoverdamped problems where the Rayleigh functional is defined only on a proper subset, and proved a minmax principle generalizing the characterization of Poincaré. Voss [92] added, more recently, the maxmin characterization corresponding to the characterization of Courant, Fischer and Weyl, and a minmax principle for nonlinear eigenproblems which are continuous in λ, but do not need to be differentiable [98].

1.4 Outline

This thesis is structured as follows. Chapter 2 shows some basic facts for nonlinear eigenvalue problems including a representation of the inverse operator in terms of the singular value decomposition and the condition number of a simple eigenvalue. Chapter 3 considers a one-sided and a two-sided (complex) Rayleigh functional but with less restricting assumptions than proposed in the literature in order to apply to a larger class of nonlinear eigenvalue problems. Local uniqueness for the two-sided Rayleigh functional will be derived as well as a first order perturbation result, which is used to show stationarity. The same analysis is done for the one-sided Rayleigh functional. We will see that the important properties of the one- and two-sided Rayleigh quotient carry over. The distance to another generalization of a Rayleigh quotient defined in [49] is analyzed.

In Chapter 4, known "one-vector" methods for nonlinear eigenvalue problems are reviewed and new "two-vector" methods, which compute eigentriples, are developed in the Rayleigh functional framework by including the results of Chapter 3. We introduce the two-sided Rayleigh functional iteration and show its locally cubically convergence, and the two-sided residual inverse iteration, which converges quadratically. The methods are compared from the theoretical point of view and tested by means of numerical examples.

Then in Chapter 5, a technique is developed to accelerate convergence for the methods presented in the second part of Chapter 4, the so-called half-step methods, obtained by this technique, are examined. Convergence improvement is shown in terms of the R-order. Numerical examples show the acceleration of convergence.

We change the type of methods to subspace expanding algorithms in Chapter 6 where various Jacobi–Davidson-type methods are considered following the previously gained knowledge on Rayleigh functionals and on one- and two-vector methods, i.e., the results of the previous chapters are included into considerations. We analyze the condition number of the operators in the asymptotic case. Different examples, including such that represent exceptional cases for the methods, are tested, as for instance problems with large condition number of the eigenvalue and with violated assumptions.

In Chapter 7 we consider complex symmetric nonlinear eigenproblems and derive an appropriate (stationary) Rayleigh functional and the complex symmetric Jacobi–Davidson algorithm. We show cubic convergence of the underlying complex symmetric Rayleigh functional iteration, and analyze the corresponding complex symmetric residual inverse iteration. Numerical examples are provided where significant differences to the standard Jacobi–Davidson method occur. A summary and outlook conclude this thesis in Chapter 8.

1.5 Motivation II

Now that we have introduced the problem, shown examples and defined special terms that we will need, we come back to the question: Why do we do this? The general motivation for the topic was shown at the beginning of this chapter. But why do we want to work especially in the directions given in the outline? Questions that will be answered are inter alia: What is the motivation for the two-sided methods—which are expected to have double costs compared to the one-sided methods? Why do we consider other versions of the well-working Jacobi–Davidson method? And why should one make so much effort on one-, resp. two-sided methods, anyway, since there are almost globally converging methods like the Jacobi–Davidson method?

To amplify diversity in mathematical research is surely necessary, but not the main reason for this research. We will rather show throughout this thesis why some methods work when others don't, which implies the importance of having more than one method in mind, and to understand what is going on from a theoretical point of view.

From an outer point of view, one could think that there exist already perfectly working methods like the nonlinear Jacobi–Davidson and the nonlinear Arnoldi method, and, if it is a polynomial problem we have to deal with, then we can even linearize it (if the polynomial is regular) and solve a linear problem, which is by common knowledge "easy". The number of recent publications regarding special types of linear problems shows, that even this is not the case, see, e.g., the thesis [75].

And still, almost all nonlinear eigenvalue problems are difficult. Often they cannot be linearized, sometimes they should not. And, what we must not forget: Inside of the nonlinear Jacobi–Davidson and Arnoldi methods, we have to solve small and dense nonlinear eigenvalue problems. In fact, JD and Arnoldi are only two different ways of updating the search space. Solving the projected (nonlinear) problem has to be done by some external method. Therefore, it is utterly important to analyze the one- and two-vector methods thoroughly. The methods that we know of are very sensitive with respect to starting values and in a large manner locally convergent. They tend to diverge for poor starting values. In this case, it is pointless to solve correction equations.

We will see that, as in the linear case the Rayleigh quotient, the one-sided Rayleigh functional is not stationary when applied to general problems. That is one motivation to consider two-vector methods, for which the generalized stationary Rayleigh functional can be used. Another motivation, which is also the reason why we have developed the generalized Jacobi–Davidson method, is that the condition number of the inverse operator in the Newton step is directly related to the condition number of the eigenvalue, whereas

this term does not occur for the generalized method. Due to both reasons, we will see that two-sided methods can even be faster than one-sided methods, although they have to solve two instead of one system. Additionally, information on the condition number of the approximated eigenvalue can be achieved, simultaneously.

Chapter 2

Preliminaries and Basic Results

This chapter gives some basic properties of nonlinear eigenvalue problems, parts of which are needed frequently throughout the remainder of this thesis.

2.1 Basic Properties

We start with showing that, like in the linear case, the algebraic simplicity of an eigenvalue implies its geometric simplicity.

Proposition 2.1 *Let λ_* be an algebraically simple eigenvalue of $T(\,.\,) : \mathbb{C} \to \mathbb{C}^{n\times n}$, i.e., $\delta(\lambda_*) = 0$ and $\dot\delta(\lambda_*) \neq 0$, where $\delta(\lambda) := \det(T(\lambda))$. Then, the eigenvalue λ_* is geometrically simple, i.e.,* $\operatorname{rank}(T(\lambda_*)) = n - 1$.

Proof. We prove the assertion indirectly. Let λ_* be algebraically simple, but $\operatorname{rank}(T(\lambda_*)) = n - d$, where $d \geq 2$. Then, there exist permutation matrices P_l, P_r of size $n \times n$ such that

$$P_l T(\lambda_*) P_r^T =: \tilde T(\lambda_*) = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix},$$

where $T_{11} \in \mathbb{C}^{(n-d)\times(n-d)}$ is nonsingular. The blockwise decomposition

$$L(\lambda)\tilde T(\lambda)R(\lambda) = D(\lambda) := \begin{bmatrix} T_{11}(\lambda) & 0 \\ 0 & T_{22}(\lambda) - T_{21}(\lambda)T_{11}^{-1}(\lambda)T_{12}(\lambda) \end{bmatrix},$$

with

$$L(\lambda) = \begin{bmatrix} I_{n-d} & 0 \\ -T_{21}(\lambda)T_{11}^{-1}(\lambda) & I_d \end{bmatrix} \quad \text{and} \quad R(\lambda) = \begin{bmatrix} I_{n-d} & -T_{11}^{-1}(\lambda)T_{12}(\lambda) \\ 0 & I_d \end{bmatrix},$$

of $\tilde{T}(\lambda) = P_l T(\lambda) P_r^T$ exists, if λ is sufficiently close to λ_*, which implies that T_{11} is nonsingular. Since $T(\lambda) = P_l^T \tilde{T}(\lambda) P_r = P_l^T L(\lambda)^{-1} D(\lambda) R(\lambda)^{-1} P_r$, the block diagonal matrix $D(\lambda)$ is equivalent to $T(\lambda)$. Henceforth, the rank of $D(\lambda_*)$ equals the rank of $T(\lambda_*)$, which is $n-d$. By construction, we have $\operatorname{rank}(T_{11}) = n-d$, which implies that the lower block in D is zero, i.e., $S(\lambda_*) := T_{22}(\lambda_*) - T_{21}(\lambda_*) T_{11}^{-1}(\lambda_*) T_{12}(\lambda_*) = 0$. Moreover,

$$\delta(\lambda) = \det T(\lambda) = \det(P_l) \det(L^{-1}) \det(D(\lambda)) \det(R^{-1}) \det(P_r) = \sigma \det(T_{11}(\lambda)) \det(S(\lambda)),$$

with $\sigma \in \{1, -1\}$, since $\det(P_l) = \pm 1$, $\det(P_r) = \pm 1$ and $\det(L^{-1}) = \det(R^{-1}) = 1$. Therefore, differentiating with respect to λ, and inserting λ_*, yields

$$\dot{\delta}(\lambda_*) = \sigma \left[\dot{\delta}_{11}(\lambda_*) \delta_{22}(\lambda_*) + \delta_{11}(\lambda_*) \dot{\delta}_{22}(\lambda_*) \right],$$

where $\delta_{11}(\lambda) = \det(T_{11}(\lambda))$ and $\delta_{22}(\lambda) = \det(S(\lambda))$. We already know that $\delta_{22}(\lambda_*) = 0$, because $S(\lambda_*) = 0$. It remains to show $\dot{\delta}_{22}(\lambda_*) = 0$. To see this, write

$$\delta_{22}(\lambda) = \sum_{\rho} (-1)^{\nu(\rho)} s_{1,\rho(1)} s_{2,\rho(2)} \dots s_{d,\rho(d)},$$

where $s_{j,\rho(j)}$ are the elements of S, $\rho = \{\rho(1), \dots, \rho(d)\}$ are permutations of $j = \{1, \dots, d\}$, and $\nu(\rho)$ states the corresponding sign. Then,

$$\dot{\delta}_{22}(\lambda_*) = \sum_{\rho} (-1)^{\nu(\rho)} \left\{ \dot{s}_{1,\rho(1)} s_{2,\rho(2)} \dots s_{d,\rho(d)} + \dots + s_{1,\rho(1)} \dots s_{d-1,\rho(d-1)} \dot{s}_{d,\rho(d)} \right\} = 0,$$

since the $s_{l,\rho(l)}$ are zero in $S(\lambda_*)$, for $j = 1, \dots d$. This means that $\dot{\delta}(\lambda_*) = 0$, which contradicts the assumption that λ_* is algebraically simple. □

The following proposition characterizes algebraically simple eigenvalues, cf. [78].

Proposition 2.2 *Let λ_* be a geometrically simple eigenvalue of $T(\,.\,) : \mathbb{C} \to \mathbb{C}^{n \times n}$, i.e., $\det T(\lambda_*) = 0$ and $\dim \ker T(\lambda_*) = 1$, $\ker T(\lambda_*) = \operatorname{span}\{x_*\}$, $\ker T(\lambda_*)^H = \operatorname{span}\{y_*\}$, with $\|x_*\| = \|y_*\| = 1$, and let T be differentiable. Then we have*

$$\lambda_* \;\; \textit{is algebraically simple} \iff y_*^H \dot{T}(\lambda_*) x_* \neq 0.$$

Proof. Consider the system

$$C(\lambda) \begin{bmatrix} x \\ \mu \end{bmatrix} := \begin{bmatrix} T(\lambda) & y_* \\ x_*^H & 0 \end{bmatrix} \begin{bmatrix} x \\ \mu \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{2.1}$$

Now set $\lambda = \lambda_*$. Multiplying the upper block by y_*^H yields $y_*^H T(\lambda_*) x + y_*^H y_* \mu = 0$, hence, $\mu = \mu(\lambda_*) = 0$. Using this, the upper block now reads as $T(\lambda_*) x = 0$, hence, $x = x_* \alpha$,

and the normalization condition $x_*^H x = x_*^H x_* \alpha = 1$ gives $x = x(\lambda_*) = x_*$. Therefore, the system is uniquely solvable with $x = x_*$, $\mu = \mu_* = 0$, and $C(\lambda_*)$ is nonsingular. This implies unique solvability also for λ close to λ_*. From the definition of μ by (2.1) we get

$$\mu(\lambda) = e_{n+1}^T C(\lambda)^{-1} e_{n+1}. \tag{2.2}$$

An alternative representation is delivered by applying Cramer's rule to (2.1), namely

$$\mu = \mu(\lambda) = e_{n+1}^T C(\lambda)^{-1} e_{n+1} = \frac{\begin{vmatrix} T(\lambda) & 0 \\ x_*^H & 1 \end{vmatrix}}{\begin{vmatrix} T(\lambda) & y_* \\ x_*^H & 0 \end{vmatrix}} = \frac{\det T(\lambda)}{\det C(\lambda)} =: \frac{\delta(\lambda)}{\gamma(\lambda)}.$$

Hence, $\mu(\lambda)$ is a scaled value of $\delta(\lambda) := \det T(\lambda)$, and $\delta(\lambda) = \mu(\lambda)\gamma(\lambda)$ and $\dot{\delta}(\lambda_*) = \dot{\mu}(\lambda_*)\gamma(\lambda_*) + \mu(\lambda_*)\dot{\gamma}(\lambda_*) = \dot{\mu}(\lambda_*)\gamma(\lambda_*)$. Since $\gamma(\lambda_*) \neq 0$, the eigenvalue is algebraically simple if and only if $\dot{\mu}(\lambda_*) \neq 0$. Differentiating equation (2.2) and inserting λ_* yields

$$\dot{\mu}(\lambda_*) = -e_{n+1}^T C(\lambda_*)^{-1} \dot{C}(\lambda_*) C(\lambda_*)^{-1} e_{n+1} = -\begin{bmatrix} y_* \\ 0 \end{bmatrix}^H \begin{bmatrix} \dot{T}(\lambda_*) & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_* \\ 0 \end{bmatrix} = -{y_*}^H \dot{T}(\lambda_*) x_*,$$

which proves the assertion. Here we have used that the left factor equals

$$C(\lambda_*)^{-H} e_{n+1} = \begin{bmatrix} y \\ \bar{\nu} \end{bmatrix} = \begin{bmatrix} y(\lambda_*) \\ \bar{\nu}(\lambda_*) \end{bmatrix} = \begin{bmatrix} y_* \\ 0 \end{bmatrix}.$$

□

Similar propositions can be found in [3, 61].

Next, we want to list some differences between nonlinear and linear eigenvalue problems. Let (λ_*, x_*, y_*) denote an eigentriple for $T(\lambda)$, with an algebraically simple eigenvalue λ_*. Then

- $y_*^H x_* \neq 0$, which is the simplicity condition in the linear case, does not hold in general. Consider the following example

  $$(A - \lambda B)x = \left(\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \right) x = 0,$$

  which has the two simple eigenvalues $\lambda_1 = 0$, $\lambda_2 = 1$. The right and left eigenvectors for λ_1 are given by $x_1 = [1\ 0]^T$, $y_1 = [0\ 1]^T$, hence, $y_1^T x_1 = 0$.

- The principle of biorthogonality [39, p. 59], i.e., that left and right eigenvectors for different eigenvalues are orthogonal, does not hold in general. This follows also from the previous point. For nonlinear eigenvalue problems this principle of biorthogonality holds in few cases with respect to the generalized scalar product defined in (1.19), as shown in [30] for overdamped problems.

- The number of eigenvalues does not need to be equal to n. So, polynomial problems of degree m and order n have up to $m \times n$ eigenvalues, whereas for instance

$$T(\lambda)x = \begin{pmatrix} 2 - \mathrm{e}^{\lambda} & 0 & & \\ 0 & 1 & \ddots & \\ & \ddots & \ddots & 0 \\ & & 0 & 1 \end{pmatrix} x = 0 \tag{2.3}$$

  has only one finite eigenvalue $\lambda_* = \ln(2)$ for all $n \geq 1$.

One of the main problems we have to deal with when analyzing nonlinear eigenvalue problems and the essential problem for constructing algorithms is that no analogon to the *Schur form* for matrices exists, in general. For an arbitrary matrix $A \in \mathbb{C}^{n\times n}$ the Schur form is given by $AX = X\Lambda$, where $X \in \mathbb{C}^{n\times n}$ is unitary, and $\Lambda \in \mathbb{C}^{n\times n}$ is upper triangular with the eigenvalues of A on its diagonal, cf. [39, p. 79]. Linear eigenvalue problem solvers use the Schur factorization when more than one eigenvalue is requested. We have no equivalent tool in the general nonlinear case. In [56] locking and restarting for quadratic eigenvalue problems is discussed and the Schur form for a quadratic problem is computed as the Schur form of the corresponding linearized problem $Ax = \lambda Bx$, where B is supposed to be nonsingular. For polynomial problems a Schur form of the linearization can always be computed if the linearization exists, i.e., if the matrix A_m in (1.7) is nonsingular, but it is costly and may be badly conditioned [53, 54].

Besides, there are different strategies to find more than one eigenvalue. One is to go through different starting pairs one after the other, which have been obtained by solving the *linearized* problem $Ax = \mu Bx$, coming from the truncated Taylor series for $T(\lambda)$ with respect to some initial value λ_0

$$\Big[T(\lambda_0) + (\lambda - \lambda_0)\dot{T}(\lambda_0)\Big]\, x = 0, \tag{2.4}$$

with $A = T(\lambda_0)$, $B = -\dot{T}(\lambda_0)$. The smallest eigenvalue of this problem is of interest, and the new eigenvalue approximation is given by $\lambda_1 = \lambda_0 + \mu$. For eigenvalues far away from λ_0 it will be necessary to compute new starting sets by solving the corresponding linear problem. However, this approach cannot avoid that an eigenpair is found several times.

Another strategy, which prevents that the same eigenvalue is found more than once in exact arithmetic, is to transform the problem into a new problem where the previously computed eigenvalue is mapped to some point outside of the region of interest, as for instance, infinity or zero. When the problem is changed, say $T(\lambda) \mapsto \tilde{T}(\lambda)$, such that $\lambda(T(\cdot)) = \lambda(\tilde{T}(\cdot)) \setminus \{\lambda_k\} \cup \{\infty\}$, where λ_k is an already computed eigenvalue of $T(\cdot)$, then we speak of a *nonequivalence deflation*, cf. [29]. This deflation can only be obtained when left and right eigenvectors are computed, see [41] for such an algorithm in a Jacobi–Davidson framework.

We will not discuss approaches for finding more than one eigenpair in this thesis.

2.2 Assumptions

In order to increase readability of the following chapters, we introduce two assumptions on the nonlinear eigenvalue problem $T(\lambda)$, that summarize several initial assumptions. We distinguish between T being real-valued and T being complex-valued.

2.2.1 Real-valued Problems

Definition 2.3 ***Assumption*** **($\mathcal{A}_{\mathbb{R}}$)** *Let $D \subset \mathbb{R}$ be an open set. We say that $T : D \to \mathbb{R}^{n\times n}$ satisfies Assumption ($\mathcal{A}_{\mathbb{R}}$) if the following conditions hold:*

(i) There exists $\lambda_* \in D$ and $\tau_* > 0$ such that

$$S_* := \bar{S}(\lambda_*, \tau_*) = \{\lambda \in \mathbb{R} : \ |\lambda - \lambda_*| \leq \tau_*\} \subset D.$$

(ii) T is differentiable on S_* and $\dot{T}$ is Lipschitz continuous on S_*, i.e.,

$$\|\dot{T}(\lambda) - \dot{T}(\mu)\| \leq L|\lambda - \mu|, \tag{2.5}$$

for all $\lambda, \mu \in S_*$ with $L > 0$.

(iii) λ_* is an algebraically simple eigenvalue of T, this means that λ_* is geometrically simple, i.e., there exist nonzero vectors x_*, y_* such that

$$\ker(T(\lambda_*)) = \operatorname{span}\{x_*\}, \quad \ker(T(\lambda_*)^H) = \operatorname{span}\{y_*\},$$

and $\dot{\delta}(\lambda_*) \neq 0$, where $\delta(\lambda) = \det(T(\lambda))$, which is equivalent to

$$\alpha := y_*^H \dot{T}(\lambda_*) x_* \neq 0. \tag{2.6}$$

Remark 2.4 *If assumption* $(\mathcal{A}_{\mathbb{R}})$ *holds, then* T *is bounded and Lipschitz continuous on* S_**, since for all* $\lambda, \mu \in S_*$

$$\begin{aligned} \|\dot{T}(\lambda)\| &\leq \|\dot{T}(\lambda_*)\| + L|\lambda - \lambda_*| \leq \|\dot{T}(\lambda_*)\| + L\tau_* =: M_1, && (2.7)\\ \|T(\lambda) - T(\mu)\| &= \|\int_0^1 \dot{T}(\lambda + t\mu)(\lambda - \mu)\,dt\| \leq M_1|\lambda - \mu|, && (2.8)\\ \|T(\lambda)\| &\leq \|T(\lambda_*)\| + M_1\tau_* =: M_0. && (2.9) \end{aligned}$$

In the case of complex problems we require complex differentiability of T, which is equivalent to T being holomorphic on some domain. This already implies existence of all derivatives of T and is a stronger condition than continuous differentiability in the real case.

2.2.2 Complex-valued Problems

Definition 2.5 ***Assumption*** $(\mathcal{A}_{\mathbb{C}})$ *Let* $D \subset \mathbb{C}$ *be an open set. We say that* $T : D \to \mathbb{C}^{n\times n}$ *satisfies Assumption* $(\mathcal{A}_{\mathbb{C}})$*, if the following conditions hold:*

(i) There exists $\lambda_* \in D$ *and* $\tau_* > 0$ *such that*

$$S_* := \bar{S}(\lambda_*, \tau_*) = \{\lambda \in \mathbb{C} : \ |\lambda - \lambda_*| \leq \tau_*\} \subset D.$$

(ii) T *is holomorphic on an open neighborhood* $\tilde{S}_*$ *of* S_**.*

(iii) λ_* *is an algebraically simple eigenvalue of* T*, this means that* λ_* *is geometrically simple, i.e., there exist nonzero vectors* x_**,* y_* *such that*

$$\ker(T(\lambda_*)) = \operatorname{span}\{x_*\}, \quad \ker(T(\lambda_*)^H) = \operatorname{span}\{y_*\},$$

and $\dot{\delta}(\lambda_*) \neq 0$*, where* $\delta(\lambda) = \det(T(\lambda))$*, which is equivalent to* $\alpha := y_*^H \dot{T}(\lambda_*)x_* \neq 0$*.*

Remark 2.6 *In contrast to the real case, see Remark 2.4, boundedness of* $T(\lambda)$ *in* S_* *implies boundedness of* $T^{(l)}(\lambda)$ *in every compact subdomain* S_{**} *of* S_**. The corresponding bound depends only on* M_0*,* S_{**} *and* l*. To see this, let* δ *be the distance of* S_{**} *and the boundary of* S_**. Let* $\lambda_0 \in S_{**}$*,* $r < \delta$ *and let* K_0 *be a circle around* λ_0 *with radius* r*. Then by the Cauchy integral formula we have*

$$\|T^{(l)}(\lambda_0)\| = \Big\|\frac{l!}{2\pi \mathrm{i}} \int_{K_0} \frac{T(\zeta)}{(\zeta - \lambda_0)^{l+1}} d\zeta\Big\| \leq \frac{l!}{2\pi}\frac{M_0}{r^{l+1}} 2\pi r = \frac{l! M_0}{r^l}.$$

With $r \to \delta$ the desired result follows, i.e.,

$$\|T^{(l)}(\lambda_0)\| \leq \frac{l! M_0}{\delta^l}, \tag{2.10}$$

for all $\lambda_0 \in S_{**}$, cf. [10]. Therefore, bound (2.5) *and the inequalities* (2.7)–(2.9) *hold for all $\lambda, \mu \in S_*$ with*

$$L := \max\{\|\ddot{T}(\lambda)\| : \lambda \in S_*\}. \tag{2.11}$$

Figure 2.1 presents the local situation around λ_*.

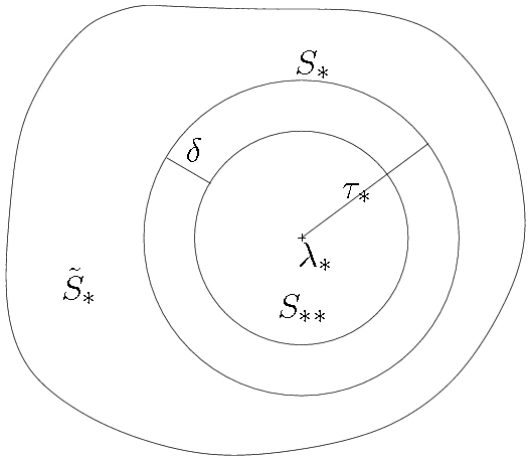

Figure 2.1: *Local situation around λ_**

Throughout this thesis we will assume that the desired eigenvalue λ_* is simple. Then, the *singular value decomposition* (SVD) for $T(\lambda_*)$ has the form

$$T(\lambda_*) = Y \Sigma X^H = [\, Y_1 \mid y_* \,] \left[\begin{array}{c|c} \Sigma_1 & 0 \\ \hline 0^T & 0 \end{array}\right] [\, X_1 \mid x_* \,]^H, \tag{2.12}$$

where $\Sigma_1 = \mathrm{diag}(\sigma_1, ..., \sigma_{n-1}) \in \mathbb{R}^{(n-1)\times(n-1)}$ is nonsingular with $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{n-1} > 0$, and the matrices $Y, X \in \mathbb{C}^{n\times n}$, containing the left and right singular vectors, are unitary.

The SVD of $T(\lambda_*)$ is a useful tool for handling nonlinear eigenvalue problems since we do not have analogons for Schur or eigendecompositions of matrices.

2.3 Eigenvalue Condition Numbers

It is a well-known fact for linear problems $Ax = \lambda x$, that simple eigenvalues of A are locally differentiable functions of A, and moreover, for a small perturbation E, there

exists a unique $\tilde{\lambda}$ which is an eigenvalue of $\tilde{A} = A + E$ such that

$$\tilde{\lambda} = \lambda_* + \frac{y_*^H E x_*}{y_*^H x_*} + \mathcal{O}(\|E\|^2),$$

cf. [85, p. 185 f.], where (λ_*, x_*, y_*) is an eigentriple for A with simple eigenvalue λ_*. From this equation the condition number of the eigenvalue is derived as

$$\kappa(\lambda_*, A) = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \sup_{\|E\| \leq \varepsilon} |\tilde{\lambda} - \lambda_*| = \frac{\|x_*\| \|y_*\|}{|y_*^H x_*|} = \frac{1}{|\cos(\measuredangle(\mathrm{span}\{x_*\}, \mathrm{span}\{y_*\}))|},$$

where $E = \varepsilon \frac{y_* x_*^H}{\|x_*\| \|y_*\|}$ delivers equivalence, see, e.g., [47]. Hence, the eigenvalue is well-conditioned for normal matrices with $\kappa = 1$, since $y_* = x_*$. Whereas for nonnormal matrices $\kappa(\lambda_*, A)$ may become arbitrarily large, when x_* and y_* are almost orthogonal.

For polynomial eigenvalue problems $P(\lambda)x = \sum_{l=0}^{m} \lambda^l A_l x = 0$, a normwise relative condition number for a nonzero simple eigenvalue can be defined by

$$\begin{aligned} \kappa_{rel}(\lambda_*, P) \;=\; \limsup_{\epsilon \to 0} \Bigg\{ & \frac{|\Delta\lambda_*|}{\epsilon |\lambda_*|} : (P(\lambda_* + \Delta\lambda_*) + \Delta P(\lambda_* + \Delta\lambda_*))(x_* + \Delta x_*) = 0, \\ & \|\Delta A_l\| \leq \epsilon \|E_l\|,\, l = 0, \ldots m \Bigg\}, \end{aligned}$$

where the matrices E_l, $l = 0, \ldots, m$ are arbitrary with constant $\|E_l\|$ and represent tolerances against which the perturbations ΔA_l to A_l are measured and where $\Delta P(\lambda) = \lambda^m \Delta A_m + \ldots \lambda \Delta A_1 + \Delta A_0$, cf. [86]. It was also shown in [86] that this condition number can be determined as

$$\kappa_{rel}(\lambda_*, P) = \sum_{l=0}^{m} |\lambda_*^l| \|E_l\| \frac{\|x_*\| \|y_*\|}{|\lambda_*| |y_*^H \dot{T}(\lambda_*) x_*|}. \tag{2.13}$$

The denominator is nonzero since $\lambda_* \neq 0$ is algebraically simple, cf. Proposition 2.2.

For general nonlinear problems we have to use a different ansatz, since we do not want to suppose any structure in the problem. We introduce a vector of perturbation parameters $\varepsilon = (\varepsilon_l) \in \mathbb{C}^d$, $l = 1, \ldots, d$, and consider the perturbed operator $T(\lambda, \varepsilon)$, where $T(\lambda) := T(\lambda, 0)$ for all $\lambda \in \mathbb{C}$.

Lemma 2.7 *Let $D \subset \mathbb{C}$ and $E \subset \mathbb{C}^d$ be open sets. Let $T(\,.\,) : D \times E \to \mathbb{C}^{n \times n}$ be continuously differentiable, and let λ_* be a simple eigenvalue of $T(\,.\,, 0)$ and x_*, y_* be the corresponding eigenvectors with unit norm. Let $\tau_* > 0$, $\varepsilon_* > 0$ be such that $\bar{S}(\lambda_*, \tau_*) \subset D$ and $\bar{S}(0, \varepsilon_*) \subset E$. Then, the first order perturbation expansion of the eigenvalue is given*

by

$$\lambda(\varepsilon)-\lambda_* \;=\; -\frac{y_*^H\dfrac{\partial T}{\partial\varepsilon}(\lambda_*,0)[\varepsilon]x_*}{y_*^H\dot{T}(\lambda_*,0)x_*}+o(\|\varepsilon\|) \;=\; -\sum_{l=1}^{d}\varepsilon_l\frac{y_*^H\dfrac{\partial T}{\partial\varepsilon_l}(\lambda_*,0)x_*}{y_*^H\dot{T}(\lambda_*,0)x_*}+o(\|\varepsilon\|). \quad (2.14)$$

The normwise condition number for λ_* is given by

$$\kappa(\lambda_*):=\limsup_{\|\varepsilon\|\to 0}\frac{|\lambda(\varepsilon)-\lambda_*|}{\|\varepsilon\|_2}=\left\|\frac{\partial\lambda}{\partial\varepsilon}(0)\right\|_2 \;=\; \sqrt{\sum_{l=1}^{d}\left|\frac{y_*^H\frac{\partial T}{\partial\varepsilon_l}(\lambda_*,0)x_*}{y_*^H\dot{T}(\lambda_*,0)x_*}\right|^2}. \quad (2.15)$$

Proof. In order to obtain a characterization for an eigenvalue which allows a differentiation, we consider system (2.1) again but additionally depending on ε, i.e.,

$$C(\lambda,\varepsilon)\begin{bmatrix}x\\ \mu\end{bmatrix}:=\begin{bmatrix}T(\lambda,\varepsilon) & y_*\\ x_*^H & 0\end{bmatrix}\begin{bmatrix}x\\ \mu\end{bmatrix}=\begin{bmatrix}0\\ 1\end{bmatrix}.$$

The bordered matrix $C(\lambda,\varepsilon)$ is nonsingular for (λ,ε) close to $(\lambda_*,0)$, cf. the proof of Proposition 2.2. Hence, the linear systems

$$C(\lambda,\varepsilon)\begin{bmatrix}x\\ \mu\end{bmatrix}=\begin{bmatrix}0\\ 1\end{bmatrix} \quad\text{and}\quad C(\lambda,\varepsilon)^H\begin{bmatrix}y\\ \nu\end{bmatrix}=\begin{bmatrix}0\\ 1\end{bmatrix} \quad (2.16)$$

can be solved uniquely for $x=x(\lambda,\varepsilon)$, $\mu=\mu(\lambda,\varepsilon)$, $y=y(\lambda,\varepsilon)$ and $\nu=\nu(\lambda,\varepsilon)$, if ε is sufficiently small. Inserting the exact eigenvalue gives $x(\lambda_*,0)=x_*$, $\mu(\lambda_*,0)=0$, $y(\lambda_*,0)=y_*$, $\nu(\lambda_*,0)=0$. We can rewrite (2.16) as

$$\begin{bmatrix}x(\lambda,\varepsilon)\\ \mu(\lambda,\varepsilon)\end{bmatrix}=C^{-1}e_{n+1},\quad\text{and}\quad\begin{bmatrix}y(\lambda,\varepsilon)\\ \nu(\lambda,\varepsilon)\end{bmatrix}=C^{-H}e_{n+1},$$

which yields

$$\mu(\lambda,\varepsilon)=e_{n+1}^TC(\lambda,\varepsilon)^{-1}e_{n+1}.$$

By differentiating this equation we obtain

$$\frac{\partial\mu}{\partial\lambda}(\lambda,\varepsilon) \;=\; -e_{n+1}^TC(\lambda,\varepsilon)^{-1}\begin{bmatrix}\dot{T}(\lambda,\varepsilon) & 0\\ 0 & 0\end{bmatrix}C(\lambda,\varepsilon)^{-1}e_{n+1}=-y(\lambda,\varepsilon)^H\dot{T}(\lambda,\varepsilon)x(\lambda,\varepsilon).$$

Hence, at $(\lambda_*,0)$ one has $\frac{\partial\mu}{\partial\lambda}(\lambda_*,0)=-y_*^H\dot{T}(\lambda_*,0)x_*=-\alpha\neq 0$. This means that $\mu(\lambda,\varepsilon)=0$ can be solved locally uniquely for $\lambda=\lambda(\varepsilon)$ and $\mu(\lambda(\varepsilon),\varepsilon)\equiv 0$, which implies

$$\frac{\partial\mu}{\partial\lambda}\frac{\partial\lambda}{\partial\varepsilon}+\frac{\partial\mu}{\partial\varepsilon}\equiv 0. \quad (2.17)$$

The derivative of μ with respect to ε is determined as follows

$$\begin{aligned}\frac{\partial\mu}{\partial\varepsilon}(\lambda,\varepsilon)[\sigma] &= -e_{n+1}^T C(\lambda,\varepsilon)^{-1}\begin{bmatrix}\frac{\partial T}{\partial\varepsilon}(\lambda,\varepsilon)[\sigma] & 0\\ 0 & 0\end{bmatrix}C(\lambda,\varepsilon)^{-1}e_{n+1}\\ &= -y(\lambda,\varepsilon)^H\frac{\partial T}{\partial\varepsilon}(\lambda,\varepsilon)[\sigma]x(\lambda,\varepsilon),\end{aligned}$$

and at the particular solution we have

$$\frac{\partial\mu}{\partial\varepsilon}(\lambda_*,0)[\varepsilon] = -y_*^H\frac{\partial T}{\partial\varepsilon}(\lambda_*,0)[\varepsilon]x_* = -\sum_{l=1}^{d}\varepsilon_l\left(y_*^H\frac{\partial T}{\partial\varepsilon_l}(\lambda_*,0)x_*\right).$$

In detail, we can write

$$\frac{\partial\mu}{\partial\varepsilon}(\lambda_*,0) = -\left(y_*^H\frac{\partial T}{\partial\varepsilon_1}(\lambda_*,0)x_*,\ldots,y_*^H\frac{\partial T}{\partial\varepsilon_d}(\lambda_*,0)x_*\right)\in\mathbb{C}^{1\times d}.$$

Now, we can rearrange (2.17) and end up with

$$\frac{\partial\lambda}{\partial\varepsilon}(\lambda_*,0)[\varepsilon] = -\frac{\frac{\partial\mu}{\partial\varepsilon}(\lambda_*,0)[\varepsilon]}{\frac{\partial\mu}{\partial\lambda}(\lambda_*,0)} = -\frac{y_*^H\frac{\partial T}{\partial\varepsilon}(\lambda_*,0)[\varepsilon]x_*}{y_*^H\dot{T}(\lambda_*,0)x_*},$$

which implies (2.14). Furthermore, equation (2.14) gives the estimate

$$|\lambda(\varepsilon)-\lambda_*| \le \sum_{l=1}^{d}|\varepsilon_l|\frac{|y_*^H\frac{\partial T}{\partial\varepsilon_l}(\lambda_*,0)x_*|}{|y_*^H\dot{T}(\lambda_*)x_*|}+o(\|\varepsilon\|),$$

and also

$$|\lambda(\varepsilon)-\lambda_*| = \left|\frac{\partial\lambda}{\partial\varepsilon}(0)\varepsilon\right|+o(\|\varepsilon\|) \le \left\|\frac{\partial\lambda}{\partial\varepsilon}(0)\right\|_2\|\varepsilon\|_2+o(\|\varepsilon\|),$$

with equality for an appropriate choice of ε, and (2.15) follows. □

Let us discuss special cases of this result. At first, consider the polynomial eigenvalue problem (1.3) with a parameterized perturbation $T(\lambda,\varepsilon)=\sum_{l=0}^{m}\lambda^l(A_l+\varepsilon_l E_l)$. Since $\partial T/\partial\varepsilon_l(\lambda,\varepsilon)=\lambda^l E_l$, equation (2.14) yields

$$\lambda(\varepsilon)-\lambda_* = -\sum_{l=1}^{d}\varepsilon_l\frac{\lambda_*^l y_*^H E_l x_*}{y_*^H\dot{T}(\lambda_*)x_*}+o(\|\varepsilon\|).$$

If we set $\varepsilon_0=\|\varepsilon\|_\infty$, then we get $\|\varepsilon_l E_l\|=|\varepsilon_l|\|E_l\|\le\varepsilon_0\|E_l\|$, and

$$\frac{|\lambda(\varepsilon)-\lambda_*|}{\|\varepsilon\|_\infty} = \frac{|\lambda(\varepsilon)-\lambda_*|}{\varepsilon_0} \le \frac{\|\varepsilon\|_\infty}{\|\varepsilon\|_\infty}\sum_{l=1}^{d}\frac{|\lambda_*^l||y_*^H E_l x_*|}{|y_*^H\dot{T}(\lambda_*)x_*|},$$

where we have used the duality of the 1-norm and the ∞-norm. The condition number of λ_* is given by

$$\limsup_{\varepsilon\to 0} \frac{|\lambda(\varepsilon)-\lambda_*|}{\|\varepsilon\|_\infty} = \sum_{l=1}^{d} \frac{|\lambda_*^l||y_*^H E_l x_*|}{|y_*^H \dot{T}(\lambda_*)x_*|} = \sum_{l=1}^{d} \frac{|\lambda_*^l|\|E_l\|\|y_*\|\|x_*\|}{|y_*^H \dot{T}(\lambda_*)x_*|},$$

which is equivalent to the relative bound (2.13). The upper bounds are reached for $E_l = \frac{\varepsilon_l y_* x_*^H}{\|x_*\|\|y_*\|}$, if all E_l are admitted to have constant norm.

As second example consider $T(\lambda) = A + \lambda I + \exp(-\lambda)I$, cf. [101]. We perturb all involved matrices differently $T(\lambda,\varepsilon) = A + \varepsilon_1 E_1 + \lambda(I+\varepsilon_2 E_2) + \exp(-\lambda)(I+\varepsilon_3 E_3)$. The first order perturbation expansion of the eigenvalue, corresponding to equation (2.14), is given by

$$\lambda(\varepsilon) - \lambda_* \;=\; -\frac{1}{y_*^H \dot{T}(\lambda_*)x_*}\left(\varepsilon_1 y_*^H E_1 x_* + \varepsilon_2 \lambda_* y_*^H E_2 x_* + \varepsilon_3 \exp(-\lambda_*) y_*^H E_3 x_*\right) + o(\|\varepsilon\|),$$

provided that λ_* is simple. This implies the bound

$$|\lambda(\varepsilon)-\lambda_*| \le \frac{|\varepsilon_1||y_*^H E_1 x_*| + |\varepsilon_2||\lambda_*||y_*^H E_2 x_*| + |\varepsilon_3||\exp(-\lambda_*)||y_*^H E_3 x_*|}{|y_*^H \dot{T}(\lambda_*)x_*|} + o(\|\varepsilon\|).$$

Then, the condition number of λ_* is given by

$$\kappa(\lambda_*) = \frac{1}{|y_*^H \dot{T}(\lambda_*)x_*|}\sqrt{|y_*^H E_1 x_*|^2 + |\lambda_*|^2|y_*^H E_2 x_*|^2 + \exp(-\lambda_*)^2|y_*^H E_3 x_*|^2}.$$

2.4 Representation of the Inverse Operator

Representations of the *resolvent* $(A-\lambda I)^{-1}$, for an arbitrary matrix A, can be derived by several available decompositions, as e.g., the spectral decomposition, Schur decomposition and singular value decomposition. It turns out that in this case the norm of the inverse shifted operator depends on the *(eigenvalue) gap*, that is $(\min_{\lambda_i\in\lambda(A)}\{\lambda_i-\lambda\})^{-1}$. We want to analyze the nonlinear case here.

Lemma 2.8 *Let $T(\lambda)$ be holomorphic, resp. twice continuously differentiable, on an open neighborhood of $S_* = \bar{S}(\lambda_*,\tau_*)$, $\tau_*>0$, and let (λ_*,x_*,y_*) be an eigentriple with simple eigenvalue λ_*. Then there exists a radius $0<\tau_0\le\tau_*$ such that, for all $\tau=\lambda-\lambda_*$ with $0<|\tau|\le\tau_0$, we have*

$$T(\lambda)^{-1} = \frac{1}{\lambda-\lambda_*} X\hat{T}(\lambda)^{-1}Y^H = \frac{1}{\lambda-\lambda_*}\left[\frac{x_* y_*^H}{y_*^H \dot{T}(\lambda_*)x_*} + \mathcal{O}(\tau)\right], \tag{2.18}$$

with X, Y as defined in (2.12) *and*

$$\hat{T}(\lambda)^{-1} = \begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{\alpha} \end{bmatrix} + \tau \begin{bmatrix} \Sigma_1^{-1} & -\frac{1}{\alpha}\Sigma_1^{-1} Y_1^H \dot{T}_* x_* \\ -\frac{1}{\alpha} y_*^H \dot{T}_* X_1 \Sigma_1^{-1} & \frac{1}{\alpha^2}\left(\frac{2}{y_*^H \ddot{T}(\zeta) x_*} + y_*^H \dot{T}_* X_1 \Sigma_1^{-1} Y_1^H \dot{T}_* x_* \right) \end{bmatrix} + \mathcal{O}(\tau^2), \tag{2.19}$$

with $\alpha = y_*^H \dot{T}(\lambda_*) x_*$ and $\dot{T}_* = \dot{T}(\lambda_*)$.

Proof. Let $\tau_0 = \min\{1, \tau_1, \tau_2,\}$, where $\tau_1 := \|\Sigma_1^{-1}\|^{-1}(\|\dot{T}(\lambda_*)\| + L/2)^{-1}$, with L defined in (2.11), resp. (2.5), and $\tau_2 := |\alpha|(\frac{L}{2} + (\|\dot{T}_*\| + \frac{L}{2})^2 \|\Sigma_1^{-1}\|)^{-1}$. The Taylor series of $T(\lambda)$ truncated after the second term is given by

$$T(\lambda) = T(\lambda_*) + \tau \dot{T}(\lambda_*) + \frac{\tau^2}{2} \ddot{T}(\zeta),$$

where $\zeta \in S(\lambda_*, \tau)$. Inserting the singular value decomposition (2.12) gives

$$T(\lambda) = Y \left\{ \begin{bmatrix} \Sigma_1 & 0 \\ 0^T & 0 \end{bmatrix} + \tau \begin{bmatrix} k_{11} & k_{12} \\ k_{21} & \alpha \end{bmatrix} + \frac{\tau^2}{2} \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} \right\} X^H := Y \hat{T}(\lambda) X^H,$$

where $k_{11} = Y_1^H \dot{T}(\lambda_*) X_1$, $k_{12} = Y_1^H \dot{T}(\lambda_*) x_*$, $k_{21} = y_*^H \dot{T}(\lambda_*) X_1$, $\alpha = y_*^H \dot{T}(\lambda_*) x_*$, and $m_{11} = Y_1^H \ddot{T}(\zeta) X_1$, $m_{12} = Y_1^H \ddot{T}(\zeta) x_*$, $m_{21} = y_*^H \ddot{T}(\zeta) X_1$, $m_{22} = y_*^H \ddot{T}(\zeta) x_*$. Now, write $\hat{T}(\lambda) := Y^H T(\lambda) X$ as

$$\hat{T}(\lambda) = \begin{bmatrix} \Sigma_1 + \tau k_{11} + \frac{\tau^2}{2} m_{11} & \tau k_{12} + \frac{\tau^2}{2} m_{12} \\ \tau k_{21} + \frac{\tau^2}{2} m_{21} & \tau\alpha + \frac{\tau^2}{2} m_{22} \end{bmatrix} =: \begin{bmatrix} \hat{T}_{11} & \hat{T}_{12} \\ \hat{T}_{21} & \hat{T}_{22} \end{bmatrix}.$$

The upper left block $\hat{T}_{11}$ is nonsingular if $\|\Sigma_1^{-1}\| \|\tau k_{11} + \frac{\tau^2}{2} m_{11}\| < 1$ by the perturbation lemma, cf. [46, p. 128]. Since $\tau < \tau_1 := \|\Sigma_1^{-1}\|^{-1}(\|\dot{T}(\lambda_*)\| + L/2)^{-1}$ and $\tau < 1$, we have

$$\begin{aligned} \|\Sigma_1^{-1}\| \|\tau k_{11} + \frac{\tau^2}{2} m_{11}\| &\leq |\tau| \|\Sigma_1^{-1}\| \left(\|\dot{T}(\lambda_*)\| + \frac{|\tau|}{2} L \right) \\ &\leq |\tau| \|\Sigma_1^{-1}\| \left(\|\dot{T}(\lambda_*)\| + \frac{L}{2} \right) = \frac{|\tau|}{\tau_1} < 1. \end{aligned}$$

Hence, $\hat{T}_{11}$ is nonsingular and the inverse can be determined as

$$\hat{T}_{11}^{-1} = (\Sigma_1 + \tau k_{11} + \frac{\tau^2}{2} m_{11})^{-1} = \Sigma_1^{-1} - \tau \Sigma_1^{-1} k_{11} \Sigma_1^{-1} + \mathcal{O}(\tau^2).$$

Now we are in a position to form the Schur complement H in order to show nonsingularity of $\hat{T}(\lambda)$, resp. $T(\lambda)$. To be more precise, let

$$\begin{aligned} H &:= \hat{T}_{22} - \hat{T}_{21} \hat{T}_{11}^{-1} \hat{T}_{12} \\ &= \tau \left[\alpha + \frac{\tau}{2} m_{22} - \tau (k_{21} + \frac{\tau}{2} m_{21}) \left(\Sigma_1 + \tau k_{11} + \frac{\tau^2}{2} m_{11} \right)^{-1} (k_{12} + \frac{\tau}{2} m_{12}) \right] \\ &= \tau[\alpha + \tau E], \end{aligned}$$

where

$$E = \frac{1}{2}m_{22} - (k_{21} + \frac{\tau}{2}m_{21})\left(\Sigma_1 + \tau k_{11} + \frac{\tau^2}{2}m_{11}\right)^{-1}(k_{12} + \frac{\tau}{2}m_{12}).$$

Then, $\hat{T}(\lambda)$ is nonsingular if and only if H is nonsingular, or rather nonzero in our case, cf. [39, p. 22 f]. Without loss of generality we assume that $\alpha > 0$, otherwise replace x_* by $-x_*$. Then, provided that $|\tau| > 0$, it suffices to show that $|\tau||E| < \alpha$. Since $|\tau| < 1$, we have

$$\begin{aligned} |E| &\leq \frac{1}{2}\|m_{22}\| + \|k_{21} + \frac{\tau}{2}m_{21}\|\|(\Sigma_1 + \tau k_{11} + \frac{\tau^2}{2}m_{11})^{-1}\|\|k_{12} + \frac{\tau}{2}m_{12}\| \\ &\leq \frac{L}{2} + \left(\|\dot{T}_*\| + \frac{\tau L}{2}\right)^2 \frac{\|\Sigma_1^{-1}\|}{1 - \|\Sigma_1^{-1}\|\|\tau k_{11} + \frac{\tau^2}{2}m_{11}\|} \\ &< \frac{L}{2} + \left(\|\dot{T}_*\| + \frac{L}{2}\right)^2 \|\Sigma_1^{-1}\| \;=\; \frac{\alpha}{\tau_2}. \end{aligned}$$

Hence, $|\tau||E| < \tau_2|E| < \tau_2\frac{\alpha}{\tau_2} = \alpha$. It follows that H is nonsingular and we are able to determine the inverse of $\hat{T}(\lambda)$, as

$$\hat{T}(\lambda)^{-1} = \begin{bmatrix} \hat{T}_{11}^{-1} + \hat{T}_{11}^{-1}\hat{T}_{12}H^{-1}\hat{T}_{21}\hat{T}_{11}^{-1} & -\hat{T}_{11}^{-1}\hat{T}_{12}H^{-1} \\ -H^{-1}\hat{T}_{21}\hat{T}_{11}^{-1} & H^{-1} \end{bmatrix} =: \begin{bmatrix} [\hat{T}^{-1}]_{11} & [\hat{T}^{-1}]_{12} \\ [\hat{T}^{-1}]_{21} & H^{-1} \end{bmatrix},$$

see [39, p. 18]. In detail we have

$$H^{-1} \;=\; \frac{1}{\tau\alpha}\left(1 - \frac{\tau E}{\alpha} + \frac{\tau^2 E^2}{\alpha^2} + \mathcal{O}(\tau^3)\right) = \frac{1}{\tau\alpha}\left(1 - \frac{\tau}{\alpha}\left(\frac{m_{22}}{2} - k_{21}\Sigma_1^{-1}k_{12}\right) + \mathcal{O}(\tau^2)\right),$$

and hence,

$$\begin{aligned} [\hat{T}^{-1}]_{11} &= \Sigma_1^{-1} + \mathcal{O}(\tau), \\ [\hat{T}^{-1}]_{12} &= -(\Sigma_1^{-1} - \tau\Sigma_1^{-1}k_{11}\Sigma_1^{-1} + \mathcal{O}(\tau^2))(\tau k_{12} + \frac{\tau^2}{2}m_{12})H^{-1} = -\frac{1}{\alpha}\Sigma_1^{-1}k_{12} + \mathcal{O}(\tau), \\ [\hat{T}^{-1}]_{21} &= -H^{-1}(\tau k_{21} + \frac{\tau^2}{2}m_{21})(\Sigma_1^{-1} - \tau\Sigma_1^{-1}k_{11}\Sigma_1^{-1} + \mathcal{O}(\tau^2)) = -\frac{1}{\alpha}k_{21}\Sigma_1^{-1} + \mathcal{O}(\tau), \end{aligned}$$

which gives (2.19) immediately. □

The presentation (2.18) helps us understanding inverse iteration steps. Suppose, we have a vector z, with $y_*^H z \neq 0$. Applying the inverse operator $T(\lambda)^{-1}$ to z yields

$$z_+ = T(\lambda)^{-1}z = \frac{y_*^H z}{\alpha(\lambda - \lambda_*)}x_* + \mathcal{O}(1),$$

for λ sufficiently close to λ_*. This means that z_+ has a dominant component in the direction of the right eigenvector x_*, for all z with $y_*^H z \neq 0$. If we choose $z = \dot{T}(\lambda)u$,

then the condition $y_*^H z \neq 0$ holds, if (λ, u) is close to (λ_*, x_*) for λ_* simple. Then we obtain $z_+ = T(\lambda)^{-1}\dot{T}(\lambda)u = \frac{y_*^H \dot{T}(\lambda)u}{y_*^H \dot{T}(\lambda_*)x_*(\lambda-\lambda_*)}x_* + \mathcal{O}(1)$, which is the inverse iteration step from (u, λ). Another choice would be $z = v$, where v is an approximation to the left eigenvector. This ansatz is used in the primal-dual method in [77, 78].

2.5 Angles and Distances

The *angle* between two (complex) vectors x and u, denoted by $\zeta = \measuredangle(\mathrm{span}\{x\}, \mathrm{span}\{u\}) = \measuredangle(\mathrm{span}\{u\}, \mathrm{span}\{x\})$, is defined by the equation

$$\cos\zeta = \frac{|x^H u|}{\|x\|\|u\|},$$

such that $0 \leq \zeta \leq \pi/2$, i.e., ζ is the smaller one of the two possibilities.

Nonlinear eigenvalue problems typically yield nonreal eigenvalues and -vectors. In the real case, convergence of the sequence of angles of two normed vectors x_* and x_k towards zero is equivalent to $\|x_k - \alpha_k x_*\| \to 0$ with $\alpha_k = 1$ or $\alpha_k = -1$. Provided that $x_k^T x_{k-1} > 0$, the more general result $\|x_k - \alpha x_*\| \to 0$ with $\alpha = \pm 1$ holds, because this assumption guarantees that x_{k-1} and x_k have the same direction.

Lemma 2.9 *Let $x, u \in \mathbb{R}^n$, $\|x\| = \|u\| = 1$, $\zeta = \measuredangle(\mathrm{span}\{x\}, \mathrm{span}\{u\})$, and assume that $\|u - x\| < 1$. Then, we have*

$$\|u - x\| = \sqrt{\frac{2}{1+\cos\zeta}}\sin\zeta, \quad \textit{and} \quad \sin\zeta \leq \|u - x\| \leq \sqrt{2}\sin\zeta, \tag{2.20}$$

and moreover,

$$\lim_{\zeta\to 0} \frac{\|u - x\|}{\sin\zeta} = 1, \tag{2.21}$$

hence $\sin\zeta \sim \|u - x\|$ for $\zeta \to 0$.

Proof. The following equation holds for unit norm vectors in $\mathbb{C}^n$

$$\delta_0^2 := \|u - x\|^2 = \|u\|^2 + \|x\|^2 - 2\mathrm{Re}(x^H u) = 2(1 - \mathrm{Re}(x^H u)). \tag{2.22}$$

For real vectors we have $x^T u = 1 - \delta_0^2/2 > 1/2$, because we have assumed that $\delta_0 < 1$. Thus, $x^H u = x^T u = |x^T u| = \cos\zeta$ and (2.22) is equivalent to

$$\delta_0^2 = \frac{2(1-\cos\zeta)(1+\cos\zeta)}{1+\cos\zeta} = \frac{2\sin^2\zeta}{1+\cos\zeta}.$$

Taking the square root of this equation and inserting the upper and the lower bound for the cosine yields (2.20).

Equation (2.21) follows, since $\lim_{\zeta\to 0} \frac{\delta_0}{\sin\zeta} = \lim_{\zeta\to 0} \sqrt{\frac{2}{1+\cos\zeta}} = 1$. □

In the real case, a zero angle of unit norm vectors u and x implies either $u = x$, i.e., $\|u-x\| = 0$, or $u = -x$, i.e., $\|u-x\| = 2$. In the complex case, however, one can have two vectors with zero angle but an arbitrary norm difference in $[0, 1]$ as the following example shows.

Example 1 *For arbitrary $\delta_0 \in [0, 1]$ consider*

$$x = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ i \end{bmatrix}, \quad \textit{and} \quad u = \frac{1}{\sqrt{2+2a^2}} \begin{bmatrix} 1+ia \\ i-a \end{bmatrix}, \quad \textit{where} \quad a = \frac{\delta_0\sqrt{1-\delta_0^2/4}}{1-\delta_0^2/2}.$$

The vectors x and u have unit norm, and $\cos\zeta = |x^H u| = 1$, i.e., the angle $\zeta = \measuredangle(\mathrm{span}\{x\}, \mathrm{span}\{u\}) = 0$. Elementary computations yield

$$\|u-x\|^2 = 2 - 2Re(u^H x) = 2 - \frac{2}{\sqrt{1+a^2}} = 2 - 2\left(1 - \frac{\delta_0^2}{2}\right) = \delta_0^2.$$

The following Lemma addresses the problem of angles and their relation to distances in the complex case.

Lemma 2.10 *Let x, $u \in \mathbb{C}^n$, where x has unit norm and $u \neq 0$ is arbitrary. Define the angle $\zeta := \measuredangle(\mathrm{span}\{x\}, \mathrm{span}\{u\})$. Then, we have*

(i) $\sin\zeta \leq \|u-x\|$,

(ii) $\sin\zeta = \|u_0 - x\|$ *for the projection* $u_0 = P_u x = \left(\frac{uu^H}{u^H u}\right) x \in \mathrm{span}\{u\}$ *of x onto* $\mathrm{span}\{u\}$.

Proof. Set $u = x + c$. Then

$$\begin{aligned} \|u\|^2 &= \|x+c\|^2 = (x+c)^H(x+c) = 1 + 2\mathrm{Re}(x^H c) + \|c\|^2, \\ |x^H u|^2 &= |1 + x^H c|^2 = (1+x^H c)^H(1+x^H c) = 1 + 2\mathrm{Re}(x^H c) + |x^H c|^2, \end{aligned}$$

and

$$\begin{aligned} \cos^2\zeta = \frac{|x^H u|^2}{\|u\|^2} &= \frac{1 + 2\mathrm{Re}(x^H c) + |x^H c|^2 + \|c\|^2 - \|c\|^2}{1 + 2\mathrm{Re}(x^H c) + \|c\|^2} \\ &= 1 - \frac{\|c\|^2 - |x^H c|^2}{1 + 2\mathrm{Re}(x^H c) + \|c\|^2} = 1 - \delta_1^2, \end{aligned}$$

where $\delta_1^2 = \frac{\|c\|^2 - |x^H c|^2}{1+2\mathrm{Re}(x^H c)+\|c\|^2}$. Now we show that $\delta_1^2 \leq \|c\|^2$:

$$\begin{aligned}
0 &\leq \big(\|c\|^2 + \mathrm{Re}(x^H c)\big)^2 = \|c\|^4 + 2\|c\|^2\mathrm{Re}(x^H c) + (\mathrm{Re}(x^H c))^2 \\
&\leq \|c\|^4 + 2\|c\|^2\mathrm{Re}(x^H c) + |x^H c|^2 \\
\iff \quad -|x^H c|^2 &\leq \|c\|^4 + 2\|c\|^2\mathrm{Re}(x^H c) \\
\iff \quad \|c\|^2 - |x^H c|^2 &\leq \|c\|^2\big(\|c\|^2 + 2\mathrm{Re}(x^H c) + 1\big) \\
\iff \quad \delta_1^2 &\leq \|c\|^2,
\end{aligned}$$

which yields immediately that $\sin^2\zeta = 1 - \cos^2\zeta = \delta_1^2 \leq \|c\|^2 = \|u - x\|^2$, thus (i) holds. Note, that for $\|c\| \geq 1$ this bound is trivial, i.e., we may consider arbitrary u. To show the second part, set $u_0 = P_u x = \frac{u}{\|u\|}\frac{u^H x}{\|u\|}$. Then

$$\begin{aligned}
\|u_0\| &= \frac{|u^H x|}{\|u\|} = \frac{|u^H x|}{\|u\|\|x\|} = \cos\zeta, \\
x^H u_0 &= \frac{(x^H u)(u^H x)}{\|u\|^2} = \frac{|x^H u|^2}{\|u\|^2\|x\|^2} = \cos^2\zeta,
\end{aligned}$$

and

$$\|u_0 - x\|^2 = \|u_0\|^2 - 2\mathrm{Re}(x^H u_0) + \|x\|^2 = \cos^2\zeta - 2\cos^2\zeta + 1 = \sin^2\zeta.$$

□

Lemma 2.10 shows that norm distances and angles cannot be used similarly in case of complex vectors, in general. To measure the quality of an approximation to a complex vector, the angle between the two should be used instead of the distance in terms of the norm, which can be much larger. However, the following lemma states that there exists a scaling factor such that the norm distance of the scaled vector and its approximation is similar to the sine of the angle between them.

Lemma 2.11 *Let x, $u \in \mathbb{C}^n$, where $\|x\| = \|u\| = 1$, and $\zeta = \measuredangle(\mathrm{span}\{x\}, \mathrm{span}\{u\}) < \pi/2$.*

(i) Then, there exists a scalar $\beta \in \mathbb{C}$, with $|\beta| = 1$ such that for $x_0 = \beta x$ the following holds

$$\|u - x_0\| = \sqrt{\frac{2}{1+\cos\zeta}}\sin\zeta, \quad \textit{and} \quad \sin\zeta \leq \|u - x_0\| \leq \sqrt{2}\sin\zeta, \tag{2.23}$$

and moreover,

$$\lim_{\zeta\to 0}\frac{\|u - x_0\|}{\sin\zeta} = 1, \tag{2.24}$$

hence, we have $\sin\zeta \sim \|u - x_0\|$ for $\zeta \to 0$.

(ii) We have

$$\|u - x^u\| = \tan\zeta, \quad \textit{where} \quad x^u = \frac{1}{u^H x} x. \tag{2.25}$$

(iii) The equality $\|u - x_u\| = \sin\zeta$ *holds for the projection* $x_u = P_x u = x(x^H u) \in \operatorname{span}\{x\}$ *of* u *onto* $\operatorname{span}\{x\}$.

Proof. To show (i), set $\beta = \frac{|u^H x|}{u^H x}$, hence $|\beta| = 1$. Then,

$$\|u - x_0\| = \sqrt{2(1 - \cos\zeta)} = \sqrt{\frac{2}{1 + \cos\zeta}} \sin\zeta.$$

Inserting the upper and lower bounds for the cosine, i.e., $0 \leq \cos\zeta \leq 1$ yields the inequalities (2.23). Equation (2.24) follows since

$$\lim_{\zeta \to 0} \frac{\|u - x_0\|}{\sin\zeta} = \lim_{\zeta \to 0} \sqrt{\frac{2}{1 + \cos\zeta}} = 1.$$

To show (ii) consider

$$\|u - x^u\|^2 = \left\|u - \frac{x}{u^H x}\right\|^2 = \|u\|^2 - 2\mathrm{Re}\frac{u^H x}{u^H x} + \frac{\|x\|^2}{|u^H x|^2} = \frac{1}{\cos^2\zeta} - 1 = \tan^2\zeta,$$

which gives (2.25). Assertion (iii) is obtained in the same way, where we end up with $\|u - x_u\|^2 = 1 - \cos^2\zeta = \sin^2\zeta$. □

Chapter 3

Nonlinear Rayleigh Functionals

This chapter provides the basic theory on *nonlinear Rayleigh functionals*. Such functionals are the generalization of *Rayleigh quotients* for matrices. While analyzing the state of the art we show occurring problems with the existing definitions and give more general definitions, in the sense that we do not restrict ourselves to hyperbolic problems, but admit complex eigenvalues and unstructured problems. In order to get local uniqueness, however, we need to restrict the admissible domains. As there is a one-sided and a two-sided Rayleigh quotient, we introduce the *one-sided* and the *two-sided Rayleigh functional*, obtain bounds and first order perturbation results of the same kind and order as they exist for linear problems, where we can see explicitly that the condition number of the eigenvalue plays a role. Stationarity properties hold in the same way.

The chapter is concluded by an analysis of the *generalized Rayleigh quotient* proposed by Lancaster for matrix polynomials. We study the connection between the generalized Rayleigh quotient and the two-sided Rayleigh functional, and compare another variant without left eigenvector approximations with the one-sided Rayleigh functional.

3.1 Introduction and Historical Review

The Rayleigh quotient, though known since it was introduced by Lord Rayleigh [69] in 1873, still plays an enormous role in processes of eigenvalue computations, not only as self-contained Rayleigh quotient iteration but more during inner iterations of projection methods as *Jacobi–Davidson* [81] or *nonlinear Arnoldi* [94]. There, Rayleigh–Ritz equations of the form

$$U^H T(\lambda) U c = 0 \tag{3.1}$$

are generated when applying Ritz–Galerkin conditions [7], where the columns u_i of the $n \times m$ matrix U form an orthonormal basis for the search subspace.

But not only orthogonal projections are used in inner loops of eigenvalue computation methods, also oblique projections occur in form of the equations

$$V^H T(\lambda) U d = 0, \qquad U^H T(\lambda)^H V e = 0, \tag{3.2}$$

that emerge when Petrov–Galerkin conditions are applied. These equations are solved when left and right eigenvectors are computed, as is done by the *two-sided Jacobi–Davidson method* [37] and the *generalized Jacobi–Davidson method* [78]. Here, the $n \times m$ matrix V forms a basis for the left search subspace.

For one-dimensional subspaces $U = u$ and $V = v$ and a linear operator $T(\lambda) = A - \lambda I$, the solution vectors c, d and e are scalars, and (3.1), resp. the left equation of (3.2), are equivalent to

$$u^H(A - \lambda I)u = 0, \quad \text{or} \quad v^H(A - \lambda I)u = 0.$$

Rearranging for λ, yields the *Rayleigh quotient* $\lambda \equiv p(u) : \mathbb{C}^n \setminus \{0\} \to D \in \mathbb{C}$, respectively, the *two-sided or generalized Rayleigh quotient*, introduced by Ostrowski [66], $\lambda \equiv p(u, v) : \mathbb{C}^n \times \mathbb{C}^n \to D \in \mathbb{C}$, i.e.,

$$p(u) = \frac{u^H A u}{u^H u}, \qquad p(u, v) = \frac{v^H A u}{v^H u}, \tag{3.3}$$

respectively, provided that $u^H u \neq 0$, $v^H u \neq 0$ resp.

The Rayleigh quotient $p(u) \equiv p(u, A)$ has some beautiful properties [67] as there are

1. Homogeneity: $p(\alpha u, \beta A) = \beta p(u, A), \quad \alpha, \beta \neq 0,$
2. Translation Invariance: $p(u, A - \alpha I) = p(u, A) - \alpha,$
3. Boundedness: $p(u)$ gives the *field of values* of A, which is defined as

   $$\{u^H A u : \ u \in \mathbb{C}^n \text{ and } u^H u = 1\},$$

   [39, p. 321], if u ranges over all nonzero vectors in $\mathbb{C}^n$, and is closed, bounded and convex.
4. Stationarity: For *normal* A, i.e., A satisfies $AA^H = A^H A$, the eigenvectors of A yield the *stationary points* of p.
5. Minimal Residual: Given $u \neq 0$, then for any scalar θ

   $$\|(A - \theta I)u\| \geq \|(A - p(u)I)u\|, \tag{3.4}$$

   with equality only when $\theta = p(u)$.

Let us discuss the stationarity of the Rayleigh quotient. Actually we cannot speak of stationarity, since $p(u)$ is not differentiable with respect to u if u is a complex vector. However, in the literature this is rather ignored, and we will follow the nomenclature by extending the definition of stationarity to the complex case as follows:

Definition 3.1 *A function f is called stationary at z if*

$$f(z + \Delta z) - f(z) = \mathcal{O}(\|\Delta z\|^2),$$

i.e, if the first order terms in a perturbation expansion vanish identically.

This definition includes the real differentiable case where we have $\nabla f(z) = 0$, when f is stationary at z. An evaluation of the term $p(u+s)-p(u)$ for $u = x_*$ shows the stationarity of the Rayleigh functional for normal matrices, immediately. This computation shows also that the Rayleigh quotient is not stationary for nonnormal matrices. This is cured by the two-sided Rayleigh quotient $p(u, v)$ which was developed in [66]. For the two-sided Rayleigh quotient, homogeneity and translation invariance hold as well, but the boundedness property fails. A small example is given by the real symmetric matrix $A = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 2 \end{smallmatrix}\right]$. For all unit norm vectors $x \in \mathbb{C}^2$ the field of values coincides with the Rayleigh quotient $p(x)$, and it can be determined by the Rayleigh–Ritz Theorem [39, p. 176], which gives $1 \leq x^H Ax \leq 2$. Moreover, $p(u)$ will always be real. In contrast, the generalized Rayleigh quotient does not need to be real and is not included in the field of values, as, for instance, vectors $u = [1 \ -i]^H$ and $v = [-1 \ 1]^T$ yielding $p(u, v) = (3 - i)/2$ show. But if we suppose unit norm vectors u and v, then the field of values is a subset of the values that can be generated by the generalized Rayleigh quotient.

Now, consider the nonlinear eigenvalue problem

$$T(\lambda)x = 0, \tag{3.5}$$

where $T(.) : D \to \mathbb{C}^{n\times n}$ is a matrix-valued mapping, with Lipschitz continuous derivative $\dot{T}(\lambda)$, if real-valued, or holomorphic, if complex-valued. Both cases give

$$\|\dot{T}(\lambda) - \dot{T}(\mu)\| \leq L|\lambda - \mu|, \tag{3.6}$$

where $L > 0$, cf. (2.5), (2.11).

In the following u and v are used as approximations for the right and left eigenvectors x_* and y_*, $\|x_*\| = \|y_*\| = 1$, corresponding to the simple eigenvalue λ_*.

Different attempts to find generalizations of the Rayleigh quotient that work for certain nonlinear eigenvalue problems have been made, cf. [26, 45, 48, 79, 82], which provide actual

quotients. In general, some of the desirable properties of the Rayleigh quotient are lost for the generalized quotients. The approach of the Rayleigh functional is more promising.

We have reviewed the definition of the Rayleigh functional in §1.3, (1.15)–(1.18). This definition is restricted to a set of real eigenvalues for different types of problems in order to derive variational principles analogously to the well-known principles for matrices. When problems of a more general structure, which do not have subsets with real eigenvalues only, are considered, existence and convergence properties of a matching Rayleigh functional are not always clear and have to our knowledge not been shown yet in the complex case.

To make the abstract definition of a Rayleigh functional more understandable we consider some examples:

- For the generalized eigenvalue problem $T(\lambda) = \lambda B - A$, equation (1.17) gives

$$\lambda = p(u) = \frac{u^H A u}{u^H B u},$$

  provided that $u^H B u \neq 0$. This is just the Rayleigh quotient for the generalized linear problem. If B is Hermitian, then condition (1.18) requires that $u^H B u > 0$, i.e., that B is positive definite. Hence, the quotient is well-defined in this case.

- For the quadratic problem $T(\lambda) = \lambda^2 A + \lambda B + C$, equation (1.17) equals

$$u^H T(\lambda) u = \lambda^2 u^H A u + \lambda u^H B u + u^H C u = 0, \tag{3.7}$$

  which yields, provided that $u^H A u \neq 0$,

$$\lambda \equiv p(u) = \frac{1}{2u^H A u} \left[\pm \sqrt{(u^H B u)^2 - 4u^H C u (u^H A u)} - u^H B u \right], \tag{3.8}$$

  and (1.18) requires that $u^H (2p(u)A + B)u > 0$. This condition means that if we assume that A is positive definite and B is positive semidefinite, then only those $p(u)$ are taken into consideration which are real and for which $p(u) > -(u^H B u)/(u^H A u)$ holds.

- For the example given in (2.3), where $T(\lambda) = \mathrm{diag}(2 - \exp(\lambda), 1, \ldots, 1)$, the Rayleigh functional is defined as a solution of $u^H T(\lambda) u = 0$. This gives

$$\lambda \equiv p(u) = \ln\left(2 + \frac{|u_2|^2 + \ldots + |u_n|^2}{|u_1|^2}\right).$$

  Inserting the eigenvector e_1 yields the exact eigenvalue $\lambda_* = \ln(2)$.

Definition (1.15)–(1.18) is problematic, because without restrictions on the domain of $p(u)$, the Rayleigh functional does not need to be unique. And even if we restrict the admissible set of vectors, we will in general have more than one value satisfying the defining equation (1.17). Consider, for instance, a polynomial problem of degree m. Then, $\lambda = p(u)$ is defined through the polynomial equation

$$\sum_{l=0}^{m} \lambda^l u^H A_l u = 0,$$

which has up to m solutions. The question which one to take is essential. Therefore, we have to restrict the range of p as well.

For the following analysis we define the set of vectors that are admissible for the Rayleigh functional p, depending on the angle to the exact eigenvector x_*

$$\mathcal{K}_\varepsilon(x_*) = \{u \in \mathbb{C}^n : \ \measuredangle(\text{span}\{u\}, \text{span}\{x_*\}) \leq \varepsilon\}, \tag{3.9}$$

where we only need the additional assumption $(\dot{T}(p(u))u, u) \neq 0$ instead of (1.18). Then, the new setting, defining the standard nonlinear Rayleigh functional, will be

$$p : u \in \mathcal{K}_\varepsilon(x_*) \longrightarrow p(u) \in \bar{S}(\lambda_*, \tau_*) \subset \mathbb{C}, \tag{3.10}$$

$$p(cu) = p(u) \quad \text{for all} \quad c \neq 0, \tag{3.11}$$

$$(T(p(u))u, u) = 0, \tag{3.12}$$

$$(\dot{T}(p(u))u, u) \neq 0. \tag{3.13}$$

for some $\varepsilon < \pi/2$ and $\tau_* > 0$. This setting is particularly suited for Hermitian problems with real eigenvalues, since then (3.13) restricts the functional to vectors u close to x_* with simple eigenvalue λ_*. We will see that we cannot prove (local) existence without the simplicity condition. However, since in general methods using the one-sided Rayleigh quotient are also used to solve non-Hermitian problems, we need to know the behavior of the Rayleigh functional applied to general problems. Then, condition (3.13) does not guarantee the simplicity of the eigenvalue.

As expected, we will show that the Rayleigh functional is not stationary for $T(\lambda)$ when left and right eigenvectors are different.

Remark 3.2 *The original definition of Rayleigh functionals was based on Hermitian, mostly positive definite, matrices. Neglecting this condition means that we admit indefinite A, in case of the quadratic operator $T(\lambda) = \lambda^2 A + \lambda B + C$, hence, it may happen that $u^H A u$ is zero. Then, equation* (3.8) *will not be well-defined. Therefore, we have to consider equation* (3.7) *first, in any case. If $u^H A u = 0$, then* (3.7) *reduces to*

$$u^H T(p(u))u = p(u)u^H B u + u^H C u = 0,$$

and the Rayleigh functional can be determined uniquely as $p(u) = -(u^H Cu)/(u^H Bu)$. *Since we assume that* (3.13) *holds, i.e, that* $2\lambda u^H Au + u^H Bu \neq 0$, *the case where* $u^H Au = 0$ and $u^H Bu = 0$ *is excluded.*

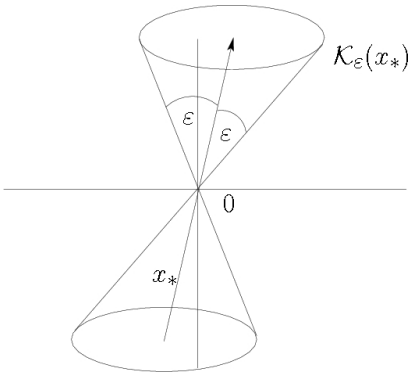

Figure 3.1: *The double circular cones* $\mathcal{K}_\varepsilon(x_*)$ *for real vectors with vertex* 0 *and vertex angle* 2ε.

Unfortunately, the minimal residual property (3.4) of the Rayleigh quotient is lost for the Rayleigh functional, in general. Consider, for instance the example in (2.3) with $x^T = (1, -1, \ldots, -1)$ and $\theta = \ln(2)$. Then, the Rayleigh functional $\lambda = p(x)$ is determined from $x^T T(\lambda)x = n + 1 - \exp(\lambda) = 0$, i.e., we have $p(x) = \ln(n+1)$. Hence, the norm of the residual with Rayleigh functional is greater than the norm of the residual with T at θ for $n > 2$, since in this case $\|T(p(x))x\| = \sqrt{n(n-1)} > \|T(\theta)x\| = \sqrt{n-1}$, in contrast to the behavior of the Rayleigh quotient for matrices.

Ostrowski [66] designed the two-sided Rayleigh quotient in order to have a stationary Rayleigh quotient for nonnormal matrices. In the same manner we define the generalized or two-sided nonlinear Rayleigh functional by

$$p : (u, v) \in \mathcal{K}_\varepsilon(x_*) \times \mathcal{K}_\varepsilon(y_*) \longrightarrow p(u, v) \in \bar{S}(\lambda_*, \tau_*) \subset \mathbb{C},$$

$$p(cu, dv) = p(u, v) \quad \text{for all} \quad c \neq 0, d \neq 0, \tag{3.14}$$

$$(T(p(u, v))u, v) = 0, \tag{3.15}$$

$$(\dot{T}(p(u, v))u, v) \neq 0. \tag{3.16}$$

for some $\varepsilon < \pi/2$, $\tau_* > 0$, which will also be shown to be stationary. Let us point out that $\varepsilon < \pi/2$ implies $u \neq 0$, $v \neq 0$, since any vector is orthogonal to the zero vector. For general problems, condition (3.16) restricts the functional to vectors close to eigenvectors corresponding to a simple eigenvalue.

Remark 3.3 *For the quadratic problem* (1.2), *equation* (3.15) *equals*

$$v^H T(p(u, v))u = p(u, v)^2 v^H Au + p(u, v)v^H Bu + v^H Cu = 0, \tag{3.17}$$

which yields

$$p(u,v) = \frac{1}{2v^HAu}\left[\pm\sqrt{(v^HBu)^2 - 4v^HCu(v^HAu)} - v^HBu\right],$$

provided that $v^HAu \neq 0$. *If* $v^HAu = 0$, *however, then* $p(u,v)$ *cannot be determined by this equation. In analogy to the case described in Remark 3.2, equation* (3.17) *reduces to* $p(u,v)v^HBu + v^HCu = 0$, *and we have the unique Rayleigh functional* $p(u,v) = -(v^HCu)/(v^HBu)$. *Condition* (3.16) *guarantees* $v^HBu \neq 0$, *then.*

Notice that $(\cdot,\cdot)$, defined by $(x,y) = y^Hx$, is a sesquilinear form, i.e., linear in the first argument, i.e., $(cx,y) = c(x,y)$, and antilinear in the second argument, i.e., $(x,cy) = \bar{c}(x,y)$ for all $c \in \mathbb{C}$.

The aim of this chapter is to analyze and characterize these Rayleigh functionals, in the first place with respect to general problems, then we will specify the results to certain structured problems.

We will show that a unique solution $p = p(u,v) \in \bar{S}(\lambda_*, \tau_0)$ of (3.15) approximating a simple eigenvalue λ_* exists, provided that $(u,v) \in \mathcal{K}_\varepsilon(x_*) \times \mathcal{K}_\varepsilon(y_*)$ for sufficiently small ε. One of the main results in this chapter will be the bound for the distance of the Rayleigh functional and the exact eigenvalue, which is quadratic in the corresponding eigenvector angles. Section 3.2, in general, deals with the generalized functional $p(u,v)$, Section 3.3 with the standard (one-sided) functional $p(u)$, which does not include information on the left eigenvector. Another main result is given by the first order perturbation expansion for $p(u,v)$, see Theorem 3.15, which implies stationarity of the functional, see Definition 3.1 and Lemma 3.18, and corresponding statements for the standard functional. Although all statements are made for complex $T(\lambda)$ assuming it to be holomorphic on some open disk, which implies existence of arbitrary many derivatives, it suffices for real problems that $\dot{T}(\lambda)$ is Lipschitz continuous, see (3.6).

Note, that Lancaster [49, p. 71] defines a generalization of the Rayleigh quotient for matrix polynomials in a different way by

$$p_L(\lambda,u,v) = \lambda - \frac{v^HT(\lambda)u}{v^H\dot{T}(\lambda)u}, \tag{3.18}$$

and proposes a generalized Rayleigh quotient iteration converging with quadratic order. If $p_L(\lambda_k,u,v)$ is viewed as new approximate eigenvalue λ_{k+1}, then (3.18) is obtained by performing one Newton step for the scalar operator

$$g(\lambda,u,v) := (T(\lambda)u,v) = 0 \tag{3.19}$$

with respect to λ from the current approximation λ_k for fixed u, v.

The connection of (3.18) and the generalized Rayleigh functional is discussed in Section 3.4. The distance to the exact eigenvalue is analyzed as well as the distance of $p_L(\lambda, u, v)$ and $p(u, v)$, for given λ. Note that for linear problems $T(\lambda) = A - \lambda I$, equation (1.17) yields the left expression of (3.3), whilst (3.15) and (3.18) yield the right one.

We will denote the angles of approximations and eigenvectors as follows

$$\xi := \measuredangle(\operatorname{span}\{u\}, \operatorname{span}\{x_*\}), \quad \eta := \measuredangle(\operatorname{span}\{v\}, \operatorname{span}\{y_*\}). \tag{3.20}$$

3.2 Existence and Stationarity of the Generalized Rayleigh Functional

3.2.1 Real-valued Problems

The first part of the analysis is done assuming that we have real eigenvectors and real eigenvalues. This assumption is different and in some sense more restrictive than definiteness, respectively hyperbolicity, which assure real eigenvalues and have been established for the generalized eigenvalue problem [85], for the quadratic eigenvalue problem [87], and for polynomial problems of arbitrary degree [55].

The assumption to have real vectors enables us to apply the implicit function theorem for vectors (u, v) in a neighborhood of (x_*, y_*), to show existence of a solution of (3.15). In this context, equation (3.19) reads as $g(p, u, v) = v^T T(p)u = 0$. The derivatives are given by $\partial_p g = v^T \dot{T}(p)u$, $\partial_u g = v^T T(p)$ and $\partial_v g = u^T T(p)^T$. At the solution $u = x_*$, $v = y_*$, $p = \lambda_*$ we have $\partial_p g(\lambda_*, x_*, y_*) = y_*^T \dot{T}(\lambda_*)x_* \neq 0$, hence $p(u, v)$ exists and is differentiable in a neighborhood of (x_*, y_*) and is locally uniquely defined by $g(p, u, v) = 0$. Thus, we obtain $g(p(u, v), u, v) \equiv 0$. Differentiation with respect to u gives

$$\partial_p g \partial_u p + \partial_u g = 0, \quad \text{i.e.,} \quad \partial_u p = -(\partial_p g)^{-1} \partial_u g = -\frac{v^T T(p)}{v^T \dot{T}(p)u},$$

and with respect to v

$$\partial_p g \partial_v p + \partial_v g = 0, \quad \text{i.e.,} \quad \partial_v p = -(\partial_p g)^{-1} \partial_v g = -\frac{u^T T(p)^T}{v^T \dot{T}(p)u},$$

and we get the Taylor expansion

$$p(u + s, v + t) = p(u, v) - \frac{v^T T(p)s}{v^T \dot{T}(p)u} - \frac{u^T T(p)^T t}{v^T \dot{T}(p)u} + R \tag{3.21}$$

with remainder R, $|R| \leq K(\|s\| + \|t\|)^2$ for some $K > 0$. The following Lemma shows stationarity in this setting even in the sense that directional derivatives vanish.

Lemma 3.4 *Let $\lambda_* \in D$ be a simple eigenvalue of $T(\,.\,) : D \subset \mathbb{R} \to \mathbb{R}^{n\times n}$, let $\dot{T}(\,.\,)$ be Lipschitz continuous on D, and let x_*, y_* be the corresponding real right and left eigenvectors. Assume that (u, v) are sufficiently close to (x_*, y_*). Then, the real Rayleigh functional p is stationary at the eigenvectors (x_*, y_*) with $\lambda_* = p(x_*, y_*)$.*

Proof. Since λ_* is simple, we have $y_*^T \dot{T}(\lambda_*) x_* \neq 0$. In addition, the vectors (u, v) need to be sufficiently close to (x_*, y_*) in order to apply the implicit function theorem. Then, the Taylor expansion (3.21) holds. Setting $u = x_*$, $v = y_*$ in (3.21) proves the assertion. □

However, to assume real vectors is very restrictive and we want to avoid this assumption in what follows. In this case, partial differentiability of g with respect to v is not given—the Cauchy–Riemann differential equations do not hold for functionals $v \mapsto v^H a$ with $a \in \mathbb{C}^n$. Even in case of the linear problem $T(\lambda) = A - \lambda I$, where $v^H(A - p(u, v)I)u = 0$ yields the well-known generalized Rayleigh quotient $p(u, v) = v^H Au/v^H u$, the quotient is not differentiable with respect to v. On the other hand, considering $p(u, v)^H$ makes it possible to differentiate with respect to v but not with respect to u. We need a representation which is uniform in both variables making this a more complicated issue.

3.2.2 Complex-valued Problems

Local existence

The implicit function theorem does not apply, since $g(p, u, v) = v^H T(p)u$ is not differentiable with respect to v. Therefore existence of $p(u, v)$ must be shown in another way. Using the Banach fixed point theorem provides the desired result and gives a bound for the distance of the Rayleigh functional and the exact eigenvalue in terms of the angles of the corresponding eigenvectors and approximations as defined in (3.20), immediately. As we have discussed in §2.5, bounds in terms of angles are more relevant, (respectively they are the proper measure,) than bounds in terms of norms, especially in the case of complex vectors.

Theorem 3.5 *Let λ_* be a simple eigenvalue of (3.5), x_*, y_* be the corresponding eigenvectors with unit norm and let $T(\lambda)$ be holomorphic on an open neighborhood of $\bar{S}(\lambda_*, \tau_*)$, with $\tau_* > 0$. Then there exist constants $0 < \tau_0 < \tau_*$, $0 < \varepsilon_0 < \pi/2$ such that, for*

all $(u,v) \in \mathcal{K}_{\varepsilon_0}(x_*) \times \mathcal{K}_{\varepsilon_0}(y_*)$*, there exists a unique* $p = p(u,v) \in S_0 := \bar{S}(\lambda_*, \tau_0)$ *with* $g(p,u,v) = (T(p(u,v))u, v) = 0$*. Moreover, one has*

$$|p(u,v) - \lambda_*| \quad \leq \quad \frac{8}{3} \frac{\|T(\lambda_*)\|}{|y_*^H \dot{T}(\lambda_*) x_*|} \tan\eta \, \tan\xi. \tag{3.22}$$

Proof. The outline of the proof is as follows. We derive a fixed point equation $G(p,u,v) = p$ for the Rayleigh functional p by introducing several new variables and apply the Banach fixed point theorem [32, p. 35] to this characterization. Therefore, we need to show contractivity of G and $G(S_0) \subset S_0$. The upper bound (3.22) is obtained as a byproduct.

First, we specify the constants. Let $\alpha = y_*^H \dot{T}(\lambda_*) x_*$, which is nonzero due to the simplicity of λ_*, and

$$\tau_0 \quad := \quad \min\left\{ \tau_*, \tau_1 := \frac{4}{17} \frac{|\alpha|}{L} \right\}, \tag{3.23}$$

$$\varepsilon_0 \quad := \quad \min\left\{ \varepsilon_1 := \arctan \frac{2}{9} \frac{|\alpha|}{\|\dot{T}(\lambda_*)\|}, \, \varepsilon_2 := \arctan \frac{4\tau_0 \, |\alpha|}{20 \|T(\lambda_*)\|} \right\}, \tag{3.24}$$

with L from (3.6). We use the singular value decomposition for $T(\lambda_*)$ as defined in (2.12)

$$T(\lambda_*) = Y \Sigma X^H = [\, Y_1 \mid y_* \,] \left[\begin{array}{c|c} \Sigma_1 & 0 \\ \hline 0^T & 0 \end{array} \right] [\, X_1 \mid x_* \,]^H,$$

and linearize T by

$$T(p) = T(\lambda_*) + (p - \lambda_*) \dot{T}(\lambda_*) + R, \tag{3.25}$$

with remainder $R = R(p, \lambda_*)$, where $\|R\| \leq \frac{L}{2} |p - \lambda_*|^2$. The decomposition

$$u = X_1 u_1 + x_* u_2 = [X_1 \; x_*] \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \qquad v = Y_1 v_1 + y_* v_2 = [Y_1 \; y_*] \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}, \tag{3.26}$$

yields

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} X_1^H u \\ x_*^H u \end{bmatrix}, \quad \text{with} \quad \|u_1\| = \|X_1^H u\| = \|u\| \sin\xi, \quad |u_2| = |x_*^H u| = \|u\| \cos\xi, \tag{3.27}$$

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} Y_1^H v \\ y_*^H v \end{bmatrix}, \quad \text{with} \quad \|v_1\| = \|Y_1^H v\| = \|v\| \sin\eta, \quad |v_2| = |y_*^H v| = \|v\| \cos\eta.$$

In these terms we have

$$\begin{aligned} g(p,u,v) &= v_2{}^H u_2 \left[y_*^H + \left(\frac{v_1}{v_2}\right)^H Y_1^H \right] \Big[T(\lambda_*) + (p-\lambda_*)\dot{T}(\lambda_*) + R \Big] \left[x_* + X_1 \frac{u_1}{u_2} \right] \\ &= v_2{}^H u_2 \Bigg\{ y_*^H T_* x_* + y_*^H T_* X_1 \frac{u_1}{u_2} + \left(\frac{v_1}{v_2}\right)^H Y_1^H T_* x_* + \left(\frac{v_1}{v_2}\right)^H Y_1^H T_* X_1 \frac{u_1}{u_2} + \\ &\quad (p-\lambda_*) \left[y_*^H \dot{T}_* x_* + \left(\frac{v_1}{v_2}\right)^H Y_1^H \dot{T}_* x_* + y_*^H \dot{T}_* X_1 \frac{u_1}{u_2} + \left(\frac{v_1}{v_2}\right)^H Y_1^H \dot{T}_* X_1 \frac{u_1}{u_2} \right] \\ &\quad + \left[y_*^H + \left(\frac{v_1}{v_2}\right)^H Y_1^H \right] R \left[x_* + X_1 \frac{u_1}{u_2} \right] \Bigg\}, \end{aligned}$$

with $T_* \equiv T(\lambda_*)$. For the sake of readability, we introduce the following variables

$$\beta = \beta(u,v) := \left(\frac{v_1}{v_2}\right)^H Y_1^H \dot{T}_* x_* + y_*^H \dot{T}_* X_1 \frac{u_1}{u_2} + \left(\frac{v_1}{v_2}\right)^H Y_1^H \dot{T}_* X_1 \left(\frac{u_1}{u_2}\right), \tag{3.28}$$

$$\gamma = \gamma(u,v) := \left(\frac{v_1}{v_2}\right)^H \Sigma_1 \frac{u_1}{u_2}, \quad \hat{u} := x_* + X_1 \frac{u_1}{u_2} \quad \text{and} \quad \hat{v} := y_* + Y_1 \frac{v_1}{v_2}.$$

With these abbreviations the equation $g(p,u,v) = 0$, which we want to solve, is equivalent to

$$v_2^H u_2 \left\{ \gamma + (p-\lambda_*)[\,\alpha + \beta\,] + \hat{v}^H R \hat{u} \right\} = 0.$$

Since $v_2^H u_2 \neq 0$ this is equivalent to

$$\gamma + (p-\lambda_*)[\,\alpha + \beta\,] + \hat{v}^H R \hat{u} = 0. \tag{3.29}$$

In order to solve for the linear part of p, we first have to make sure that $\alpha + \beta$ is nonzero.

Proposition 3.6 *Under the assumptions of Theorem 3.5 and with β from (3.28), let $\varepsilon_1 \in (0, \pi/2)$ be defined by $\tan\varepsilon_1 = \frac{2}{9}\frac{|y_*^H \dot{T}(\lambda_*) x_*|}{\|\dot{T}(\lambda_*)\|} =: \frac{2}{9}\, q_*$. Then*

$$|\beta(u,v)| \leq \frac{1}{2}\,|\alpha| \tag{3.30}$$

holds for all $(u,v) \in \mathcal{K}_{\varepsilon_1}(x_*) \times \mathcal{K}_{\varepsilon_1}(y_*)$.

Proof. Taking the estimate $\tan\varepsilon_1 = \frac{2}{9} q_* \leq \frac{2}{9} < \frac{1}{4}$ into account, we obtain

$$\begin{aligned} |\beta| &\leq \left\| \frac{v_1}{v_2} \right\| \|Y_1\| \|\dot{T}_*\| \|x_*\| + \left\| \frac{u_1}{u_2} \right\| \|\dot{T}_*\| \|X_1\| \|y_*\| + \left\| \frac{v_1}{v_2} \right\| \left\| \frac{u_1}{u_2} \right\| \|Y_1\| \|\dot{T}_*\| \|X_1\| \\ &= \|\dot{T}_*\| \left[\tan\eta + \tan\xi + \tan\eta \tan\xi \right] \\ &\leq \|\dot{T}_*\| \left[2\tan\varepsilon_1 + \frac{1}{4}\tan\varepsilon_1 \right] \leq \|\dot{T}_*\| \frac{9}{4} \tan\varepsilon_1 = \frac{1}{2}|\alpha|. \qquad \square \end{aligned}$$

With (3.24) the assumptions are satisfied and inequality (3.30) implies that $|\alpha+\beta| \geq |\alpha| - |\beta| \geq \frac{1}{2}|\alpha| > 0$, hence we can rearrange (3.29) for p and derive the fixed point equation

$$p = \lambda_* - \frac{1}{\alpha+\beta(u,v)} \left\{ \gamma(u,v) + \hat{v}^H R(p,\lambda_*)\hat{u} \right\} =: G(p,u,v), \tag{3.31}$$

to which we want to apply the Banach fixed point theorem. We start by showing contractivity of G.

Proposition 3.7 *Under the assumptions of Theorem 3.5 and with L from* (3.6)*, let $\tau_1 := \frac{4}{17}\frac{|\alpha|}{L}$ and $\tau_0 := \min\{\tau_*, \tau_1\}$. Then, for G as defined in* (3.31)*, there exists a constant $\kappa < 1$ such that for all $(u,v) \in \mathcal{K}_{x_*}(\varepsilon_1) \times \mathcal{K}_{y_*}(\varepsilon_1)$ we have*

$$|G(p,u,v) - G(\mu,u,v)| \leq \kappa |p-\mu| \quad \textit{for all} \;\; p,\, \mu \in S_0 := \bar{S}(\lambda_*, \tau_0).$$

Proof. Considering (3.31) we get

$$|G(p,u,v) - G(\mu,u,v)| \;\leq\; \frac{\left|\hat{v}^H (R(p,\lambda_*) - R(\mu,\lambda_*))\hat{u}\right|}{|\alpha+\beta|} \leq \frac{2\|\hat{v}\|\|\hat{u}\|\|R(p,\lambda_*) - R(\mu,\lambda_*)\|}{|\alpha|}.$$

Since the summands of $\hat{u} = x_* + X_1 \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$ are orthogonal we have $\|\hat{u}\| = \sqrt{1+\tan^2\xi} \leq \sqrt{1+\tan^2\varepsilon_1} \leq \sqrt{\frac{17}{16}}$. An analogous inequality is obtained for $\hat{v}$. By an elementary extension of Taylors formula (see, e.g., [76]) we derive

$$\|R(p,\lambda_*) - R(\mu,\lambda_*)\| \leq \frac{L}{2}\big(|p-\lambda_*| + |\mu-\lambda_*|\big)|p-\mu| \leq L\tau_0|p-\mu|,$$

and altogether

$$|G(p,u,v) - G(\mu,u,v)| \leq \frac{2}{|\alpha|}\frac{17}{16}L\tau_0|p-\mu| \leq \frac{2}{|\alpha|}\frac{17}{16}L\tau_1|p-\mu| = \frac{1}{2}|p-\mu|,$$

i.e., G is contractive on S_0 with constant $\kappa = 1/2$. □

Next we show that G maps S_0 onto S_0.

Proposition 3.8 *Assume $\mathcal{A}_{\mathbb{C}}$. Let $\tan\varepsilon_2 = \sqrt{\frac{\tau_0\,|\alpha|}{4\|T(\lambda_*)\|}}$, $\varepsilon_0 := \min\{\varepsilon_1, \varepsilon_2\}$. Then, for G as defined in* (3.31)*,*

$$G(S_0) \subset S_0 \quad \textit{is satisfied for all} \;\; (u,v) \in \mathcal{K}_{\varepsilon_0}(x_*) \times \mathcal{K}_{\varepsilon_0}(y_*).$$

Proof. If we prove that

$$|G(\lambda_*,u,v) - \lambda_*| \leq \frac{1}{2}\,\tau_0, \tag{3.32}$$

then $G(S_0) \subset S_0$ holds, since then we have

$$|G(p,u,v) - \lambda_*| \leq |G(p,u,v) - G(\lambda_*,u,v)| + |G(\lambda_*,u,v) - \lambda_*| \leq \frac{1}{2}|p-\lambda_*| + \frac{1}{2}\tau_0 \leq \tau_0$$

for $p \in S_0$. Thus, it suffices to show (3.32). From (3.31) we obtain

$$\begin{aligned} |G(\lambda_*,u,v) - \lambda_*| &= \frac{\left|\gamma(u,v) + \hat{v}^H R(\lambda_*,\lambda_*)\hat{u}\right|}{|\alpha + \beta(u,v)|} = \frac{|\gamma(u,v)|}{|\alpha+\beta(u,v)|} \leq \frac{2}{|\alpha|}\tan\xi\tan\eta\|T(\lambda_*)\| \\ &\leq \frac{2}{|\alpha|}\tan^2\varepsilon_2\|T(\lambda_*)\| = \frac{1}{2}\tau_0. \qquad \square \end{aligned}$$

Hence, all conditions required for applying the fixed point theorem hold and the existence of a unique solution $p = p(u,v)$ in S_0 such that $g(p,u,v) = 0$ for $(u,v) \in \mathcal{K}_{\varepsilon_0}(x_*) \times \mathcal{K}_{\varepsilon_0}(y_*)$ follows.

It remains to show inequality (3.22), which can be done by considering (3.31)

$$\begin{aligned} |p-\lambda_*| &= \frac{|\gamma(u,v) + \hat{v}^H R(p,\lambda_*)\hat{u}|}{|\alpha+\beta|} \\ &\leq \frac{2}{|\alpha|}\left(|\gamma(u,v)| + \|\hat{v}\|\|\hat{u}\|\|R(p,\lambda_*)\|\right) \\ &\leq \frac{2}{|\alpha|}\left(\|T(\lambda_*)\|\tan\xi\tan\eta + \frac{17}{16}\frac{L}{2}|p-\lambda_*|^2\right). \end{aligned} \tag{3.33}$$

According to $|p-\lambda_*|^2 \leq \tau_0|p-\lambda_*|$ we have $\left(1 - \frac{17L\tau_0}{16|\alpha|}\right)|p-\lambda_*| \leq \frac{2\|T(\lambda_*)\|}{|\alpha|}\tan\xi\tan\eta$ and additionally $\frac{17L\tau_0}{16|\alpha|} \leq \frac{17L\tau_1}{16|\alpha|} = \frac{1}{4}$ which yields

$$|p-\lambda_*| \leq \frac{8}{3}\frac{\|T(\lambda_*)\|}{|\alpha|}\tan\xi\tan\eta,$$

i.e., the bound (3.22). $\square$

Remark 3.9 *Theorem 3.5 in general does not hold for eigenvalues that are not simple. In this case it may happen that $\alpha = y_*^H \dot{T}(\lambda_*)x_*$ is zero, hence α can be zero in some neighborhood $\alpha+\beta$ of α, and the fixed point characterization* (3.31) *is not admissible.*

In order to evaluate the sharpness of the bound (3.22), we take a closer look at (3.33) at the limit $\xi, \eta \to 0+$, by considering the components

$$\begin{aligned} \frac{|\gamma|}{\tan\xi\tan\eta} &= \frac{\left|\left(\frac{v_1}{v_2}\right)^H \Sigma_1 \frac{u_1}{u_2}\right|}{\tan\xi\tan\eta} \leq \|\Sigma_1\| = \|T(\lambda_*)\|, \\ |\alpha+\beta\,| &= |y_*^H\dot{T}(\lambda_*)x_* + \mathcal{O}(\tan\xi + \tan\eta + \tan\xi\tan\eta)| \longrightarrow |\alpha|, \\ \frac{|\hat{v}^H R\hat{u}|}{\tan\xi\tan\eta} &= \mathcal{O}\left(\frac{(p-\lambda_*)^2}{\tan\xi\tan\eta}\right) = \mathcal{O}(\tan\xi\tan\eta), \end{aligned}$$

of the equivalent formulation

$$\frac{|p-\lambda_*|}{\tan\xi\tan\eta} = \frac{\left|\frac{\gamma}{\tan\xi\tan\eta} + \frac{\hat{v}^H R\hat{u}}{\tan\xi\tan\eta}\right|}{|\alpha+\beta|}$$

of equation (3.33), which yields

$$\limsup_{\xi,\eta\to 0+} \frac{|p-\lambda_*|}{\tan\xi\tan\eta} \leq \frac{\|T(\lambda_*)\|}{|\alpha|}.$$

Equality is achieved if $\frac{v_1}{v_2} = (\tan\eta)e_1$ and $\frac{u_1}{u_2} = (\tan\xi)e_1$, where e_1 denotes the first unit vector. This means that the constant $8/3$ in the estimate (3.22) can asymptotically be replaced by 1, and the order in the angles is correct. To be more precise, the factor $8/3$ could be replaced by $1/(1-\omega)$, with $\omega = \mathcal{O}(\tan\xi + \tan\eta)$, which reflects the asymptotic case and holds also in the linear case, cf. [77].

Note that the constant $\kappa = 1/2$ in the proof of Proposition 3.7 can be chosen arbitrarily in $(0,1)$. But since all other constants depend on κ they have to be chosen appropriately.

Remark 3.10 *Suppose that $\xi < \pi/3$ and $\eta < \pi/3$. Then, one has $\tan\xi \leq 2\sin\xi$ and $\tan\eta \leq 2\sin\xi$, and the bound* (3.22) *implies*

$$|p(u,v)-\lambda_*| \leq K\sin\xi\sin\eta, \tag{3.34}$$

where $K = 32\|T(\lambda_*)\|/(3|y_*^H\dot{T}(\lambda_*)x_*|)$. We have shown in Lemma 2.10 that $\sin\xi \leq \|u-x_*\|$. Hence, with inequality (3.34) *we obtain*

$$|p(u,v)-\lambda_*| \leq K\|u-x_*\|\|v-y_*\|. \tag{3.35}$$

Notice that the norm terms may be large if the vectors have wrong scaling behavior. Rescaling of x_* and y_* helps in this case, cf. Lemma 2.11: Define $x_u = x_*(x_*^H u)$, $y_v = y_*(y_*^H v)$. Then we have $\|u-x_u\| = \sin\xi$, $\|v-y_v\| = \sin\eta$, hence (3.34) *gives*

$$|p(u,v)-\lambda_*| \leq K\sin\xi\sin\eta = K\|u-x_u\|\|v-y_v\|.$$

On the other hand, if $\|u-x_*\|$ and $\|v-y_*\|$ are smaller than $\sin(\pi/3) = \sqrt{3}/2$, then the angles ξ, η are smaller than $\pi/3$, too, and bound (3.34) *holds.*

A different kind of existence result, though without a bound, is given in [55] for polynomial eigenvalue problems. There, the set of all roots of all possible functions $(T(\lambda)u,u)$, $(u \neq 0)$ is called the *root domain* of the operator-valued function $T(\lambda) = \sum_{l=0}^{m} \lambda^l A_l$, where $A_l \in \mathbb{C}^{n\times n}$, $(l = 0,\ldots,m)$.

Theorem 3.11 *[55, p. 143] Suppose that the condition*

$$\inf_{\lambda\in\Gamma,\,\|u\|=1} |(T(\lambda)u,u)| > 0,$$

holds for the operator $T(\lambda)=\sum_{l=0}^{m}\lambda^l A_l$ *on the circle* $\Gamma=\{\lambda:\ |\lambda-\lambda_0|=\tau\}$*, and there are* l *roots of the polynomial* $(T(\lambda)u,u)$*,* $(u\neq 0)$ *in the disk* $G^+=\{\lambda:\ |\lambda-\lambda_0|<\tau\}$*. Then* $T(\lambda)$ *has a spectral divisor of order* l *whose spectrum coincides with the part of the spectrum of* $T(\lambda)$ *lying in* G^+*.*

Note, that a *monic* matrix polynomial $B(\lambda)=\sum_{l=0}^{m}\lambda^l B_l$, i.e., a polynomial with leading coefficient $B_m=I$, is called a *spectral divisor* of $T(\lambda)$ if there exists a matrix polynomial $C(\lambda)$ such that $T(\lambda)=C(\lambda)B(\lambda)$, and $B(\lambda)$ and $C(\lambda)$ have disjoint spectra.

Perturbation Expansion

The existence result given by Theorem 3.5 enables us to derive a first order perturbation expansion and the corresponding first order bound for $p(u,v)$. Preliminary to this, we prove two auxiliary propositions.

Proposition 3.12 *Let* $(u,v)\in\mathcal{K}_{\varepsilon_0/2}(x_*)\times\mathcal{K}_{\varepsilon_0/2}(y_*)$*,* $0\le\delta_u,\delta_v\le\delta_0:=\frac{\cos\left(\frac{\varepsilon_0}{2}\right)-\cos(\varepsilon_0)}{1+\cos(\varepsilon_0)}$ *with* $0<\varepsilon_0<\pi/2$*, where*

$$\delta_u := \|s\|/\|u\|, \quad \delta_v := \|t\|/\|v\|, \tag{3.36}$$

for $s,t\in\mathbb{C}^n$*. Then* $(u+s,v+t)\in\mathcal{K}_{\varepsilon_0}(x_*)\times\mathcal{K}_{\varepsilon_0}(y_*)$*. Note that* $\delta_0<1$*, and* $\delta_0\sim\frac{3}{16}\varepsilon_0^2$ *for* $\varepsilon_0\to 0$*.*

Proof. Let $\tilde{\xi}:=\measuredangle(\mathrm{span}\{x_*\},\mathrm{span}\{u+s\})$. For $\delta_0<1$ we have $u+s\neq 0$, since $\|u+s\|\ge\|u\|-\|s\|\ge\|u\|(1-\delta_0)>0$. It follows that

$$\begin{aligned}
\cos\tilde{\xi} &= \frac{|(u+s)^H x_*|}{\|u+s\|} \ge \frac{|u^H x_*+s^H x_*|}{(1+\delta_0)\|u\|} \ge \frac{|u^H x_*|}{(1+\delta_0)\|u\|} - \frac{|s^H x_*|}{(1+\delta_0)\|u\|} \\
&\ge \frac{1}{1+\delta_0}\left(\cos\xi-\delta_0\right) \ge \frac{1}{1+\delta_0}\left(\cos\left(\frac{\varepsilon_0}{2}\right)-\delta_0\right) \\
&= \frac{1}{1+\frac{\cos\left(\frac{\varepsilon_0}{2}\right)-\cos(\varepsilon_0)}{1+\cos(\varepsilon_0)}}\left(\cos\left(\frac{\varepsilon_0}{2}\right)-\frac{\cos\left(\frac{\varepsilon_0}{2}\right)-\cos(\varepsilon_0)}{1+\cos(\varepsilon_0)}\right) \;=\; \cos\varepsilon_0,
\end{aligned}$$

i.e., $\tilde{\xi}\le\varepsilon_0<\pi/2$. Moreover, inserting the power series of the cosine yields

$$\delta_0=\frac{1-\frac{\varepsilon_0^2}{8}-1+\frac{\varepsilon_0^2}{2}}{2+\frac{\varepsilon_0^2}{2}}+\mathcal{O}(\varepsilon_0^4)\sim\frac{3}{16}\varepsilon_0^2,$$

for $\varepsilon_0 \to 0$.

The result for $\tilde{\eta} := \measuredangle(\mathrm{span}\{y_*\}, \mathrm{span}\{v+t\})$ can be shown in the same way. □

Proposition 3.13 *Under the assumptions and with the constants τ_0 and ε_0 of Theorem 3.5, let $(u,v) \in \mathcal{K}_{\varepsilon_0}(x_*) \times \mathcal{K}_{\varepsilon_0}(y_*)$. Then we have*

$$|v^H \dot{T}(\lambda_*)u| \;\geq\; \frac{|\alpha|}{2}\|u\|\|v\| \cos\eta \cos\xi, \tag{3.37}$$

and

$$|\chi_1| := |v^H \dot{T}(p(u,v))u| \;\geq\; \frac{4|\alpha|}{17}\|u\|\|v\|, \tag{3.38}$$

for $p(u,v) \in S_0 = \bar{S}(\lambda_*, \tau_0)$.

Proof. With the notation as in Theorem 3.5 we have

$$\begin{aligned} v^H \dot{T}(\lambda_*)u &= v_2^H u_2 \left[y_*^H + \left(\frac{v_1}{v_2}\right)^H Y_1^H \right] \dot{T}(\lambda_*) \left[x_* + X_1 \frac{u_1}{u_2} \right] \\ &= v_2^H u_2 \left[y_*^H \dot{T}_* x_* + \left(\frac{v_1}{v_2}\right)^H Y_1^H \dot{T}_* x_* + y_*^H \dot{T}_* X_1 \frac{u_1}{u_2} + \left(\frac{v_1}{v_2}\right)^H Y_1^H \dot{T}_* X_1 \frac{u_1}{u_2} \right] \\ &= v_2^H u_2 (\alpha + \beta), \end{aligned}$$

hence, $|v^H \dot{T}(\lambda_*)u| \geq \|u\|\|v\| \cos\eta \cos\xi\, [|\alpha| - |\beta|]$, and (3.37) follows from Proposition 3.6. Now, consider $\chi_1 = v^H \dot{T}(p)u = v^H \dot{T}(\lambda_*)u + v^H[\dot{T}(p) - \dot{T}(\lambda_*)]u$ with $p \equiv p(u,v)$. Then

$$\begin{aligned} |v^H \dot{T}(p)u| &\geq |v^H \dot{T}(\lambda_*)u| - L|p - \lambda_*|\|u\|\|v\| \\ &\geq (\frac{1}{2}\cos\eta\cos\xi|\alpha| - L\tau_0)\|u\|\|v\| \\ &\geq (\frac{1}{2}\cos^2\varepsilon_0|\alpha| - L\frac{4|\alpha|}{17L})\|u\|\|v\| \\ &\geq (\frac{8}{17}|\alpha| - \frac{4}{17}|\alpha|)\|u\|\|v\| = \frac{4}{17}|\alpha|\|u\|\|v\|, \end{aligned}$$

since by Theorem 3.5 we have $\tau_0 \leq \frac{4|\alpha|}{17L}$ and $\tan\varepsilon_0 \leq 1/4$, and hence

$$\cos\varepsilon_0 = \frac{1}{\sqrt{1+\tan^2\varepsilon_0}} \geq \frac{4}{\sqrt{17}}.$$

□

Remark 3.14 *With the notation and assumptions of Theorem 3.5, we can specify the general constants M_0 and M_1 given in* (2.9) *and* (2.7) *in this setting, and obtain*

$$\begin{aligned}\|\dot{T}(p)\| &= \|\dot{T}(\lambda_*) + [\dot{T}(p) - \dot{T}(\lambda_*)]\| \\ &\leq \|\dot{T}(\lambda_*)\| + L|p-\lambda_*| \leq \|\dot{T}(\lambda_*)\| + L\frac{8\|\dot{T}(\lambda_*)\|}{3|\alpha|}\tan\xi\tan\eta \\ &\leq \|\dot{T}(\lambda_*)\|\left\{1 + \frac{8L}{3|\alpha|}\tan^2\varepsilon_0\right\} \leq \|\dot{T}(\lambda_*)\|\left\{1 + \frac{L}{6|\alpha|}\right\} =: \hat{M}_1,\end{aligned}$$

since by Proposition 3.6, $\tan\varepsilon_0 < 1/4$*, which implies that*

$$\begin{aligned}\|T(p)\| &= \|T(\lambda_*) + [T(p) - T(\lambda_*)]\| \\ &\leq \|T(\lambda_*)\| + \|\dot{T}(p)\||p-\lambda_*| \leq \|T(\lambda_*)\| + \hat{M}_1\frac{8\|\dot{T}(\lambda_*)\|}{3|\alpha|}\tan\xi\tan\eta \\ &\leq \|T(\lambda_*)\| + \frac{\hat{M}_1\|\dot{T}(\lambda_*)\|}{6|\alpha|} =: \hat{M}_0.\end{aligned}$$

Finally, we are able to derive an expansion of the complex Rayleigh functional $p(u,v)$, which replaces the Taylor expansion (3.21) in the real case.

Theorem 3.15 *For τ_0, ε_0 specified in Theorem 3.5, let λ_* be a simple eigenvalue of* (3.5)*, x_*, y_* be the corresponding eigenvectors with unit norm and let $T(\lambda)$ be holomorphic on an open neighborhood of $\bar{S}(\lambda_*,\tau_*)$, with $\tau_* > 0$. Let $(u,v) \in \mathcal{K}_{\varepsilon_0/2}(x_*) \times \mathcal{K}_{\varepsilon_0/2}(y_*)$. Then, there exists δ_0 with $0 \leq \delta_u, \delta_v \leq \delta_0$ and $K > 0$ such that $(u+s, v+t) \in K_{\varepsilon_0}(x_*) \times \mathcal{K}_{\varepsilon_0}(y_*)$ and $p(u+s,v+t)$ exists uniquely in $\bar{S}(\lambda_*,\tau_0)$ and we have*

$$p(u+s,v+t) = p(u,v) - \frac{v^H T(p(u,v))s + t^H T(p(u,v))u}{v^H\dot{T}(p(u,v))u} - \rho(s,t), \tag{3.39}$$

with

$$|\rho(s,t)| \leq K(\delta_u + \delta_v)^2,$$

and moreover,

$$|p(u+s,v+t) - p(u,v)| \leq \frac{34\hat{M}_0}{|\alpha|}(\delta_u + \delta_v + \delta_u\delta_v). \tag{3.40}$$

Proof. Existence follows by Proposition 3.12 and Theorem 3.5. We will show inequality (3.40) first, then (3.39), and the upper bound for ρ at last.

Set $\mu = p(u+s, v+t) - p(u,v)$, $p \equiv p(u,v)$ and let $T(p+\mu) = T(p) + \dot{T}(p)\mu + R$ where $\|R\| \leq \frac{L}{2}\mu^2$, and L as in (3.6). Then

$$\begin{aligned} 0 &= (v+t)^H T(p+\mu)(u+s) \\ &= (v+t)^H \left\{ T(p) + \dot{T}(p)\mu + R \right\} (u+s) \\ &= v^H T(p)u + v^H \dot{T}(p)\mu u + v^H T(p)s + v^H \dot{T}(p)\mu s + t^H T(p)u \\ &\quad + t^H \dot{T}(p)\mu u + t^H T(p)s + t^H \dot{T}(p)\mu s + (v+t)^H R(u+s), \end{aligned}$$

and

$$\mu = -\frac{l + t^H T s + \hat{R}}{\chi_1 + \hat{\beta}}, \tag{3.41}$$

with $l := v^H T(p)s + t^H T(p)u$ representing the linear part, χ_1 as in (3.38),

$$\hat{\beta} := v^H \dot{T} s + t^H \dot{T} u + t^H \dot{T} s, \tag{3.42}$$

and $\hat{R} = (v+t)^H R(u+s)$. Since $\alpha \neq 0$, $u \neq 0$, $v \neq 0$, the term in the denominator of μ is nonzero due to the following proposition.

Proposition 3.16 *Under the assumptions of Theorem 3.15, and with χ_1 as in* (3.38)*, $\hat{\beta}$ as in* (3.42)*, $\alpha = y_*^H \dot{T}(\lambda_*)x_*$, let $\delta_0 \leq \frac{|\alpha|}{17\|\dot{T}(p)\|}$. Then we have*

$$|\chi_1 + \hat{\beta}| \geq \frac{1}{17}|\alpha|\|u\|\|v\|.$$

Proof. With (3.38) we obtain

$$\begin{aligned} |\chi_1 + \hat{\beta}| &\geq |\chi_1| - |\hat{\beta}| \geq \frac{4}{17}|\alpha|\|u\|\|v\| - (\delta_u + \delta_v + \delta_u\delta_v)\|\dot{T}(p)\|\|u\|\|v\| \\ &\geq \left(\frac{4}{17}|\alpha| - (2+\delta_0)\delta_0\|\dot{T}(p)\| \right) \|u\|\|v\| \\ &\geq \left(\frac{4}{17}|\alpha| - 3\delta_0\|\dot{T}(p)\| \right) \|u\|\|v\| \geq \frac{1}{17}|\alpha|\|u\|\|v\|. \end{aligned}$$

□

Equation (3.41) implies that

$$|\mu| \leq \frac{\|u\|\|v\|\hat{M}_0(\delta_u + \delta_v + \delta_u\delta_v) + \frac{L}{2}|\mu|^2\|u\|\|v\|(1+\delta_u)(1+\delta_v)}{\frac{|\alpha|}{17}\|u\|\|v\|}. \tag{3.43}$$

Suppose that $|\mu| \leq \hat{\tau} := \frac{2|\alpha|}{34L(1+\delta_u)(1+\delta_v)}$. Hence $|\mu|^2 \leq \hat{\tau}|\mu|$, which, inserted in (3.43), gives

$$\left(1 - \frac{17L}{2|\alpha|}\hat{\tau}(1+\delta_u)(1+\delta_v) \right) |\mu| \leq 17\frac{\hat{M}_0}{|\alpha|}(\delta_u + \delta_v + \delta_u\delta_v),$$

yielding (3.40), i.e., $|\mu| \leq 34\hat{M}_0(\delta_u + \delta_v + \delta_u\delta_v)/|\alpha|$. According to

$$\delta_u + \delta_v + \delta_u\delta_v = \left(1 + \frac{\delta_v}{2}\right)\delta_u + \left(1 + \frac{\delta_u}{2}\right)\delta_v \leq \left(1 + \frac{\delta_0}{2}\right)(\delta_u + \delta_v),$$

we also obtain

$$|\mu| \leq 34\frac{\hat{M}_0}{|\alpha|}\left(1 + \frac{\delta_0}{2}\right)(\delta_u + \delta_v). \tag{3.44}$$

Now, let $q := \hat{\beta}/\chi_1$. Then $\chi_1 + \hat{\beta} = \chi_1[1+q]$ and due to $1/(1+q) = 1 - q/(1+q)$, equation (3.41) yields

$$\mu = -\frac{l + t^H Ts + \hat{R}}{\chi_1}\left[1 - \frac{q}{1+q}\right],$$

hence

$$\rho = -\mu - \frac{l}{\chi_1} = \frac{1}{\chi_1}\left\{t^H Ts + \hat{R} - \frac{q}{1+q}\left[l + t^H Ts + \hat{R}\right]\right\},$$

which is equivalent to (3.39). It remains to show that $|\rho(s,t)| \leq K(\delta_u + \delta_v)^2$.

We analyze the absolute value of q first

$$|q| \;=\; \frac{|\hat{\beta}|}{|\chi_1|} \leq \frac{17\hat{M}_1}{4|\alpha|}(\delta_u + \delta_v + \delta_u\delta_v) \leq \frac{17\hat{M}_1}{4|\alpha|}\left(1 + \frac{\delta_0}{2}\right)(\delta_u + \delta_v).$$

Suppose that $\delta_0 \leq \frac{2|\alpha|}{51\hat{M}_1} \leq \frac{|\alpha|}{17\|\hat{T}(p)\|}$. Then $|q| \leq 4|\alpha|/17\hat{M}_1(2+\delta_0)\delta_0 \leq \frac{1}{2}$. Furthermore, we have

$$|1+q| = \frac{|\chi_1 + \hat{\beta}|}{|\chi_1|} \geq \frac{|\alpha|\|u\|\|v\|}{17\hat{M}_1\|u\|\|v\|} = \frac{|\alpha|}{17\hat{M}_1},$$

and

$$(1+\delta_u)(1+\delta_v) \leq 1 + 2\delta_0 + \delta_0^2 \leq 1 + \frac{4|\alpha|}{51\hat{M}_1} + \frac{4|\alpha|^2}{51^2\hat{M}_1^2} =: Q_1.$$

Altogether we obtain the following estimate

$$\begin{aligned}
|\rho| \;&=\; \frac{1}{|\chi_1|}\Bigg\{\|u\|\|v\|\hat{M}_0\delta_u\delta_v + (1+\delta_u)(1+\delta_v)\frac{L}{2}|\mu|^2\|u\|\|v\| + \frac{\|u\|\|v\|\hat{M}_0(\delta_u+\delta_v)|q|}{|1+q|} \\
&\quad + \frac{|q|}{|1+q|}\|u\|\|v\|\hat{M}_0\delta_u\delta_v + \frac{|q|}{|1+q|}\frac{L}{2}|\mu|^2\|u\|\|v\|(1+\delta_u)(1+\delta_v)\Bigg\} \\
&\leq\; \frac{17}{|\alpha|}\Bigg\{\hat{M}_0\delta_u\delta_v + \frac{17\hat{M}_0\hat{M}_1}{2|\alpha|}\delta_u\delta_v + \frac{34^2Q_1L\hat{M}_0^2}{|\alpha|^2}\left(1 + \frac{\delta_0}{2}\right)^2(\delta_u+\delta_v)^2 \\
&\quad + \frac{17^2\hat{M}_0\hat{M}_1^2}{|\alpha|^2}\left(1 + \frac{\delta_0}{2}\right)(\delta_u+\delta_v)^2 + \frac{17\cdot 34^2Q_1L\hat{M}_0^2\hat{M}_1}{4|\alpha|^3}\left(1 + \frac{\delta_0}{2}\right)^2(\delta_u+\delta_v)^2\Bigg\} \\
&=\; \frac{17}{|\alpha|}\left\{K_1(\delta_u+\delta_v)^2 + K_2\delta_u\delta_v\right\} \\
&\leq\; K(\delta_u+\delta_v)^2,
\end{aligned}$$

where $K_1 := \frac{17^2\hat{M}_0}{|\alpha|^2}\left(1+\frac{\delta_0}{2}\right)\left\{4Q_1L\hat{M}_0\left(1+\frac{\delta_0}{2}\right)+\hat{M}_1^2+\frac{17Q_1L\hat{M}_0\hat{M}_1}{|\alpha|}\left(1+\frac{\delta_0}{2}\right)\right\}$, and $K_2 := \hat{M}_0\left\{1+\frac{17\hat{M}_1}{2|\alpha|}\right\}$ and $K := K_1\left\{1+\frac{K_2}{2K_1}\right\}^2$. □

Theorem 3.15 provides an expansion of the complex generalized Rayleigh functional which has the same structure as the real expansion defined in (3.21), because there we have $u^TT(p)^Tt = t^TT(p)u$. Moreover, the complex functional $p(u,v)$ is linear in s but antilinear in t. The expansion also shows that a perturbation of the Rayleigh functional is of a relative kind, in that perturbations in p do only depend on relative perturbations in u and v.

Remark 3.17 *Theorem 3.15 also implies relative Lipschitz continuity of p in both arguments, i.e.,*

$$|p(u+s,v+t)-p(u,v)| \le \frac{34\hat{M}_0}{|\alpha|}\left(1+\frac{\delta_0}{2}\right)\left(\frac{\|s\|}{\|u\|}+\frac{\|t\|}{\|v\|}\right),$$

see inequality (3.40) *and its deduction* (3.44).

Lemma 3.18 *Let λ_* be a simple eigenvalue of* (3.5)*, x_*, y_* be the corresponding eigenvectors with unit norm and let $T(\lambda)$ be holomorphic on an open neighborhood of $\bar{S}(\lambda_*,\tau_*)$. Then, the complex generalized Rayleigh functional p is stationary at (x_*,y_*) and $\lambda_* = p(x_*,y_*)$.*

Proof. For vectors $(u,v) \in \mathcal{K}_{\varepsilon_0/2}(x_*) \times \mathcal{K}_{\varepsilon_0/2}(x_*)$, the assumptions of Theorem 3.15 hold, hence equation (3.39) is valid. Inserting $u = x_*$, $v = y_*$ in (3.39) gives

$$p(x_*+s,y_*+t)-p(x_*,y_*) = \mathcal{O}((\|s\|+\|t\|)^2),$$

which shows the stationarity. □

Remark 3.19 *In general one will have differentiability of $p(u,v)$ with respect to u and of $p(u,v)^H$ with respect to v but no uniform results.* Parlett *[67], however, also uses the term stationary for the complex two-sided Rayleigh quotient $\frac{v^HAu}{v^Hu}$, i.e., in the linear case. As he did, we could have developed the analysis by means of the directional derivatives*

$$p^1(u,v;s) \equiv \lim_{\epsilon\to 0}\frac{[p(u+\epsilon s,v)-p(u,v)]}{\epsilon}, \qquad p^1(u,v;t) \equiv \lim_{\epsilon\to 0}\frac{[p(u,v+\epsilon t)-p(u,v)]}{\epsilon},$$

for arbitrary complex s, t, but real ϵ. On the other hand, with the fixed point equation we were able to derive the bound (3.22) *which is in terms of angles. It seems that such a bound cannot be derived starting from a perturbation expansion with absolute perturbations.*

3.3 The Standard Nonlinear Rayleigh Functional

Although the complex nonlinear Rayleigh functional $p(u)$, defined by (3.10)–(3.13), and its subspace generalization can be found in various algorithms for nonlinear eigenvalue problems, there is to our knowledge no existence analysis. Such an analysis is, therefore, provided in this section, where problems having equivalent left and right eigenvectors are considered first, i.e., according to Table 1.1, real symmetric problems and Hermitian problems with real eigenvalues. Notice that since we do only consider one eigenvalue, the property $x_* = y_*$ can be restricted to the eigenvectors corresponding to this single λ_*. Global characterizations as presented in Table 1.1 are not necessary but sufficient.

General problems will be tackled afterwards. As expected, the Rayleigh functional is not stationary in this case. The analysis is done using the same techniques as before and follows immediately from the previous results.

3.3.1 Structured Problems

Suppose that $T(\lambda_*) = T(\lambda_*)^H$, which, e.g., is the case if $T(\lambda_*)$ is real symmetric or Hermitian at the real eigenvalue λ_*. Then we have $y_* = x_*$, and instead of the singular value decomposition of $T(\lambda_*)$ we can work with the spectral decomposition

$$T(\lambda_*) = X\Lambda X^H = [\, X_1 \mid x_* \,] \left[\begin{array}{c|c} \Lambda_1 & 0 \\ \hline 0^T & 0 \end{array}\right] [\, X_1 \mid x_* \,]^H , \tag{3.45}$$

where X is unitary and Λ_1 is nonsingular and diagonal. Note that $|\Lambda_1| = \Sigma_1$ elementwise.

The adapted version of Theorem 3.5 is given by the following corollary.

Corollary 3.20 *Let λ_* be a simple eigenvalue of (3.5), where $T(\lambda_*) = T(\lambda_*)^H$. Suppose that x_* is the corresponding eigenvector with unit norm and let $T(\lambda)$ be holomorphic on an open neighborhood of $\bar{S}(\lambda_*, \tau_*)$, with $\tau_* > 0$. Then there exist constants $0 < \tau_0 < \tau_*$, $0 < \varepsilon_0 < \pi/2$ such that for all $u \in \mathcal{K}_{\varepsilon_0}(x_*)$ there exists a unique $p = p(u) \in S_0 := \bar{S}(\lambda_*, \tau_0)$ with $g(p, u, u) \equiv g(p, u) = u^H T(p(u))u = 0$. Moreover, one has*

$$|p(u) - \lambda_*| \quad \leq \quad \frac{8}{3} \, \frac{\|T(\lambda_*)\|}{|x_*^H \dot{T}(\lambda_*) x_*|} \tan^2 \xi.$$

Proof. Since x_* is left and right eigenvector, we only need to substitute y_* by x_* in the proof of Theorem 3.5 and take (3.45) instead of (2.12), i.e., we substitute Y_1 by X_1 and

Σ_1 by Λ_1. The new $\tilde{\alpha} = x_*^H \dot{T}(\lambda_*) x_*$ is automatically nonzero, because λ_* is simple. □

In analogy to linear Hermitian problems, we give a bound for the distance of Rayleigh quotient and exact eigenvalue.

Corollary 3.21 *Let A be a Hermitian matrix and (λ_*, x_*) an eigenpair of A and suppose that $u \in \mathbb{C}^n$ with $\measuredangle(\mathrm{span}\{u\}, \mathrm{span}\{x_*\}) < \pi/2$. Then the Rayleigh quotient $p(u) = \frac{u^H A u}{u^H u}$ satisfies the inequality*

$$|p(u) - \lambda_*| \leq \frac{\|A - \lambda_* I\|}{\|u\|^2} \frac{\|(I - uu^H)x_*\|^2}{\|u^H x_*\|^2} \equiv \frac{\|T(\lambda_*)\|}{\|u\|^2} \tan^2 \xi.$$

Proof. With

$$(A - \lambda_* I)uu^H x_* \;=\; (A - \lambda_* I)(x_* - (I - uu^H)x_*) = -(A - \lambda_* I)(I - uu^H)x_*,$$

we also have $(uu^H x_*)^H (A - \lambda_* I) = -x_*^H (I - uu^H)^H (A - \lambda_* I)$ and, since p is homogeneous, $p(u) = p(uu^H x_*)$, and

$$\begin{aligned}
|p(u) - \lambda_*| &= \left| \frac{(uu^H x_*)^H A uu^H x_*}{(uu^H x_*)^H uu^H x_*} - \lambda_* \right| \\
&= \left| \frac{(uu^H x_*)^H (A - \lambda_* I) uu^H x_*}{x_*^H uu^H uu^H x_*} \right| \\
&= \left| \frac{x_*^H (I - uu^H)^H (A - \lambda_* I)(I - uu^H) x_*}{u^H u (u^H x_*)^2} \right| \\
&\leq \frac{\|A - \lambda_* I\|}{\|u\|^2} \frac{\|(I - uu^H)x_*\|^2}{|u^H x_*|^2} \;=\; \frac{\|T(\lambda_*)\|}{\|u\|^2} \tan^2 \xi.
\end{aligned}$$

□

The bound for the nonlinear version of the Rayleigh quotient is a nice generalization of this bound, since for the linear Hermitian problem $\tilde{\alpha}$ reduces to $|x_*^H \dot{T}(\lambda_*) x_*| = |-x_*^H x_*| = 1$.

3.3.2 General Problems

We want to consider the standard version of the Rayleigh functional but for general $T(\lambda)$, which means that there is no information about the left eigenvector included. The main difference to previous results is that the assumption that λ_* is algebraically simple can be neglected, i.e., $y_*^H \dot{T}(\lambda_*) x_* \neq 0$ is not needed, but instead we need to impose that

$$\tilde{\alpha} := x_*^H \dot{T}(\lambda_*) x_* \neq 0, \tag{3.46}$$

which is a different assumption when left and right eigenvectors are different, see (3.13). The bound for the distance of functional and eigenvalue is linear in the angle as expected.

Corollary 3.22 *Let λ_* be an eigenvalue of (3.5), let x_* be the corresponding eigenvector with unit norm and let $T(\lambda)$ be holomorphic on an open neighborhood of $\bar{S}(\lambda_*, \tau_*)$, with $\tau_* > 0$. Suppose that (3.46) holds. Then there exist constants $0 < \tau_0 < \tau_*$, $0 < \varepsilon_0 < \pi/2$ such that for all $u \in \mathcal{K}_{\varepsilon_0}(x_*)$ there exists a unique $p = p(u) \in S_0 := \bar{S}(\lambda_*, \tau_0)$ such that $g(p,u) = (T(p(u))u, u) = 0$. Moreover, one has*

$$|p - \lambda_*| \quad \leq \quad \frac{10}{3} \, \frac{\|T(\lambda_*)\|}{|x_*^H \dot{T}(\lambda_*) x_*|} \tan \xi. \tag{3.47}$$

Proof. The proof goes along the lines of the proof of Theorem 3.5, but with additional terms arising from $x_*^H T(\lambda_*)$, where we had $y_*^H T(\lambda_*) = 0$ before.

First, we specify the constants. For $\tilde{\alpha}$ as in (3.46), let

$$\tau_0 \quad := \quad \min \left\{ \tau_*, \tau_1 := \frac{4}{17} \frac{|\tilde{\alpha}|}{L} \right\}, \tag{3.48}$$

$$\varepsilon_0 \quad := \quad \min \left\{ \varepsilon_1 := \arctan \frac{2}{9} \frac{|\tilde{\alpha}|}{\|\dot{T}(\lambda_*)\|}, \, \varepsilon_2 := \arctan \frac{4\tau_0 \, |\tilde{\alpha}|}{20 \|T(\lambda_*)\|} \right\}. \tag{3.49}$$

With (2.12), (3.25) and

$$u = X_1 u_1 + x_* u_2 = [X_1 \; x_*] \begin{bmatrix} u_1 \\ u_2 \end{bmatrix},$$

we have

$$\begin{aligned} g(p,u) \quad &= \quad u_2^H u_2 \left[x_*^H + \left(\frac{u_1}{u_2} \right)^H X_1^H \right] \left[T(\lambda_*) + (p - \lambda_*) \dot{T}(\lambda_*) + R \right] \left[x_* + X_1 \frac{u_1}{u_2} \right] \\ &= \quad u_2^H u_2 \Bigg\{ x_*^H T_* x_* + x_*^H T_* X_1 \frac{u_1}{u_2} + \left(\frac{u_1}{u_2} \right)^H X_1^H T_* x_* + \left(\frac{u_1}{u_2} \right)^H X_1^H T_* X_1 \frac{u_1}{u_2} \\ &\quad + (p - \lambda_*) \left[x_*^H \dot{T}_* x_* + \left(\frac{u_1}{u_2} \right)^H X_1^H \dot{T}_* x_* + x_*^H \dot{T}_* X_1 \frac{u_1}{u_2} + \left(\frac{u_1}{u_2} \right)^H X_1^H \dot{T}_* X_1 \frac{u_1}{u_2} \right] \\ &\quad + \left[x_*^H + \left(\frac{u_1}{u_2} \right)^H X_1^H \right] R \left[x_* + X_1 \frac{u_1}{u_2} \right] \Bigg\}. \end{aligned}$$

We set $\hat{u} := x_* + X_1 \frac{u_1}{u_2}$ as before, and

$$\tilde{\beta} = \tilde{\beta}(u) := \left(\frac{u_1}{u_2} \right)^H X_1^H \dot{T}_* x_* + x_*^H \dot{T}_* X_1 \frac{u_1}{u_2} + \left(\frac{u_1}{u_2} \right)^H X_1^H \dot{T}_* X_1 \left(\frac{u_1}{u_2} \right), \tag{3.50}$$

$$\tilde{\gamma} = \tilde{\gamma}(u) := x_*^H T_* X_1 \frac{u_1}{u_2} + \left(\frac{u_1}{u_2} \right)^H X_1^H T_* X_1 \frac{u_1}{u_2}. \tag{3.51}$$

Note, that $\tilde{\gamma}$ has an additional term compared to γ from equation (3.29). With this notation and since $u_2^H u_2 \neq 0$, the equation $g(p,u) = 0$ is equivalent to

$$\tilde{\gamma} + (p - \lambda_*)[\,\tilde{\alpha} + \tilde{\beta}\,] + \hat{u}^H R\hat{u} \;=\; 0. \tag{3.52}$$

In order to solve for the linear part of p we first have to guarantee that $\tilde{\alpha} + \tilde{\beta}$ is nonzero. Therefore, we prove the following proposition.

Proposition 3.23 *Under the assumptions of Corollary 3.22, let $\varepsilon_1 > 0$ and let (3.46) hold. Decompose $u \in \mathcal{K}_{\varepsilon_1}(x_*)$ according to (3.26) and (3.27). If $\tan\varepsilon_1 = \frac{2}{9}\frac{|\tilde{\alpha}|}{\|\dot{T}(\lambda_*)\|} =: \frac{2}{9}\,q_*$, then*

$$|\tilde{\beta}(u)| \leq \frac{1}{2}\,|\tilde{\alpha}| \tag{3.53}$$

holds for all $u \in \mathcal{K}_{x_*}(\varepsilon_1)$, with $\tilde{\beta}$ as in (3.50).

Proof. Taking the estimate $\tan\varepsilon_1 = \frac{2}{9}q_* \leq \frac{2}{9} < \frac{1}{4}$ into account, we obtain

$$\begin{aligned}
|\tilde{\beta}| \;&\leq\; \left\|\frac{u_1}{u_2}\right\| \|X_1\|\|\dot{T}_*\|\|x_*\| + \left\|\frac{u_1}{u_2}\right\| \|\dot{T}_*\|\|X_1\|\|x_*\| + \left\|\frac{u_1}{u_2}\right\|^2 \|\dot{T}_*\|\|X_1\|^2 \\
&=\; \|\dot{T}_*\|\left[2\tan\xi + \tan^2\xi\right] \\
&\leq\; \|\dot{T}_*\|\left[2\tan\varepsilon_1 + \frac{1}{4}\tan\varepsilon_1\right] \leq \|\dot{T}_*\|\frac{9}{4}\tan\varepsilon_1 = \frac{1}{2}|\tilde{\alpha}|. \qquad \square
\end{aligned}$$

Suppose that $u \in \mathcal{K}_{x_*}(\varepsilon_1)$, with ε_1 from Proposition 3.23. Inequality (3.53) then implies that $|\tilde{\alpha} + \tilde{\beta}| \geq |\tilde{\alpha}| - |\tilde{\beta}| \geq \frac{1}{2}|\tilde{\alpha}| > 0$, i.e., we can solve (3.52) for p and derive the fixed point form

$$p = \lambda_* - \frac{1}{\tilde{\alpha} + \tilde{\beta}(u)}\left\{\tilde{\gamma}(u) + \hat{u}^H R(p,\lambda_*)\hat{u}\right\} =: \tilde{G}(p,u) \tag{3.54}$$

to which we can apply the Banach fixed point theorem [32, p. 35]. We start by showing contractivity of $\tilde{G}$.

Proposition 3.24 *Under the assumptions of Corollary 3.22, let $\tau_1 := \frac{4}{17}\frac{|\tilde{\alpha}|}{L}$ and $\tau_0 := \min\{\tau_*, \tau_1\}$. Then, for $\tilde{G}$ as defined in (3.54), there exists a constant $\kappa < 1$ such that for all $u \in \mathcal{K}_{x_*}(\varepsilon_1)$ the inequality*

$$|\tilde{G}(p,u) - \tilde{G}(\mu,u)| \leq \kappa|p - \mu|$$

holds for all p, $\mu \in S_0 := \bar{S}(\lambda_*, \tau_0)$.

Proof. As in the proof of Proposition 3.7 we get $\|\hat{u}\| \leq \sqrt{\frac{17}{16}}$, and $\|R(p,\lambda_*) - R(\mu,\lambda_*)\| \leq L\tau_0|p-\mu|$. Considering (3.54) yields

$$\begin{aligned} |\tilde{G}(p,u) - \tilde{G}(\mu,u)| &\leq \frac{\left|\hat{u}^H(R(p,\lambda_*) - R(\mu,\lambda_*))\hat{u}\right|}{|\tilde{\alpha}+\tilde{\beta}|} \\ &\leq \frac{2}{|\tilde{\alpha}|}\|\hat{u}\|^2\|R(p,\lambda_*) - R(\mu,\lambda_*)\| \\ &\leq \frac{2}{|\tilde{\alpha}|}\frac{17}{16}L\tau_0|p-\mu| \leq \frac{2}{|\tilde{\alpha}|}\frac{17}{16}L\tau_1|p-\mu| = \frac{1}{2}|p-\mu|, \end{aligned}$$

i.e., $\tilde{G}$ is contractive on S_0 with coefficient $\kappa = 1/2$. □

Since (3.48) holds, the assumptions of this proposition are satisfied, and $\tilde{G}$ is a contraction. It remains to show that $\tilde{G}$ maps onto S_0.

Proposition 3.25 *Under the assumptions of Corollary 3.22, let* $\tan\varepsilon_2 = \frac{4\tau_0\,|\tilde{\alpha}|}{20\|T(\lambda_*)\|}$, ε_0 *as defined in* (3.49). *Then, for* $\tilde{G}$ *as defined in* (3.54),

$$\tilde{G}(S_0) \subset S_0 \quad \textit{is satisfied for all } u \in \mathcal{K}_{\varepsilon_0}(x_*).$$

Proof. It suffices to show (3.32) with $\tilde{G}$ instead of G. Again (3.54) yields

$$\begin{aligned} |\tilde{G}(\lambda_*,u) - \lambda_*| &= \frac{\left|\tilde{\gamma}(u) + \hat{u}^H R(\lambda_*,\lambda_*)\hat{u}\right|}{|\tilde{\alpha}+\tilde{\beta}(u)|} = \frac{|\tilde{\gamma}(u)|}{|\tilde{\alpha}+\tilde{\beta}(u)|} \leq \frac{2}{|\tilde{\alpha}|}\|T(\lambda_*)\|(\tan\xi + \tan^2\xi) \\ &\leq \frac{2}{|\tilde{\alpha}|}\|T(\lambda_*)\|\tan\varepsilon_2(1+\tan\varepsilon_0) \leq \frac{10}{4|\tilde{\alpha}|}\|T(\lambda_*)\|\tan\varepsilon_2 = \frac{1}{2}\tau_0. \end{aligned}$$

□

Hence, all conditions required for applying the fixed point theorem hold and the existence of a unique solution $p = p(u)$ in S_0 such that $g(p,u) = 0$ follows provided that $u \in \mathcal{K}_{\varepsilon_0}(x_*)$.

To derive the error estimate (3.47), we consider (3.54) and get

$$\begin{aligned} |p-\lambda_*| &= \frac{|\tilde{\gamma}(u) + \hat{u}^H R(p,\lambda_*)\hat{u}|}{|\tilde{\alpha}+\tilde{\beta}|} \\ &\leq \frac{2}{|\tilde{\alpha}|}\left(|\tilde{\gamma}(u)| + \|\hat{u}\|^2\|R(p,\lambda_*)\|\right) \\ &\leq \frac{2}{|\tilde{\alpha}|}\left(\|T(\lambda_*)\|(\tan\xi + \tan^2\xi) + \frac{17}{16}\frac{L}{2}|p-\lambda_*|^2\right) \\ &\leq \frac{2}{|\tilde{\alpha}|}\left(\|T(\lambda_*)\|\frac{5}{4}\tan\xi + \frac{17}{16}\frac{L}{2}\tau_0|p-\lambda_*|\right), \end{aligned} \tag{3.55}$$

or equivalently

$$\left(1 - \frac{17}{16|\tilde{\alpha}|} L\tau_0\right) |p - \lambda_*| \;\leq\; \frac{10\|T(\lambda_*)\|}{4|\tilde{\alpha}|} \tan\xi.$$

Finally, $\frac{17}{16|\tilde{\alpha}|} L\tau_0 \leq \frac{17}{16|\tilde{\alpha}|} L\tau_1 = \frac{1}{4}$ yields (3.47). □

This result describes the analogon of the inequality

$$|p(u) - \lambda_*| \;\leq\; \frac{\|A - \lambda_* I\|}{\|u\|} \tan\xi$$

for linear problems $T(\lambda) = A - \lambda I$, which is proved in the same way as Corollary 3.21. To assess the sharpness of the bound (3.47), we consider the limit of (3.55) for $\xi \to 0+$. In order to do this we consider the equivalent formulation

$$\frac{|p - \lambda_*|}{\tan\xi} = \frac{\left|\frac{\tilde{\gamma}(u)}{\tan\xi} + \frac{\hat{u}^H R \hat{u}}{\tan\xi}\right|}{|\tilde{\alpha} + \tilde{\beta}|}$$

of equation (3.55). Investigating the individual components

$$\begin{aligned}
\frac{|\tilde{\gamma}|}{\tan\xi} &= \frac{\left|x_*^H T_* X_1 \frac{u_1}{u_2} + \left(\frac{u_1}{u_2}\right)^H X_1^H T_* X_1 \frac{u_1}{u_2}\right|}{\tan\xi} \\
&\leq (1 + \tan\xi)\|\Sigma_1\| = (1 + \tan\xi)\|T(\lambda_*)\| \longrightarrow \|T(\lambda_*)\|, \\
|\tilde{\alpha} + \tilde{\beta}\,| &= |x_*^H \dot{T}(\lambda_*) x_* + \mathcal{O}(2\tan\xi + \tan^2\xi)| \longrightarrow |\tilde{\alpha}|, \\
\frac{|\hat{u}^H R \hat{u}^H|}{\tan\xi} &= \mathcal{O}\left(\frac{(p - \lambda_*)^2}{\tan\xi}\right) = \mathcal{O}(\tan\xi),
\end{aligned}$$

yields

$$\limsup_{\xi \to 0+} \frac{|p - \lambda_*|}{\tan\xi} \leq \frac{\|T(\lambda_*)\|}{|\tilde{\alpha}|},$$

where equality is achieved if $\frac{u_1}{u_2} = (\tan\xi) e_1$. This means that the constant 10/3 in the estimate (3.47) can asymptotically be replaced by 1.

3.3.3 Perturbation Expansion

We are interested in a representation for the perturbed one-sided $p(u)$ analogous to (3.39) for the generalized functional. Propositions 3.12 and 3.13 can be reused; adjustments are given by the following corollaries.

Corollary 3.26 *Let $u \in \mathcal{K}_{\varepsilon_0/2}(x_*)$, $0 \leq \delta_u \leq \delta_0 := \frac{\cos\left(\frac{\varepsilon_0}{2}\right) - \cos(\varepsilon_0)}{1 + \cos(\varepsilon_0)}$, where $\delta_u := \|s\|/\|u\|$ for $s \in \mathbb{C}^n$. Then $(u + s) \in \mathcal{K}_{\varepsilon_0}(x_*)$. Note that $\delta_0 < 1$, and $\delta_0 \sim \frac{3}{16}\varepsilon_0^2$ for $\varepsilon_0 \to 0$.*

Proof. The proof is exactly the same as the proof of Proposition 3.12, where we can leave out the second part. □

Corollary 3.27 *Under the assumptions and with the constants τ_0 and ε_0 of Corollary 3.22, let $u \in \mathcal{K}_{\varepsilon_0}(x_*)$. Then we have*

$$|u^H \dot{T}(\lambda_*)u| \;\geq\; \frac{|\tilde{\alpha}|}{2}\|u\|^2 \cos^2 \xi \tag{3.56}$$

and

$$|\chi_2| := |u^H \dot{T}(p(u))u| \;\geq\; \frac{4|\tilde{\alpha}|}{17}\|u\|^2, \tag{3.57}$$

where $p(u) \in S_0$.

Proof. With the notation as in Corollary 3.22 we have

$$\begin{aligned} u^H \dot{T}(\lambda_*)u &= u_2^H u_2 \left[x_*^H + \left(\frac{u_1}{u_2}\right)^H X_1^H\right] \dot{T}(\lambda_*) \left[x_* + X_1 \frac{u_1}{u_2}\right] \\ &= u_2^H u_2 \left[x_*^H \dot{T}_* x_* + \left(\frac{u_1}{u_2}\right)^H X_1^H \dot{T}_* x_* + x_*^H \dot{T}_* X_1 \frac{u_1}{u_2} + \left(\frac{u_1}{u_2}\right)^H X_1^H \dot{T}_* X_1 \frac{u_1}{u_2}\right] \\ &= u_2^H u_2 (\tilde{\alpha} + \tilde{\beta}), \end{aligned}$$

hence, $|u^H \dot{T}(\lambda_*)u| \geq \|u\|^2 \cos^2 \xi \left[|\tilde{\alpha}| - |\tilde{\beta}|\right]$, and (3.56) follows with Proposition 3.23. Now, consider $\chi_2 = u^H \dot{T}(p)u = u^H \dot{T}(\lambda_*)u + u^H[\dot{T}(p) - \dot{T}(\lambda_*)]u$ with $p \equiv p(u)$. Thus

$$\begin{aligned} |u^H \dot{T}(p)u| &\geq |u^H \dot{T}(\lambda_*)u| - L|p - \lambda_*|\|u\|^2 \\ &\geq (\frac{1}{2}\cos^2 \xi |\tilde{\alpha}| - L\tau_0)\|u\|^2 \\ &\geq (\frac{1}{2}\cos^2 \varepsilon_0 |\tilde{\alpha}| - L\frac{4|\tilde{\alpha}|}{17L})\|u\|^2 \\ &\geq (\frac{8}{17}|\tilde{\alpha}| - \frac{4}{17}|\tilde{\alpha}|)\|u\|^2 = \frac{4}{17}|\tilde{\alpha}|\|u\|^2, \end{aligned}$$

since with Corollary 3.22 we have $\tau_0 \leq \frac{4|\tilde{\alpha}|}{17L}$ and $\tan \varepsilon_0 \leq 1/4$, hence $\cos \varepsilon_0 = \frac{1}{\sqrt{1+\tan^2 \varepsilon_0}} \geq \frac{4}{\sqrt{17}}$ holds. □

Theorem 3.28 *For τ_0, ε_0 specified in Corollary 3.22, let x_* be an eigenvector with unit norm corresponding to the eigenvalue λ_* for $T(\lambda)x = 0$, and let $T(\lambda)$ be holomorphic on an open neighborhood of $\bar{S}(\lambda_*, \tau_*)$, with $\tau_* > 0$. Suppose that $x_*^H \dot{T}(\lambda_*)x_* \neq 0$ and*

$u \in \mathcal{K}_{\varepsilon_0/2}(x_*)$. Then there exist δ_0 with $0 \leq \delta_u = \|s\|/\|u\| \leq \delta_0$ and $K > 0$ such that $(u+s) \in K_{\varepsilon_0}(x_*)$ and $p(u+s)$ exists uniquely in $\bar{S}(\lambda_*, \tau_0)$, and we have

$$p(u+s) = p(u) - \frac{u^H T(p(u))s + s^H T(p(u))u}{u^H \dot{T}(p(u))u} - \rho(s), \tag{3.58}$$

with

$$|\rho(s)| \leq K\delta_u^2,$$

and moreover,

$$|p(u+s) - p(u)| \leq \frac{34\hat{M}_0}{|\tilde{\alpha}|}(2\delta_u + \delta_u^2). \tag{3.59}$$

Proof. Existence follows by Corollary 3.26 and Theorem 3.22. We will show inequality (3.59) first, then (3.58) and the upper bound for ρ.

Set $\mu = p(u+s) - p(u)$, $p \equiv p(u)$ and let $T(p+\mu) = T(p) + \dot{T}(p)\mu + R$, where $\|R\| \leq \frac{L}{2}\mu^2$, L from (3.6). Thus,

$$\begin{aligned}
0 &= (u+s)^H T(p+\mu)(u+s) \\
&= (u+s)^H \left\{ T(p) + \dot{T}(p)\mu + R \right\} (u+s) \\
&= u^H T(p)u + u^H \dot{T}(p)\mu u + u^H T(p)s + u^H \dot{T}(p)\mu s + s^H T(p)u \\
&\quad + s^H \dot{T}(p)\mu u + s^H T(p)s + s^H \dot{T}(p)\mu s + (u+s)^H R(u+s)
\end{aligned}$$

and

$$\mu = -\frac{l + s^H T s + \check{R}}{\chi_2 + \check{\beta}}, \tag{3.60}$$

with $l := u^H T(p)s + s^H T(p)u$ representing the linear part, χ_2 as in (3.57) and

$$\begin{aligned}
\check{\beta} &:= u^H \dot{T} s + s^H \dot{T} u + s^H \dot{T} s, \\
\check{R} &:= (u+s)^H R(u+s).
\end{aligned} \tag{3.61}$$

Since we have assumed that $\tilde{\alpha} \neq 0$, and $u \neq 0$, the denominator of μ is nonzero due to the following proposition.

Proposition 3.29 *Under the assumptions of Theorem 3.28, and with χ_2 as in* (3.57)*, $\check{\beta}$ as in* (3.61)*, $\tilde{\alpha} = x_*^H \dot{T}(\lambda_*)x_*$, let $\delta_0 \leq \frac{|\tilde{\alpha}|}{17\|\dot{T}(p)\|}$. Then we have*

$$|\chi_2 + \check{\beta}| \geq \frac{1}{17}|\tilde{\alpha}|\|u\|^2.$$

Proof. With (3.57) we obtain

$$\begin{aligned} |\chi_2 + \check{\beta}| &\geq |\chi_2| - |\check{\beta}| \geq \frac{4}{17}|\tilde{\alpha}|\|u\|^2 - (2\delta_u + \delta_u^2)\|\dot{T}(p)\|\|u\|^2 \\ &\geq \left(\frac{4}{17}|\tilde{\alpha}| - (2+\delta_0)\delta_0\|\dot{T}(p)\|\right)\|u\|^2 \\ &\geq \left(\frac{4}{17}|\tilde{\alpha}| - 3\delta_0\|\dot{T}(p)\|\right)\|u\|^2 \geq \frac{1}{17}|\tilde{\alpha}|\|u\|^2. \end{aligned}$$

□

Equation (3.60) implies that

$$|\mu| \leq \frac{17(\hat{M}_0(2\delta_u + \delta_u^2) + \frac{L}{2}|\mu|^2(1+\delta_u)^2)}{|\tilde{\alpha}|}. \tag{3.62}$$

Suppose that $|\mu| \leq \check{\tau} := \frac{|\tilde{\alpha}|}{17L(1+\delta_u)^2}$. Therefore, we have $|\mu|^2 \leq \check{\tau}|\mu|$, which, inserted in (3.62), yields

$$\left(1 - \frac{17L}{2|\tilde{\alpha}|}\check{\tau}(1+\delta_u)^2\right)|\mu| \leq 17\frac{\hat{M}_0}{|\tilde{\alpha}|}(2\delta_u + \delta_u^2),$$

which gives (3.59), i.e., $|\mu| \leq 34\hat{M}_0(2\delta_u + \delta_u^2)/|\tilde{\alpha}|$. According to $2\delta_u + \delta_u^2 \leq (2+\delta_0)\delta_u$ we also have

$$|\mu| \leq 34\frac{\hat{M}_0}{|\tilde{\alpha}|}(2+\delta_0)\delta_u. \tag{3.63}$$

Now, let $q := \check{\beta}/\chi_2$. Then $\chi_2 + \check{\beta} = \chi_2[1+q]$ and due to $1/(1+q) = 1 - q/(1+q)$, equation (3.60) yields

$$\mu = -\frac{l + s^H T s + \check{R}}{\chi_2}\left[1 - \frac{q}{1+q}\right],$$

hence

$$\rho = -\mu - \frac{l}{\chi_2} = \frac{1}{\chi_2}\left\{s^H T s + \check{R} - \frac{q}{1+q}\left[l + s^H T s + \check{R}\right]\right\},$$

which is equivalent to (3.58). It remains to show that $|\rho(s)| \leq K(\delta_u)^2$.

Let us analyze the absolute value of q

$$|q| = \frac{|\check{\beta}|}{|\chi_2|} \leq \frac{17\hat{M}_1}{4|\tilde{\alpha}|}(2\delta_u + \delta_u^2) \leq \frac{17\hat{M}_1}{4|\tilde{\alpha}|}(2+\delta_0)\,\delta_u.$$

Suppose that $\delta_0 \leq \frac{2|\tilde{\alpha}|}{51\hat{M}_1} \leq \frac{|\tilde{\alpha}|}{17\|\dot{T}(p)\|}$. Then $|q| \leq 17\hat{M}_1(2+\delta_0)\delta_0/4|\tilde{\alpha}| \leq 1/2$. Furthermore, we have

$$|1+q| = \frac{|\chi_2 + \check{\beta}|}{|\chi_2|} \geq \frac{|\tilde{\alpha}|\|u\|^2}{17\hat{M}_1\|u\|^2} = \frac{|\tilde{\alpha}|}{17\hat{M}_1}$$

and $(1+\delta_u)^2 \leq (1+\delta_0)^2 \leq (1+2|\tilde{\alpha}|/51\hat{M}_1)^2 =: Q_1$. Altogether we derive the following estimate

$$\begin{aligned}
|\rho| &= \frac{1}{|\chi_2|}\Bigg\{\|u\|^2\hat{M}_0\delta_u^2 + (1+\delta_u)^2\frac{L}{2}|\mu|^2\|u\|^2 + \frac{2\|u\|^2\hat{M}_0\delta_u|q|}{|1+q|} \\
&\quad + \frac{|q|}{|1+q|}\|u\|^2\hat{M}_0\delta_u^2 + \frac{|q|}{|1+q|}\frac{L}{2}|\mu|^2\|u\|^2(1+\delta_u)^2\Bigg\} \\
&\leq \frac{17}{|\chi_2|}\Bigg\{\hat{M}_0\delta_u^2 + \frac{34^2Q_1L\hat{M}_0^2}{|\tilde{\alpha}|^2}(2+\delta_0)^2\delta_u^2 + \frac{17^2\hat{M}_0\hat{M}_1^2}{|\tilde{\alpha}|^2}(2+\delta_0)4\delta_u^2 \\
&\quad + \frac{17\hat{M}_0\hat{M}_1}{2|\tilde{\alpha}|}\delta_u^2 + \frac{17\cdot 34^2Q_1L\hat{M}_0^2\hat{M}_1}{4|\tilde{\alpha}|^3}(2+\delta_0)^2\delta_u^2\Bigg\} \\
&\leq K\delta_u^2,
\end{aligned}$$

where

$$K = \frac{17\hat{M}_0}{|\chi_2|}\left\{1 + \frac{17\hat{M}_1}{2|\tilde{\alpha}|} + \frac{4\cdot 17^2\hat{M}_1^2}{|\tilde{\alpha}|^2}(2+\delta_0) + \frac{34^2Q_1L\hat{M}_0}{|\tilde{\alpha}|^2}\left(1 + \frac{17\hat{M}_1}{4|\tilde{\alpha}|}\right)(2+\delta_0)^2\right\}.$$

□

A straightforward conclusion is the following statement about stationarity.

Lemma 3.30 *Under the assumptions of Theorem 3.28, the Rayleigh functional applied to an eigenvector is stationary only if left and right eigenvector are equal.*

Proof. Equation (3.58) implies

$$p(u+s) - p(u) = -\frac{u^HT(p(u))s + s^HT(p(u))u}{u^H\dot{T}(p(u))u} - \rho(s).$$

Because $\rho(s)$ is of order $\mathcal{O}(\|s\|^2)$, the assertion follows if $u^HT(p(u)) = 0$ and $T(p(u))u = 0$, i.e., if $u = x_*$ is right and left eigenvector corresponding to $p(x_*) = \lambda_*$. □

Inequality (3.59) immediately yields relative Lipschitz continuity of $p(u)$, i.e.,

$$|p(u+s) - p(u)| \leq \frac{34\hat{M}_0(2+\delta_0)}{|x_*^H\dot{T}(\lambda_*)x_*|}\frac{\|s\|}{\|u\|}.$$

3.4 The Relationship between the Generalized Rayleigh Quotient and the Rayleigh Functional

3.4.1 Two-sided Quotient and Functional

Now that we have seen the properties of different versions of Rayleigh functionals, we want to know how the Lancaster approximation p_L given by (3.18) acts, and how both are related to each other. For our purpose we define the domain of p_L as follows

$$p_L : \ (\lambda, u, v) \in \ S_0 \times \mathcal{K}_{\varepsilon_0}(x_*) \times \mathcal{K}_{\varepsilon_0}(y_*) \ \mapsto \ p_L(\lambda, u, v) = \lambda - \frac{v^H T(\lambda) u}{v^H \dot{T}(\lambda) u} \in \mathbb{C},$$

where S_0 and ε_0 are as defined in Theorem 3.5. This restriction guarantees that the denominator $v^H \dot{T}(\lambda) u$ is nonzero in the neighborhood of values λ and vectors u, v corresponding to the simple eigenvalue λ_*. One observes that p_L is also homogeneous, i.e.,

$$p_L(\lambda, cu, dv) = p_L(\lambda, u, v) \qquad \text{for all} \ \ c \neq 0, \ d \neq 0.$$

Stationarity of the generalized Rayleigh quotient is proved in [48] by showing that the difference $p_L(\lambda + \delta\lambda, u + s, v + t) - p_L(\lambda, u, v)$ is zero to first order.

For the generalized Rayleigh functional, the bound given in (3.22) is of theoretic interest—the practical application is more or less limited to polynomial problems. In general, a direct solution of $g(p, u, v) := v^H T(p) u = 0$ is not possible. However, a good approximation will be given by a Newton step for the equation with respect to p. As mentioned before, this leads to the generalized Rayleigh quotient p_L with $p = \lambda$, or equivalently,

$$p_L(\lambda, u, v) = \lambda - \frac{g(\lambda, u, v)}{\dot{g}(\lambda, u, v)}. \tag{3.64}$$

The distance of the approximation p_L and the exact eigenvalue λ_* is determined by Theorem 3.32, but beforehand we prove the following auxiliary lemma.

Lemma 3.31 *Under the assumptions and with the constants of Theorem 3.5 and with* $\alpha = y_*^H \dot{T}(\lambda_*) x_*$, *we have*

$$|\dot{g}(\lambda, u, v)| \geq \frac{|\alpha|}{4} \|u\| \|v\| \cos\xi \cos\eta$$

for all $\lambda \in S_0$ *and all* $(u, v) \in \mathcal{K}_{\varepsilon_0}(x_*) \times \mathcal{K}_{\varepsilon_0}(y_*)$.

Proof. With the notation as in the proof of Theorem 3.5, we can rewrite the derivative as follows

$$\begin{aligned}\dot{g}(\lambda,u,v) &= (v_2^H u_2)\hat{v}^H\dot{T}(\lambda)\hat{u} = (v_2^H u_2)\hat{v}^H[\dot{T}(\lambda_*)+(\dot{T}(\lambda)-\dot{T}(\lambda_*))]\hat{u}\\ &= (v_2^H u_2)\left(\alpha+\beta+\hat{v}^H[\dot{T}(\lambda)-\dot{T}(\lambda_*)]\hat{u}\right),\end{aligned}$$

where β is defined in equation (3.28). Since the assumptions of Theorem 3.5 hold, we can use Proposition 3.6, which gives $|\beta| \le |\alpha|/2$. With $\|\dot{T}(\lambda)-\dot{T}(\lambda_*)\| \le L|\lambda-\lambda_*|$, and since we know from the proof of Theorem 3.5 that $\tau_0 \le \tau_1 = \frac{4|\alpha|}{17L}$ and $\|\hat{u}\| = \|\hat{v}\| = \sqrt{\frac{17}{16}}$, we end up with the desired inequality

$$\begin{aligned}|\dot{g}(\lambda,u,v)| &\ge |v_2^H u_2|\left(|\alpha|-|\beta|-\|\hat{u}\|\|\hat{v}\|\,\|\dot{T}(\lambda)-\dot{T}(\lambda_*)\|\right)\\ &\ge \|u\|\|v\|\cos\xi\cos\eta\left(\frac{|\alpha|}{2}-\frac{17}{16}L\frac{4}{17}\frac{|\alpha|}{L}\right) = \frac{|\alpha|}{4}\|u\|\|v\|\cos\xi\cos\eta.\end{aligned}$$

□

Theorem 3.32 *Under the assumptions and with the constants of Theorem 3.5, the iterate* (3.64) *is well-defined for all* $\lambda \in S_0$ *and all* $(u,v) \in \mathcal{K}_{\varepsilon_0}(x_*)\times\mathcal{K}_{\varepsilon_0}(y_*)$*, and we have*

$$|p_L(\lambda,u,v)-\lambda_*| \le \frac{4\|T(\lambda_*)\|}{|\alpha|}\tan\xi\tan\eta + \frac{2L}{|\alpha|}\frac{|\lambda-\lambda_*|^2}{\cos\xi\cos\eta}. \tag{3.65}$$

Proof. According to Lemma 3.31 we have $|\dot{g}(\lambda,u,v)| > 0$, i.e., the generalized Rayleigh quotient p_L is well-defined. The representation $g(\lambda_*,u,v) = g(\lambda)+\dot{g}(\lambda)(\lambda_*-\lambda)+v^H R(\lambda)u$, with remainder $R(\lambda)$ from (3.25) and $\|R(\lambda)\| \le \frac{L}{2}|\lambda-\lambda_*|^2$ implies

$$\begin{aligned}|g(\lambda,u,v)+\dot{g}(\lambda)(\lambda_*-\lambda)| &= |g(\lambda_*,u,v)-v^H R(\lambda)u|\\ &\le |g(\lambda_*,u,v)| + \|u\|\|v\|\|R(\lambda)\|\\ &\le \|T(\lambda_*)\|\|u\|\|v\|\sin\xi\sin\eta + \|u\|\|v\|\frac{L}{2}|\lambda-\lambda_*|^2.\end{aligned}$$

The definition (3.64) yields

$$p_L-\lambda_* = \lambda-\lambda_* - \frac{g(\lambda,u,v)}{\dot{g}(\lambda,u,v)} = -\frac{1}{\dot{g}(\lambda,u,v)}\left[g(\lambda,u,v)-\dot{g}(\lambda,u,v)(\lambda-\lambda_*)\right].$$

Thus, with Lemma 3.31 we obtain

$$\begin{aligned}|p_L(\lambda,u,v)-\lambda_*| &\le \frac{1}{|\dot{g}(\lambda,u,v)|}\,|g(\lambda,u,v)+\dot{g}(\lambda,u,v)(\lambda_*-\lambda)|\\ &\le \frac{4\|T(\lambda_*)\|}{|\alpha|}\tan\xi\tan\eta + \frac{2L}{|\alpha|}\frac{|\lambda-\lambda_*|^2}{\cos\xi\cos\eta},\end{aligned}$$

which proves the assertion. □

Bound (3.65) consists of two terms. The first one has the order of the bound for the generalized Rayleigh functional, but the second one is an additional nonzero term that depends on the radius of the disk S_0. Since $\tan\varepsilon_0 \leq 1/4$ with ε_0 from Theorem 3.5, an upper bound for the second term is $L|\lambda - \lambda_*|^2/8|\alpha|$.

Now we can determine the distance between the two approximations given by the generalized Rayleigh quotient $p_L(\lambda, u, v)$ and the generalized Rayleigh functional $p(u, v)$.

Corollary 3.33 *Under the assumptions of Theorem 3.32, we have*

$$|p(u,v) - p_L(\lambda,u,v)| \leq \frac{22\|T(\lambda_*)\|}{3|\alpha|}\tan\xi\tan\eta + \frac{2L}{|\alpha|}\frac{|\lambda-\lambda_*|^2}{\cos\xi\cos\eta},$$

for all $\lambda \in S_0$ and all $(u, v) \in \mathcal{K}_{\varepsilon_0}(x_*) \times \mathcal{K}_{\varepsilon_0}(y_*)$.

Proof. Since $|p(u,v) - p_L| \leq |p(u,v) - \lambda_*| + |p_L - \lambda_*|$, the estimate follows with (3.22) and (3.65). □

Note that the bounds given in this section correspond to a p_L that is defined for all kinds of nonlinear eigenvalue problems, in contrast to the definition of Lancaster in [49], which was made for matrix polynomials. The main difference in the definitions of the two generalized Rayleigh-type updates is that the Newton update p_L depends not only on the vectors u and v, but also on the previous eigenvalue approximation λ. On the other hand, the Rayleigh functional gives a new eigenvalue approximation without knowledge of the previous localization of the eigenvalue, as the Rayleigh quotient for matrices does.

3.4.2 One-sided Quotient and Functional

In [49] it is assumed that approximations for left and right eigenvectors are available at the same time. Only few methods—especially those needed in applications—however, compute the left eigenvector simultaneously, because this causes an increase in the costs while the additional information is seldom needed.

From this point of view an eigenvalue update is required that is based on the right eigenvector only. As we have seen before, the standard functional $p \equiv p(u)$ derived as solution of $u^H T(p(u))u = 0$ gives such an approximation. A reasonable way, in the general case, for solving this equation is to compute one Newton step. This leads to

$$p_N(\lambda, u) = \lambda - \frac{u^H T(\lambda)u}{u^H \dot{T}(\lambda)u},$$

with $\lambda \in S_0$, where p_N is defined by

$$p_N : \ (\lambda, u) \in \ S_0 \times \mathcal{K}_{\varepsilon_0}(x_*) \mapsto p_N \in \mathbb{C},$$

with S_0 and ε_0 as defined in Corollary 3.22, respectively Corollary 3.20. In the general case, this restriction cannot guarantee that the denominator $u^H \dot{T}(\lambda) u$ is nonzero in this neighborhood. In fact, we need $\tilde{\alpha} \neq 0$ from (3.46) instead of $\alpha \neq 0$ which is given when we approximate a simple eigenvalue. For Hermitian problems with real eigenvalues and for real symmetric problems these coincide, i.e., $\tilde{\alpha} = \alpha$, since $y_* = x_*$. So, for problems where left and right eigenvectors are different from each other, we can neglect the simplicity condition but have to assume that $\tilde{\alpha} \neq 0$.

Again, p_N is homogeneous, i.e.,

$$p_N(\lambda, cu) = p_N(\lambda, u) \qquad \text{for all} \ \ c \neq 0,$$

but not stationary, except for $T(\lambda_*) = T(\lambda_*)^H$.

For further analysis, we write p_N as

$$p_N = \lambda - \frac{g(\lambda, u)}{\dot{g}(\lambda, u)}, \tag{3.66}$$

where $g(\lambda, u) \equiv g(\lambda, u, u)$ defined in (3.19). We proceed as in the previous section by determining the sharpness of the approximation p_N to the exact eigenvalue λ_*.

Corollary 3.34 *Under the assumptions and with the constants of Corollary 3.22, we have* $|\dot{g}(\lambda, u)| \geq \frac{|\tilde{\alpha}|}{4} \|u\|^2 \cos^2 \xi$ *for all* $\lambda \in S_0$ *and all* $u \in \mathcal{K}_{\varepsilon_0}(x_*)$.

Proof. With the notation as in Corollary 3.22 we can rewrite the derivative as follows

$$\begin{aligned} \dot{g}(\lambda, u) &= (u_2^H u_2) \hat{u}^H \dot{T}(\lambda) \hat{u} = (u_2^H u_2) \hat{u}^H [\dot{T}(\lambda_*) + (\dot{T}(\lambda) - \dot{T}(\lambda_*))] \hat{u} \\ &= (u_2^H u_2) \left(\tilde{\alpha} + \tilde{\beta} + \hat{u}^H [\dot{T}(\lambda) - \dot{T}(\lambda_*)] \hat{u} \right), \end{aligned}$$

with $\tilde{\beta}$ from equation (3.50). Since the assumptions of Corollary 3.22 hold, we can use Proposition 3.23 giving us $|\tilde{\beta}| \leq |\tilde{\alpha}|/2$. With $\|\dot{T}(\lambda) - \dot{T}(\lambda_*)\| \leq L|\lambda - \lambda_*|$, and since we know from the proof of Corollary 3.22 that $\tau_0 \leq \tau_1 = \frac{4|\tilde{\alpha}|}{17L}$ and $\|\hat{u}\| = \sqrt{\frac{17}{16}}$, we end up with the desired inequality

$$\begin{aligned} |\dot{g}(\lambda, u)| &\geq |u_2^H u_2| \left(|\tilde{\alpha}| - |\tilde{\beta}| - \|\hat{u}\|^2 \, \|\dot{T}(\lambda) - \dot{T}(\lambda_*)\| \right) \\ &\geq \|u\|^2 \cos^2 \xi \left(\frac{|\tilde{\alpha}|}{2} - \frac{17}{16} L \frac{4}{17} \frac{|\tilde{\alpha}|}{L} \right) = \frac{|\tilde{\alpha}|}{4} \|u\|^2 \cos^2 \xi. \end{aligned}$$

□

Corollary 3.35 *Under the assumptions and with the constants of Corollary 3.22, the iterate* (3.66) *is well-defined for all* $\lambda \in S_0$ *and all* $u \in \mathcal{K}_{\varepsilon_0}(x_*)$*, and we have*

$$|p_N(\lambda, u) - \lambda_*| \leq \frac{5\|T(\lambda_*)\|}{|\tilde{\alpha}|} \tan\xi + \frac{2L}{|\tilde{\alpha}|} \frac{|\lambda - \lambda_*|^2}{\cos^2\xi}. \tag{3.67}$$

Proof. According to Corollary 3.34, $|\dot{g}(\lambda, u)| > 0$, i.e., p_N is well-defined. We have

$$g(\lambda_*, u) = g(\lambda) + \dot{g}(\lambda)(\lambda_* - \lambda) + u^H R(\lambda) u,$$

with remainder $R(\lambda)$ from (3.25) and $\|R(\lambda)\| \leq \frac{L}{2}|\lambda - \lambda_*|^2$. Note that $g(\lambda_*, u) = \tilde{\gamma}$ from (3.51). Hence,

$$\begin{aligned} |g(\lambda, u) + \dot{g}(\lambda)(\lambda_* - \lambda)| &= |g(\lambda_*, u) - u^H R(\lambda) u| \\ &\leq |g(\lambda_*, u)| + \|u\|^2 \|R(\lambda)\| \\ &\leq \|T(\lambda_*)\| \|u\|^2 (\sin\xi\cos\xi + \sin^2\xi) + \|u\|^2 \frac{L}{2} |\lambda - \lambda_*|^2. \end{aligned}$$

The definition (3.66) of p_N gives

$$p_N - \lambda_* = \lambda - \lambda_* - \frac{g(\lambda, u)}{\dot{g}(\lambda, u)} = -\frac{1}{\dot{g}(\lambda, u)} \left[g(\lambda, u) - \dot{g}(\lambda, u)(\lambda - \lambda_*) \right],$$

thus, with Lemma 3.34 we obtain

$$\begin{aligned} |p_N - \lambda_*| &\leq \frac{1}{|\dot{g}(\lambda, u)|} |g(\lambda, u) + \dot{g}(\lambda, u)(\lambda_* - \lambda)| \\ &\leq \frac{4\|T(\lambda_*)\|}{|\tilde{\alpha}|} (\tan\xi + \tan^2\xi) + \frac{2L}{|\tilde{\alpha}|} \frac{|\lambda - \lambda_*|^2}{\cos^2\xi} \\ &\leq \frac{4\|T(\lambda_*)\|}{|\tilde{\alpha}|} (\tan\xi + \tan\xi\tan\varepsilon_0) + \frac{2L}{|\tilde{\alpha}|} \frac{|\lambda - \lambda_*|^2}{\cos^2\xi}, \end{aligned}$$

which gives (3.67), since $\tan\varepsilon_0 \leq 1/4$. □

The second order in the angles of bound (3.65) is not carried over to bound (3.67)—which reflects also the behavior of the corresponding Rayleigh functional.

Finally, we can determine the distance between the two approximations given by the Rayleigh quotient p_N and the one-sided Rayleigh functional $p(u)$.

Corollary 3.36 *Under the assumptions of Corollary 3.35, we have*

$$|p(u) - p_N(\lambda, u)| \leq \frac{25\|T(\lambda_*)\|}{3|\tilde{\alpha}|} \tan\xi + \frac{2L}{|\tilde{\alpha}|} \frac{|\lambda - \lambda_*|^2}{\cos^2\xi},$$

for all $\lambda \in S_0$ *and all* $u \in \mathcal{K}_{\varepsilon_0}(x_*)$*.*

Proof. Since $|p(u) - p_N| \leq |p(u) - \lambda_*| + |p_N - \lambda_*|$ the estimate follows with (3.47) and (3.67). □

3.5 Conclusion

We have considered a consistent generalization of the (two-sided) Rayleigh quotient for nonlinear eigenvalue problems, in particular, a standard and a generalized Rayleigh functional, for which we have shown local unique existence, first order perturbation results and error bounds in the vectors as well as stationarity.

These results depend heavily on the algebraic simplicity of the considered eigenvalue, implying that $y_*^H \dot{T}(\lambda_*)x_* \neq 0$. In case of the standard functional this condition is replaced by $x_*^H \dot{T}(\lambda_*)x_* \neq 0$, which is a different condition if working on problems where $T(\lambda_*) \neq T(\lambda_*)^H$.

The generalized Rayleigh functional is stationary and a quadratic error bound in the angles of corresponding eigenvectors and approximations has been derived.

The one-sided functional is not stationary in the general case where $T(\lambda_*) \neq T(\lambda_*)^H$. The bound for the distance of the functional and the exact eigenvalue is only linear in the angle.

The difference to the generalized Rayleigh quotient p_L, which is closely related to the Rayleigh functional, has been analyzed, as well as this difference in the case of one-sided terms.

Altogether, we have seen that the Rayleigh functional reflects the properties of the Rayleigh quotient in the case of nonlinear eigenvalue problems.

Chapter 4

Newton-type Methods

This chapter provides a summary of methods that are related to the Newton method applied to the extended system (1.11) and alternative variants with emphasis on general nonlinear eigenvalue problems. At first, the Newton step for (1.11) itself is analyzed and three different realizations are given. We will find conditions such that it is well-defined and corresponding bounds. Convergence of the inverse iteration method [5] follows immediately. The method of inverse iteration belongs to the class of *one-vector* methods. We say that a method is of *one-vector-type*, if only one vector—in our case typically the right eigenvector—is approximated by a search space consisting of one vector. This means that new corrections are added to the old approximations. We will also review the well-known residual inverse iteration method [61] and the method of successive linear problems [72].

Chapter 3 motivates methods that compute the two-sided Rayleigh functional, since this promises a higher order of convergence. However, its computation requires left eigenvector approximations. Therefore, we introduce *two-vector methods* in Section 4.2. Two-vector methods are doubled one-vector methods that iterate left and right eigenvector simultaneously. So, they are generalizations of the one-vector methods. We introduce the *two-sided Rayleigh functional iteration*, which is a generalized inverse iteration with generalized Rayleigh functional. We will show its locally cubic convergence, which reflects the behavior of the two-sided Rayleigh quotient iteration for matrices, cf. [66]. Furthermore, we generalize the residual inverse iteration and obtain the new *two-sided residual inverse iteration*. Using the generalized Rayleigh functional, we obtain a higher order of convergence. We will also discuss other versions as the alternating Rayleigh functional iteration and the generalized Rayleigh functional iteration, which is based on a method in [78].

In general, one has at most quadratic convergence for the one-vector and at most cubic convergence for the two-vector iterations.

We will summarize the results of this chapter in Table 4.1, and compare the methods numerically afterwards.

Besides Newton-type methods that work on the extended problem—which will be discussed here—there is a class of methods that consider the characteristic equation

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

see [58] and references therein. We will not discuss this approach.

When analyzing iterative methods, the terms Q- and *R-order of convergence* are used to describe the type and order of convergence of the involved sequences. They are defined as follows, cf. [76, p. 81f].

Definition 4.1 *A zero-sequence $\{\varepsilon_k\}$ with nonnegative real entries is said to be convergent at least of Q-order $\kappa \geq 1$, if there exists an index k_0 and a constant $\bar{Q} > 0$ such that*

$$\varepsilon_{k+1} \leq \bar{Q}\varepsilon_k^\kappa, \quad \text{for all} \quad k \geq k_0.$$

If $\kappa = 1$, we assume that $\bar{Q} < 1$, additionally.

Definition 4.2 *A zero-sequence $\{\varepsilon_k\}$ with nonnegative real entries is said to be convergent at least of R-order $\kappa \geq 1$, if there exists an index k_0 and constants $0 < \bar{R} < 1$, $\bar{E} > 0$ such that*

$$\varepsilon_k \leq \bar{E}[\bar{R}]^{\kappa^k}, \quad \text{for all} \quad k \geq k_0.$$

If $\kappa = 1$, the exponent κ^k is changed to k.

Definition 4.3 *A sequence $\{x_k\}$ with limit x_* is called at least convergent of order κ, if the corresponding sequence $\{\varepsilon_k = \|x_k - x_*\|\}$ of error norms satisfies this property.*

Moreover, sequences converging at least with order κ will shortly be called to be *convergent of order κ*, in general. Let us point out that the bounds $\bar{E}[\bar{R}]^{\kappa^k}$ themselves build a sequence that converges with Q-order κ, i.e., $\{\varepsilon_k\}$ converges with R-order κ, if and only if there is a majorizing sequence which converges with Q-order κ. Therefore, Q-order κ implies R-order κ, but the opposite is not true in general.

4.1 Methods for Approximating One Eigenpair

One of the standard techniques for linear eigenvalue problems—especially in case an eigenvalue approximation is available—is the method of inverse iteration, which is closely related to the Newton method. We want to analyze the nonlinear case in detail, here.

As already mentioned in the Introduction (1.11)–(1.14), the application of Newton's method is based on the extended system

$$F_w(x,\lambda) = \begin{pmatrix} T(\lambda)x \\ w^H x - 1 \end{pmatrix} = 0, \tag{4.1}$$

where the normalizing vector $w \in \mathbb{C}^n$ has to be chosen such that $w^H x_* \neq 0$. Obviously, equation (4.1) is solved by $(x_*^w = x_*/w^H x_*, \lambda_*)$. The first derivative of F_w at (u,λ) is given by the Jacobian

$$\partial F_w(u,\lambda) = \begin{bmatrix} T(\lambda) & \dot{T}(\lambda)u \\ w^H & 0 \end{bmatrix}. \tag{4.2}$$

At the solution (x_*^w, λ_*) the Jacobian is nonsingular, if λ_* is simple and $w^H x_* \neq 0$. Then, the Newton method is well-defined and Q-quadratically convergent, provided that (u,λ) is sufficiently close to the solution. A more precise formulation will be given later. If $T(\,.\,)$ is twice continuously differentiable on S_*, then the second derivative of F_w is given by

$$\partial^2 F_w(x,\lambda)\left[\begin{pmatrix} s \\ \mu \end{pmatrix}, \begin{pmatrix} s_1 \\ \mu_1 \end{pmatrix}\right] = \begin{bmatrix} \dot{T}(\lambda)\mu_1 s + \dot{T}(\lambda)s_1\mu + \ddot{T}(\lambda)x\mu\mu_1 \\ 0 \end{bmatrix},$$

and in particular,

$$\partial^2 F_w(x,\lambda)\begin{bmatrix} s \\ \mu \end{bmatrix}^2 = \begin{bmatrix} 2\dot{T}(\lambda)s\mu + \ddot{T}(\lambda)x\mu^2 \\ 0 \end{bmatrix} = \mu \begin{bmatrix} 2\dot{T}(\lambda)s + \ddot{T}(\lambda)x\mu \\ 0 \end{bmatrix}. \tag{4.3}$$

The important property of this derivative is the fact that μ can be extracted from the sum. Later on, we will see that this implies the cubic convergence of the two-sided Rayleigh functional iteration, although the second derivative of F_w is left out for the actual computation of the Newton step: The Newton step $(u_k,\lambda_k) \mapsto (u_{k+1},\lambda_{k+1})$, where $u_{k+1} = u_k + s_k$, $\lambda_{k+1} = \lambda_k + \mu_k$, is given by the system

$$\begin{bmatrix} T(\lambda_k) & \dot{T}(\lambda_k)u_k \\ w^H & 0 \end{bmatrix}\begin{bmatrix} s_k \\ \mu_k \end{bmatrix} = -\begin{bmatrix} T(\lambda_k)u_k \\ 0 \end{bmatrix}, \tag{4.4}$$

for the Newton corrections (s_k,μ_k), which becomes

$$\begin{aligned} T(\lambda_k)s_k + \mu_k\dot{T}(\lambda_k)u_k &= -T(\lambda_k)u_k \\ w^H s_k &= 0, \end{aligned}$$

provided that u_k satisfies the normalizing condition $w^H u_k = 1$. This definition is recommended for practical implementation. For theoretical analysis it is more convenient to use the to (4.4) equivalent linear system

$$\begin{bmatrix} T(\lambda_k) & \dot{T}(\lambda_k)u_k \\ w^H & 0 \end{bmatrix}\begin{bmatrix} u_{k+1} \\ \mu_k \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \iff \begin{aligned} T(\lambda_k)u_{k+1} + \mu_k\dot{T}(\lambda_k)u_k &= 0 \\ w^H u_{k+1} &= 1. \end{aligned} \tag{4.5}$$

In this formulation we observe that for $k \geq 1$ we have $w^H u_k = 1$, i.e., except for the initial u_0, all subsequent u_k are normalized automatically. So it seems reasonable to choose u_0 such that $w^H u_0 = 1$, too, i.e., the iterates u_k start and stay in the hyperplane $w^H x = 1$.

If $T(\lambda_k)$ is nonsingular, which is the case for all $\lambda_k \neq \lambda_*$ close to λ_*, cf. Lemma 2.8, then the upper block of (4.5) can be solved for u_{k+1} and gives

$$u_{k+1} = -\mu_k T(\lambda_k)^{-1} \dot{T}(\lambda_k) u_k.$$

This can be written as $u_{k+1} = -\mu_k x_{k+1}$, where

$$x_{k+1} = T(\lambda_k)^{-1} \dot{T}(\lambda_k) u_k \quad \Longleftrightarrow \quad T(\lambda_k) x_{k+1} = \dot{T}(\lambda_k) u_k.$$

Inserting u_{k+1} into the lower block of (4.5) yields $w^H u_{k+1} = -\mu_k w^H x_{k+1} = 1$, hence $\mu_k = -1/w^H x_{k+1}$, and finally

$$\lambda_{k+1} = \lambda_k - \frac{1}{w^H x_{k+1}} = \lambda_k - \frac{w^H u_k}{w^H x_{k+1}}, \tag{4.6}$$

since $w^H u_k = 1$. For a linear problem $T(\lambda) = A - \lambda I$, we have $x_{k+1} = -(A - \lambda_k I)^{-1} u_k$, and the step $(u_k, \lambda_k) \mapsto (u_{k+1}, \lambda_{k+1}) = (\mu_k (A - \lambda_k I)^{-1} u_k, \lambda_k + \mu_k)$ is called an inverse iteration step from (u_k, λ_k), with

$$\lambda_{k+1} = \lambda_k - \frac{w^H u_k}{w^H x_{k+1}} = \frac{w^H A u_{k+1}}{w^H u_{k+1}},$$

i.e., λ_{k+1} is the two-sided Rayleigh quotient to u_{k+1} and w. In this case λ_{k+1} does not depend on the previous λ_k, but only on u_{k+1} and w.

Summarizing, we have found three different ways to realize the Newton step $(u_k, \lambda_k) \mapsto (u_{k+1}, \lambda_{k+1})$. For simplicity of notation we omit the index k and replace the index $k+1$ by "+".

Proposition 4.4 *For given $w \in \mathbb{C}^n$ with $w^H x_* \neq 0$ and (u, λ) with $w^H u = 1$, $\lambda \neq \lambda_*$, the Newton successor $(u_+ = u + s, \lambda_+ = \lambda + \mu)$ for equation (4.1) can be determined by one of the following three linear systems:*

(V_1) *Solve*

$$\begin{bmatrix} T(\lambda) & \dot{T}(\lambda) u \\ w^H & 0 \end{bmatrix} \begin{bmatrix} s \\ \mu \end{bmatrix} = - \begin{bmatrix} T(\lambda) u \\ 0 \end{bmatrix}, \tag{4.7}$$

for (s, μ), set $(u_+ = u + s, \lambda_+ = \lambda + \mu)$.

(V_2) *Solve*

$$\begin{bmatrix} T(\lambda) & \dot{T}(\lambda)u \\ w^H & 0 \end{bmatrix} \begin{bmatrix} u_+ \\ \mu \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \tag{4.8}$$

for (u_+, μ), *set* $\lambda_+ = \lambda + \mu$.

(V_3) *Solve*

$$T(\lambda)x_+ = \dot{T}(\lambda)u, \tag{4.9}$$

for x_+. *Compute* $\mu = -1/w^H x_+$, *set* $(u_+ = x_+/w^H x_+, \lambda_+ = \lambda + \mu)$.

In exact arithmetic all three versions deliver the unique Newton corrections. In practical issues (4.7) should be used, since computing the corrections s avoids rounding errors compared to the case where the whole new vector u_+ is determined.

Next, we come to the actual analysis. We will use the Newton step (4.8), since this is the most convenient formulation for this purpose. However, we start with a characterization based on general borderings, which can be used later for different matrices.

Lemma 4.5 *Let* $D \subset \mathbb{C}$ *be an open set, and let* (λ_*, x_*, y_*) *be an eigentriple with geometrically simple eigenvalue* λ_* *for* $T(\,.\,) : D \to \mathbb{C}^{n\times n}$, *where* $\|x_*\| = \|y_*\| = 1$. *Define the bordered matrix*

$$C(\lambda) := C(\lambda, a, b) = \begin{bmatrix} T(\lambda) & a \\ b^H & 0 \end{bmatrix}. \tag{4.10}$$

Then,

(i) $C(\lambda_*)$ *is nonsingular if and only if*

$$|y_*^H a| > 0, \quad \textit{and} \quad |b^H x_*| > 0. \tag{4.11}$$

(ii) If (4.11) *holds, then*

$$\|C(\lambda_*)^{-1}\| \le \frac{\sqrt{\|\Sigma_1^{-1}\|^2\|a\|^2\|b\|^2 + |y_*^H a|^2 + |b^H x_*|^2}}{|y_*^H a||b^H x_*|} = M_2(a, b), \tag{4.12}$$

with Σ_1 *from* (2.12).

Proof. Since λ_* is geometrically simple, we have $\operatorname{rank} T(\lambda_*) = n - 1$. Using the SVD (2.12), we decompose

$$\begin{aligned} a \equiv a := YY^H a = Y_1 a_1 + y_* a_2, \quad a_1 = Y_1^H a, \quad a_2 = y_*^H a, \\ b \equiv b := XX^H b = X_1 b_1 + x_* b_2, \quad b_1 = X_1^H b, \quad b_2 = x_*^H b, \end{aligned}$$

and obtain

$$\hat{C}_* = \hat{C}(\lambda_*) := \left[\begin{array}{c|c} Y^H & 0 \\ \hline 0^T & 1 \end{array}\right] C(\lambda_*) \left[\begin{array}{c|c} X & 0 \\ \hline 0^T & 1 \end{array}\right] = \left[\begin{array}{c|c|c} \Sigma_1 & 0 & a_1 \\ \hline 0^T & 0 & a_2 \\ \hline b_1^H & b_2^H & 0 \end{array}\right].$$

Now, $\hat{C}_*$ and hence $C(\lambda_*)$ is nonsingular, if and only if the Schur complement

$$\begin{bmatrix} 0 & a_2 \\ b_2^H & 0 \end{bmatrix} - \begin{bmatrix} 0^T \\ b_1^H \end{bmatrix} \Sigma_1^{-1} \left[0 \,|\, a_1\right] = \begin{bmatrix} 0 & a_2 \\ b_2^H & -b_1^H \Sigma_1^{-1} a_1 \end{bmatrix}$$

of Σ_1 in $\hat{C}_*$ is nonsingular, i.e., if and only if $|a_2| = |y_*^H a| > 0$ and $|b_2| = |x_*^H b| > 0$.

Now, let (4.11) hold. Then block elimination gives

$$\hat{C}_*^{-1} = \left[\begin{array}{c|c|c} \Sigma_1^{-1} & -\Sigma_1^{-1} a_1 a_2^{-1} & 0 \\ \hline -b_1^H b_2^{-H} \Sigma_1^{-1} & b_1^H b_2^{-H} \Sigma_1^{-1} a_1 a_2^{-1} & b_2^{-H} \\ \hline 0^T & a_2^{-1} & 0 \end{array}\right] = \left[\begin{array}{c|c} B_{11} & B_{12} \\ \hline B_{21}^H & 0 \end{array}\right],$$

where $B_{11} = \begin{bmatrix} I \\ -b_1^H b_2^{-H} \end{bmatrix} \Sigma_1^{-1} \left[\, I \,|\, -a_1 a_2^{-1} \right]$, $B_{12} = \begin{bmatrix} 0 \\ b_2^{-H} \end{bmatrix}$ and $B_{21} = \begin{bmatrix} 0 \\ a_2^{-H} \end{bmatrix}$.

This yields

$$\|C(\lambda_*)^{-1}\| = \|\hat{C}_*^{-1}\| \leq \sqrt{\|B_{11}\|^2 + \|B_{12}\|^2 + \|B_{21}\|^2}.$$

According to

$$\left\| \left[I \,\middle|\, -\frac{b_1}{b_2} \right] \right\| = \sqrt{1 + \left\| \frac{b_1}{b_2} \right\|^2} = \frac{\|b\|}{|b_2|}, \quad \left\| \left[I \,\middle|\, -\frac{a_1}{a_2} \right] \right\| = \sqrt{1 + \left\| \frac{a_1}{a_2} \right\|^2} = \frac{\|a\|}{|a_2|},$$

we have

$$\|B_{11}\| \leq \frac{\|\Sigma_1^{-1}\| \|a\| \|b\|}{|a_2||b_2|}.$$

Because of $\|B_{12}\| = 1/|b_2|$ and $\|B_{21}\| = 1/|a_2|$ we end up with

$$\|C(\lambda_*)^{-1}\| = \|\hat{C}_*^{-1}\| \leq \frac{\sqrt{\|\Sigma_1^{-1}\|^2 \|a\|^2 \|b\|^2 + |a_2|^2 + |b_2|^2}}{|a_2||b_2|},$$

which is (4.12). □

Note that the assumption that λ_* is geometrically simple is sufficient in the first place, since condition (4.11) is added. We will see that for the Newton step, condition (4.11)

holds in the neighborhood of vectors corresponding to an algebraically simple eigenvalue (and $w^H x_* \neq 0$ holds by assumption). The assumption that λ_* is algebraically simple, then implies its geometric simplicity, cf. Proposition 2.1.

We can use Lemma 4.5 to show nonsingularity of the Jacobian $\partial F_w(u,\lambda) = C(\lambda, \dot{T}(\lambda)u, w)$ in (4.7), defined in (4.10), at the solution. Under certain assumptions we derive statements on the uniform boundedness of the inverse matrices $C(\lambda)^{-1}$ for u tending towards x_* and for λ sufficiently close to λ_*.

Lemma 4.6 *Let $D \subset \mathbb{C}$ be an open set, and let (λ_*, x_*, y_*) be an eigentriple with algebraically simple eigenvalue λ_* for $T(\,.\,) : D \to \mathbb{C}^{n\times n}$, continuously differentiable, where $\|x_*\| = \|y_*\| = 1$, and $S_* = \bar{S}(\lambda_*, \tau_*) \subset D$ for some $\tau_* > 0$. Let $\dot{T}(\lambda)$ be Lipschitz continuous with constant $L > 0$, and let M_1 be the Lipschitz constant for $T(\lambda)$. Let $0 \leq \omega < \pi/2$, and $\omega := \measuredangle(\mathrm{span}\{w\}, \mathrm{span}\{x_*\})$ for w with $\|w\| = 1$. Then for all u with $\|u - x_*^w\| \leq \min\{\delta_1(\omega), \delta_2(\omega)/\sqrt{2}\}$, where $x_*^w = x_*/w^H x_*$,*

$$\delta_1(\omega) := \frac{1}{\sqrt{2}\cos\omega}, \quad \delta_2(\omega) := \frac{|y_*^H \dot{T}(\lambda_*)x_*|\cos^2\omega}{2\sqrt{M_1^2 + 4L^2 + \|\dot{T}(\lambda_*)\|^2}\sqrt{(1+\|\Sigma_1^{-1}\|^2)\|\dot{T}(\lambda_*)\|^2 + 1}},$$

and all λ with $|\lambda - \lambda_*| \leq \tau := \min\left\{\tau_*, \delta_2(\omega)/\sqrt{2}\right\}$, the matrix $C(\lambda, \dot{T}(\lambda)u, w)$ defined in (4.10) is nonsingular and

$$\|C(\lambda, \dot{T}(\lambda)u, w)^{-1}\| \leq 2M_2(\omega) = \frac{2\sqrt{(1+\|\Sigma_1^{-1}\|^2)\|\dot{T}(\lambda_*)\|^2 + 1}}{|y_*^H \dot{T}(\lambda_*)x_*|\cos\omega}. \tag{4.13}$$

Proof. Condition (4.11) gives the nonsingularity of $C(\lambda_*, \dot{T}(\lambda_*)x_*^w, w)$, since due to the algebraic simplicity of λ_* we have $\alpha = y_*^H \dot{T}(\lambda_*)x_* \neq 0$, and $|w^H x_*| = \cos\omega > 0$ holds by assumption. The bound (4.12) already yields

$$\begin{aligned}\|C_*^{-1}\| := \|C(\lambda_*, \dot{T}(\lambda_*)x_*^w, w)^{-1}\| &\leq \frac{\sqrt{\|\Sigma_1^{-1}\|^2\|\dot{T}(\lambda_*)x_*^w\|^2 + |y_*^H\dot{T}(\lambda_*)x_*^w|^2 + \cos^2\omega}}{|y_*^H\dot{T}(\lambda_*)x_*^w|\cos\omega} \\ &\leq \frac{\sqrt{(1+\|\Sigma_1^{-1}\|^2)\|\dot{T}(\lambda_*)\|^2 + 1}}{|y_*^H\dot{T}(\lambda_*)x_*|\cos\omega} = M_2(\omega),\end{aligned}$$

since $\|x_*^w\| = 1/\cos\omega$. We want to derive such a bound for (u,λ) in the neighborhood of (λ_*, x_*^w). Consider

$$\delta C := C(\lambda, \dot{T}(\lambda)u, w) - C(\lambda_*, \dot{T}(\lambda_*)x_*^w, w) = \begin{bmatrix} T(\lambda) - T(\lambda_*) & \dot{T}(\lambda)u - \dot{T}(\lambda_*)x_*^w \\ 0 & 0 \end{bmatrix}.$$

If $\|\delta C\|\|C_*^{-1}\| \leq 1/2$, then $C(\lambda) = C(\lambda, \dot{T}(\lambda)u, w)$ is nonsingular and $\|C(\lambda)^{-1}\| \leq 2\|C_*^{-1}\|$ by the perturbation lemma [46, p. 128]. We have

$$\begin{aligned}\|\delta C\| &\leq \sqrt{\|T(\lambda) - T(\lambda_*)\|^2 + \|\dot{T}(\lambda)u - \dot{T}(\lambda_*)x_*^w\|^2} \\ &\leq \sqrt{M_1^2|\lambda - \lambda_*|^2 + (\|\dot{T}(\lambda)u - \dot{T}(\lambda_*)u\| + \|\dot{T}(\lambda_*)u - \dot{T}(\lambda_*)x_*^w\|)^2} \\ &\leq \sqrt{M_1^2|\lambda - \lambda_*|^2 + (L|\lambda - \lambda_*|\|u\| + \|\dot{T}(\lambda_*)\|\|u - x_*^w\|)^2}.\end{aligned}$$

Define $z = \left(\begin{smallmatrix} u \\ \lambda \end{smallmatrix}\right)$ and $z_* = \left(\begin{smallmatrix} x_*^w \\ \lambda_* \end{smallmatrix}\right)$, so $|\lambda - \lambda_*| \leq \|z - z_*\|$ and $\|u - x_*^w\| \leq \|z - z_*\|$ holds. With

$$\|u\| \leq \|x_*^w\| + \|u - x_*^w\| \leq \frac{1}{\cos\omega} + \delta_1(\omega) = \left(1 + \frac{1}{\sqrt{2}}\right)\frac{1}{\cos\omega} \leq \frac{2}{\cos\omega}, \tag{4.14}$$

this implies

$$\|\delta C\| \leq \sqrt{M_1^2 + \left(\frac{2L}{\cos\omega}\right)^2 + \|\dot{T}(\lambda_*)\|^2}\|z - z_*\|.$$

By construction, we have

$$\|z - z_*\| \leq \sqrt{\|u - x_*^w\|^2 + |\lambda - \lambda_*|^2} \leq \sqrt{\left(\frac{\delta_2(\omega)}{\sqrt{2}}\right)^2 + \left(\frac{\delta_2(\omega)}{\sqrt{2}}\right)^2} = \delta_2(\omega),$$

which gives now

$$\|\delta C\|\|C_*^{-1}\| \leq \sqrt{M_1^2 + \left(\frac{2L}{\cos\omega}\right)^2 + \|\dot{T}(\lambda_*)\|^2}\delta_2(\omega)M_2(\omega) \leq 1/2,$$

and the assertions follow. □

Lemma 4.6 provides the assumptions we need in order to have a well-defined Newton step with a nonsingular Jacobian, but still depending on w. Hence, we are able to estimate the quality of the Newton corrected approximations, which is done in the following theorem.

Theorem 4.7 *Under the assumptions and with the constants of Lemma 4.6, let $\ddot{T}(\,.\,)$ be twice continuously differentiable on S_*. Then, the Newton step $(u, \lambda) \mapsto (u_+, \lambda_+) = (u + s, \lambda + \mu)$, with $w^H u = 1$, for equation (4.1) is well-defined, and the errors satisfy the inequality*

$$\left\| \begin{bmatrix} u_+ - x_*^w \\ \lambda_+ - \lambda_* \end{bmatrix} \right\| \leq |\lambda - \lambda_*| \left\{ K_1\|u - x_*^w\| + K_2|\lambda - \lambda_*| \right\}, \tag{4.15}$$

with $x_*^w = x_*/w^H x_*$ and $K_1 = 2M_1M_2(\omega)$, $K_2 = 2LM_2(\omega)/\cos\omega$.

Proof. In Lemma 4.6 we have already shown the existence of the inverse Jacobian $\partial F_w(u,\lambda)^{-1} \equiv C(\lambda, \dot{T}(\lambda)u, w)^{-1}$, and that $\|\partial F_w(u,\lambda)^{-1}\| \le 2M_2(\omega)$ holds uniformly provided that ε_0 is sufficiently small and $|w^H x_*| \ge \cos\omega > 0$.

Define $z = \binom{u}{\lambda}$, $z_+ = \binom{u_+}{\lambda_+}$ and $z_* = \binom{x_*^w}{\lambda_*}$, hence $F_w(z_*) = 0$. The Newton step for (4.1) starting with z reads as $z_+ = z - \partial F_w(z)^{-1} F_w(z)$. Adding the solution vector z_* to both sides of this equation yields

$$\begin{aligned} z_+ - z_* &= z - z_* - \partial F_w(z)^{-1} F_w(z) \\ &= \partial F_w(z)^{-1}\{F_w(z_*) - F_w(z) - \partial F_w(z)(z_* - z)\} \\ &= \partial F_w(z)^{-1} \int_0^1 (1-t)\partial^2 F_w(z + t(z_* - z))(z_* - z)^2 dt \\ &= \partial F_w(z)^{-1} \int_0^1 (1-t)\partial^2 F_w(z(t))(z_* - z)^2 dt, \end{aligned}$$

where $z(t) = \begin{pmatrix} x(t) \\ \lambda(t) \end{pmatrix} = \begin{pmatrix} u + t(x_*^w - u) \\ \lambda + t(\lambda_* - \lambda) \end{pmatrix}$. See, e.g., [76, p. 36] for the integral representation of the second order remainder in Taylor's formula. Using (4.3) we obtain

$$\partial^2 F_w(x(t),\lambda(t))(z_* - z)^2 = (\lambda_* - \lambda) \begin{bmatrix} 2\dot{T}(\lambda(t))(x_*^w - u) + \ddot{T}(\lambda(t))x(t)(\lambda_* - \lambda) \\ 0 \end{bmatrix}.$$

Hence, we have

$$z_+ - z_* = \partial F_w(z)^{-1} \int_0^1 (1-t)(\lambda_* - \lambda) \begin{bmatrix} 2\dot{T}(\lambda(t))(x_*^w - u) + \ddot{T}(\lambda(t))x(t)(\lambda_* - \lambda) \\ 0 \end{bmatrix} dt,$$

which, with $\mu = \lambda - \lambda_*$, leads to

$$\begin{aligned} \|z_+ - z_*\| &\le \|\partial F_w(z)^{-1}\||\mu| \int_0^1 (1-t)\left[2\|\dot{T}(\lambda(t))\|\|u - x_*^w\| + \|\ddot{T}(\lambda(t))\|\|x(t)\||\mu|\right] dt \\ &\le 2M_2(\omega)|\mu| \left\{ 2M_1\|u - x_*^w\| \int_0^1 (1-t)dt + |\mu|\frac{L}{\cos\omega} \int_0^1 (1-t)(2-t)dt \right\} \\ &\le 2M_2(\omega)|\mu| \left\{ M_1\|u - x_*^w\| + |\mu|\frac{L}{\cos\omega} \right\}, \end{aligned}$$

since, with (4.14), we have $\|x(t)\| = \|tx_*^w + (1-t)u\| \le t\|x_*\| + (1-t)\|u\| \le \frac{2-t}{\cos\omega}$. □

The convergence of (u_k, λ_k) to (x_*^w, λ_*) follows from inequality (4.15) for sufficiently good starting values with $\|u_0 - x_*^w\| \le \delta_0$, $|\lambda_0 - \lambda_*| \le \delta_0$, where δ_0 is sufficiently small. Indeed,

then we have $\|u_1 - x_*^w\| \leq (K_1 + K_2)\delta_0^2$, $|\lambda_0 - \lambda_*| \leq (K_1 + K_2)\delta_0^2$. If $(K_1 + K_2)\delta_0 \leq 1/2$, this implies $\|u_1 - x_*^w\| \leq \frac{1}{2}\delta_0$, $|\lambda_1 - \lambda_*| \leq \frac{1}{2}\delta_0$. Hence, by induction we obtain

$$\|u_k - x_*^w\| \leq \left(\frac{1}{2}\right)^k \delta_0, \quad |\lambda_k - \lambda_*| \leq \left(\frac{1}{2}\right)^k \delta_0.$$

One could ask, what influence the normalization vector w has for convergence and computational matters. Suppose we have a second normalization vector $\tilde{w}$ and $(\tilde{u}_+, \tilde{\mu})$ is the solution of (4.8) with $\tilde{w}$ instead of w. Then, since both $x_*^w = x_*/w^H x_*$ and $x_*^{\tilde{w}} = x_*/\tilde{w}^H x_*$ are multiples of x_*, they have the same angle with u. Moreover, since the pair $(\tilde{u}_+, \tilde{\mu})$ satisfies the full rank upper block equation $T(\lambda)\tilde{u}_+ + \dot{T}(\lambda)\tilde{u}\tilde{\mu} = 0$ of (4.8), as (u_+, μ) does, one has $\tilde{u}_+ = \gamma u_+$, $\tilde{\mu}_+ = \gamma\mu_+$ with some scalar $\gamma \neq 0$. The lower block in (4.8) is given by $\tilde{w}^H\tilde{u}_+ = \tilde{w}^H(\gamma u_+) = 1$, hence $\gamma = 1/(\tilde{w}^H u_+)$. Therefore, the angles of u_+ and $\tilde{u}_+$ with respect to x_* are the same and the normalized vectors satisfy

$$\frac{\tilde{u}_+}{\|\tilde{u}_+\|} = \frac{\gamma}{\|\gamma\|}\frac{u_+}{\|u_+\|} = \beta\frac{u_+}{\|u_+\|},$$

where $|\beta| = 1$, i.e., they are the same up to the phase factor β.

The bound provided by Theorem 4.7 is an important step pointing in the direction of cubic convergence of the Newton method with Rayleigh functional update for Hermitian problems with real eigenvalues *(Rayleigh functional iteration)*, because of its product structure. In this case, we do not use the λ-update from the Newton step but the Rayleigh functional $p(u)$, which gives a quadratic bound in the angles, cf. [71] for the real case. Likewise, we will formulate a two-vector inverse iteration, where the left eigenvector is iterated simultaneously, with two-sided Rayleigh functional update, which will be called the two-sided Rayleigh functional iteration. Then, the Rayleigh functional $p(u, v) \equiv \lambda$ yields the quadratic bound (3.22) for general problems, which implies an order three bound for the Newton correction.

Remark 4.8 *In standard Newton theory, cf. [64, 76], the function F_w is supposed to be differentiable with Lipschitz continuous first order derivative ∂F_w and Lipschitz constant $L_F > 0$, which guarantees a second order remainder*

$$\|F_w(z + \Delta z) - F_w(z) - \partial F_w(z)\Delta z\| \leq \frac{L_F}{2}\|\Delta z\|^2,$$

cf. [76, p. 32]. Then, the error bounds become $\|z_+ - z_*\| \leq Q\|z - z_*\|^2$, with $Q = M_F L_F/2$, where $\|\partial F_w(z)^{-1}\| \leq M_F$ holds uniformly in a neighborhood of z_*. In our case, this gives

$$\left\|\begin{bmatrix} u_{k+1} - x_*^w \\ \lambda_{k+1} - \lambda_* \end{bmatrix}\right\| \leq Q\left\|\begin{bmatrix} u_k - x_*^w \\ \lambda_k - \lambda_* \end{bmatrix}\right\|^2,$$

which implies

$$\left.\begin{array}{c}\|u_{k+1}-x_*^w\|\\ |\lambda_{k+1}-\lambda_*|\end{array}\right\} \;\leq\; Q\left\{\|u_k-x_*^w\|+|\lambda_k-\lambda_*|\right\}^2.$$

Obviously, separate convergence results for the sequences $\{u_k\}$ *and* $\{\lambda_k\}$ *cannot be derived from this bound. The special product form of the bound* (4.15) *comes from the specific form of the upper block* $T(\lambda)x$ *in* $F_w(x,\lambda)$*, which leads to the structured second order derivative* (4.3).

Note that in the linear case $T(\lambda) = A - \lambda I$, the term $\ddot{T}(\lambda)$ vanishes, i.e., inequality (4.15) becomes

$$\left\|\begin{bmatrix}u_+ - x_*^w\\ \lambda_+ - \lambda_*\end{bmatrix}\right\| \leq K_1|\lambda-\lambda_*|\|u-x_*^w\|.$$

This implies that both sequences $\{u_k\}$ and $\{\lambda_k\}$ converge Q-superlinearly. The estimate (4.15) has already been used by Unger in 1950 [90] for the linear case, where he also mentioned that the theory applies to nonlinear problems.

The standard formulation of the inverse iteration is given in Algorithm 1, cf. [58], which uses the Newton step (4.9) for both u_{k+1} and λ_{k+1}. The inverse iteration for linear problems was introduced by Wielandt in 1944 as a method for computing eigenfunctions of linear operators and turned into a method for computing eigenvectors of matrices by Wilkinson, see [42] for an extensive review and supplement.

Algorithm 1 *Inverse iteration*

Input: (λ_0, u_0), *normalization vector* w *such that* $w^H u_0 = 1$
for $k = 0, 1, 2, ...$
S1: *Solve* $T(\lambda_k)x_{k+1} = \dot{T}(\lambda_k)u_k$
S2: $\lambda_{k+1} = \lambda_k - (w^H u_k)/(w^H x_{k+1})$
S3: *Normalize* $u_{k+1} = (x_{k+1})/(w^H x_{k+1})$

Algorithm 1, if properly implemented, converges locally quadratically, cf. [5], which follows also from our previous analysis that immediately gives the following theorem.

Theorem 4.9 *Let* $D \subset \mathbb{C}$ *be an open set, and let* (λ_*, x_*, y_*) *be an eigentriple with algebraically simple eigenvalue* λ_* *for* $T(\,.\,) : D \to \mathbb{C}^{n\times n}$*, twice continuously differentiable, where* $\|x_*\| = \|y_*\| = 1$*, and* $S_* = \bar{S}(\lambda_*, \tau_*) \subset D$ *for some* $\tau_* > 0$*. Let* $0 \leq \omega < \pi/2$*, and* $\omega := \measuredangle(\mathrm{span}\{w\}, \mathrm{span}\{x_*\})$ *for* w *with* $\|w\| = 1$*. Then, there exist constants* $\delta_0 > 0$*,*

$K_1(\omega) > 0$, $K_2(\omega) > 0$, such that for all (u_0, λ_0) with $\|u_0 - x_*^w\| \le \delta_0$ and $|\lambda_0 - \lambda_*| \le \delta_0$, where $x_*^w = x_*/w^H x_*$, the inverse iteration in Algorithm 1 is well-defined and converges Q-quadratically to the solution (x_*^w, λ_*), since

$$\left\| \begin{bmatrix} u_{k+1} - x_*^w \\ \lambda_{k+1} - \lambda_* \end{bmatrix} \right\| \le Q(\omega) \left\| \begin{bmatrix} u_k - x_*^w \\ \lambda_k - \lambda_* \end{bmatrix} \right\|^2 \le \left(\frac{1}{2}\right)^{k+1} \sqrt{2}\delta_0 \tag{4.16}$$

holds with $Q(\omega) = \sqrt{K_1(\omega)^2 + K_2(\omega)^2}$. Moreover, we have

$$\left.\begin{matrix} \|u_{k+1} - x_*^w\| \\ |\lambda_{k+1} - \lambda_*| \end{matrix}\right\} \quad \le \quad |\lambda_k - \lambda_*| \left\{ K_1(\omega)\|u_k - x_*^w\| + K_2(\omega)|\lambda_k - \lambda_*| \right\}.$$

Proof. Set $\delta_0 = \min\left\{\tau_*, \delta_1(\omega), \delta_2(\omega)/\sqrt{2}, 1/(2\sqrt{2}Q(\omega))\right\}$. Inequality (4.15) implies (4.16) with $Q(\omega) = \sqrt{K_1(\omega)^2 + K_2(\omega)^2}$, with K_1, K_2 from Theorem 4.7. Furthermore, we obtain $Q\|z_0 - z_*^w\| \le Q\sqrt{2}\delta_0 \le \frac{1}{2}$, hence $\|z_1 - z_*^w\| \le Q\|z_0 - z_*^w\|^2 \le \frac{1}{2}\|z_0 - z_*^w\|$, and the right inequality in (4.16), which shows convergence to the solution, follows by induction. The second bound has been shown in Theorem 4.7. It provides Q-superlinear convergence of the sequence $\{\lambda_k\}$. □

If λ_0 is fixed, then Algorithm 1 yields the classical inverse iteration in case of linear problems, where S2 is neglected. If λ is updated, then Algorithm 1 yields the Rayleigh quotient iteration with the Rayleigh quotient $\lambda_{k+1} = \frac{w^H A x_{k+1}}{w^H x_{k+1}}$, for the linear operator $T(\lambda) = A - \lambda I$. For linear Hermitian problems, cubic convergence can be achieved with the one-sided Rayleigh quotient $\lambda_{k+1} = \frac{x_{k+1}^H A x_{k+1}}{x_{k+1}^H x_{k+1}}$ instead of S2, cf. [66, 67].

Corresponding to the Rayleigh quotient iteration for linear problems, in [71] a Rayleigh functional iteration was proposed and analyzed for real symmetric problems with real eigenvalues. Locally cubic convergence was shown for this case. The Rayleigh functional iteration differs from Algorithm 1 only in S2, where S2 is replaced by $\lambda_{k+1} = p(x_{k+1})$ such that $x_{k+1}^H T(\lambda_{k+1}) x_{k+1} = 0$. If we set $w = w_k = u_k$ such that u_k has unit norm in every iteration step, then we can use the one-sided Rayleigh functional $p(x_{k+1})$ also for general problems, since then we have $|p(x_{k+1}) - \lambda_*| \le K \tan \xi_{k+1} = K\|u_{k+1} - x_*^w\|$ by (2.25) and (3.47). Thus, if inequality (4.15) holds, then it implies quadratic convergence as well. However, we have not considered the inverse Jacobian for the case $w = u$, yet.

If the normalization parameter w is chosen as $w \equiv w_k = T(\lambda_k)^H v_{k+1}$, where v_{k+1} is an approximation to the left eigenvector, as was suggested in [72], then S2 becomes

$$\lambda_{k+1} = \lambda_k - \frac{v_{k+1}^H T(\lambda_k) u_{k+1}}{v_{k+1}^H \dot{T}(\lambda_k) u_{k+1}}, \tag{4.17}$$

which is just the Rayleigh quotient proposed in [49], cf. the definition of p_L in (3.18). In general, no approximation to the left eigenvector will be at hand. Methods that compute the left eigenvector simultaneously, are discussed in §4.2. Unless the left eigenvector is iterated as well, different normalizations are as good as any other choice, generally speaking.

Suppose that we are in the k-th iteration step of the inverse iteration and solve the linear system by computing a factorization of $T \equiv T(\lambda_k)$. Then, in order to reduce costs, we would like to keep this factorization for a while to solve the subsequent systems that occur in the next iteration steps. If λ_k is sufficiently close to λ_*, then this is feasible, as Lemma 2.8 shows, since we have

$$x_{k+1} = T(\lambda_k)^{-1}\dot{T}(\lambda_k)u_k = \frac{y_*^H\dot{T}(\lambda_k)u_k}{y_*^H\dot{T}(\lambda_*)x_*(\lambda_k - \lambda_*)}x_* + \mathcal{O}(1),$$

i.e., a dominant component in the direction of x_*. However, this presentation holds only for sufficiently good λ_k. If this is not the case, then the linear problem $Tx = \gamma\dot{T}(\lambda_{k+l})x$, for $l \geq 1$, is solved in subsequent iterations, which will often lead to divergence. Thus, in general at least in the first iteration steps a new factorization for $T(\lambda_k)$ is required. For medium sized problems this is feasible, for large problems it is in general not due to the high storage and time requirements. Then, iterative solvers must be used, typically in combination with appropriate preconditioners.

In order to overcome the costs that arise in the computation of the eigenvector approximation, Neumaier [61] proposed the so-called *residual inverse iteration method* that allows an iteration with a fixed matrix $T(\lambda_0)$ for several or possibly all iterations steps. The idea is as follows:

Assume, that $T(\lambda)$ is twice continuously differentiable. Then the Newton step (1.14) gives

$$\begin{aligned} x_k - x_{k+1} &= x_k + (\lambda_{k+1} - \lambda_k)T(\lambda_k)^{-1}\dot{T}(\lambda_k)x_k \\ &= T(\lambda_k)^{-1}(T(\lambda_k) + (\lambda_{k+1} - \lambda_k)\dot{T}(\lambda_k))x_k \\ &= T(\lambda_k)^{-1}T(\lambda_{k+1})x_k + \mathcal{O}(|\lambda_{k+1} - \lambda_k|^2). \end{aligned}$$

Neglecting the second order term yields

$$x_{k+1} = x_k - T(\lambda_k)^{-1}T(\lambda_{k+1})x_k = T(\lambda_k)^{-1}[T(\lambda_k) - T(\lambda_{k+1})]x_k,$$

which shows the relation to the standard inverse iteration step S1 of Algorithm 1. The iteration converges still if λ_k in $T(\lambda_k)^{-1}$ is replaced by a fixed shift. The actual iteration takes places on the residual of the current approximation, which explains the name of the method.

Algorithm 2 *Residual inverse iteration*

Input: (λ_0, u_0), *normalization vector* w *such that* $w^H u_0 = 1$
for $k = 0, 1, 2, ...$
S1: *Solve* $w^H T(\lambda_0)^{-1} T(\lambda_{k+1}) u_k = 0$ *for* λ_{k+1}
S2: *Compute the residual* $r_k = T(\lambda_{k+1}) u_k$
S3: *Solve* $T(\lambda_0) s_k = r_k$ *for* s_k
S4: *Set* $x_{k+1} = u_k - s_k, \quad u_{k+1} = x_{k+1}/w^H x_{k+1}$

To have as little costs as possible the shift λ_0 is fixed in Algorithm 2. During the process it is advisable to update the shift every now and then, particularly when the rate of convergence is not satisfying.

The eigenvalue update in S1 seems unusual and brings the generalized Rayleigh functional back to mind. In fact, the term $y_0^H = w^H T(\lambda_0)^{-1}$ is an approximation to the left eigenvector y_*, since

$$y_0 = T(\lambda_0)^{-H} w \sim \frac{(x_*^H w) y_*}{x_*^H \dot{T}(\lambda_*)^H y_* (\lambda_0 - \lambda_*)},$$

is a step in the direction of y_*, if λ_0 is sufficiently close to λ_*, cf. Lemma 2.8. The corresponding Newton step would deliver $y_0 = T(\lambda_0)^{-H} \dot{T}(\lambda_0)^H w$, but the first update is good enough. However, since the vector w is fixed, the term $y_0 = T(\lambda_0)^{-H} w$ is not as good as an iterated left eigenvector approximation, and the eigenvalue update is not as good as the generalized Rayleigh functional. A cheap improvement of Algorithm 2 is to save the last $y_j^H = y_{j-1}^H T(\lambda_{k+1})^{-1}$, which is updated whenever the shift is updated to the actual λ_{k+1}, starting with $y_0^H = w^H T(\lambda_0)^{-1}$, and take y_j^H instead of $w^H T(\lambda_0)^{-1}$ in S1. The normalization in S4 can be replaced by $u_{k+1} = x_{k+1}/\|x_{k+1}\|$. If the shift and y_j are constantly updated, then the eigenvalue update in S2 will be close to the generalized Rayleigh functional, the approximation quality of which is quadratic in the angles.

If $T(\lambda)$ is twice continuously differentiable, λ_* is algebraically simple and x_* is normalized by $w^H x_* = 1$, then the residual inverse iteration converges for all (λ_0, u_0) sufficiently close to (λ_*, x_*), and we have

$$\frac{\|x_{k+1} - x_*\|}{\|x_k - x_*\|} = \mathcal{O}(|\lambda_0 - \lambda_*|) \quad \text{and} \quad |\lambda_{k+1} - \lambda_*| = \mathcal{O}(\|x_k - x_*\|), \tag{4.18}$$

see [61]. It is also stated in [61] that updating the shift λ_0 by the most recent value of λ_k implies quadratic convergence, and in case of Hermitian problems with real eigenvalues—where S1 in Algorithm 2 would be replaced by the Rayleigh functional $u_k^H T(\lambda_{k+1}) u_k = 0$,

causing $|\lambda_{k+1} - \lambda_*| = \mathcal{O}(\|x_k - x_*\|^2)$—cubic convergence. We cannot follow this argument. Cubic convergence would be given if

$$\|x_{k+1} - x_*\| = \mathcal{O}(\|x_k - x_*\|^3).$$

But, setting $\lambda_0 = \lambda_k$ gives $|\lambda_k - \lambda_*| = \mathcal{O}(\|x_{k-1} - x_*\|^t)$, and inserting this in the left equation of (4.18) yields

$$\|x_{k+1} - x_*\| = \mathcal{O}(\|x_k - x_*\|\|x_{k-1} - x_*\|^t), \tag{4.19}$$

where $t = 1$ in general, respectively $t = 2$ in case $x_* = y_*$. Particularly, setting $\lambda_0 = \lambda_{k+1}$ implies $x_{k+1} = 0$ in exact arithmetic, hence, the best update that we are allowed to use is λ_k.

The R-order of the sequence $\{x_k\}$ with permanent update is $\kappa = (1 + \sqrt{1 + 4t})/2$, which is the positive root of the equation $\kappa^2 = \kappa + t$ corresponding to (4.19), i.e., we have $\kappa = (1 + \sqrt{5})/2 \approx 1.62$ for $t = 1$, and $\kappa = 2$ for $t = 2$.

The third algorithm in this class of methods is the *method of successive linear problems*, which was introduced by Ruhe [72]. If $T(\lambda)$ is twice continuously differentiable, then

$$T(\lambda + \mu) = T(\lambda) + \mu\dot{T}(\lambda) + \frac{\mu^2}{2}R(\lambda, \mu),$$

where the remainder $R(\lambda, \mu)$ is bounded by

$$\|R(\lambda, \mu)\| \leq \sup_{|\xi|<|\mu|} \|\ddot{T}(\lambda + \xi)\|.$$

Discarding R yields Algorithm 3, cf. equation (2.4). We have to solve a linear generalized eigenproblem in every iteration step (successively), which may be costly depending on the structure and the size of the matrices.

Algorithm 3 *Method of successive linear problems*

Input: λ_0
for $k = 0, 1, 2, ...$
S1: *Solve the linear eigenproblem* $T(\lambda_k)x = -\mu\dot{T}(\lambda_k)x$
S2: *Choose the eigenvalue* μ *smallest in modulus*
S3: $\lambda_{k+1} = \lambda_k + \mu$

However, in contrast to the other presented methods, the method of successive linear problems gives the user more control on the approximated eigenvalue, since linear eigenproblem solvers give several eigenpairs back, in general. If the problem is of moderate size,

the QZ decomposition can be used to solve the linear problem. When all eigenvalues of the linearized problem are computed, one can be sure to choose always the smallest eigenvalue in absolute value, and in that it is (almost) guaranteed that the final eigensolution is indeed the smallest eigenvalue. This cannot be guaranteed for Algorithms 1 and 2, which are strongly dependent on the starting values. However, for genuinely nonlinear problems and a bad starting value, the eigenvalues of the linear problem will be far away from the eigenvalues of the nonlinear problem, and the method may fail. It is a locally convergent method anyhow.

Algorithm 3 converges quadratically to the simple eigenvalue λ_*, if $T(\lambda)$ is twice continuously differentiable, $\dot{T}(\lambda_*)$ is nonsingular, and zero is an algebraically simple eigenvalue of $\dot{T}(\lambda_*)^{-1}T(\lambda_*)$, cf. [72]. The last condition can be neglected, since it follows already from the algebraic simplicity of λ_*: The linear problem $\dot{T}(\lambda_*)^{-1}T(\lambda_*)x = \lambda x$ can be rewritten into $(T(\lambda_*) - \lambda\dot{T}(\lambda_*))x = 0$, for which $(0, x_*, y_*)$ is an eigentriple. Proposition 2.2 states, that zero is an algebraically simple eigenvalue, if $-y_*^H\dot{T}(\lambda_*)x_* \neq 0$, which is equivalent to the algebraic simplicity of λ_* for $T(\lambda)x = 0$.

If the nonlinear eigenproblem allows a variational characterization of the eigenvalues, see §1.3, then a good choice will be the *safeguarded iteration method* [100].

4.2 Methods for Approximating One Eigentriple

The analysis in Chapter 3 enables us to analyze methods using the generalized Rayleigh functional, which can be used when the left eigenvector is computed as well, and consequently have a higher order of convergence.

Starting with the inverse iteration method coupled with the nonlinear generalized Rayleigh functional we discuss a variety of related methods that have different advantages. In fact, all variants that are known for linear problems can be transferred, as there are, for instance, the alternating Rayleigh quotient iteration, the two-sided Rayleigh quotient iteration [67], and a generalized Rayleigh quotient iteration [77].

Another approach is to generalize the simplified Newton methods of Section 4.1, such that they compute eigentriples, i.e., a system for the left eigenvector is added. The two remaining methods from the last section can be reformulated in this way: The residual inverse iteration method and the method of successive linear problems. Generalizing the residual inverse iteration is reasonable, since there is only little extra cost for the second system, where the factorization of $T(\sigma)$ used in the first system can be reused (conjugately transposed) for the second system. Then, with Chapter 3, applying the generalized Rayleigh

functional yields some gain in the convergence rate, see §4.2.2.

However, it seems to make no sense to construct a generalization of the method of successive linear problems. The eigenvalue update depends only on the solution of one linear eigenproblem, we cannot include the information gained by solving an adjoint linear system according to the left eigenvector. Therefore, the effort for iterating the left eigenvector simultaneously doubles the cost but does not accelerate convergence, unless there is an efficient method for linear problems that gives the left eigenvector almost for free.

4.2.1 Two-sided Rayleigh Functional Iteration

In this section we follow the approach given in [71], where the inverse iteration method of Algorithm 1 for the eigenvector is combined with the computation of the nonlinear Rayleigh functional $p(u)$. When we compute left eigenvector approximations, we are able to compute the generalized nonlinear Rayleigh functional $p(u, v)$, instead of the one-sided Rayleigh functional $p(u)$, which promises a higher order in the angles ξ, η, defined in (3.20), provided we are sufficiently close to the solution. We end up with an algorithm that, applied to linear problems, coincides with Ostrowski's two-sided Rayleigh quotient iteration for nonnormal matrices [66].

Algorithm 4 *Two-sided Rayleigh functional iteration (RFI)*

Input: (λ_0, u_0, v_0) *where* $u_0^H u_0 = v_0^H v_0 = 1$
for $k = 0, 1, 2, ...$
S1: *Solve* $T(\lambda_k)x_{k+1} = \dot{T}(\lambda_k)u_k$, *set* $u_{k+1} = x_{k+1}/\|x_{k+1}\|$
S2: *Solve* $T(\lambda_k)^H y_{k+1} = \dot{T}(\lambda_k)^H v_k$, *set* $v_{k+1} = y_{k+1}/\|y_{k+1}\|$
S3: *Solve* $v_{k+1}^H T(\lambda_{k+1})u_{k+1} = 0$ *for* λ_{k+1}

The costs for the two-sided Rayleigh functional iteration are similar as those of the inverse iteration in Algorithm 1. It is overly pessimistic to claim that two-sided methods have double cost, in general:

1. First of all, if the first linear system in the k-th iteration is solved by factorizing $T(\lambda_k)$, then the same factorization can be used for the second system with $T(\lambda_k)^H$ by taking the conjugate transpose. Compared to Algorithm 1, theoretically, only two more backward solves with triangular matrices per iteration are needed.

2. For large problems a preconditioned iterative solver is needed to solve the systems. Then, at least the preconditioner can be reused for the second system.

3. The information obtained by the left eigenvector iterates leads to better approximations in the eigenvalues, implying a faster convergence of the method. Moreover, the generalized Rayleigh functional is stationary in contrast to the one-sided Rayleigh functional, respectively the Newton update, which is used to obtain eigenvalue updates for the inverse iteration. Hence, in general, we will solve less than two times the linear systems that need to be solved by the inverse iteration.

4. Using parallel processing, the linear systems can be solved simultaneously. The easiest way is to handle each system on one processor. Then, two-sided methods are faster than one-sided methods, due to the third point, if there is no additional parallel processing of matrix-vector multiplications etc.

Since we assume no special structure of T, this is the most general statement that we can make.

Remark 4.10 *In practice, matrix-vector multiplications will be handled appropriately. Consider for instance the polynomial problem* (1.3) *of degree m and order n. Then the computation of $T(\lambda)x$ will be done through computing $m+1$ matrix-vector multiplications and adding the resulting vectors ($(m+1)n^2+2mn$ flops), rather than adding $m+1$ matrices and performing one matrix-vector multiplication afterwards ($(2m+1)n^2$ flops). All single matrix-vector multiplications can be send to separate processors.*

We want to prove convergence of the two-sided Rayleigh functional iteration as stated in Algorithm 4. Since the normalizing conditions are chosen differently from the ones corresponding to the Newton step (4.9) and the inverse iteration method in Algorithm 1, we have to show nonsingularity of the Jacobian and the uniform boundedness of the norms of the inverse Jacobians close to the solution with $w = u$.

Lemma 4.11 *Let $D \subset \mathbb{C}$ be an open set, and let (λ_*, x_*, y_*) be an eigentriple with algebraically simple eigenvalue λ_* for $T(\,.\,) : D \to \mathbb{C}^{n\times n}$, continuously differentiable, where $\|x_*\| = \|y_*\| = 1$, and $\bar{S}(\lambda_*, \tau_*) \subset D$ for some $\tau_* > 0$. Let $\dot{T}(\lambda)$ be Lipschitz continuous with constant $L > 0$, and let M_1 be the Lipschitz constant for $T(\lambda)$. Let $\tan\varepsilon_0 = \frac{|\alpha|}{2}\|\dot{T}(\lambda_*)\|$. Then for all u with $\|u\| = 1$, $\xi := \measuredangle(\operatorname{span}\{u\}, \operatorname{span}\{x_*\}) \le \varepsilon_0$, and all λ with*

$$|\lambda - \lambda_*| \le \tau := \min\left\{\tau_*, \frac{1}{2M_2^0\sqrt{M_1^2 + L^2}}\right\},$$

where

$$M_2^0 := \frac{\sqrt{(1 + \|\Sigma_1^{-1}\|^2)\|\dot{T}(\lambda_*)\|^2 + 1}}{\cos^2\varepsilon_0\,|y_*^H\dot{T}(\lambda_*)x_*|}, \tag{4.20}$$

with Σ_1 from (2.12)*, the matrix $C(\lambda, \dot{T}(\lambda)u, u)$ defined in* (4.10) *is nonsingular and*

$$\|C(\lambda, \dot{T}(\lambda)u, u)^{-1}\| \leq \frac{M_2^0}{1 - M_2^0|\lambda - \lambda_*|\sqrt{M_1^2 + L^2}} \leq 2M_2^0. \tag{4.21}$$

Proof. With u_1, u_2 from (3.27), we have

$$|y_*^H \dot{T}(\lambda_*)u| = |y_*^H \dot{T}(\lambda_*)[x_* u_2 + X_1 u_1]| \geq \cos\varepsilon_0[|\alpha| - \|\dot{T}(\lambda_*)\| \tan\varepsilon_0] \geq \cos\varepsilon_0 \frac{|\alpha|}{2} > 0, \tag{4.22}$$

since $\alpha = y_*^H \dot{T}(\lambda_*)x_* \neq 0$ due to the algebraic simplicity of λ_*, and $|u^H x_*| = \cos\varepsilon_0 > 0$ holds with the definition of ε_0. Thus, condition (4.11) is satisfied, hence $C(\lambda_*)$ is nonsingular. The bound (4.12) already yields

$$\|C(\lambda_*, \dot{T}(\lambda_*)u, u)^{-1}\| \leq \frac{\sqrt{\|\Sigma_1^{-1}\|^2 \|\dot{T}(\lambda_*)\|^2 + |y_*^H \dot{T}(\lambda_*)u|^2 + |u^H x_*|^2}}{|y_*^H \dot{T}(\lambda_*)u||u^H x_*|} = M_2(\dot{T}(\lambda_*)u, u),$$

an upper bound of which can be specified to M_2^0 given in (4.20), using $|u^H x_*| \leq \|u\|\|x_*\| = 1$ and (4.22). Consider

$$\delta C := C(\lambda, \dot{T}(\lambda)u, w) - C(\lambda_*, \dot{T}(\lambda_*)u, w) = \begin{bmatrix} T(\lambda) - T(\lambda_*) & (\dot{T}(\lambda) - \dot{T}(\lambda_*))u \\ 0 & 0 \end{bmatrix}.$$

The perturbation lemma [46, p. 128] yields the nonsingularity of $C(\lambda)$ and the bound $\|C(\lambda)^{-1}\| \leq \frac{1}{1-\kappa}\|C_*^{-1}\|$, if $\|\delta C\|\|C_*^{-1}\| \leq \kappa < 1$. Assuming that $|\lambda - \lambda_*| \leq \tau$, we obtain

$$\begin{aligned} \|\delta C\|\|C_*^{-1}\| &\leq \sqrt{\|T(\lambda) - T(\lambda_*)\|^2 + \|(\dot{T}(\lambda) - \dot{T}(\lambda_*))u\|^2}\|C_*^{-1}\| \\ &\leq M_2^0\sqrt{M_1^2 + L^2}|\lambda - \lambda_*| =: \kappa \leq \frac{1}{2} < 1, \end{aligned}$$

which implies (4.21). □

Inequality (4.15) is stated in terms of norm distances, but for further analysis we would like to have such a bound in terms of angles.

Corollary 4.12 *Under the assumptions and with the constants of Theorem 4.7, set $w = u$ with $\|u\| = 1$, and define $\xi_+ = \measuredangle(\mathrm{span}\{u_+\}, \mathrm{span}\{x_*\})$. Then we have*

$$\sin\xi_+ \leq |\lambda - \lambda_*| \left\{K_1 \tan\xi + K_2|\lambda - \lambda_*|\right\}. \tag{4.23}$$

Proof. Since $x_*^w = x_*/w^H x_*$, we have

$$\|u - x_*^w\|^2 = \left\|u - \frac{x_*}{w^H x_*}\right\|^2 = \|u\|^2 - 2\mathrm{Re}\frac{u^H x_*}{w^H x_*} + \frac{\|x_*\|^2}{|w^H x_*|^2}.$$

If we set $w = u$ where $\|u\| = 1$, then this yields $\|u - x_*^u\|^2 = \frac{1}{\cos^2\xi} - 1 = \tan^2\xi$, hence

$$\tan\xi = \|u - x_*^u\|. \tag{4.24}$$

On the other hand we have

$$\|u_+ - x_*^u\| = \|x_*^u\| \left\| \frac{u_+}{\|x_*^u\|} - \frac{x_*^u}{\|x_*^u\|} \right\| = \frac{1}{\cos\xi} \left\| \frac{u_+}{\|x_*^u\|} - \frac{x_*^u}{\|x_*^u\|} \right\|.$$

Using Lemma 2.10, we end up with

$$\sin\xi_+ \leq \left\| \frac{u_+}{\|x_*^u\|} - \frac{x_*^u}{\|x_*^u\|} \right\| = \cos\xi \|u_+ - x_*^u\| \leq \|u_+ - x_*^u\|,$$

which implies (4.23) by inserting the terms in (4.15). □

Altogether, we end up with the following convergence theorem for the two-sided Rayleigh functional iteration.

Theorem 4.13 *Let $D \subset \mathbb{C}$ be an open set and let $T : D \to \mathbb{C}^{n\times n}$ be twice continuously differentiable (holomorphic) in D. Let $\lambda_* \in D$ be an algebraically simple eigenvalue of T with corresponding right and left eigenvectors x_*, y_*, with $\|x_*\| = \|y_*\| = 1$, and let $\tau_* > 0$ be such that $S_* := \{\lambda \in \mathbb{C} : |\lambda - \lambda_*| \leq \tau_*\} \subset D$. Then, there exist constants $0 < \varepsilon_0 < \pi/3$, $0 < \tau_0 \leq \tau_*$ and $K_0^a > 0, \tilde{K}_0^a > 0$, $C_0^a > 0$, $K_0^d > 0, \tilde{K}_0^d > 0$, $C_0^d > 0$, $K > 0$, $Q_0 > 0$ with $\tau_0 Q_0 \leq 1/2$, such that the two-sided Rayleigh functional iteration in Algorithm 4 is well-defined for every initial triplet (λ_0, u_0, v_0), $(u_0, v_0) \in \mathcal{K}_{\varepsilon_0}(x_*) \times \mathcal{K}_{\varepsilon_0}(y_*)$, $\|u_0\| = \|v_0\| = 1$, and $\lambda_0 \in \bar{S}(\lambda_*, \tau_0)$. Moreover, it converges in the following sense*

$$\lim_{k\to\infty} \lambda_k = \lambda_*, \quad \lim_{k\to\infty} \xi_k = \lim_{k\to\infty} \eta_k = 0,$$

where $\xi_k = \measuredangle(\mathrm{span}\{u_k\}, \mathrm{span}\{x_*\})$, $\eta_k = \measuredangle(\mathrm{span}\{v_k\}, \mathrm{span}\{y_*\})$, and the rate of convergence is characterized by

$$\sin\xi_{k+1} \leq \|u_{k+1} - x_*^{u_k}\| \leq \begin{cases} \frac{1}{2}\sin\xi_k, \\ K_0^a \sin^2\xi_k \sin\eta_k, \\ K_0^d \|u_k - x_*^{u_k}\|^2 \|v_k - y_*^{v_k}\|, \end{cases} \tag{4.25}$$

$$\sin\eta_{k+1} \leq \|v_{k+1} - y_*^{v_k}\| \leq \begin{cases} \frac{1}{2}\sin\eta_k, \\ \tilde{K}_0^a \sin^2\eta_k \sin\xi_k, \\ \tilde{K}_0^d \|v_k - y_*^{v_k}\|^2 \|u_k - x_*^{u_k}\|, \end{cases} \tag{4.26}$$

and

$$|\lambda_{k+1} - \lambda_*| \leq \begin{cases} 4K\left(\frac{1}{2}\right)^{2^{k+1}}, \\ C_0^a |\lambda_k - \lambda_*|^2 \sin\xi_k \sin\eta_k, \\ C_0^d |\lambda_k - \lambda_*|^2 \|u_k - x_*^{u_k}\| \|v_k - y_*^{v_k}\|, \end{cases} \tag{4.27}$$

where $x_*^{u_k} = x_*/u_k^H x_*$, $y_*^{v_k} = x_*/v_k^H y_*$. *The R-order is at least 3.*

Proof. Set $\tau_0 := \min\left\{\tau_*, \frac{1}{2M_2^0\sqrt{M_1^2+L^2}}, \frac{1}{2Q_0}\right\}$, with L from (2.11), M_1 from (2.7), M_2^0 from (4.20) and $Q_0 = \max\{2K_1 + 4KK_2\sin\eta_k, 2K_1 + 4KK_2\sin\xi_k\}$, with $K = \frac{8\|T(\lambda_*)\|}{3|y_*^H\dot{T}(\lambda_*)x_*|}$. The constants K_1 and K_2 from Theorem 4.7 can be simplified to $K_1 = 2M_1M_2^0$, $K_2 = LM_2^0$. We prepared the statements for this proof as far as possible and have to merge the pieces now. In the k-th step we use $w = w_k = u_k$, such that we have $\|u_0\| = \ldots = \|u_k\| = \|u_{k+1}\| = \ldots = \|x_*\| = 1$. In this way, the normalizing assumptions of Lemma 4.11 and Theorem 4.7 are satisfied, and $w^H x_* = x_*^H x_* = 1$. For $\tan\varepsilon_0 = \frac{|\alpha|}{4}\|\dot{T}(\lambda_*)\|$, the assumptions of Lemma 4.11 hold, hence the Jacobian $C(\lambda, \dot{T}(\lambda)u, u)$ is nonsingular, and its inverse is bounded by $2M_2^0$, see (4.21). Therefore, the Newton step is well-defined. Since the assumptions of Theorem 3.5 are satisfied as well, we can use bound (3.22), i.e., $p(u_k, v_k) \leq K\tan\xi_k\tan\eta_k$. Requiring $\varepsilon_0 < \pi/3$ implies

$$\sin\xi_{k+1} \leq |\lambda_k - \lambda_*|\left\{2K_1\sin\xi_k + K_2|\lambda_k - \lambda_*|\right\}, \tag{4.28}$$

from inequality (4.23). This implies

$$\sin\xi_{k+1} \leq \tau_0\sin\xi_k\left\{2K_1 + 4KK_2\sin\eta_k\right\} \leq \tau_0 Q_0\sin\xi_k \leq \frac{1}{2}\sin\xi_k,$$

since $|\lambda_0 - \lambda_*| \leq \tau_0 \leq 1/2Q_0$, i.e., we have at least superlinear convergence. Inequality (4.28) coupled with bound (3.34) yields

$$\sin\xi_{k+1} \leq 4K\sin\xi_k\sin\eta_k\left\{2K_1\sin\xi_k + 4KK_2\sin\xi_k\sin\eta_k\right\}.$$

Hence, we have $\sin\xi_{k+1} \leq K_0^a\sin\xi_k^2\sin\eta_k$, with $K_0^a = 4K(2K_1 + 4KK_2\sin\eta_k)$.

All previous results can be equally obtained for the Newton step for the left eigenvector. The estimates hold, since the 2-norm is invariant under conjugate transposing, i.e., $\|\,.\,\| = \|(\,.\,)^H\|$. Hence, we obtain $\sin\eta_{k+1} \leq \tilde{K}_0^a\sin\eta_k^2\sin\xi_k$ with $\tilde{K}_0^a = 4K(2K_1 + 4KK_2\sin\xi_k)$, and

$$\sin\eta_{k+1} \leq \tau_0\sin\eta_k\left\{2K_1 + 4KK_2\sin\xi_k\right\} \leq \tau_0 Q_0\sin\eta_k \leq \frac{1}{2}\sin\eta_k.$$

In order to derive the bounds in norm distances we use (3.22) and (4.24), i.e., $\tan\xi_k = \|u_k - x_*^{u_k}\|$, to obtain $|\lambda_k - \lambda_*| \leq K\|u_k - x_*^{u_k}\|\|v_k - y_*^{v_k}\|$. Thus, inequality (4.15) immediately yields the result with $K_0^d = K(K_1 + KK_2\|v_k - y_*^{v_k}\|)$ and $\tilde{K}_0^d = K(K_1 + KK_2\|u_k - x_*^{u_k}\|)$.

The eigenvalue inequality (4.27) comes from inserting (4.28) in the bound for the Rayleigh functional $p(u_{k+1}, v_{k+1}) \leq K\tan\xi_{k+1}\tan\eta_{k+1}$, where the constants are given by $C_0^a = K_0^a\tilde{K}_0^a/(4K)$ and $C_0^d = 4K_0^d\tilde{K}_0^d/K$. In detail, inequality (4.28) implies that

$$|\lambda_{k+1} - \lambda_*| \leq C_0^a|\lambda_k - \lambda_*|^2\sin\xi_k\sin\eta_k \leq 4K\left(\frac{1}{2}\right)^{2^{k+1}}.$$

The inequalities (4.25), (4.26) and (4.27) provide Q-superquadratic convergence of the sequences $\{\xi_k\}$, $\{\eta_k\}$ and $\{\lambda_k\}$. To show the R-order 3, we analyze the corresponding matrix

$$M = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 1 & 2 \end{bmatrix}$$

of indices of $\sin \xi_k$, $\sin \eta_k$ and $|\lambda_k - \lambda_*|$ on the right hand sides of (4.25), (4.26) and (4.27). The largest eigenvalue of M is $\lambda_* = 3$, the corresponding eigenvector of which is positive, which proves the assertion. □

Algorithm 4 shows the natural generalization of the two-sided Rayleigh quotient iteration [66] to nonlinear problems. Moreover, the locally cubic convergence of the two-sided Rayleigh quotient iteration applied to linear problems carries over to the nonlinear formulation with Rayleigh functional update. Theorem 4.13 also implies the locally cubic convergence of the Rayleigh functional iteration for Hermitian problems with real eigenvalues, where $\lambda_k = p(u_k)$ is used, since $x_* = y_*$, which was shown in [71].

As already mentioned, a version of Algorithm 4, that computes corrections instead of updated approximations, could yield better results, because of less rounding errors. Therefore, we will give an appropriate algorithm using bordered matrices. Since we have to solve a dual system for the left eigenvector, e.g.,

$$\begin{bmatrix} T(\lambda)^H & \dot{T}(\lambda)^H v \\ \tilde{w}^H & 0 \end{bmatrix} \begin{bmatrix} v_+ \\ \nu \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

with some normalization vector $\tilde{w}$, we would like to use $C(\lambda, \dot{T}(\lambda)u, w)^H$, in order to save the second factorization when solving two linear systems. Therefore, we change the normalization conditions for both vectors to $w = \dot{T}(\lambda)^H v$ and $\tilde{w} = \dot{T}(\lambda)u$. The new matrix is given by

$$C(\lambda, \dot{T}(\lambda)u, \dot{T}(\lambda)^H v) := \begin{bmatrix} T(\lambda) & \dot{T}(\lambda)u \\ v^H \dot{T}(\lambda) & 0 \end{bmatrix}, \tag{4.29}$$

the adjoint of which characterizes the second linear system of Algorithm 4, which can be reformulated as presented in Algorithm 5.

Now, the matrix $C(\lambda_*, \dot{T}(\lambda_*)u, \dot{T}(\lambda_*)^H v)$ is nonsingular, if $y_*^H \dot{T}(\lambda_*)u \neq 0$ and $v^H \dot{T}(\lambda_*)x_* \neq 0$, which holds for sufficiently good approximations (u, v) to right and left eigenvectors corresponding to an algebraically simple eigenvalue λ_*. The result of Theorem 4.13 is obtained in the same way, only the constants differ slightly.

If we replace the Rayleigh functional computation in S3 of Algorithm 4 by the generalized Rayleigh quotient $\lambda_{k+1} = p_L(\lambda_k, u_{k+1}, v_{k+1}) = \lambda_k - \frac{v_{k+1}^H T(\lambda_k) u_{k+1}}{v_{k+1}^H \dot{T}(\lambda_k) u_{k+1}}$, then (4.23) becomes

$$\sin \xi_{k+1} \leq \left(\frac{4\|T(\lambda_*)\|}{|\alpha|} \tan \xi_k \tan \eta_k + \frac{2L}{|\alpha|} \frac{|\lambda_{k-1} - \lambda_*|^2}{\cos \xi_k \cos \eta_k} \right) \{K_1 \tan \xi_k + K_2 |\lambda_k - \lambda_*|\},$$

using (3.65). Hence, the cubic convergence is lost.

Algorithm 5 *Two-sided Rayleigh functional iteration (bordered version)*

Input: (λ_0, u_0, v_0) *where* $u_0^H u_0 = v_0^H v_0 = 1$

for $k = 0, 1, 2, ...$

S1: *Set* $C_k = C(\lambda_k, \dot{T}(\lambda_k) u_k, \dot{T}(\lambda_k)^H v_k) = \begin{bmatrix} T(\lambda_k) & \dot{T}(\lambda_k) u_k \\ v_k^H \dot{T}(\lambda_k) & 0 \end{bmatrix}$

S2: *Solve* $C_k \begin{bmatrix} s_k \\ \mu_k \end{bmatrix} = - \begin{bmatrix} T(\lambda_k) u_k \\ 0 \end{bmatrix}$, *set* $u_{k+1} = \dfrac{u_k + s_k}{\|u_k + s_k\|}$

S3: *Solve* $C_k^H \begin{bmatrix} t_k \\ \nu_k \end{bmatrix} = - \begin{bmatrix} T(\lambda_k)^H v_k \\ 0 \end{bmatrix}$, *set* $v_{k+1} = \dfrac{v_k + t_k}{\|v_k + t_k\|}$

S4: *Solve* $v_{k+1}^H T(\lambda_{k+1}) u_{k+1} = 0$ *for* λ_{k+1}

Since we have seen in Lemma 2.8 that applying $T(\lambda)^{-1}$ to any vector is a step in the direction of the eigenvector, we propose a simplified version of the two-sided Rayleigh functional iteration, which avoids the derivative in the linear systems. In some cases this might save matrix-vector multiplications. Quadratic convergence is proved, but the assumptions on the vectors change. We formulate the algorithm in the bordered matrix presentation, see Algorithm 6.

The following convergence analysis is obtained differently from the one for Algorithm 4, since the matrix $C(\lambda, u, v)$ no longer represents the Jacobian in the Newton step for the extended system (4.1). We will consider the bordered matrix formulation that is equivalent to S2 in Algorithm 6.

$$C(\lambda, u, v) \begin{bmatrix} x \\ \mu \end{bmatrix} := \begin{bmatrix} T(\lambda) & u \\ v^H & 0 \end{bmatrix} \begin{bmatrix} x \\ \mu \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \iff \begin{array}{r} T(\lambda)x + \mu u = 0 \\ v^H x = 1. \end{array} \tag{4.30}$$

Lemma 4.5 immediately provides the nonsingularity of $C(\lambda_*, u, v)$ if $y_*^H u \neq 0$ and $v^H x_* \neq 0$. Vectors that satisfy these restrictions can be described by $\mathcal{K}_\varepsilon$, see (3.9), in a cross-over way, i.e., $C(\lambda_*, u, v)$ is nonsingular if $u \in \mathcal{K}_\varepsilon(y_*)$ and $v \in \mathcal{K}_\varepsilon(x_*)$, cf. condition (4.11).

Algorithm 6 *Simplified (two-sided) Rayleigh functional iteration (sRFI)*

Input: (λ_0, u_0, v_0) *where* $u_0^H u_0 = v_0^H v_0 = 1$

for $k = 0, 1, 2, ...$

S1: *Set* $C_k = C(\lambda_k, u_k, v_k) = \begin{bmatrix} T(\lambda_k) & u_k \\ v_k^H & 0 \end{bmatrix}$

S2: *Solve* $C_k \begin{bmatrix} s_k \\ \mu_k \end{bmatrix} = - \begin{bmatrix} T(\lambda_k)u_k \\ 0 \end{bmatrix}$, *set* $u_{k+1} = \dfrac{u_k + s_k}{\|u_k + s_k\|}$

S3: *Solve* $C_k^H \begin{bmatrix} t_k \\ \nu_k \end{bmatrix} = - \begin{bmatrix} T(\lambda_k)^H v_k \\ 0 \end{bmatrix}$, *set* $v_{k+1} = \dfrac{v_k + t_k}{\|v_k + t_k\|}$

S4: *Solve* $v_{k+1}^H T(\lambda_{k+1}) u_{k+1} = 0$ *for* λ_{k+1}

Corollary 4.14 *Assume* $(\mathcal{A}_{\mathbb{C}})$. *Then, for all* ε_1 *with* $0 < \varepsilon_1 < \pi/2$, *there exists a radius* τ_0, $0 < \tau_0 \leq \tau_*$, *such that for all* $(u, v) \in \mathcal{K}_{\varepsilon_1}(y_*) \times \mathcal{K}_{\varepsilon_1}(x_*)$ *with* $\|u\| \leq \delta_0$, $\|v\| \leq \delta_0$, $\delta_0 > 0$, *and all* $\lambda, \mu \in S_0 := \bar{S}(\lambda_*, \tau_0)$ *we have*

(i) $C(\lambda, u, v)$ *is nonsingular and its inverse is bounded by*

$$\|C(\lambda, u, v)^{-1}\| \leq \frac{M_2^1}{1 - M_1 M_2^1 |\lambda - \lambda_*|} \leq 2M_2^1, \tag{4.31}$$

with

$$M_2^1 := \frac{\sqrt{\|\Sigma_1^{-1}\|\delta_0^2 + 2}}{\delta_0 \cos^2 \varepsilon_1}. \tag{4.32}$$

(ii) $C(\,\cdot\,, u, v)^{-1}$ *is Lipschitz continuous, i.e.,*

$$\|C(\lambda, u, v)^{-1} - C(\mu, u, v)^{-1}\| \leq \; L_{C^{-1}} |\lambda - \mu|,$$

where $L_{C^{-1}} = 4M_1 (M_2^1)^2$.

Proof. Lemma 4.5 provides the nonsingularity of $C(\lambda_*, u, v)$ for all $(u, v) \in \mathcal{K}_{\varepsilon_1}(y_*) \times \mathcal{K}_{\varepsilon_1}(x_*)$. Applying inequality (4.12) yields

$$\|C(\lambda_*, u, v)^{-1}\| \leq \frac{\sqrt{\|\Sigma_1^{-1}\|^2 \|u\|^2 \|v\|^2 + |y_*^H u|^2 + |v^H x_*|^2}}{|y_*^H u||v^H x_*|} = M_2(u, v).$$

Since $\|u\| \leq \delta_0$, $\|v\| \leq \delta_0$, and $|y_*^H u| \geq \delta_0 \cos \varepsilon_0$, $|v^H x_*| \geq \delta_0 \cos \varepsilon_0$, we have $\|C(\lambda_*, u, v)^{-1}\| \leq M_2^1$, with M_2^1 defined in (4.32).

Set

$$\tau_0 := \min\left\{\tau_*, \frac{1}{2M_1M_2^1}\right\},$$

and consider

$$\delta C := C(\lambda,u,v) - C(\lambda_*,u,v) = \begin{bmatrix} T(\lambda)-T(\lambda_*) & 0 \\ 0 & 0 \end{bmatrix}.$$

The perturbation lemma [46, p. 128] yields the nonsingularity of $C(\lambda,u,v)$ and the bound $\|C(\lambda,u,v)^{-1}\| \le \frac{1}{1-\kappa}\|C(\lambda_*,u,v)^{-1}\|$, if $\|\delta C\|\|C(\lambda_*,u,v)^{-1}\| \le \kappa < 1$. Assuming that $|\lambda-\lambda_*| \le \tau_0$, we obtain

$$\|\delta C\|\|C(\lambda_*,u,v)^{-1}\| \le \sqrt{\|T(\lambda)-T(\lambda_*)\|^2}\|C(\lambda_*,u,v)^{-1}\| \le M_1M_2^1|\lambda-\lambda_*| =: \kappa \le \frac{1}{2} < 1,$$

which implies (4.31), i.e., assertion (i).

Moreover, for all $\lambda,\mu \in S_0$ and fixed u, v one obtains

$$\begin{aligned}\|C(\lambda)^{-1} - C(\mu)^{-1}\| &= \|C(\lambda)^{-1}\left[C(\lambda)-C(\mu)\right]C(\mu)^{-1}\| \\ &\le \|C(\lambda)^{-1}\|\|C(\mu)^{-1}\|\|C(\lambda)-C(\mu)\| \\ &\le 4M_1(M_2^1)^2|\lambda-\mu|,\end{aligned}$$

which proves the Lipschitz continuity of $C(\cdot,u,v)^{-1}$. □

Under the assumptions of Corollary 4.14, equation (4.30) has the unique solution

$$\hat{x}(\lambda) := \begin{bmatrix} x(\lambda) \\ \mu(\lambda) \end{bmatrix} = C(\lambda,u,v)^{-1}e_{n+1}.$$

In particular, setting $\lambda = \lambda_*$ yields

$$\hat{u}_* := \begin{bmatrix} u_* \\ \mu_* \end{bmatrix} := C(\lambda_*,u,v)^{-1}e_{n+1} = \begin{bmatrix} x_*/v^Hx_* \\ 0 \end{bmatrix}. \tag{4.33}$$

Lemma 4.15 *Assume* $(\mathcal{A}_{\mathbb{C}})$. *Then, for all* ε_1 *with* $0 < \varepsilon_1 < \pi/2$, *there exists a radius* τ_0, $0 < \tau_0 \le \tau_*$, *such that for all* $(u,v) \in \mathcal{K}_{\varepsilon_1}(y_*) \times \mathcal{K}_{\varepsilon_1}(x_*)$ *we have*

$$\sin\xi_+ \;\le\; \frac{2M_1M_2^1}{|x_*^Hv|}\,|\lambda-\lambda_*|, \tag{4.34}$$

where $\xi_+ := \xi_+(\lambda) = \measuredangle(\mathrm{span}\{x(\lambda)\}, \mathrm{span}\{x_*\})$.

Proof. Since the assumptions of Corollary 4.14 are satisfied, the bound (4.31) holds. Setting $x = u_* + \delta u$ and using (4.33) yields

$$\begin{aligned}\|\delta u\| &= \|x - u_*\| \leq \|\hat{x} - \hat{u}_*\| = \|C(\lambda)^{-1} e_{n+1} - C(\lambda_*)^{-1} e_{n+1}\| \\ &= \|C(\lambda)^{-1} \left[C(\lambda) - C(\lambda_*)\right] C(\lambda_*)^{-1} e_{n+1}\| \leq \|C(\lambda)^{-1}\| \|C(\lambda) - C(\lambda_*)\| \|\hat{u}_*\| \\ &\leq \frac{1}{|x_*^H v|} 2 M_1 M_2^1 |\lambda - \lambda_*| =: \zeta,\end{aligned}$$

with $C(\lambda) \equiv C(\lambda, u, v)$. Moreover, we have

$$\|x\|^2 = x^H x = \frac{1}{|x_*^H v|^2} + 2\mathrm{Re}\left(\frac{x_*^H \delta u}{x_*^H v}\right) + \|\delta u\|^2 \geq 1,$$

and because of $x_*^H x = x_*^H (u_* + \delta u) = x_*^H u_* + x_*^H \delta u = \frac{1}{v^H x_*} + x_*^H \delta u$, we also have

$$\begin{aligned}|x_*^H x|^2 &= (x_*^H x)(\overline{x_*^H x}) = \left(\frac{1}{v^H x_*} + x_*^H \delta u\right)\left(\frac{1}{x_*^H v} + \overline{x_*^H \delta u}\right) \\ &= \frac{1}{|x_*^H v|^2} + 2\mathrm{Re}\left(\frac{x_*^H \delta u}{x_*^H v}\right) + |x_*^H \delta u|^2.\end{aligned}$$

Thus, we end up with

$$\begin{aligned}\cos^2 \xi_+ = \left[\frac{|x_*^H x|}{\|x_*\| \|x\|}\right]^2 &= \frac{\frac{1}{|x_*^H v|^2} + 2\mathrm{Re}\left(\frac{x_*^H \delta u}{x_*^H v}\right) + |x_*^H \delta u|^2 + \|\delta u\|^2 - \|\delta u\|^2}{\frac{1}{|x_*^H v|^2} + 2\mathrm{Re}\left(\frac{x_*^H \delta u}{x_*^H v}\right) + \|\delta u\|^2} \\ &= 1 - \frac{\|\delta u\|^2 - |x_*^H \delta u|^2}{\|x\|^2} =: 1 - \omega,\end{aligned}$$

and eventually $\sin^2 \xi_+ = 1 - \cos^2 \xi_+ = \omega = \frac{\|\delta u\|^2 - |x_*^H \delta u|^2}{\|x\|^2} \leq \left(\frac{\|\delta u\|}{\|x\|}\right)^2 \leq \zeta^2$, which proves the assertion. □

Again, similar results may be generated for the angle between left eigenvector and its approximation, applying the techniques to $T(\lambda)^H$. In combination with Theorem 3.5 we end up with the following convergence theorem.

Theorem 4.16 *Let $D \subset \mathbb{C}$ be an open set and let $T : D \to \mathbb{C}^{n \times n}$ be holomorphic in D. Let $\lambda_* \in D$ be an algebraically simple eigenvalue of T, and let $\tau_* > 0$ be such that $S_* := \{\lambda \in \mathbb{C} : |\lambda - \lambda_*| \leq \tau_*\} \subset D$. Then for every ε_0, $0 < \varepsilon_0 < \pi/3$, there exists a radius τ_0, $0 < \tau_0 \leq \tau_*$ and constants $K_1^a > 0$, $C_1 > 0$ with $C_1 \tau_0 \leq 1/2$, such that the simplified Rayleigh quotient iteration of Algorithm 6 is well-defined for every initial triplet (λ_0, u_0, v_0) with $(u_0, v_0) \in \mathcal{K}_{\varepsilon_0}(y_*) \times \mathcal{K}_{\varepsilon_0}(x_*)$, $\lambda_0 \in S_0 = \bar{S}(\lambda_*, \tau_0)$, and converges in the following sense*

$$\lim_{k \to \infty} \lambda_k = \lambda_*, \quad \lim_{k \to \infty} \xi_k = \lim_{k \to \infty} \eta_k = 0,$$

where $\xi_k = \measuredangle(\operatorname{span}\{u_k\}, \operatorname{span}\{x_*\})$, $\eta_k = \measuredangle(\operatorname{span}\{v_k\}, \operatorname{span}\{y_*\})$. *The rate of convergence is characterized by*

$$|\lambda_{k+1} - \lambda_*| \leq C_1 |\lambda_k - \lambda_*|^2 \tag{4.35}$$

and

$$\left.\begin{matrix} \sin\xi_{k+1} \\ \sin\eta_{k+1} \end{matrix}\right\} \leq K_1^a \sin\eta_k \sin\xi_k. \tag{4.36}$$

Proof. Take τ as defined in the proof of Corollary 4.14. Requiring that $\varepsilon_0 < \pi/3$ implies $\sin\xi \leq \sqrt{3}/2$, hence $\tan\xi \leq 2\sin\xi$. Lemma 4.15 and Theorem 3.5, in particular (3.34) yield

$$\sin\xi_{k+1} \leq \frac{2M_1M_2^1}{|b_2|}\,|\lambda_k - \lambda_*| \leq \frac{64M_1M_2^1\|T(\lambda_*)\|}{3|b_2||\alpha|}\,\sin\xi_k \sin\eta_k,$$

and, therefore,

$$|\lambda_{k+1} - \lambda_*| \leq \frac{32\|T(\lambda_*)\|}{3|\alpha|}\sin\xi_{k+1}\sin\eta_{k+1} \leq \frac{128(M_1M_2^1)^2\|T(\lambda_*)\|}{3|b_2|^2|\alpha|}\,|\lambda_k - \lambda_*|^2.$$

Setting $C_1 := \frac{128(M_1M_2^1)^2\|T(\lambda_*)\|}{3|b_2|^2|\alpha|}$ and $K_1^a := \frac{64M_1M_2^1\|T(\lambda_*)\|}{3|b_2||\alpha|}$ gives (4.35) and (4.36). □

We have analyzed the two-sided Rayleigh functional iteration and a simplified version. The cubic convergence of the two-sided Rayleigh functional iteration is lost, when we neglect the derivative in the inverse iteration step. Furthermore, the imposed assumptions are different. The difference lies in the domain of admissible vectors. For the two-sided Rayleigh functional iteration, we assumed that u, v satisfy

$$|v^H\dot{T}(\lambda_*)x_*| > 0 \quad \text{and} \quad |u^H\dot{T}(\lambda_*)^H y_*| > 0,$$

which is related to the simplicity condition of the eigenvalue, in that for vectors u, v close to x_*, y_* corresponding to a simple eigenvalue, the inequalities are satisfied. A completely different assumption is needed for the proof of convergence of the simplified Rayleigh functional iteration, namely $(u, v) \in \mathcal{K}_{\varepsilon_0}(y_*) \times \mathcal{K}_{\varepsilon_0}(x_*)$, i.e.,

$$|v^H x_*| > 0 \quad \text{and} \quad |u^H y_*| > 0,$$

which is not naturally given in the neighborhood of eigenvectors corresponding to a simple eigenvalue, see the example in §2.1. However, this may be an advantage if an ill-conditoned eigenvalue is sought, where $y_*^H\dot{T}(\lambda_*)x_*$ is close to zero.

Note that if $T(\lambda) = A - \lambda I$, then Algorithms 4 and 5 are equivalent to the two-sided Rayleigh quotient iteration and both convergence locally cubically.

4.2.2 Two-sided Residual Inverse Iteration

If applied to a Hermitian eigenproblem with real eigenvalues, respectively a real symmetric eigenproblem, the residual inverse iteration [61] in Algorithm 2 was shown to converge with R-order 2 towards the eigenvector corresponding to the simple eigenvalue λ_*, where the eigenvalue update λ_{k+1} is retrieved as the nonlinear Rayleigh functional, which is a solution of $u_{k+1}^H T(\lambda_{k+1}) u_{k+1} = 0$, see §4.1.

Since we have derived techniques and knowledge regarding the two-sided Rayleigh functional in Chapter 3, it is straightforward to formulate a generalization of the residual inverse iteration method, which computes eigentriples. Then, we solve a system for left eigenvector approximations additionally, simultaneously to the system for the right eigenvector approximations. The new eigenvalue approximation is obtained as two-sided Rayleigh functional.

Algorithm 7 *Two-sided residual inverse iteration*

Input: (λ_0, u_0, v_0), *normalization vectors* w_x, w_y, *such that* $w_x^H u_0 = 1$, $w_y^H v_0 = 1$

for $k = 0, 1, 2, ...$

S1: *Solve* $v_k^H T(\lambda_{k+1}) u_k = 0$ *for* λ_{k+1}

S2: *Compute the residuals* $r_k^x = T(\lambda_{k+1}) u_k, \quad r_k^y = T(\lambda_{k+1})^H v_k$

S3: *Solve* $T(\lambda_0) s_k = r_k^x$ *for* s_k

S4: *Solve* $T(\lambda_0)^H t_k = r_k^y$ *for* t_k

S5: *Set* $x_{k+1} = u_k - s_k, \quad u_{k+1} = x_{k+1} / w_x^H x_{k+1}$

$y_{k+1} = v_k - t_k, \quad v_{k+1} = y_{k+1} / w_y^H y_{k+1}$

The two-sided, resp. generalized, residual inverse iteration converges as follows.

Theorem 4.17 *Let* λ_* *be a simple eigenvalue of* $T(\lambda)$, *and suppose that* $T(\lambda)$ *is twice continuously differentiable,* x_* *is the corresponding right eigenvector normalized by* $w_x^H x_* = 1$, y_* *is the corresponding left eigenvector normalized by* $w_y^H y_* = 1$. *Then the two-sided residual inverse iteration converges for all* λ_0 *sufficiently close to* λ_*, *if*

$$\xi_0 = \measuredangle(\mathrm{span}\{u_0\}, \mathrm{span}\{x_*\}) \leq \pi/3, \quad \eta_0 = \measuredangle(\mathrm{span}\{v_0\}, \mathrm{span}\{y_*\}) \leq \pi/3,$$

and we have

$$\frac{\|u_{k+1} - x_*\|}{\|u_k - x_*\|} = \mathcal{O}(|\lambda_0 - \lambda_*|), \quad \frac{\|v_{k+1} - y_*\|}{\|v_k - y_*\|} = \mathcal{O}(|\lambda_0 - \lambda_*|), \tag{4.37}$$

$$|\lambda_{k+1} - \lambda_*| = \mathcal{O}(\|u_k - x_*\| \|v_k - y_*\|). \tag{4.38}$$

Moreover, if the shift λ_0 is updated by λ_k in every iteration, then the R-order of the two-sided residual inverse iteration method is at least 2, and we have

$$\|u_{k+1} - x_*\| = \mathcal{O}(\|u_k - x_*\|\|u_{k-1} - x_*\|\|v_{k-1} - y_*\|), \tag{4.39}$$
$$\|v_{k+1} - y_*\| = \mathcal{O}(\|v_k - y_*\|\|u_{k-1} - x_*\|\|v_{k-1} - y_*\|). \tag{4.40}$$

Proof. See [61] for the proof of the equation (4.37), which implies (4.38), since u and v do not depend on each other, when we take the algorithm applied to $T(\lambda)^H$. The bound for the eigenvalue $\lambda_{k+1} = p(u_k, v_k)$ has been shown in Theorem 3.5, where the normwise bound (3.35) holds if the initial angles ξ_0, η_0 are smaller than $\pi/3$. If the shift is regularly updated by λ_k, then we have $|\lambda_k - \lambda_*| = \mathcal{O}(\|u_{k-1} - x_*\|\|v_{k-1} - y_*\|)$, which implies (4.39) and (4.40).

In order to show the statement on the R-order of convergence we define the terms $\gamma = |\lambda_k - \lambda_*|$,$\gamma_+ = |\lambda_{k+1} - \lambda_*|$, $\delta = \|u_k - x_*\|$, $\delta_+ = \|u_{k+1} - x_*\|$, $\varepsilon = \|v_k - y_*\|$, $\varepsilon_+ = \|v_{k+1} - y_*\|$. Then the iterates in the two-sided residual inverse iteration method satisfy the following inequalities

$$\begin{aligned} \gamma_+ &\leq 4K\delta\varepsilon, \\ \delta_+ &\leq K_2^d\gamma\delta, \\ \varepsilon_+ &\leq K_2^d\gamma\varepsilon, \end{aligned}$$

with $K = \frac{8\|T(\lambda_*)\|}{3|y_*^H \dot{T}(\lambda_*)x_*|}$, $K_2^d > 0$. The matrix of exponents of γ, δ, ε is given by

$$\begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix},$$

the largest eigenvalue of which equals 2, where the corresponding eigenvector is positive in all components. This means, as shown in [74], that the R-order of the two-sided residual inverse iteration method is at least 2. □

If the shift λ_0 is not updated, then the two-sided residual inverse iteration has the same R-order of convergence as the (one-sided) but constantly updated residual inverse iteration, i.e., the R-order is $(1 + \sqrt{5})/2$ for general T, since in this case the sequences are in some sense uncoupled and the new information is not used.

Since we do not need an extra factorization when we solve the system corresponding to the left eigenvector in S4 of Algorithm 7, there is not too much extra cost. Note that a convergence enhancement in the vectors regarding the (one-sided) residual inverse iteration can only be achieved when the shift λ_0 is updated once in a while. Otherwise the vectors will converge linearly.

4.2.3 Alternating Rayleigh Functional Iteration

PARLETT [67] proposed an alternating Rayleigh quotient iteration which converges globally but only linearly. Transfering this idea to nonlinear eigenproblems means that only one vector is iterated by applying alternatingly $T(\lambda)$ and $T(\lambda)^H$ on the iterates. Thus, the algorithm stops either with an approximated left or an approximated right eigenvector and the missing one can be obtained by computing at least one more correction. In fact, this method belongs to the class of one-sided (one-vector) methods of §4.1. The key difference to the two-sided methods is that it uses only one vector for the iteration. However, we define it in a way, such that two systems per outer iteration are solved, and, the extra cost in order to obtain the remaining eigenvector is the same as for other two-sided methods, i.e., to solve at least one linear system. Algorithm 8 pesents the nonlinear alternating Rayleigh functional iteration.

Algorithm 8 *Alternating Rayleigh functional iteration (ARFI)*

Input: (λ_0, u_0) *where* $u_0^H u_0 = 1$

for $k = 0, 2, 4, \ldots$

S1: *Solve* $T(\lambda_k)x_{k+1} = \dot{T}(\lambda_k)u_k$, *set* $u_{k+1} = x_{k+1}/\|x_{k+1}\|$

S2: *Solve* $u_{k+1}^H T(\lambda_{k+1})u_{k+1} = 0$ *for* λ_{k+1}

S3: *Solve* $T(\lambda_{k+1})^H x_{k+2} = \dot{T}(\lambda_{k+1})^H u_{k+1}$, *set* $u_{k+2} = x_{k+2}/\|x_{k+2}\|$

S4: *Solve* $u_{k+2}^H T(\lambda_{k+2})u_{k+2} = 0$ *for* λ_{k+2}

Local existence of the complex Rayleigh functional $p(u)$ defined by $u^H T(p(u))u = 0$ and an inequality for $|p(u) - \lambda_*|$ was shown in Section 3.3 for general $T(\lambda)$. Notice that PARLETT's proof of convergence cannot be adapted, since the minimal residual property of the Rayleigh quotient (3.4) does not hold for the Rayleigh functional, in general, cf. Chapter 3.

In [37] the idea of the alternating Rayleigh quotient iteration is applied to the framework of the Jacobi–Davidson method which gives the so-called alternating Jacobi–Davidson method. This method is more robust with respect to nonnormal problems than the standard Jacobi–Davidson method, cf. Chapter 6.

Though Rayleigh functional iterations can only be proved to converge locally, in practice they converge almost always—with the constraint, that one might not be able to predict the region of the outcome. It was shown in [8] that the set of points for which the Rayleigh quotient iteration diverges for real symmetric matrices is of measure zero. However, in [14] it was observed that for nonnormal matrices the Rayleigh quotient iteration can behave

chaotically. The alternating Rayleigh functional iteration may behave better in some examples than the inverse iteration or Rayleigh functional iteration, but the (expected) linear convergence makes it a less favorable method, cf. Section 4.5 for examples.

4.2.4 Generalized Rayleigh Functional Iteration

In [78] a generalized Rayleigh quotient iteration (GRQI) for nonlinear and nonnormal eigenvalue problems was introduced, which is based on singularity theory and a former algorithm for linear eigenproblems, see [77]. In the nonlinear method the two linear systems

$$\begin{bmatrix} T(\lambda) & v \\ u^H & 0 \end{bmatrix} \begin{bmatrix} s \\ \mu \end{bmatrix} = \begin{bmatrix} -T(\lambda)u \\ 0 \end{bmatrix}, \qquad \begin{bmatrix} T(\lambda)^H & u \\ v^H & 0 \end{bmatrix} \begin{bmatrix} t \\ \nu \end{bmatrix} = \begin{bmatrix} -T(\lambda)^H v \\ 0 \end{bmatrix} \tag{4.41}$$

are solved, where s and t are corrections for the right and left eigenvector approximations u and v. The scalar μ is viewed as unfolding parameter, depending on λ, i.e., $\mu \equiv \mu(\lambda)$. A Newton step for μ leads to the update formula

$$\lambda_{k+1} = \lambda_k - \frac{\mu_k}{(v_k + t_k)^H \dot{T}(\lambda_k)(u_k + s_k)}, \tag{4.42}$$

in terms of the actual iterates in the k-th iteration step. Locally quadratic convergence is shown in [78]. The method is presented in Algorithm 9.

Algorithm 9 *Generalized Rayleigh quotient iteration (GRQI)*

Input: (λ_0, u_0, v_0) *where* $u_0^H u_0 = v_0^H v_0 = 1$

for $k = 0, 1, 2, \ldots$

S1: *Set* $C_k = C_k(\lambda_k, v_k, u_k) = \begin{bmatrix} T(\lambda_k) & v_k \\ {u_k}^H & 0 \end{bmatrix}$

S2: *Solve* $C_k \begin{bmatrix} s_k \\ \mu_k \end{bmatrix} = - \begin{bmatrix} T(\lambda_k)u_k \\ 0 \end{bmatrix}$, *set* $u_{k+1} = \dfrac{u_k + s_k}{\|u_k + s_k\|}$

S3: *Solve* $C_k^H \begin{bmatrix} t_k \\ \bar{\nu}_k \end{bmatrix} = - \begin{bmatrix} T(\lambda_k)^H v_k \\ 0 \end{bmatrix}$, *set* $v_{k+1} = \dfrac{v_k + t_k}{\|v_k + t_k\|}$

S4: *Compute* $\lambda_{k+1} = \lambda_k - \dfrac{\mu_k}{(v_k + t_k)^H \dot{T}(\lambda_k)(u_k + s_k)}$

But what was the motivation behind this counterintuitive bordering? The generalized Rayleigh quotient iteration was constructed in order to minimize the asymptotic condition

number of the matrix $C_k \approx C(\lambda_*)$ close to the solution λ_*. To see this, consider the (first) linear system of the two-sided Rayleigh functional iteration, or equivalently, of the inverse iteration. The normalization condition plays an arbitrary role, so let

$$Z_* := \begin{bmatrix} T(\lambda_*) & \dot{T}(\lambda_*)x_* \\ w^H & 0 \end{bmatrix}$$

be the matrix defining the linear system at the solution (λ_*, x_*, y_*), where w is the normalization vector for x. With the singular value decomposition (2.12) for $T(\lambda_*)$, we define the matrix Z by

$$Z := \begin{bmatrix} Y^H & 0 \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} T(\lambda_*) & \dot{T}(\lambda_*)x_* \\ w^H & 0 \end{bmatrix} \begin{bmatrix} X & 0 \\ 0^T & 1 \end{bmatrix} = \begin{bmatrix} \Sigma_1 & 0 & Y_1^H \dot{T}(\lambda_*)x_* \\ 0^T & 0 & y_*^H \dot{T}(\lambda_*)x_* \\ w^H X_1 & w^H x_* & 0 \end{bmatrix},$$

which has the same singular values as Z_*. Replacing $y_*^H \dot{T}(\lambda_*)x_*$, which is nonzero since λ_* is assumed to be simple, in the n-th row by 0, yields the singular matrix

$$\hat{Z} := \begin{bmatrix} \Sigma_1 & 0 & Y_1^H \dot{T}(\lambda_*)x_* \\ 0^T & 0 & 0 \\ w^H X_1 & w^H x_* & 0 \end{bmatrix},$$

i.e., the smallest singular value $\hat{\sigma}_{n+1}$ of $\hat{Z}$ is zero. Perturbation theory of singular values, see [46], gives

$$\sigma_{n+1} = |\sigma_{n+1} - \hat{\sigma}_{n+1}| \leq \|Z - \hat{Z}\| = |y_*^H \dot{T}(\lambda_*)x_*|,$$

where σ_{n+1} is the smallest singular value of Z, and therefore also of Z_*. On the other hand, we have also $\sigma_{n+1} \leq |w^H x_*|$. The definition of the spectral norm then yields

$$\left\| \begin{bmatrix} T(\lambda_*) & \dot{T}(\lambda_*)x_* \\ w^H & 0 \end{bmatrix}^{-1} \right\| \geq \frac{1}{\min\{|y_*^H \dot{T}(\lambda_*)x_*|, |w^H x_*|\}}. \tag{4.43}$$

Hence, we expect numerical problems for the computation of ill-conditioned eigenvalues, i.e., when α becomes small, cf. §2.3. The reason for this dependency lies in the upper left vector of Z_*, which for ill-conditioned eigenvalues is almost orthogonal to the missing direction $\operatorname{span}\{y_*\}$ in the range $\operatorname{im} T(\lambda_*) = \operatorname{span}\{y_*\}^\perp$. Therefore, the left eigenvector y_* is the vector of choice for the upper left corner of the matrix Z_*, in order to circumvent the possible growth of the inverse Jacobian.

As was shown in [78], the matrices used in the generalized Rayleigh quotient iteration in S2 and S3 of Algorithm 9 do not depend on $\alpha = y_*^H \dot{T}(\lambda_*)x_*$, i.e., the generalized Rayleigh quotient iteration is not sensitive with respect to the eigenvalue.

However, the algebraic simplicity condition is necessary, since the matrices in (4.41) would be singular at λ_* otherwise.

A new variant of the generalized Rayleigh quotient iteration is shown in Algorithm 10. The same systems are solved, and the difference to Algorithm 9 is the computation of the Rayleigh functional in S4 instead of the Newton step for $\mu(\lambda)$. Both algorithms solve the second system with the conjugate transposed matrix of the first system, which allows numerical advantages compared to the case with two different matrices.

Algorithm 10 *Generalized Rayleigh functional iteration (GRFI)*

Input: (λ_0, u_0, v_0) *where* $u_0^H u_0 = v_0^H v_0 = 1$

for $k = 0, 1, 2, ...$

S1: *Set* $C_k = C(\lambda_k, v_k, u_k) = \begin{bmatrix} T(\lambda_k) & v_k \\ {u_k}^H & 0 \end{bmatrix}$

S2: *Solve* $C_k \begin{bmatrix} s_k \\ \mu_k \end{bmatrix} = - \begin{bmatrix} T(\lambda_k) u_k \\ 0 \end{bmatrix}$, *set* $u_{k+1} = \dfrac{u_k + s_k}{\|u_k + s_k\|}$

S3: *Solve* $C_k^H \begin{bmatrix} t_k \\ \bar{\nu}_k \end{bmatrix} = - \begin{bmatrix} T(\lambda_k)^H v_k \\ 0 \end{bmatrix}$, *set* $v_{k+1} = \dfrac{v_k + t_k}{\|v_k + t_k\|}$

S4: *Solve* ${v_{k+1}}^H T(\lambda_{k+1}) u_{k+1} = 0$ *for* λ_{k+1}

Convergence is of quadratic order, as the following theorem shows.

Theorem 4.18 *Let $D \subset \mathbb{C}$ be an open set and let $T : D \to \mathbb{C}^{n\times n}$ be holomorphic on D. Let $\lambda_* \in D$ be an algebraically simple eigenvalue of T, and let $\tau_* > 0$ be such that $S_* := \{\lambda \in \mathbb{C} : |\lambda - \lambda_*| \le \tau_*\} \subset D$. Then for every ε_0, $0 < \varepsilon_0 < \pi/3$, there exists a radius $\tau_0 > 0$ and constants $C_3 > 0$, $K_3^a > 0$ with $C_3\tau_0 \le 1/2$, such that the generalized Rayleigh functional iteration in Algorithm 10 is well-defined for every initial triplet (λ_0, u_0, v_0) with $(u_0, v_0) \in \mathcal{K}_{x_*}(\varepsilon_0) \times \mathcal{K}_{y_*}(\varepsilon_0)$, $\lambda_0 \in \bar{S}(\lambda_*, \tau_0)$, and converges in the following sense*

$$\lim_{k\to\infty} \lambda_k = \lambda_*, \quad \lim_{k\to\infty} \xi_k = \lim_{k\to\infty} \eta_k = 0,$$

where $\xi_k = \measuredangle(\mathrm{span}\{u_k\}, \mathrm{span}\{x_*\})$, $\eta_k = \measuredangle(\mathrm{span}\{v_k\}, \mathrm{span}\{y_*\})$. The rate of convergence is characterized by

$$|\lambda_{k+1} - \lambda_*| \le C_3 |\lambda_k - \lambda_*|^2$$

and

$$\left.\begin{matrix} \sin\xi_{k+1} \\ \sin\eta_{k+1} \end{matrix}\right\} \le K_3^a \sin\eta_k \sin\xi_k. \tag{4.44}$$

Proof. In [78], Theorem 7, the quadratic convergence is shown for the angles between eigenvector approximations and eigenvectors for Algorithm 9 and the bound

$$\{\sin\xi_{k+1}, \sin\eta_{k+1}\} \;\leq\; \tilde{K}_3|\lambda_k - \lambda_*| \tag{4.45}$$

for $\tilde{K}_3 > 0$ is provided.

Requiring $\varepsilon_0 < \pi/3$ implies $\tan\xi_k \leq 2\sin\xi_k$, and with Theorem 3.5, the quadratic order of the Rayleigh functional $\lambda_{k+1} \equiv p(u_{k+1}, v_{k+1})$ follows, i.e.,

$$|\lambda_{k+1} - \lambda_*| \leq \frac{32\|T(\lambda_*)\|}{3|\alpha|}\sin\eta_{k+1}\sin\xi_{k+1} \leq C_3|\lambda_k - \lambda_*|^2,$$

with $C_3 = \tilde{K}_3^2\frac{32\|T(\lambda_*)\|}{3|\alpha|}$. Inequality (4.44) follows with (4.45), and $K_3^a = \tilde{K}_3\frac{32\|T(\lambda_*)\|}{3|\alpha|}$. □

4.3 Theoretical Comparison of the Methods

This section gives an overview of the methods we have seen so far and compares them with respect to storage and complexity. At first we summarize abbreviations for the algorithms, which will be used below.

II	inverse iteration	Algorithm 1
RII	residual inverse iteration	Algorithm 2
MOSLP	method of successive linear problems	Algorithm 3
RFI	two-sided Rayleigh functional iteration	Algorithm 4, 5
sRFI	simplified Rayleigh functional iteration	Algorithm 6
TRII	two-sided residual inverse iteration	Algorithm 7
ARFI	alternating Rayleigh functional iteration	Algorithm 8
GRFI	generalized Rayleigh functional iteration	Algorithm 10

We start by listing the methods which can be written in bordered matrix form, i.e., we leave out the residual iteration methods and the method of successive linear problems. The following table shows the structure of the remaining methods, where the bordered systems are shown, followed by the update step for the eigenvalue, where λ_+ is the new approximation.

II

$$\begin{bmatrix} T(\lambda) & \dot{T}(\lambda)u \\ u^H & 0 \end{bmatrix}\begin{bmatrix} s \\ \mu \end{bmatrix} \;=\; -\begin{bmatrix} T(\lambda)u \\ 0 \end{bmatrix}$$

$$\lambda_+ = \lambda - \frac{(u+s)^H T(\lambda)(u+s)}{(u+s)^H\dot{T}(\lambda)(u+s)}$$

sRFI

$$\begin{bmatrix} T(\lambda) & u \\ v^H & 0 \end{bmatrix} \begin{bmatrix} s \\ \mu \end{bmatrix} = - \begin{bmatrix} T(\lambda)u \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} T(\lambda)^H & v \\ u^H & 0 \end{bmatrix} \begin{bmatrix} t \\ \nu \end{bmatrix} = - \begin{bmatrix} T(\lambda)^H v \\ 0 \end{bmatrix}$$

$$\lambda_+ : (v+t)^H T(\lambda_+)(u+s) = 0$$

RFI

$$\begin{bmatrix} T(\lambda) & \dot{T}(\lambda)u \\ v^H\dot{T}(\lambda) & 0 \end{bmatrix} \begin{bmatrix} s \\ \mu \end{bmatrix} = - \begin{bmatrix} T(\lambda)u \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} T(\lambda)^H & \dot{T}(\lambda)^H v \\ u^H\dot{T}(\lambda)^H & 0 \end{bmatrix} \begin{bmatrix} t \\ \nu \end{bmatrix} = - \begin{bmatrix} T(\lambda)^H v \\ 0 \end{bmatrix}$$

$$\lambda_+ : (v+t)^H T(\lambda_+)(u+s) = 0$$

ARFI

$$\begin{bmatrix} T(\lambda) & \dot{T}(\lambda)u \\ u^H\dot{T}(\lambda) & 0 \end{bmatrix} \begin{bmatrix} s \\ \mu \end{bmatrix} = - \begin{bmatrix} T(\lambda)u \\ 0 \end{bmatrix}$$

$$\lambda_+ : (u+s)^H T(\lambda_+)(u+s) = 0$$

$$u_+ = (u+s)/\|u+s\|$$

$$\begin{bmatrix} T(\lambda_+)^H & \dot{T}(\lambda_+)^H u_+ \\ u_+^H\dot{T}(\lambda_+)^H & 0 \end{bmatrix} \begin{bmatrix} s \\ \mu \end{bmatrix} = - \begin{bmatrix} T(\lambda_+)^H u_+ \\ 0 \end{bmatrix}$$

$$\lambda_{++} : (u_+ + s)^H T(\lambda_{++})(u_+ + s) = 0$$

GRFI

$$\begin{bmatrix} T(\lambda) & v \\ u^H & 0 \end{bmatrix} \begin{bmatrix} s \\ \mu \end{bmatrix} = - \begin{bmatrix} T(\lambda)u \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} T(\lambda)^H & u \\ v^H & 0 \end{bmatrix} \begin{bmatrix} t \\ \nu \end{bmatrix} = - \begin{bmatrix} T(\lambda)^H v \\ 0 \end{bmatrix}$$

$$\lambda_+ : (v+t)^H T(\lambda_+)(u+s) = 0.$$

Table 4.1 presents an overview of most of the methods that have been discussed in this chapter. We list requirements on $T(\lambda)$ and other conditions in the third column, then the number of matrix-vector multiplications needed per iteration—where we view the multplication of $T(\lambda)$ and a vector as one matrix-vector multiplication—the number of preconditioners theoretically needed per iteration, and the order of convergence for eigenvalue and -vector approximations.

Moreover, since all methods assume that a simple eigenvalue is computed, this assumption is not listed below. Differentiability assumptions on $T(\lambda)$ are given for the real case, and may be replaced by assuming that $T(\lambda)$ is holomorphic on some domain in the complex case. We gave initial conditions only with respect to the starting vectors. Additionally,

	assumptions on		costs per iteration		convergence rates	
Method	$T(\lambda)$	starting vectors	# MVM	# Prec	$\lvert\lambda_* - \lambda_{k+1}\rvert =$	$\varepsilon^u_{k+1} = \varepsilon^v_{k+1} =$
II	$\ddot{T}(\lambda)$ continuous	$\lvert y_*^H \dot{T}(\lambda_*) u_0\rvert > 0$	$1 + \mathcal{O}(LinSys)$	1	$\mathcal{O}(\varepsilon^u_k \lvert\lambda_* - \lambda_k\rvert + \lvert\lambda_* - \lambda_k\rvert^2)$	$\mathcal{O}(\varepsilon^u_k \lvert\lambda_* - \lambda_k\rvert + \lvert\lambda_* - \lambda_k\rvert^2)$
RFI	$\ddot{T}(\lambda)$ continuous	$\lvert y_*^H \dot{T}(\lambda_*) u_0\rvert > 0$ $\lvert x_*^H \dot{T}(\lambda_*)^H v_0\rvert > 0$	$2(1 + \mathcal{O}(LinSys))$	1	$\mathcal{O}(\lvert\lambda_* - \lambda_k\rvert^2 \varepsilon^u_k \varepsilon^v_k)$ $= \mathcal{O}((\varepsilon^u_k)^3 (\varepsilon^v_k)^3)$	$\mathcal{O}((\varepsilon^u_k)^2 \varepsilon^v_k)$ $\mathcal{O}(\varepsilon^u_k (\varepsilon^v_k)^2)$
sRFI	$\dot{T}(\lambda)$ Lipschitz continuous	$\lvert y_*^H u_0\rvert > 0$ $\lvert x_*^H v_0\rvert > 0$	$1 + 2\mathcal{O}(LinSys)$	1	$\mathcal{O}(\lvert\lambda_* - \lambda_k\rvert^2)$ $= \mathcal{O}(\varepsilon^u_{k+1} \varepsilon^v_{k+1})$	$\mathcal{O}(\varepsilon^u_k \varepsilon^v_k)$
RII	$\ddot{T}(\lambda)$ continuous	$w_x^H u_0 = 1$	$1 + \mathcal{O}(LinSys)$	0	$\mathcal{O}(\varepsilon^u_k)$	$\mathcal{O}(\lvert\lambda_0 - \lambda_*\rvert \varepsilon^u_k)$
updated RII	$\ddot{T}(\lambda)$ continuous	$w_x^H u_0 = 1$	$1 + \mathcal{O}(LinSys)$	1	$\mathcal{O}(\varepsilon^u_k)$	$\mathcal{O}(\varepsilon^u_{k-1} \varepsilon^u_k)$
TRII	$\ddot{T}(\lambda)$ continuous	$w_x^H u_0 = 1$ $w_y^H v_0 = 1$	$2(1 + \mathcal{O}(LinSys))$	0	$\mathcal{O}(\varepsilon^u_k \varepsilon^v_k)$	$\mathcal{O}(\lvert\lambda_0 - \lambda_*\rvert \varepsilon^u_k)$ $\mathcal{O}(\lvert\lambda_0 - \lambda_*\rvert \varepsilon^v_k)$
updated TRII	$\ddot{T}(\lambda)$ continuous	$w_x^H u_0 = 1$ $w_y^H v_0 = 1$	$2(1 + \mathcal{O}(LinSys))$	1	$\mathcal{O}(\varepsilon^u_k \varepsilon^v_k)$	$\mathcal{O}(\varepsilon^u_{k-1} \varepsilon^v_{k-1} \varepsilon^u_k)$ $\mathcal{O}(\varepsilon^u_{k-1} \varepsilon^v_{k-1} \varepsilon^v_k)$
MOSLP	$\ddot{T}(\lambda)$ continuous $\dot{T}(\lambda_*)$ nonsingular		$\mathcal{O}(LinEVP)$	1	$\mathcal{O}(\lvert\lambda_* - \lambda_k\rvert^2)$	$\mathcal{O}((\varepsilon^u_k)^2)$
GRFI	$\dot{T}(\lambda)$ Lipschitz continuous	$\lvert x_*^H u_0\rvert > 0$ $\lvert y_*^H v_0\rvert > 0$	$2(1 + \mathcal{O}(LinSys))$	1	$\mathcal{O}(\lvert\lambda_* - \lambda_k\rvert^2)$ $= \mathcal{O}(\varepsilon^u_{k+1} \varepsilon^v_{k+1})$	$\mathcal{O}(\varepsilon^u_k \varepsilon^v_k)$

Table 4.1: *Summary of methods that compute eigenpairs (λ_*, x_*) or -triples (λ_*, x_*, y_*), with assumptions, number of matrix-vector multiplications and maximal number of preconditioners per iteration, convergence properties, when a simple eigenvalue is sought, i.e., $y_*^H \dot{T}(\lambda_*) x_* \neq 0$. Notation: $\varepsilon^u_k := \measuredangle(\mathrm{span}\{u_k\}, \mathrm{span}\{x_*\})$, $\varepsilon^v_k := \measuredangle(\mathrm{span}\{v_k\}, \mathrm{span}\{y_*\})$, or $\varepsilon^u_k := \|u_k - x_*\|$, $\varepsilon^v_k := \|v_k - y_*\|$, respectively.*

we have to keep in mind that the methods are locally convergent, hence the starting triple (pair), including the eigenvalue approximation, has to be sufficiently close to the solution. For the method of successive linear problems only the starting value λ_0 has to be provided.

The number of matrix-vector multiplications is given in terms of $T(\lambda)$, where we assume that matrix-vector multiplications with respect to $\dot{T}(\lambda)$ are roughly of the same order, and that $T(\lambda_k)u$ provides $T(\lambda_{k+1})u$ with an extra cost of $\mathcal{O}(n)$. The expression $\mathcal{O}(LinSys)$ refers to the cost of solving the associated linear system, with respect to the preconditioner, which includes several more matrix-vector multiplications if done iteratively, or $\mathcal{O}(n^2)$ in case of a direct solution with the available preconditioner, for example an LU-type decomposition.

Note that the convergence rates in terms of angles for the inverse iteration method given in Table 4.1 have not been obtained, yet. However, if u_k is brought to unit norm as in Algorithm 4, then equation (4.24) holds, and the terms in angles follow in the same way as in case of the two-sided Rayleigh functional iteration, cf. Theorem 4.13.

4.4 Computation of the Rayleigh Functional

Chapter 3 provides the theorical results in terms of bounds and properties for the concept of nonlinear Rayleigh functionals. However, the bound (3.22) may often be more of theoretical interest, because the practical application to compute the actual functional is limited to polynomial problems, if efficient methods for finding roots of the Rayleigh functional defining equation (3.15) are available. In general, a direct solution of this equation is not possible. Obviously, one Newton step for $g(\lambda, u, v) := v^H T(\lambda)u = 0$ with $\lambda \in S_0$ defined by (3.18), i.e., the generalized Rayleigh quotient p_L, will give a good approximation, the sharpness of which is determined by Theorem 3.32.

Let us consider the quadratic operator $T(\lambda) = \lambda^2 A + \lambda B + C$. As already mentioned, the generalized Rayleigh functional can be determined explicitly from equation (3.17) as

$$p(u, v) = \frac{1}{2v^H Au}\left[\pm\sqrt{(v^H Bu)^2 - 4v^H Cu(v^H Au)} - v^H Bu\right], \tag{4.46}$$

provided that $v^H Au \neq 0$. Setting $v = u$ yields the one-sided Rayleigh functional $p(u)$.

A major difference is that this form does not depend on the previously computed eigenvalue approximation, but only on the eigenvector approximations. So, unless the eigenvectors are in the region of convergence, i.e., $(u, v) \in \mathcal{K}_{\varepsilon_0}(x_*) \times \mathcal{K}_{\varepsilon_0}(y_*)$, cf. Theorem 3.5, we cannot be sure, whether $p(u, v)$ is a proper approximation. Moreover, the selection of a solution of (4.46) seems problematic when no information about the desired eigenvalue is available,

since one of the roots is spurious.

Figure 4.1 shows what will happen in most cases when the eigenvalue update is taken as one of the two roots of the Rayleigh functional from the beginning of the algorithm, although the distance of starting value and exact eigenvalue is only 0.07: The iteration stagnates. A promising strategy, outlined in Algorithm 11, is to perform the Newton

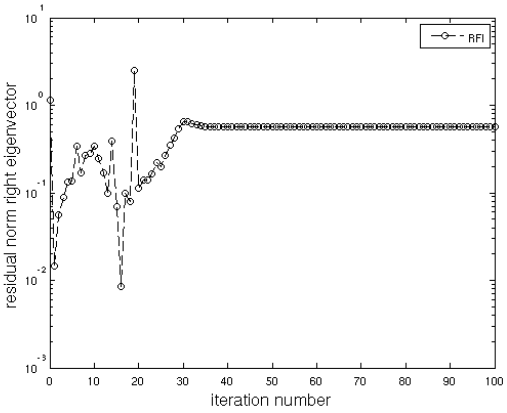

Figure 4.1: *Stagnation of the two-sided Rayleigh functional iteration when the eigenvalue update is obtained as Rayleigh functional from the beginning, applied to Example 2, see below.*

step, i.e., to compute the generalized Rayleigh quotient p_L, as long as the (outer) residual norm is greater than some initially set constant $\epsilon \ll 1$ to get a proper approximation, and solve directly in all subsequent iterations. Choose from the two values by examining the distance to the previous approximation, and take the one which is closer.

Algorithm 11 *Direct solution of Rayleigh functional equation*

Input: $(\lambda_{k-1}, u, v),\ T(\lambda) = \lambda^2 A + \lambda B + C$

if $\min\{\|T(\lambda_{k-1})u\|, \|v^H T(\lambda_{k-1})\|\} \geq \epsilon, \qquad \lambda_k = \lambda_{k-1} - \dfrac{v^H T(\lambda_{k-1})u}{v^H \dot{T}(\lambda_{k-1})u}$

else $h_1 = \dfrac{1}{2v^H Au}\left[\sqrt{(v^H Au)^2 - 4v^H Cu(v^H Au)} - v^H Bu\right]$

$h_2 = \dfrac{1}{2v^H Au}\left[-\sqrt{(v^H Bu)^2 - 4v^H Cu(v^H Au)} - v^H Bu\right]$

if $|\lambda_{k-1} - h_1| < |\lambda_{k-1} - h_2|, \qquad \lambda_k = h_1$

else $\lambda_k = h_2$

Output: λ_k

As the if-step in Algorithm 11 suggests, it is not sufficient to have a good approximation for the right eigenvector, but we need to have both vectors u and v to be sufficiently

good approximations to right and left eigenvector. In examples sometimes one eigenvector converges faster than the other due to initial vectors. Then the Newton update has to be chosen.

Remark 4.19 *Algorithm 11 can also be formulated for one-sided methods with one-sided Rayleigh functional by setting $v = u$.*

In [7] a Rayleigh quotient $x^H T(\lambda)x = 0$ is considered for the quadratic eigenvalue problem. It is proposed to compute the residuals $r_{\lambda_k} = (\lambda_k^2 A + \lambda_k B + C)u$ for $k = 1, 2$ in order to decide which of the roots is the sought approximation. An additional matrix-vector multiplication has to be evaluated here. For poor starting values this approach may lead to wrong decisions, however.

In practice it turned out that we can never guarantee for all cases and problems that the selection of the Rayleigh functional works, even if we use a small tolerance. Therefore, it is advisable to take the Newton update p_L everytime.

A third strategy to get an update of the Rayleigh functional for a quadratic problem, is to linearize it first—if possible—and take the Rayleigh quotient for the linearized problem. One possible linearization if A is nonsingular is

$$(\mathcal{A} + \lambda\mathcal{B})\tilde{x} = 0, \quad \mathcal{A} = \begin{bmatrix} C & 0 \\ 0 & -A^H \end{bmatrix}, \quad \mathcal{B} = \begin{bmatrix} B & A \\ A^H & 0 \end{bmatrix}, \quad \tilde{x} = \begin{bmatrix} x \\ \lambda x \end{bmatrix}.$$

Let $\tilde{u} = \left[\begin{smallmatrix} u_1 \\ u_2 \end{smallmatrix}\right]$, $\tilde{v} = \left[\begin{smallmatrix} v_1 \\ v_2 \end{smallmatrix}\right]$ be approximations for right and left eigenvector of this extended system. Then the generalized Rayleigh quotient for the generalized linear problem of double size is given by

$$\lambda \;=\; \frac{\tilde{v}^H \mathcal{A}\tilde{u}}{\tilde{v}^H \mathcal{B}\tilde{u}} \;=\; \frac{v_1^H C u_1 - v_2^H A^H u_2}{v_1^H B u_1 + v_1^H A u_2 + v_2^H A^H u_1}.$$

Note, that 4 matrix-vector multiplications have to be evaluated here if A is not Hermitian. This equation is a direct update for the linearized problem. If we do not want to linearize, we could set $u_2 = \lambda u_1$, $v_2 = \lambda v_1$ ending up with

$$\lambda \;=\; \frac{v_1^H C u_1 - \lambda^2 v_1^H A^H u_1}{v_1^H B u_1 + \lambda v_1^H A u_1 + \lambda v_1^H A^H u_1},$$

which can be iterated by inserting the old value of λ into the right hand side. If A is nonsingular, so is $\mathcal{B}$, and we can transform the generalized linear problem into the special linear eigenproblem

$$\mathcal{B}^{-1}\mathcal{A}x = \lambda x, \quad \text{where} \quad \mathcal{B}^{-1}\mathcal{A} = \begin{bmatrix} 0 & -I \\ A^{-1}C & A^{-1}B \end{bmatrix},$$

for which the two-sided Rayleigh quotient iteration with $\lambda = v^H(\mathcal{B}^{-1}\mathcal{A})u/v^Hu$ converges locally cubically [66, 67].

4.5 Numerical Experiments

So far, we have discussed various methods and different suggestions that are based on theoretical aspects. In the following we present numerical examples. Due to better comparability and potentially more accurate corrections, we implemented the algorithms, for which it was feasible, in the bordered matrix presentation as is outlined in §4.3.

The task of preconditioning appears to be relevant for all methods, since one or two systems of equations are to be solved in every step. One could think of different strategies and problems here. First, concerning the whole process, i.e., which method to take to tackle the systems in combination with a—for this method optimal—preconditioner, and second, having chosen ones favorite, how often and when exactly are new preconditioners computed.

Naturally, data changes more rapid during the first iterations, thus an advisable strategy is to compute preconditioners at the beginning and next if $|\lambda_{k+1} - \lambda_k| > \varepsilon$ for some given tolerance ε.

The implemented algorithms use preconditioned GMRes to solve the linear systems, where the preconditioners are obtained by an incomplete LU factorization (MATLAB: luinc) of the bordered matrix C, such that $C \approx LU$, with options $options.droptol = 0.005$, $options.udiag = 1$, i.e., the entries smaller in magnitude than the local drop tolerance (the product of the drop tolerance 0.005 and the norm of the corresponding column of C) are dropped from the appropriate factor, and any zeros on the diagonal of the upper triangular factor are replaced by the local drop tolerance, see the MATLAB help.

In case of the two-sided Rayleigh functional iteration we used $C^H \approx U^H L^H$ for solving the dual equation, and the same is done for the other two-vector iterations, such that every method computes at most one preconditioner, respectively decomposition, in every iteration. In practice one would use much less preconditioners, but the performance of the methods depends heavily on the existence of a new preconditioner and we could make every method look fast, if we know how to tune it best. Hence, to have a fair comparison, we compute preconditioners at every iteration step.

Experiments were run under MATLAB 7.4 on an Intel(R) Pentium(R) 4 CPU 3.20 GHz with 1 GB RAM.

We start with three quadratic eigenproblems, the first is a nice one for all of the methods,

the next two present exceptional cases.

Example 2 QEP$_1$ *Consider the quadratic eigenvalue problem*

$$T(\lambda)x = (\lambda^2 A + \lambda B + C)x = 0,$$

where

$$A = I,\ B = \begin{bmatrix} 2 & 0 & a/2 & & \\ 0 & \ddots & \ddots & \ddots & \\ -a/2 & \ddots & \ddots & \ddots & a/2 \\ & \ddots & \ddots & \ddots & 0 \\ & & -a/2 & 0 & 2 \end{bmatrix},\ C = \begin{bmatrix} 2 & -1+a & & & \\ -1-a & \ddots & \ddots & & \\ & \ddots & \ddots & -1+a & \\ & & -1-a & 2 & \end{bmatrix}$$

for some parameter $a > 0$.

For $a = 0$ the matrices are symmetric, but we are interested in the nonsymmetric case. If $a > 0$ increases, then the nonnormality of C increases, whilst B is skew-symmetric for any a, i.e., we have $B = -B^T$. Figure 4.2 shows two plots of eigenvalues for different a for a size 100 problem. For increasing a, the eigenvalues are moving apart, making it more difficult to compute them accurately.

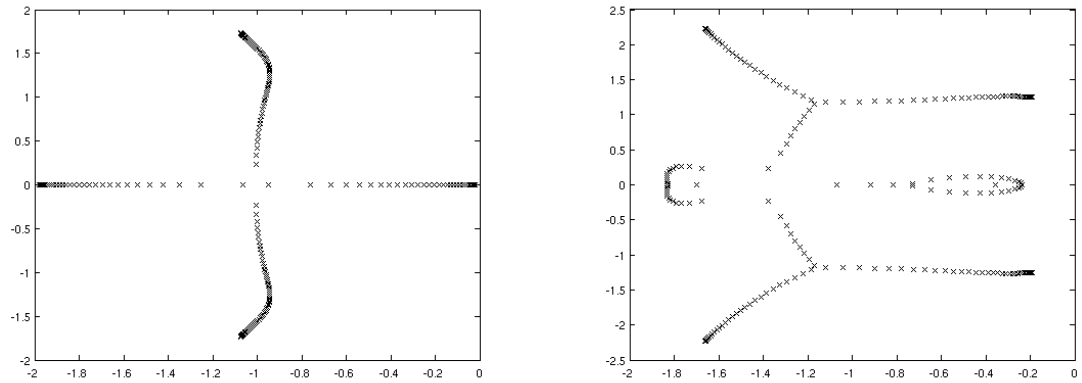

Figure 4.2: *Eigenvalues in the complex plane of QEP$_1$ of order $n = 100$ for $a = 0.2$ on the left hand side and $a = 1$ on the right hand side.*

In subsequent tables the following headline abbreviations will be used:

$\lambda_{it} - \lambda_*$ = forward error of the final eigenvalue approximation (if available)

it	=	number of iterations
$\|res_u\|$	=	$\|T(\lambda_{it})u_{it}\|$ where u_{it} is the final right eigenvector approximation
$\|res_v\|$	=	$\|T(\lambda_{it})^H v_{it}\|$ where v_{it} is the final left eigenvector approximation
gm_1	=	number of inner solves for the linear system corresponding to the right eigenvector
gm_2	=	number of inner solves for the linear system corresponding to the left eigenvector
gm_1/it	=	average number of inner solves per outer iteration for the first linear system
gm_2/it	=	average number of inner solves per outer iteration for the second linear system
cpu	=	consumed time in seconds

We compute the Rayleigh functional as shown in Algorithm 11 with $\epsilon = 1e-8$. We hope to ensure with this condition that both iterated vectors are in the region of convergence for the Rayleigh functional. The termination tolerance for the residual norm was set to $1e-10$.

At first, we ran the methods for the case $a = 0.2$ without using an iterative solver and solved the linear systems directly with exact LU factors. Results are shown in Table 4.2, and correspond to the upper figure in Figure 4.3, which shows residual norms. Note that

	$\|\lambda_{it} - \lambda_*\|$	it	$\|res_u\|$	$\|res_v\|$	cpu
sRFI	3.9e-06	9	3.6e-15	3.7e-15	0.05
RFI	3.9e-06	9	3.7e-16	3.2e-16	0.05
ARFI	3.9e-06	27	5.9e-11	1.3e-10	0.12
GRFI	3.9e-06	18	4.0e-15	4.1e-15	0.07
II	3.9e-06	13	5.2e-12		0.11
RII	2.6e-01	15	6.2e-12		0.04
TRII	3.9e-06	14	5.2e-13	5.2e-13	0.04
MOSLP	3.9e-06	4	1.3e-14		0.32

Table 4.2: QEP_1 , $a = 0.2$, $n = 100$, *direct solver.*

the residual inverse iteration converges not to the smallest eigenvalue, but to a nearby one which is indicated by $|\lambda_{it} - \lambda_*| = 0.26$. The smallest number of iterations was needed by the method of successive linear problems. But, on the other hand, solving linear eigenproblems completely, takes a lot of time. The two-sided Rayleigh functional iteration and its simplified variant give almost the same results, they need less iterations than the other

methods (except for the MOSLP) and are approximately the fastest together with the two-sided residual iteration method.

The same example was run with the iterative solver preconditioned GMRes, cf. Table4.3. New preconditioners are computed in every iteration. Again, the residual inverse iteration finishes with another than the smallest eigenvalue of the problem, and the two-sided residual inverse iteration converges to the next closest eigenvalue. All methods, except the method of successive linear problems—for which nothing changed—are somewhat slower. This may be due to the excellent performance of the backslash routine incorporated in MATLAB, which was used in the previous sample run, where a lot of work has been done in the past few years. For large systems where decompositions are too large to be kept in memory, however, iterative methods may be the only option. The almost linear con-

	$\lvert\lambda_{it}-\lambda_*\rvert$	it	$\lVert res_u\rVert$	$\lVert res_v\rVert$	gm_1	gm_2	$gm1/it$	gm_2/it	cpu
sRFI	3.9e-06	9	5.5e-14	4.4e-15	64	67	7.1	7.4	0.18
RFI	3.9e-06	9	2.6e-14	2.5e-15	63	67	7.0	7.4	0.18
ARFI	4.0e-06	33	9.2e-10	3.9e-03	187	227	5.7	6.9	0.59
GRFI	3.9e-06	19	5.9e-10	5.9e-10	80	80	4.2	4.2	0.25
II	3.9e-06	15	9.0e-12		134		8.9		0.18
RII	1.8e-01	12	1.8e-12		26		2.2		0.25
TRII	1.4e-03	8	2.8e-12	5.7e-12	47	52	5.9	6.5	0.31
MOSLP	3.9e-06	4	1.1e-14						0.33

Table 4.3: QEP_1 , $a = 0.2$, *iterative solver.*

vergence of the alternating Rayleigh functional iteration is reflected in the corresponding Figure 4.3. The fastest methods are the two-sided Rayleigh functional iteration, the simplified version and the inverse iteration.

The method of successive linear problems is especially suited for small and dense problems, where for instance the eig-function of MATLAB can be used to determine the solutions of the linear eigenproblem. For larger problems one needs to think of other methods than using a full QZ decomposition. The eigs-function, which was made for sparse linear problems, cannot be used for general linear problems $A - \lambda B$, where B is not symmetric positive definite. But that is just the case we consider.

The second example is of different nature. Since we are interested in the performance of the algorithms with respect to ill-conditioned eigenvalues, we created the following quadratic eigenvalue problem, where the term $\alpha = y_*^H \dot{T}(\lambda_*)x_*$ can be chosen arbitrarily small.

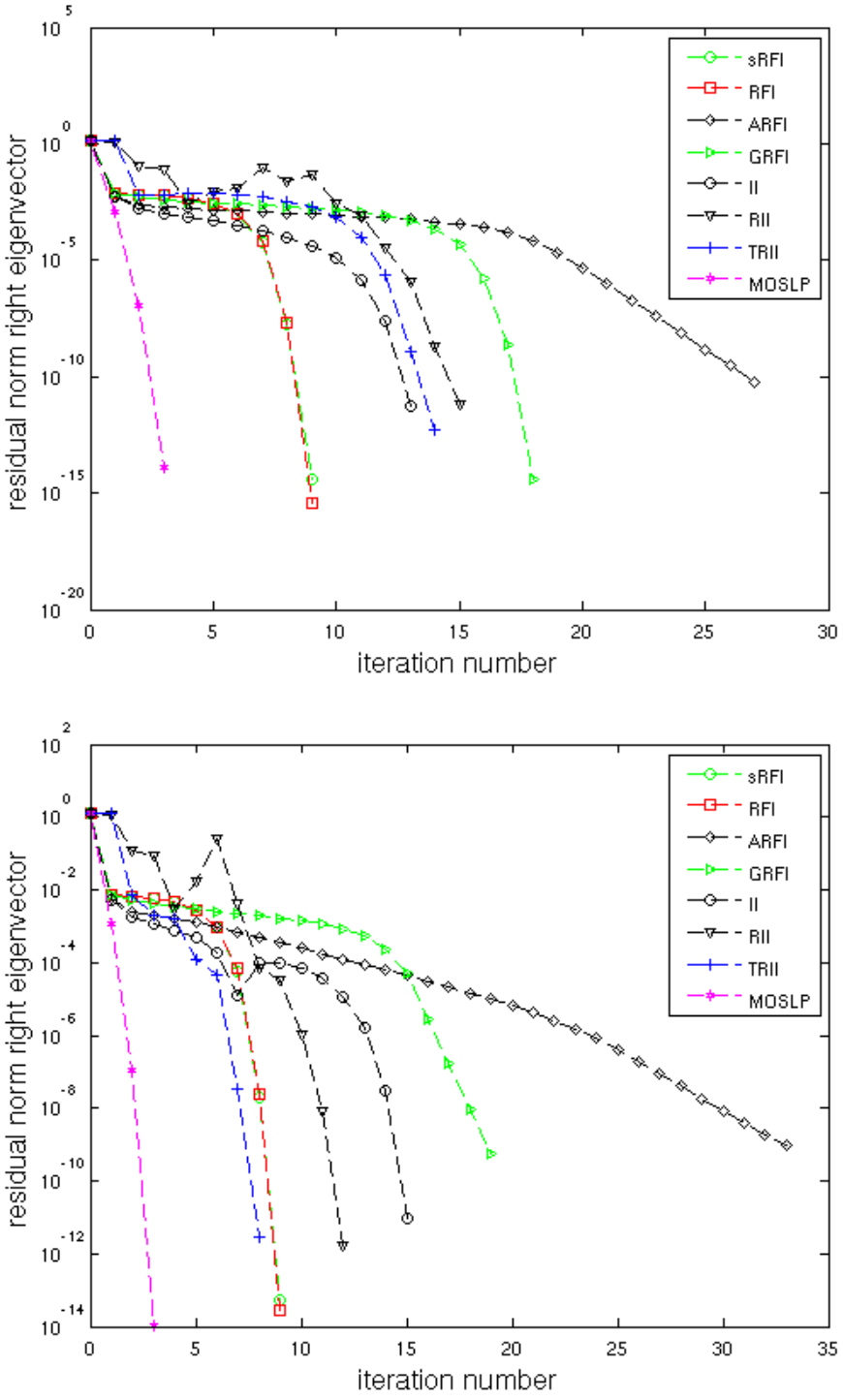

Figure 4.3: *Residual norms of displayed methods applied to the quadratic eigenvalue problem QEP_1 corresponding to Tables 4.2, 4.3. The upper figure represents the results for the methods using a direct solver, the lower using an iterative solver.*

Example 3 QEP$_2$ *Let*

$$T(\lambda)x = (\lambda^2 A + \lambda B + C)x = 0.$$

We want to construct A, B and C such that $\lambda_* = 1$ and $\alpha = y_*^H \dot{T}(1)x_*$ can be chosen arbitrarily. Based on the singular value decomposition (2.12), i.e., in this case

$$T(\lambda_*) = A + B + C = Y\Sigma X^H = [Y_1|y_*] \left[\begin{array}{c|c} \Sigma_1 & \\ \hline & 0 \end{array}\right] [X_1|x_*]^H,$$

we set $x_* = e_n$ and choose $[X_1|x_*] = I_n$ for simplicity and $y_* = \frac{1}{\sqrt{n}}(1 \ldots 1)^T$. Then, setting the n-th column of B to $b_n = \frac{\alpha}{2}\sqrt{n}e_1$ and the n-th column of C to $c_n = \frac{\alpha}{4}\sqrt{n}e_2$ yields the desired dependence

$$y_*^H \dot{T}(\lambda_*)x_* = \frac{1}{\sqrt{n}}(1 \ldots 1)^T(B + 2C)e_n = \frac{1}{\sqrt{n}}(1 \ldots 1)^T(b_n + 2c_n) = \alpha.$$

The remaining parts of B and C can be chosen arbitrarily, we took random entries in the example. To compute A, we have to compute the unitary matrix Y, where the last vector of Y is the left eigenvector y_*. Therefore, we compute a Householder reflection, cf. [46], and obtain

$$Y = \left(-I_n + \frac{(y_* + e_n)(y_* + e_n)^T}{1 + \frac{1}{\sqrt{n}}}\right).$$

One can check that $Ye_n = y_*$. Then, the matrix A can be determined by rearranging the singular value decomposition which yields

$$A = Y\Sigma - B - C.$$

The singular values on the diagonal of Σ_1 are random in the interval $[100/\kappa, 100]$ with $\sigma_1 = 100$, $\sigma_{n-1} = 100/\kappa$, where κ is the condition number of Σ_1.

It may happen that $\lambda_* = 1$ is not the largest, resp. smallest, eigenvalue in absolute value. The methods of this chapter rather tend to the eigenvalue smallest in absolute value than to another, since we iterate with $T(\lambda)^{-1}$. For sufficiently good starting values and vectors, in principle every simple eigenvalue can be computed. The same holds for the Jacobi–Davidson-type methods in Chapter 6. In most cases, we observed, that $\lambda_* = 1$ will be the smallest eigenvalue in absolute value, if not we generated another set of matrices.

For moderate $\alpha = 1e-6$, the MATLAB included polyeig-function already computes the slightly inaccurate $\lambda_* = 0.99999999870220$.

In our experiments we started with $\lambda_0 = 1.1$. Table 4.4 and Figure 4.4 show the results. The inverse iteration method and the residual inverse iteration method converge to another

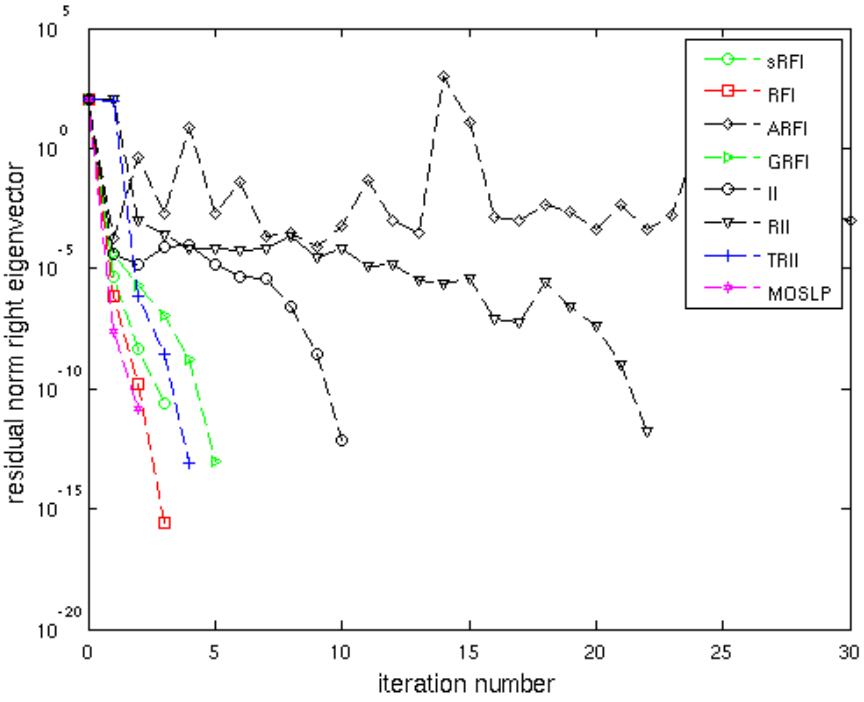

Figure 4.4: *Example* QEP_2*, starting value* $\lambda_0 = 1.1$*, order* $n = 100$*,* $\alpha = 1e-6$*.*

eigenvalue. The alternating Rayleigh functional iteration converges too slow to converge within 100 iterations. The other methods converge fast and accurate—at least in one vector.

	$\lvert\lambda_{it} - \lambda_*\rvert$	it	$\lVert res_u \rVert$	$\lVert res_v \rVert$	gm_1	gm_2	gm_1/it	gm_2/it	cpu
sRFI	6.4e-11	3	2.5e-11	3.1e-04	6	4	2.0	1.3	0.06
RFI	5.2e-13	3	2.4e-16	2.5e-10	6	6	2.0	2.0	0.07
ARFI	8.0e+00	100	1.1e-04	1.2e+02	350	870	3.5	8.7	4.51
GRFI	2.2e-16	5	1.0e-13	2.1e-06	8	8	1.6	1.6	0.09
II	4.6e+00	10	6.7e-13		78		7.8		0.14
RII	4.6e+00	22	1.6e-12		75		3.4		0.40
TRII	2.0e-15	4	8.3e-14	4.8e-07	6	5	1.5	1.2	0.08
MOSLP	1.4e-06	3	1.4e-11						0.20

Table 4.4: QEP_2*,* $n = 100$*,* $\alpha = 1e-6$*.*

Next we decreased α to $1e-12$. Then, the polyeig-function computes the eigenvalue 0.99900727660049, which is inaccurate to the forth decimal number, although this eigenvalue is well-separated from the remaining spectrum, see Figure 4.5.

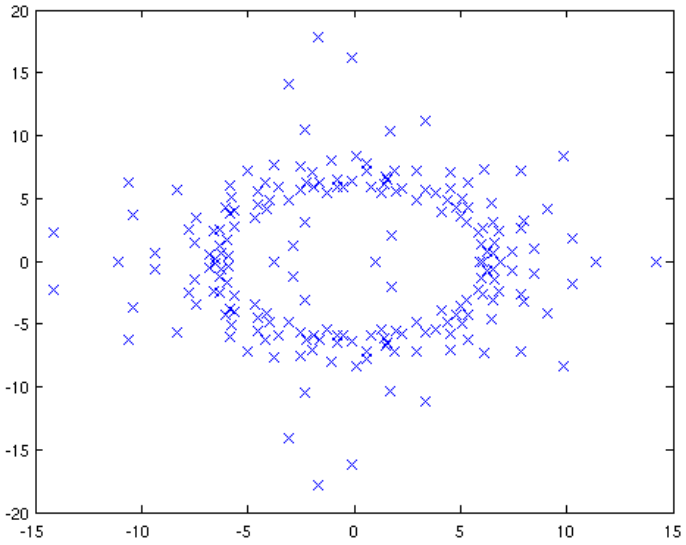

Figure 4.5: *Eigenvalues of Example QEP$_2$ for $\alpha = 1 - 12e$ computed by the* MATLAB *polyeig-function.*

For $\alpha = 1e - 14$ the nearest eigenvalue to 1 computed by polyeig in an example run was $1.2756 - 0.6414$i. Hence, we could not ensure, whether the exact eigenvalue 1 was the smallest in norm and took $\alpha = 1e - 12$ for testing the methods. The results are shown in Table 4.5 and Figure 4.6. The behavior of the two-sided Rayleigh functional iteration

	$\lvert\lambda_{it} - \lambda_*\rvert$	it	$\lVert res_u \rVert$	$\lVert res_v \rVert$	gm_1	gm_2	gm_1/it	gm_2/it	cpu
sRFI	2.5e-01	3	1.7e-12	2.0e+01	8	6	2.7	2.0	0.09
RFI	0.0e+00	4	1.0e-14	2.1e-13	8	9	2.0	2.2	0.24
ARFI	2.9e+01	100	2.6e-03	4.5e+03	338	846	3.4	8.5	4.67
GRFI	1.5e-01	3	3.5e-12	6.5e+01	8	8	2.7	2.7	0.06
II	5.2e+00	6	6.3e-11		46		7.7		0.08
RII	2.3e+01	100	4.3e+02		359		3.6		7.55
TRII	1.1e-05	3	1.2e-13	2.5e-03	5	4	1.7	1.3	0.12
MOSLP	3.9e-03	2	9.2e-14						0.11

Table 4.5: *Example QEP$_2$, $\alpha = 1e - 12$.*

is remarkable. It obtains the exact eigenvalue $\lambda_* = 1$ (in finite precision arithmetic). Second close is the two-sided residual inverse iteration method, but not that accurate. The convergence results of the other methods are even worse, although most of the methods

compute at least the right eigenvector satisfactorily.

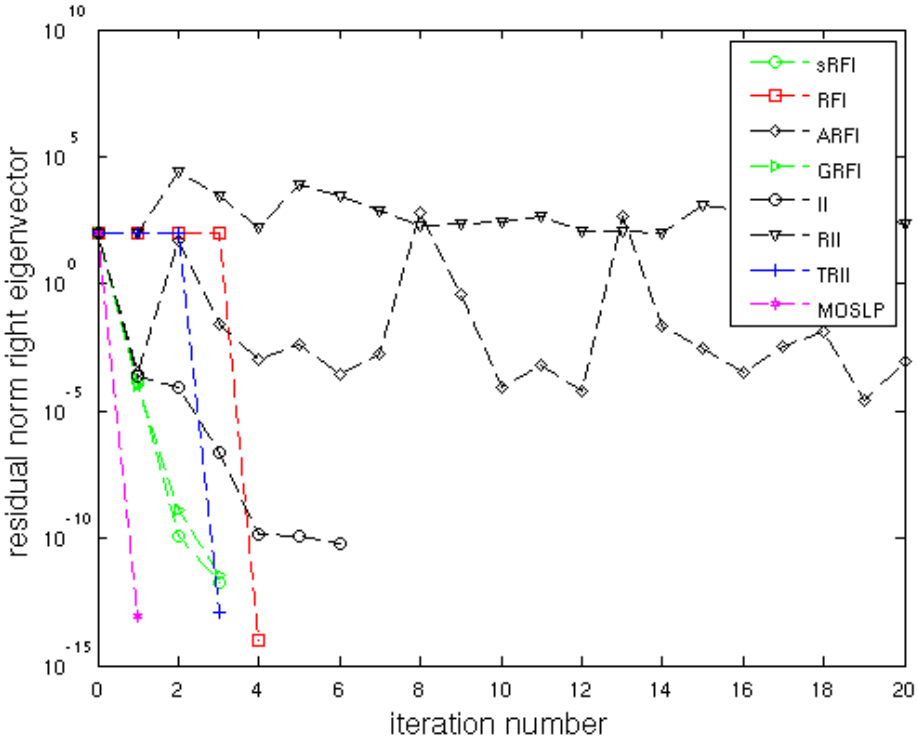

Figure 4.6: *Example QEP$_2$, starting value* $\lambda_0 = 1.1$, *order* $n = 100$, $\alpha = 1e-12$.

The next example is again a constructed one in order to analyze the performance of the methods when one of the assumptions fails. For this case, we designed a quadratic problem where $x_*^H \dot{T}(\lambda_*)x_* = 0$ holds for the simple dominant eigenvalue λ_*, i.e., $y_*^H \dot{T}(\lambda_*)x_* \neq 0$. This should not be a problem for the Newton-type methods of this chapter, but as we will see later, cf. Chapter 6, we need to assume that $x_*^H \dot{T}(\lambda_*)x_* \neq 0$ for the Jacobi–Davidson method.

Example 4 QEP$_3$ *Let* $T(\lambda) = \lambda^2 A + \lambda B + C$, *with* $A, B, C \in \mathbb{R}^{n\times n}$, *and let* $E_A, E_B, E_C \in \mathbb{R}^{(n-2)\times(n-2)}$ *be three diagonal perturbation matrices with entries* $\varepsilon_A, \varepsilon_B, \varepsilon_C$ *on the diagonal smaller then* ε, $0 < \varepsilon \ll 1$. *Define the matrices* A, B, C *by*

$$A = \left[\begin{array}{cc|c} 1 & 0 & \\ 1 & 1 & \\ \hline & & A_{22} \end{array}\right], \quad B = \left[\begin{array}{cc|c} -2 & -1 & \\ -1 & 0 & \\ \hline & & B_{22} \end{array}\right], \quad C = \left[\begin{array}{cc|c} 1 & 0 & \\ 0 & 0 & \\ \hline & & C_{22} \end{array}\right],$$

where $A_{22} = I_{n-2} + E_A$, $B_{22} = I_{n-2} + 3E_B$, $C_{22} = I_{n-2} + 2E_C$.

The exact eigenvalues of the upper left part are given by $\{1, 0, 0, 0\}$, and the whole spectrum will be

$$\lambda(T) = \{1, 0, 0, 0, (-1 + \varepsilon_A)_{n-2}, (-2 + \varepsilon_C)_{n-2}\},$$

where the last two values appear $(n-2)$-times each. The right and left eigenvectors corresponding to $\lambda_* = 1$ are $x_* = e_1$, this is the first unit vector and $y_* = e_1 + e_2$. Thus, they satisfy $y_*^H \dot{T}(\lambda_*) x_* \neq 0$, but $x_*^H \dot{T}(\lambda_*) x_* = 0$. Furthermore, in the case $n = 2$, i.e., we consider only the upper left part, we have $y_*^H \dot{T}(\lambda_*) x_* \neq 1/\sqrt{2}$ and $y_*^H \dot{T}(\lambda_*) y_* = 1$.

We run the methods for an order 100 problem. All methods approximated $\lambda = 1$, although we started with $\lambda_0 = -0.1$, but only the simplified Rayleigh functional iteration, the generalized Rayleigh functional iteration and the method of successive linear problems, which converges with a linear rate (see Figure 4.7), hit the requested tolerance $1e-10$ in the residuals (see Table 4.6). A possible explanation for this result is that the methods,

	$\lvert\lambda_{it}\rvert$	it	$\lVert res_u \rVert$	$\lVert res_v \rVert$	gm_1	gm_2	gm_1/it	gm_2/it	cpu
sRFI	1.0e-00	28	2.1e-16	2.1e-16	53	60	1.9	2.1	0.23
RFI	1.0e+00	100	1.9e-04	2.1e-04	260	399	2.6	4.0	1.02
ARFI	1.0e+00	100	1.9e-05	5.2e-06	299	402	3.0	4.0	1.69
GRFI	1.0e-00	22	1.8e-12	1.8e-12	41	41	1.9	1.9	0.18
II	1.0e+00	100	1.7e-04		313		3.1		0.52
RII	1.0e+00	100	2.0e-05		247		2.5		0.46
TRII	1.0e+00	100	2.0e-03	2.2e-03	228	300	2.3	3.0	0.84
MOSLP	1.0e+00	98	9.7e-11						10.70

Table 4.6: *QEP*$_3$, $\lambda_0 = -0.1$, $n = 100$.

which do not converge well, depend on the term $x_*^H \dot{T}(\lambda_*) x_*$ in a hidden way. Although, we did not need to assume that $x_*^H \dot{T}(\lambda_*) x_* \neq 0$ for the proof of convergence of the two-sided Rayleigh functional iteration, we will see the connection in Chapter 6. Changing λ_0 to $1e-06$ only scales Figure 4.7. With random starting vectors, most methods will not converge.

Example 5 PDDE *The discretization of a partial delay differential equation (PDDE) gives the following nonlinear eigenvalue problem*

$$(-\lambda I + A_0 + (A_1 + A_2)\exp(-\lambda))x = 0,$$

where A_0, A_1, A_2 are real symmetric matrices [43].

Setting the starting value to $\lambda_0 = 0.1$ yields a good performance for most methods, but the residual inverse iteration method and the method of successive linear problems obtain huge

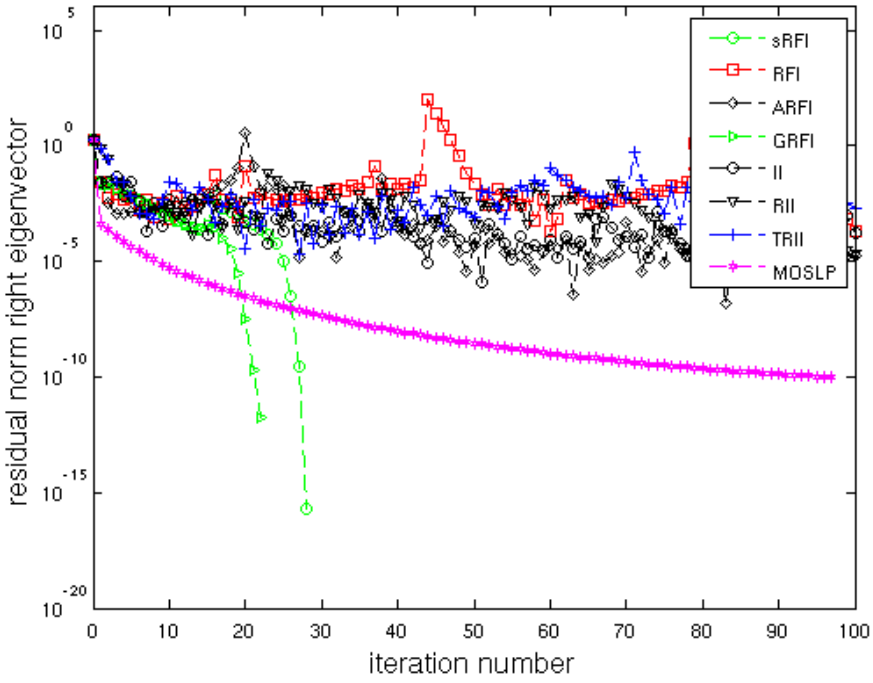

Figure 4.7: *Example QEP$_3$, starting value* $\lambda_0 = -0.1$*, order* $n = 100$*.*

residual norms, which are already achieved after few iterations and stagnate on this high level, see Figure 4.8. The other methods converge fast, but to two different eigenvalues. The column $|\lambda_{it}|$ shows the absolute value of the last eigenvalue approximation, since we have no information on exact eigenvalues.

	$\lvert\lambda_{it}\rvert$	it	$\lVert res_u \rVert$	$\lVert res_v \rVert$	gm_1	gm_2	gm_1/it	gm_2/it	cpu
sRFI	1.9e+01	5	6.2e-14	7.0e-14	12	11	2.4	2.2	0.28
RFI	1.9e+01	6	5.9e-14	7.2e-14	13	13	2.2	2.2	0.16
ARFI	1.6e+01	4	5.9e-14	7.6e-14	10	20	2.5	5.0	0.14
GRFI	1.9e+01	5	5.2e-14	9.3e-14	12	12	2.4	2.4	0.13
II	1.6e+01	7	1.1e-11		63		9.0		0.13
RII	1.8e+02	100	2.6e+79		101		1.0		0.27
TRII	1.9e+01	7	1.4e-13	7.8e-14	12	11	1.7	1.6	0.09
MOSLP	5.4e+01	100	9.6e+23						0.56

Table 4.7: *PDDE,* $\lambda_0 = 0.1$*,* $n = 50$*.*

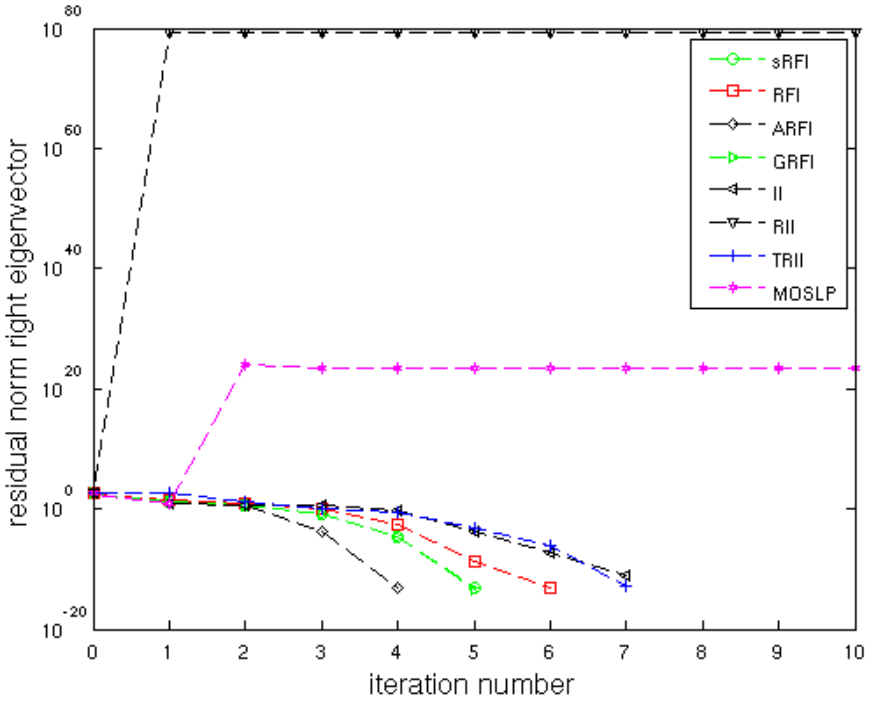

Figure 4.8: *Results for Example PDDE with starting value* $\lambda_0 = 0.1$.

4.6 Conclusion

In this chapter several methods of Newton-type for nonlinear eigenvalue problems have been discussed, which compute one eigenvalue and the corresponding eigenvector, or even left and right eigenvector. There, the Rayleigh functional theory of Chapter 3 was used and convergence analyzed. We have shown that the nonlinear two-sided Rayleigh functional iteration converges locally cubically as the two-sided Rayleigh quotient iteration for nonnormal matrices. A simplified version without $\dot{T}(\lambda)$ in the linear systems has been proposed and was shown to converge locally quadratically. This version is less sensitive with respect to the eigenvalue condition number, although the matrix is not optimally bordered in this sense. If an ill-conditioned eigenvalue is sought, it is advisable to use the generalized Rayleigh functional iteration, which converges locally quadratically as well.

The extension of the residual inverse iteration to a two-vector methods makes it possible to compute the two-sided Rayleigh functional which converges quadratically under certain assumptions.

In the experiments, it turned out that a direct application of the Rayleigh functional is more complicated in the nonlinear (polynomial) case than in the linear case. For degree m

polynomials the Rayleigh functional has up to m solutions, but it is hard to decide which one to choose. Several strategies for quadratic problems were suggested and tested. They work sometimes but not always and tend to end up in circles. The save way is to perform the Newton step, hence, to use the generalized Rayleigh quotient p_L, p_N, resp., although the cubic convergence is lost in this case.

Furthermore, we could not detect large differences in the performance of the two-sided Rayleigh functional iteration and its simplified version, in general. The plotted residual norms were either on top of each other or quite close. However, in case of the exceptional example QEP_2, the simplified RFI converged in contrast to the RFI.

We have experimented with different quadratic and one genuine nonlinear example. In general, with sufficient treatments and patience one will get every method working optimized regarding the number and way of preconditoners and the linear solver. Since these are only basic methods, they need to be adjusted depending on the environment where they are used.

The performance of the methods depends on the problem and on the starting values, but in general the inverse iteration will give good results.

Chapter 5

Half-step Methods

This chapter continues with the two-vector methods of the previous chapter—but slightly changed. We want to introduce an easy way to accelerate convergence, which can in principle be applied to all left and right eigenvector computing methods.

Most of the tested algorithms in Chapter 4 have equivalent structure in that they compute a correction for the right eigenvector, then a correction for the left eigenvector and afterwards both results are included in the Rayleigh functional which gives the corresponding new eigenvalue approximation.

We will show that convergence can be improved, in the sense that the R-order increases, if an additional Rayleigh functional is inserted in between the computation of the eigenvector updates in order to work with a more accurate matrix when solving the second system, e.g., solving the left eigenvector updating system. This is shown by writing the sequences of errors as a system of difference inequalities and using the results of J. W. Schmidt [74]. He has investigated the R-order of coupled sequences and has shown that the spectral radius of the matrix of some exponents gives a lower bound for the common R-order of the coupled sequences.

The idea of inserting an additional eigenvalue computation will be applied to the Rayleigh functional methods and to the generalized residual inverse iteration method. The methods are named with the additional term *half-step*, although one could also argument that they are *double-step* methods, instead. The solution of the first Rayleigh functional defining equation will be marked by $k + 1/2$.

Half-step methods are in some sense particularly interesting when the basic methods are not so—if costs matter, i.e., if the problem is large. We will briefly discuss the costs of two-vector methods compared to one-vector methods with respect to the order n. If we do so, the following two cases have to be distinguished:

- n moderate: For a moderate size n, in the sense that it is possible to compute a sparse LU factorization within a reasonable amount of time, only one LU factorization of the system matrix $C_k = L_k U_k$ is needed in an iteration of the two-vector methods in §4.2, since the second linear system has C_k^H as system matrix and can be solved using L_k, U_k, too. The cost for the second system reduces then to an additional forward and backward solve. Moreover, in the case that computing on several processors is available, the two linear systems can be solved completely in parallel.

  Note that even if the matrix is Hermitian it will not be definite, in general, such that a Cholesky factorization cannot be used.

  However, changing the second system as is done in half-step methods implies that new LU factors have to be retrieved for the solution of the second linear system.

- n large: If n is large, linear systems have to be solved iteratively, and it is advisible to use preconditioners. In any case, the second linear system in a two-vector method has to be computed from scratch. However, the use of a preconditioner is not only helpful but in practice it is often necessary for convergence. The standard way to obtain a preconditioner is to use an incomplete LU factorization, which is updated only if convergence decreases drastically. Therefore, changing the second system in the half-step methods is negligible with respect to the preconditioner, which can be used by taking the conjugate transpose as in case of the two-vector methods. The costs for two-vector methods and half-step two-vector methods are almost equal, only that half-step methods will be shown to converge faster, hence the costs decrease. However, half-step methods are not as easy to parallelize as the basic methods, since the second system can only be solved after the solution of the first one is at hand.

5.1 Half-step Rayleigh Functional Iteration

We start with recalling the essential parts of the two-sided Rayleigh functional iteration. Suppose we have finished the k-th iteration of Algorithm 4, i.e., we have $(u = u_k, v = v_k)$ and $\lambda = p(u, v)$. Then, the two-sided Rayleigh functional iteration can be described by the following scheme, where convergence properties are shown on the right hand side; cf. Theorem 4.13.

$$\text{RFI:} \quad \begin{cases} (\lambda, u) & \mapsto & u_+ \sim T(\lambda)^{-1}\dot{T}(\lambda)u : & \sin\xi_+ & \leq & K_0^a \sin^2\xi \sin\eta, \\ (\lambda, v) & \mapsto & v_+ \sim T(\lambda)^{-H}\dot{T}(\lambda)^H v : & \sin\eta_+ & \leq & \tilde{K}_0^a \sin\xi \sin^2\eta, \\ (u_+, v_+) & \mapsto & \lambda_+, \text{ s.t., } v_+^H T(\lambda_+)u_+ = 0 : & |\lambda_+ - \lambda_*| & \leq & 4K \sin\xi_+ \sin\eta_+, \end{cases}$$

with constants K_0^a, $\tilde{K}_0^a$, K as in the proof of Theorem 4.13. Recall the definition of the angles

$$\begin{aligned} \xi &:= \measuredangle\left(\operatorname{span}\{x_*\}, \operatorname{span}\{u\}\right), & \eta &:= \measuredangle\left(\operatorname{span}\{y_*\}, \operatorname{span}\{v\}\right), \\ \xi_+ &:= \measuredangle\left(\operatorname{span}\{x_*\}, \operatorname{span}\{u_+\}\right), & \eta_+ &:= \measuredangle\left(\operatorname{span}\{y_*\}, \operatorname{span}\{v_+\}\right). \end{aligned}$$

The scheme is to be read as follows: The pair (λ, u) is mapped to a new vector u_+, which is proportional to $T(\lambda)^{-1}\dot{T}(\lambda)u$ and the related angle ξ_+ of the vectors u_+ and x_* can be measured in terms of the angles ξ, η, i.e., $\sin\xi_+ \leq K_0^a \sin^2\xi \sin\eta$. Then, the pair (λ, v) is mapped to the new vector v_+ in the same way, but with the adjoint matrix $T(\lambda)^H$. When both new vectors have been computed, they are used to obtain a new eigenvalue approximation λ_+.

We introduce the abbreviating terms

$$\begin{array}{rclrclrcl} \delta & := & \sin\xi, & \varepsilon & := & \sin\eta, & \gamma & := & |\lambda - \lambda_*|, \\ \delta_+ & := & \sin\xi_+, & \varepsilon_+ & := & \sin\eta_+, & \gamma_+ & := & |\lambda_+ - \lambda_*|, \end{array} \tag{5.1}$$

for ease of notation, which yield the following inequalities

$$\text{RFI:} \qquad \begin{cases} \delta_+ & \leq & K_0^a \delta^2 \varepsilon, \\ \varepsilon_+ & \leq & \tilde{K}_0^a \delta \varepsilon^2, \\ \gamma_+ & \leq & 4K\delta_+\varepsilon_+ \quad \leq \quad C_0^a \gamma^2 \delta\varepsilon, \end{cases}$$

for the two-sided Rayleigh functional iteration, with C_0^a from (4.27). Apparently, the information gathered by the solution of the first correction equation, i.e., the one for the right eigenvector, is not included in the second correction equation, i.e., the one for the left eigenvector. This is changed when we insert a Rayleigh functional computation in between the two linear systems, which gives the new scheme

$$\text{HSRFI:} \qquad \begin{cases} (\lambda, u) & \mapsto & u_+ \sim T(\lambda)^{-1}\dot{T}(\lambda)u : & \sin\xi_+ & \leq & K_0^a \sin^2\xi \sin\eta, \\ (u_+, v) & \mapsto & \lambda_{1/2},\ \text{s.t.,}\ v^H T(\lambda_{1/2})u_+ = 0 : & |\lambda_{1/2} - \lambda_*| & \leq & 4K \sin\xi_+ \sin\eta, \\ (\lambda_{1/2}, v) & \mapsto & v_+ \sim T(\lambda_{1/2})^{-H}\dot{T}(\lambda_{1/2})^H v : & \sin\eta_+ & \leq & \tilde{K}_0^a \sin\xi_+ \sin^2\eta, \\ (u_+, v_+) & \mapsto & \lambda_+,\ \text{s.t.,}\ v_+^H T(\lambda_+)u_+ = 0 : & |\lambda_+ - \lambda_*| & \leq & 4K \sin\xi_+ \sin\eta_+, \end{cases}$$

provided that $\lambda = p(u, v)$, or in other words, provided that we are at least in the second iteration of the algorithm. This scheme corresponds to Algorithm 12, the so-called *half-step*

Rayleigh functional iteration. In other terms we obtain

$$\text{HSRFI:} \quad \begin{cases} \delta_+ & \leq & K_0^a \delta^2 \varepsilon, & \\ \gamma_{1/2} & \leq & 4K\delta_+\varepsilon & \leq & K_0^a \delta^2 \varepsilon^2, \\ \varepsilon_+ & \leq & \tilde{K}_0^a \delta_+ \varepsilon^2 & \leq & K_0^a \tilde{K}_0^a \delta^2 \varepsilon^3, \\ \gamma_+ & \leq & 4K\delta_+\varepsilon_+ & \leq & 4(K_0^a)^2 \tilde{K}_0^a \delta^4 \varepsilon^4. \end{cases}$$

Although the improvement is clearly visible in the exponents, it is in this form not comprehensible in terms of order statements.

Algorithm 12 *Half-step Rayleigh functional iteration (HSRFI)*

Input: (λ_0, u_0, v_0) *where* $u_0^H u_0 = v_0^H v_0 = 1$

for $k = 0, 1, 2, ...$

S1: *Solve* $\begin{bmatrix} T(\lambda_k) & \dot{T}(\lambda_k)u_k \\ v_k^H \dot{T}(\lambda_k) & 0 \end{bmatrix} \begin{bmatrix} s_k \\ \mu_k \end{bmatrix} = - \begin{bmatrix} T(\lambda_k)u_k \\ 0 \end{bmatrix},$

set $u_{k+1} = \dfrac{u_k + s_k}{\|u_k + s_k\|}$

S2: *Solve* ${v_k}^H T(\lambda_{k+1/2}) u_{k+1} = 0$ *for* $\lambda_{k+1/2} = p(u_{k+1}, v_k)$

S3: *Solve* $\begin{bmatrix} T(\lambda_{k+1/2})^H & \dot{T}(\lambda_{k+1/2})^H v_k \\ u_{k+1}^H \dot{T}(\lambda_{k+1/2})^H & 0 \end{bmatrix} \begin{bmatrix} t_k \\ \nu_k \end{bmatrix} = - \begin{bmatrix} T(\lambda_{k+1/2})^H v_k \\ 0 \end{bmatrix},$

set $v_{k+1} = \dfrac{v_k + t_k}{\|v_k + t_k\|}$

S4: *Solve* $v_{k+1}^H T(\lambda_{k+1}) u_{k+1} = 0$ *for* $\lambda_{k+1} = p(u_{k+1}, v_{k+1})$

The technique developed in [74]—which we have already used in Chapter 4—allows the analysis of the coupled sequences with respect to the R-order and gives Lemma 5.1, according to Algorithm 12. We neglect the term *two-sided* in the name of the method, since the half-step technique can be applied only to two-sided methods.

Lemma 5.1 *Under the assumptions and with the constants of Theorem 4.13, there exists* $\delta_0 > 0$*, and the half-step Rayleigh functional iteration in Algorithm 12 is well-defined for every initial triplet* (λ_0, u_0, v_0)*,* $(u_0, v_0) \in \mathcal{K}_{\varepsilon_0}(x_*) \times \mathcal{K}_{\varepsilon_0}(y_*)$*,* $\|u_0\| = \|v_0\| = 1$*, with* $\|u_0 - \frac{x_*}{v_0^H \dot{T}(\lambda_0) x_*}\| \leq \delta_0$*,* $\|v_0 - \frac{y_*}{u_0^H \dot{T}(\lambda_0)^H y_*}\| \leq \delta_0$*, and* $\lambda_0 \in \bar{S}(\lambda_*, \tau_0)$*, and converges with R-order 4.*

Proof. Theorem 4.13 was stated for the algorithm that corresponds to the inverse iteration step. With slightly different constants, it holds also for the algorithm in bordered matrix form. Hence, Algorithm 12 is well defined. Define γ, δ, ε as in (5.1). In these terms, inequality (4.28) gives $\delta_+ \leq K_0^a/(4K)\gamma\delta$, and $\varepsilon_+ \leq \tilde{K}_0^a/(4K)\gamma\varepsilon$. Then, the error seqences of the half-step Rayleigh functional iteration satisfy

$$\begin{array}{rcccl} \delta_+ & \leq & \frac{K_0^a}{4K}\gamma\delta, & & \\ \gamma_{1/2} & \leq & 4K\delta_+\varepsilon & \leq & K_0^a\gamma\delta\varepsilon, \\ \varepsilon_+ & \leq & \frac{\tilde{K}_0^a}{4K}\gamma_{1/2}\varepsilon & \leq & \frac{K_0^a\tilde{K}_0^a}{4K}\gamma\delta\varepsilon^2, \\ \gamma_+ & \leq & 4K\delta_+\varepsilon_+ & \leq & \frac{(K_0^a)^2\tilde{K}_0^a}{4K}\gamma^2\delta^2\varepsilon^2, \end{array}$$

in the k-th iteration, where $k \geq 2$. The matrix of exponents corresponding to the inequalities for δ_+, ε_+, γ_+ is given by

$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 2 \\ 2 & 2 & 2 \end{bmatrix},$$

which has spectral radius 4. Since the corresponding eigenvector is positive, the assumptions of [74, Theorem 1] are satisfied, and the sequences $\{\delta\}_n$, $\{\varepsilon\}_n$, $\{\gamma\}_n$ converge at least with R-order 4. □

For linear problems $T(\lambda)x = (A - \lambda I)x = 0$, the Rayleigh functional $p(u, v)$ reduces to the Rayleigh quotient $\frac{v^H Au}{v^H u}$, provided that $v^H u \neq 0$. Hence, the two-sided Rayleigh quotient iteration for nonnormal matrices [66, 67] can be accelerated in the same way by using Algorithm 12, and the R-order increases to 4. Moreover, there is no extra matrix-vector multiplication necessary, since for the second Rayleigh quotient in step S4 of Algorithm 12 only the left vector changes

$$\lambda_{k+1/2} = \frac{v_k^H A u_{k+1}}{v_k^H u_{k+1}} \quad \text{and} \quad \lambda_{k+1} = \frac{v_{k+1}^H A u_{k+1}}{v_{k+1}^H u_{k+1}}.$$

The same holds for the Rayleigh functional.

Notice that the half-step Rayleigh functional iteration can be regarded as an inverse iteration with two-sided Rayleigh functional update, where every second step computes a correction for the left eigenvector approximation. If we separate Algorithm 12 in two parts, where the second part is understood as next iteration step, then this method converges with R-order 2 like the inverse iteration (Algorithm 1), but delivers the left eigenvector as well. The information on the left eigenvector is particularly worthwile when the condition number of the eigenvalue is of interest. In §2.3 we have shown that the condition number $\kappa(\lambda_*)$ is proportional to $1/y_*^H \dot{T}(\lambda_*)x_*$. This means that during a two-sided method we can

estimate the condition number and check whether the obtained eigenvalue result makes sense. Another advantage of the half-step Rayleigh functional iteration compared to the inverse iteration is that the two-sided Rayleigh functional, computed by the two-sided method, is stationary in contrast to the one-sided Rayleigh functional, which is a possible λ-update for the inverse iteration, respectively the corresponding Newton step updates p_L from (3.64) and p_N from (3.66).

5.2 Half-step Generalized Rayleigh Functional Iteration

Next we apply the idea of updating the eigenvalue twice to the generalized Rayleigh functional iteration in Algorithm 10, for which the particular steps and their effects on convergence provided that $\lambda = p(u, v)$, are as follows

$$\text{GRFI:} \quad \begin{cases} (\lambda, v) & \mapsto & u_+ \sim T(\lambda)^{-1}v : & \sin\xi_+ & \leq & \tilde{K}_3|\,\lambda - \lambda_*|, \\ (\lambda, u) & \mapsto & v_+ \sim T(\lambda)^{-H}u : & \sin\eta_+ & \leq & \tilde{K}_3|\,\lambda - \lambda_*|, \\ (u_+, v_+) & \mapsto & \lambda_+,\ \text{s.t.},\ v_+^H T(\lambda_+)u_+ = 0 : & |\,\lambda_+ - \lambda_*| & \leq & 4K\sin\xi_+\sin\eta_+, \end{cases}$$

corresponding to inequality (4.45), and with $K > 0$ as before. If we insert the computation of a second Rayleigh functional in between the computation of the two vector updates we end up with

$$\text{HSGRFI:} \quad \begin{cases} (\lambda, v) & \mapsto & u_+ \sim T(\lambda)^{-1}v : & \sin\xi_+ & \leq & \tilde{K}_3|\lambda - \lambda_*|, \\ (u_+, v) & \mapsto & \lambda_{1/2},\ \text{s.t.},\ v^H T(\lambda_{1/2})u_+ = 0 : & |\lambda_{1/2} - \lambda_*| & \leq & 4K\sin\eta\sin\xi_+, \\ (\lambda_{1/2}, u_+) & \mapsto & v_+ \sim T(\lambda_{1/2})^{-H}u_+ : & \sin\eta_+ & \leq & \tilde{K}_3|\lambda_{1/2} - \lambda_*|, \\ (u_+, v_+) & \mapsto & \lambda_+,\ \text{s.t.},\ v_+^H T(\lambda_+)u_+ = 0 : & |\lambda_+ - \lambda_*| & \leq & 4K\sin\xi_+\sin\eta_+. \end{cases}$$

The algorithm is given in Algorithm 13.

The acceleration of convergence in terms of the R-order is stated in the following Lemma.

Lemma 5.2 *Under the assumptions and with the constants of Theorem 4.18, the half-step generalized Rayleigh functional iteration in Algorithm 13 is well-defined for every initial triplet (λ_0, u_0, v_0) with $(u_0, v_0) \in \mathcal{K}_{x_*}(\varepsilon_0) \times \mathcal{K}_{y_*}(\varepsilon_0)$, $\lambda_0 \in \bar{S}(\lambda_*, \tau_0)$, and the R-order of convergence is $(3 + \sqrt{5})/2 \approx 2.62$.*

Proof. Theorem (4.18) was shown for the explicitly bordered algorithm, hence we have a well-defined algorithm here. The error sequences from (5.1) satisfy the following inequalities

$$\begin{aligned} \delta_+ &\leq \tilde{K}_3\gamma, \\ \gamma_{1/2} &\leq 4K\delta_+\varepsilon &&\leq 4K\tilde{K}_3\gamma\varepsilon, \\ \varepsilon_+ &\leq \tilde{K}_3\gamma_{1/2} &&\leq 4K\tilde{K}_3^2\gamma\varepsilon, \\ \gamma_+ &\leq 4K\delta_+\varepsilon_+ &&\leq 16K^2\tilde{K}_3^3\gamma^2\varepsilon, \end{aligned}$$

due to (4.45) and (3.22). Note, that instead of the third inequality we could also use $\varepsilon_+ \leq K_3^a\delta_+\varepsilon \leq K_3^a\tilde{K}_3\gamma\varepsilon$ from (4.44), which yields just a different constant. Now, consider only the mapping $(\gamma, \varepsilon) \mapsto (\gamma_+, \varepsilon_+)$, hence the system

$$\begin{aligned} \gamma_+ &\leq Q_1\gamma^2\varepsilon, \\ \varepsilon_+ &\leq Q_2\gamma\varepsilon, \end{aligned} \tag{5.2}$$

with $Q_1 := 16K^2\tilde{K}_3^3$, $Q_2 := 4K\tilde{K}_3^2$. The matrix of exponents of system (5.2) with respect to γ, ε, is given by $\left[\begin{smallmatrix} 2 & 1 \\ 1 & 1 \end{smallmatrix}\right]$, which has spectral radius $(3+\sqrt{5})/2$. Since the corresponding eigenvector $(2/(\sqrt{5}-1), 1)^T$ is positive, the assumptions of [74, Theorem 1] are satisfied, and the sequences $\{\gamma\}_n$, $\{\varepsilon\}_n$ converge at least with R-order $(3+\sqrt{5})/2 \approx 2.62$. The same can be shown for $\{\delta\}_n$. □

Algorithm 13 *Half-step generalized Rayleigh functional iteration*

Input: (λ_0, u_0, v_0) *where* $u_0^H u_0 = v_0^H v_0 = 1$

for $k = 0, 1, 2, ...$

S1: *Set* $C_k = C(\lambda_k, u_k, v_k) = \begin{bmatrix} T(\lambda_k) & v_k \\ {u_k}^H & 0 \end{bmatrix}$

S2: *Solve* $C_k \begin{bmatrix} s_k \\ \mu_k \end{bmatrix} = -\begin{bmatrix} T(\lambda_k)u_k \\ 0 \end{bmatrix}$, *set* $u_{k+1} = \dfrac{u_k + s_k}{\|u_k + s_k\|}$

S3: *Solve* ${v_k}^H T(\lambda_{k+1/2})u_{k+1} = 0$ *for* $\lambda_{k+1/2} = p(u_{k+1}, v_k)$

S4: *Set* $C_{k+1/2} = \begin{bmatrix} T(\lambda_{k+1/2}) & v_k \\ {u_{k+1}}^H & 0 \end{bmatrix}$

S5: *Solve* $C_{k+1/2}^H \begin{bmatrix} t_k \\ \bar{\nu}_k \end{bmatrix} = -\begin{bmatrix} T(\lambda_{k+1/2})^H v_k \\ 0 \end{bmatrix}$, *set* $v_{k+1} = \dfrac{v_k + t_k}{\|v_k + t_k\|}$

S6: *Solve* $v_{k+1}^H T(\lambda_{k+1})u_{k+1} = 0$ *for* $\lambda_{k+1} = p(u_{k+1}, v_{k+1})$

The consequence of the improved order of convergence is certainly a slightly increased computational effort coming from the evaluation of the additional Rayleigh functional

on one hand and from updating the matrix C_k to $C_{k+1/2}$ on the other, which causes the preconditioner to be more inaccurate. Updating C_k can be cheap, depending on the underlying problem. Updating the Rayleigh functional involves only vector-vector multiplications, in general, since only the left vector changes.

The same procedure gives a *half-step simplified Rayleigh functional iteration*. The simplified Rayleigh functional iteration in Algorithm 6 was developed by dropping the derivatives $\dot{T}(\lambda_k)$ in the correction equations of the Rayleigh functional iteration. Since the convergence results for the sequences of errors for the (so) simplified Rayleigh functional iteration only differ in the constants compared to the results for the generalized Rayleigh functional iteration, the half-step simplified version is expected to have the same R-order as the half-step generalized Rayleigh functional iteration.

5.3 Half-step Residual Inverse Iteration

The idea of increasing convergence for generalized methods by including an additional Rayleigh functional is particularly effective for the two-sided residual inverse iteration method. For this method, linear systems are solved in general for several iterations with a fixed matrix $T(\sigma)$. When the eigenvalue approximation is updated after the first system has been solved, the matrix of the second system does not change and the preconditioner is still supposed to be a good approximation for $T(\sigma)$. For the other half-step methods, the whole matrix, resp. bordered matrix, changes depending on the new value $\lambda_{k+1/2}$.

The half-step (two-sided) residual inverse iteration method is presented in Algorithm 14.

Algorithm 14 *Half-step two-sided residual inverse iteration*

Input: (λ_0, u_0, v_0), *normalization vectors* w_x, w_y, *such that* $w_x^H u_0 = 1$, $w_y^H v_0 = 1$

for $k = 0, 1, 2, ...$

S1: *Solve* $v_k^H T(\lambda_{k+1/2}) u_k = 0$ *for* $\lambda_{k+1/2}$

S2: *Compute the residual* $r_k^x = T(\lambda_{k+1/2}) u_k$

S3: *Solve* $T(\lambda_0) s_k = r_k^x$ *for* s_k, *set* $x_{k+1} = u_k - s_k$, $u_{k+1} = x_{k+1}/w_x^H x_{k+1}$

S4: *Solve* $v_k^H T(\lambda_{k+1}) u_{k+1} = 0$ *for* λ_{k+1}

S5: *Compute the residual* $r_k^y = T(\lambda_{k+1})^H v_k$

S6: *Solve* $T(\lambda_0)^H t_k = r_k^y$ *for* t_k, *set* $y_{k+1} = v_k - t_k$, $v_{k+1} = y_{k+1}/w_y^H y_{k+1}$

An increased R-order of convergence compared to the two-sided residual inverse iteration

can only be determined if we suppose that the shift λ_0 is updated by the second last approximation every time, i.e., by λ_k in S3 and by $\lambda_{k+1/2}$ in S6 of Algorithm 14.

Lemma 5.3 *Let λ_* be a simple eigenvalue of $T(\lambda)$, and suppose $T(\lambda)$ is twice continuously differentiable, x_* is the corresponding right eigenvector normalized by $w_x^H x_* = 1$, y_* is the corresponding left eigenvector normalized by $w_y^H y_* = 1$. Then the half-step two-sided residual inverse iteration converges for all λ_0 sufficiently close to λ_*, if $\xi_0 = \measuredangle(\mathrm{span}\{u_0\}, \mathrm{span}\{x_*\}) \leq \pi/3$, $\eta_0 = \measuredangle(\mathrm{span}\{v_0\}, \mathrm{span}\{y_*\}) \leq \pi/3$. If the shift λ_0 in Algorithm 14 is constantly updated by the second last eigenvalue approximation, then the R-order of convergence is approximately 2.8.*

Proof. The first part follows with Theorem 4.17. We define the terms $\gamma = |\lambda_k - \lambda_*|$, $\gamma_{1/2} = |\lambda_{k+1/2} - \lambda_*|$,$\gamma_+ = |\lambda_{k+1} - \lambda_*|$, $\delta = \|u_k - x_*\|$, $\delta_+ = \|u_{k+1} - x_*\|$, $\varepsilon = \|v_k - y_*\|$, $\varepsilon_+ = \|v_{k+1} - y_*\|$, in order to get the following scheme for the half-step residual inverse iteration, where the inequalities have been given in Theorem 4.17

$$\begin{aligned} \gamma_{1/2} &\leq 4K\delta\varepsilon, \\ \delta_+ &\leq K_2^d\gamma\delta, \\ \gamma_+ &\leq 4K\delta_+\varepsilon \leq 4KK_2^d\gamma\delta\varepsilon, \\ \varepsilon_+ &\leq K_2^d\gamma_{1/2}\varepsilon \leq 4KK_2^d\delta\varepsilon^2, \end{aligned}$$

with $K_2^d > 0$ from the proof of Theorem (4.17), and $K > 0$ as before. The matrix of exponents of γ, δ, ε for the last three inequalities is

$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix},$$

the largest eigenvalue of which is approximately 2.8, where the corresponding eigenvector is positive in all components. This means that the R-order of the half-step two-sided residual inverse iteration method is approximately 2.8. □

This is a theoretical result, which will not be achieved since, in applications the matrix $T(\sigma)$ will be updated only if necessary. However, this rate is better than the one for the generalized Rayleigh functional iteration.

5.4 Numerical Experiments

The convergence acceleration for half-step methods arises from using the new information which is obtained by solving the first linear system immediately. Therefore, one has to wait

until this information is at hand, before proceeding with the second system is possible. The disadvantage of this principle is that the two linear systems of each iteration cannot be solved simultaneously. However, if we split the half-step methods in two parts such that only one linear system is solved per iteration, cf. the alternating Rayleigh functional iteration, then we end up with methods comparable with the one-vector methods of Section 4.1. There, the linear systems have to be solved one after the other.

The additional Rayleigh functional causes only $\mathcal{O}(n)$ costs since the matrix-vector multiplications for solving $v_k^H T(\lambda_{k+1/2})u_{k+1} = 0$ and $v_{k+1}^H T(\lambda_{k+1})u_{k+1} = 0$ stay the same.

We begin by testing the following nonsymmetric linear eigenproblem.

Example 6 2d-Convection-Diffusion *The finite differences discretization of the convection-diffusion operator*

$$-\nu\Delta u + \mu\nabla u = \lambda u \quad in \quad \Omega = [0,1]^2,$$

as defined in [33] with homogeneous Dirichlet boundary conditions, yields the linear eigenproblem $Ax = \lambda x$, where A is nonsymmetric and has 5 nonzero diagonals. The departure from normality of the operator increases as the diffusion coefficient $\nu \in \mathbb{R}$ decreases and as the convection coefficient $\mu \in \mathbb{R}^2$ increases; cf. [33]. In the following we will set $\mu_1 = \mu_2$ where $\mu = [\mu_1, \mu_2]^T$.

Table 5.1 shows the results for the discussed half-step methods applied to Example 6 with parameters $\nu = 1$ and $\mu_1 = 0.2$, compared to the results for the original methods; see p. 109 for the terms in the header of the table. Figure 5.1 shows residual norms versus

	$\lvert\lambda_{it} - \lambda_*\rvert$	it	$\lVert res_u \rVert$	$\lVert res_v \rVert$	gm_1	gm_2	gm_1/it	gm_2/it	cpu
sRFI	9.1e-11	7	5.2e-13	5.3e-13	56	56	8.0	8.0	0.74
HSsRFI	9.1e-11	5	1.3e-08	4.9e-11	38	41	7.6	8.2	0.61
RFI	9.1e-11	7	5.2e-13	5.3e-13	56	56	8.0	8.0	0.82
HSRFI	9.1e-11	5	1.3e-08	4.9e-11	38	41	7.6	8.2	0.68
GRFI	9.1e-11	9	5.1e-13	5.0e-13	70	69	7.8	7.7	1.09
HSGRFI	9.1e-11	6	1.8e-10	1.5e-12	47	49	7.8	8.2	0.81
TRII	9.1e-11	12	5.0e-13	4.7e-13	93	96	7.8	8.0	1.00
HSTRII	9.1e-11	8	4.9e-12	2.1e-12	64	59	8.0	7.4	0.62

Table 5.1: *Convection-diffusion discretization, $n = 1024$.*

iteration numbers. Since we have a linear problem, the Rayleigh functional iteration and the simplified Rayleigh functional iteration are exactly the same, the graphs lie on top

of each other. We can take the Rayleigh quotient $(v_k^H A u_k)/v_k^H u_k$ as update for λ_k for all methods.

The nonnomality of A is slight; all methods compute the smallest eigenvalue and corresponding eigenvectors accurate. All half-step methods converge faster regarding time and iterations than the original methods. Greatest savings occur for the half-step two-sided residual inverse iteration method, which is approximately 40% faster than the two-sided residual inverse iteration method.

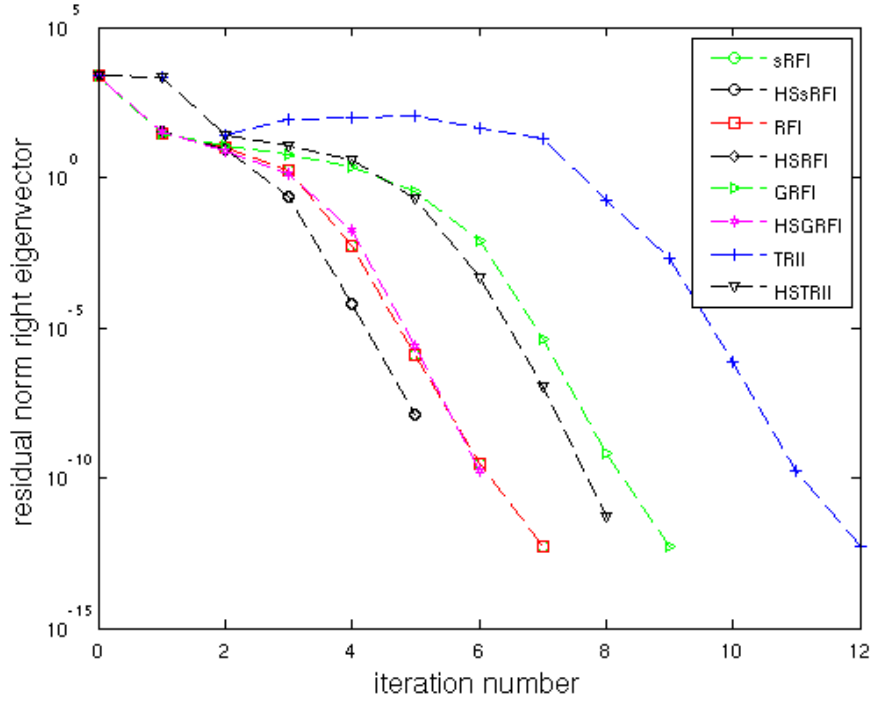

Figure 5.1: *Example convection-diffusion discretization, order* $n = 1024$.

Next, we run the unsymmetric quadratic problem given in Example 2. Results are presented in Table 5.2 and Figure 5.2. We observe faster convergence for all half-step methods again, in iteration numbers and consumed time. The generalized Rayleigh functional iteration spends 6 iterations more than its half-step variant, and stays behind its expectations. All methods converged to the smallest eigenvalue in absolute value. The simplified Rayleigh functional iteration acts like the Rayleigh functional iteration, and so do their half-step versions.

Figure 5.2 reveals also the relativity of the convergence acceleration for all methods. In particular, the difference between the iteration numbers of original and half-step method increases with the iteration number corresponding to the original method, i.e., for the

	$\lvert\lambda_{it} - \lambda_*\rvert$	it	$\lVert res_u \rVert$	$\lVert res_v \rVert$	gm_1	gm_2	gm_1/it	gm_2/it	cpu
sRFI	3.9e-06	9	3.1e-14	7.5e-16	64	67	7.1	7.4	0.19
HSsRFI	3.9e-06	7	5.5e-10	1.2e-13	49	63	7.0	9.0	0.16
RFI	3.9e-06	9	4.1e-14	8.2e-16	64	67	7.1	7.4	0.19
HSRFI	3.9e-06	7	1.4e-09	1.7e-13	48	62	6.9	8.9	0.17
GRFI	3.9e-06	20	3.1e-11	3.1e-11	84	84	4.2	4.2	0.29
HSGRFI	3.9e-06	14	2.3e-10	4.1e-11	60	72	4.3	5.1	0.24
TRII	3.9e-06	14	6.2e-15	6.5e-15	83	84	5.9	6.0	0.65
HSTRII	3.9e-06	11	1.8e-10	4.5e-13	60	57	5.5	5.2	0.52

Table 5.2: QEP_1, $a = 0.2$, $n = 100$.

two fastest methods sRFI and RFI the acceleration exhibits in an amount of 2 iterations, the slower two-sided residual inverse iteration takes 3 iterations more than its half-step version, and the slowest method, the GRFI, needs 6 iterations more than the HSGRFI.

Furthermore, for this example the two-sided residual inverse iteration converges with equal rate as the half-step generalized Rayleigh functional iteration.

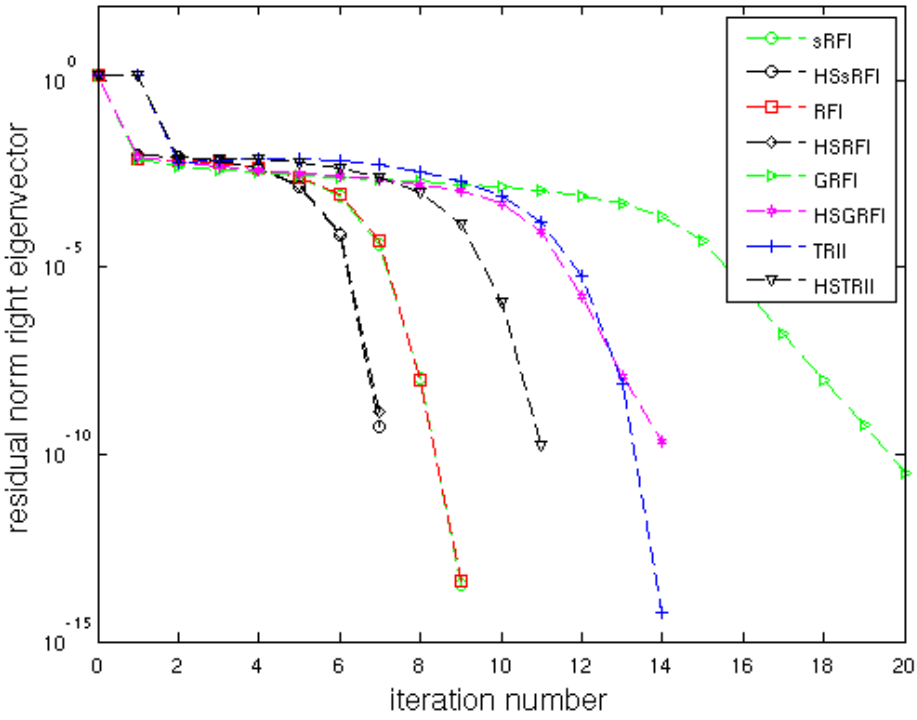

Figure 5.2: *Example* QEP_1 *for* $n = 100$.

For the partial delay differential equation discretization, see Example 5, all methods converge to the same eigenvalue, except for the diverging residual inverse iteration methods, see Table 5.3 and Figure 5.3. For this example we have already observed divergence of the RII methods in Chapter 4.

	$\lvert\lambda_{it}\rvert$	it	$\lVert res_u\rVert$	$\lVert res_v\rVert$	gm_1	gm_2	gm_1/it	gm_2/it	cpu
sRFI	1.9e+01	7	6.2e-14	7.2e-14	17	16	2.4	2.3	0.11
HSsRFI	1.9e+01	4	5.5e-14	4.8e-14	9	20	2.2	5.0	0.10
RFI	1.9e+01	7	1.0e-13	1.2e-13	17	16	2.4	2.3	0.13
HSRFI	1.9e+01	4	6.5e-14	6.8e-14	9	20	2.2	5.0	0.11
GRFI	1.9e+01	7	5.1e-14	5.9e-14	17	15	2.4	2.1	0.13
HSGRFI	1.9e+01	4	6.4e-14	7.2e-14	9	20	2.2	5.0	0.13
TRII	1.5e+01	100	1.2e+07	1.3e+07	101	101	1.0	1.0	0.49
HSTRII	2.1e+01	100	3.1e+09	3.2e+09	101	101	1.0	1.0	0.52

Table 5.3: *Example 5 PDDE,* $n = 50$.

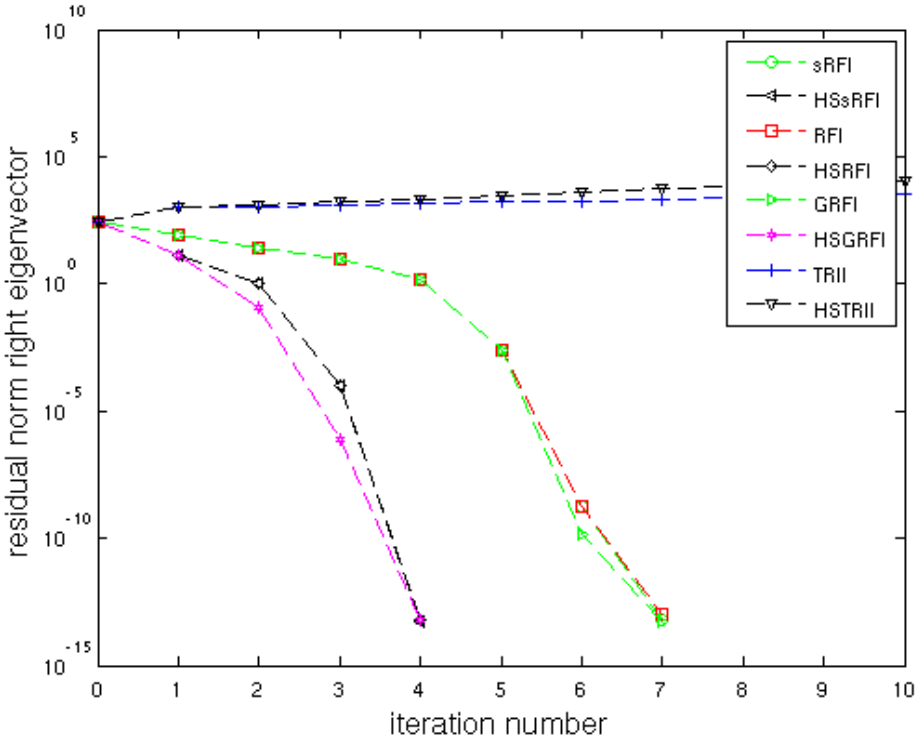

Figure 5.3: *Example PDDE with* $n = 50$.

The second column of Table 5.3 gives the absolute value of the final eigenvalue approximation, since we have no information on the exact eigenvalues. The half-step versions are in any case 3 iterations faster than the original methods. As before, the curves for the Rayleigh functional iteration and the simplified Rayleigh functional iteration lie almost on top of each other, the differences to the generalized Rayleigh functional iteration are little.

Chapter 6

Jacobi–Davidson-type Methods

6.1 Introduction

Chapters 4 and 5 discuss locally convergent Newton-type methods for nonlinear eigenvalue problems. Naturally, the methods are highly sensitive with respect to the starting values. Often, an erratic and chaotic behavior can be observed, cf. [14]. For linear problems $T(\lambda) = A - \lambda I$, it is well-known [8,9] that the basins of attraction for the Rayleigh quotient iteration may collapse around attractors when the eigenvalues of A are not well-separated. This property holds even for the case of the Grassmann–Rayleigh quotient iteration [1], i.e., the Rayleigh quotient on a p-dimensional subspace *(Grassmann manifold)*, and for the two-sided Grassmann–Rayleigh quotient iteration [2], which is based on the two-sided Rayleigh quotient iteration.

In order to avoid the possible failure of Newton-type methods, one strategy is to constantly expand a search space by collecting directions which have been already computed in the previous step, typically orthogonalized. If the starting vectors are not inside the basins of attraction, one of the subsequent vectors will hopefully be—assuming that sufficient time and disk space are available—and the method will converge, sooner or later. The idea of building a search space by an orthogonal matrix coupled with the Newton method is realized in the Jacobi–Davidson method, which brings together the ideas of Jacobi and Davidson, see [81] for a historical survey and details. The correction vectors are obtained by a (preconditioned) Newton step, and new approximations are derived by solving a small dimensional projected nonlinear eigenproblem; cf. the next section for details. The eigenvalue approximations are obtained by a nonlinear Rayleigh–Ritz step—the one-dimensional form of which is just the Rayleigh functional.

Hence, not only is the Jacobi–Davidson method related to the inverse iteration and the

Rayleigh functional iteration, it also needs methods like the inverse iteration for the computation of the projected nonlinear eigenproblems that occur in every iteration.

Therefore, one- and two-vector methods are an important basis for subspace expanding eigenvalue methods in two ways: They are essential as routines when projected nonlinear problems have to be solved, and they show ways to compute correction vectors. Basically, there are two ways to improve the approximations to eigenvectors. One is to take the residual inverse iteration update S3 in Algorithm 2, which leads to the nonlinear Arnoldi method, see [94]. The other is to take the inverse iteration step, resp. Newton step, S1 in Algorithm 1, which leads to Jacobi–Davidson-type methods. Another algorithm is given by the rational Krylov method, see [73], which can also be written as an iterative projection method [44]. There, the search space is updated by the orthogonalized residual.

This chapter summarizes different Jacobi–Davidson algorithms suited for nonlinear eigenvalue problems, which are based on the Rayleigh functional methods of Chapter 4. If necessary, adaptations are made and convergence properties with respect to included Rayleigh functionals are shown. The methods are analyzed regarding their properties in the asymptotic case and we will observe that the two-sided and the generalized variant are less (or even not) sensitive with respect to the condition number of the approximated eigenvalue.

6.2 Nonlinear Jacobi–Davidson

The Jacobi–Davidson algorithm has established itself over the last decade as one of the most promising methods in eigenvalue computations especially for large sparse systems. It was introduced in 1996 by Sleijpen and van der Vorst [81] for linear problems, generalized for generalized linear problems [22] and formulated for polynomial eigenproblems in [7]. The nonlinear version was published in 2004 by Betcke and Voss [13], see also [97].

The key idea in order to avoid the local sensitivity of the one-sided methods, is to save the computed corrections s by extending a subspace $\mathcal{U}$. Let U be the associated basis. Then, we get an approximation (λ, u) for an eigenvalue and -vector by imposing the Ritz–Galerkin condition on the residual, i.e.,

$$r := T(\lambda)u \perp \mathcal{U}. \tag{6.1}$$

Writing $u = Uc$, $c \in \mathbb{C}^k$, leads to the size $k \times k$ eigenproblem

$$U^H T(\lambda) U c = 0. \tag{6.2}$$

For $k = 1$ we can set $c = 1$ and (6.2) becomes the Rayleigh functional defining equation (3.12), when we suppose $\lambda \equiv \lambda(u)$. In this case and for sufficiently good starting values we know that the result approximates the eigenvalue with first order in terms of the angle ξ between u and x_*. If $T(\lambda_*) = T(\lambda_*)^H$ then the order is quadratic. For $k > 1$ the problem becomes more complicated, since we need to determine c and λ. It is shown in [27] that, for some i vectors with $(i < k)$, the so approximated eigenpairs $(\lambda_i, u_i \equiv Uc_i)$ converge linearly to the eigenpairs (λ_{*i}, x_{*i}) of problem (1.1). Furthermore, it is pointed out that this result is weaker than the one for linear eigenproblems where the error estimate on the eigenvalue λ_{*i} is $\mathcal{O}(|\lambda_{*i}/\lambda_{i-1}|^k)$ instead of $\mathcal{O}(|\lambda_{i-2}/\lambda_{i-1}|^k)$, which is the order in the nonlinear case.

The second essential part of the Jacobi–Davidson method is the correction equation

$$\left(I - \frac{\dot{T}(\lambda)uu^H}{u^H\dot{T}(\lambda)u}\right) T(\lambda)(I - uu^H)s = -T(\lambda)u, \tag{6.3}$$

which is solved for s in order to improve the approximation u. We will see that this equals the Newton step (4.8) with $x = u + s$, $w = u$ for the extended system (1.11). The first equation of (4.8) equals

$$T(\lambda)(u + s) + \dot{T}(\lambda)u\mu = 0. \tag{6.4}$$

After equation (6.4) is multiplied from the left with u^H, we can rearrange it with respect to μ and obtain $\mu = -(u^HT(\lambda)u + u^HT(\lambda)s)/(u^H\dot{T}(\lambda)u)$ which, inserted back in (6.4), gives

$$\left(I - \frac{\dot{T}(\lambda)uu^H}{u^H\dot{T}(\lambda)u}\right) T(\lambda)s = -\left(I - \frac{\dot{T}(\lambda)uu^H}{u^H\dot{T}(\lambda)u}\right) T(\lambda)u.$$

Since $u \perp r$, this yields (6.3) when we assume additionally that $s \perp u$. Furthermore, we have to make the additional assumption that

$$u^H\dot{T}(\lambda)u \neq 0, \tag{6.5}$$

to guarantee that the left projector is well-defined. If $x_*^H\dot{T}(\lambda_*)x_* \neq 0$, then (6.5) will be satisfied for (λ, u) close to the solution (λ_*, x_*). If the problem is Hermitian with real eigenvalues, then (6.5) is satisfied in the neighborhood of (λ_*, x_*) when λ_* is algebraically simple. A numerical example shows what happens in case $x_*^H\dot{T}(\lambda_*)x_* = 0$, where T is non-Hermitian with simple eigenvalue λ_*, cf. Section 6.6.2. The algorithm is stated in Algorithm 15.

If we are sufficiently close to the eigenvector, then we can expect quadratic convergence if the correction equation (6.3) is solved exactly. As was already observed for linear problems, cf. [81], the correction equation does not need to be solved up to high accuracy during the

first iterations, or whenever the residual norm is large. It is state of the art to decrease the tolerance for the termination of the iterative solver corresponding to the correction equation in the same way as the residual $r = T(\lambda)u$ decreases. However, in [80] it is shown, that for Krylov solvers in general, and also in the eigenvalue context of the inexact Arnoldi method, the accuracy in the computation of linear system can even be effectively relaxed, i.e., to obtain convergence, tolerances do not need to be decreased. When the correction equation is solved inexactly, then the whole method refers to an inexact Rayleigh quotient iteration, see e.g. [63]. The equivalence of a simplified preconditioned Jacobi–Davidson method and the inexact Rayleigh quotient iteration with altered preconditioner is shown in [23]. To our knowledge such results exists only for linear eigenvalue problems. In addition, see [104] for studies on the linear Jacobi–Davidson, Rayleigh quotient iteration and Newton updates, in general.

Algorithm 15 *Nonlinear Jacobi–Davidson method*

Input: *eigenvector approximation* u *with* $u^H u = 1$*, tolerance* $\epsilon > 0$

S0: *Set* $U = [\,u\,]$

for $k = 1, 2, \ldots$

S1: *Solve* $U^H T(\lambda) U c = 0$ *for* (λ, c)

S2: $u = Uc,\ r = T(\lambda)u$

S3: **if** $\|r\|/\|u\| < \epsilon$ **stop**

S4: *Compute* $s \perp u$ *from*

$$\left(I - \frac{\dot{T}(\lambda)uu^H}{u^H\dot{T}(\lambda)u}\right) T(\lambda)(I - uu^H)s \ = -r$$

S5: *MGS*(U, s)*, normalize* s*, set* $U = [\,U, s\,]$

In [51] the procedure in S2 is also called nonlinear Rayleigh–Ritz method. A nonlinear eigenproblem has to be solved in this step. In the optimal case, all eigenvalues of the $k \times k$ problem are returned, to enable a selection and force convergence to the desired eigenpair. The number of eigenpairs of the projected problem depends on the number of eigenpairs of the basic problem and the size of the search space. Precisely, the linear $k \times k$ problem has exactly k pairs, the polynomial problem of degree m size k has up to $k \times m$ pairs. The MATLAB polyeig-function can be used for polynomial eigenproblems and returns all eigenpairs. Once started, the selection mode should be kept for the duration of the whole algorithm. To obtain eigenvalues from the inner spectrum is a more complicated issue, see for instance [38].

For non-polynomial problems it is often too complicated to compute all eigenpairs. We can choose one of the methods of Chapter 4, which depend highly on the given starting values, but in general they will converge rather to the eigenvalue smallest in absolute value than to a large one. The chaotic behavior that may be observed in some cases, leads to unpredictable results. It may happen that the target eigenvalue changes in every iteration, hence the vector u changes and corrections will be computed ever for different vectors. In this case, even the expansion of the subspace does not guarantee a final convergence. Therefore, adequate starting pairs must be provided externally, for example by solving the linear problem that is generated by the truncated Taylor series close to a given shift σ, cf. (2.4).

Let us discuss the remaining steps of Algorithm 15 briefly. The correction equation in S5 can be solved by a preconditioned Krylov solver. Section 6.4 provides the details for the way of treating the preconditioner. When we have computed a correction s, then we orthogonalize it with respect to previous vectors in order to extract the new information and normalize it. The term MGS in S6 means a modified Gram–Schmidt procedure. The search space is expanded by the so gained s. Recall that for the new vector $x = u + s$, we have $\operatorname{im}[U, x] = \operatorname{im}[U, s]$, since $u \in \operatorname{im} U$. As for the bordered matrices of the one- and two-vector methods, to avoid rounding errors it is more convenient to work with the corrections s, than computing an updated $x = u + s$.

6.3 Generalized Jacobi–Davidson-type Methods

In general, the Rayleigh–Ritz step in the Jacobi–Davidson method will give approximations for the eigenvalue of first order in the angle (we have shown this for $k = 1$ and we expect the same for $k > 1$). If we want to improve these approximations, we can use the ideas and analysis of Chapter 3, which provides a higher order bound for the generalized Rayleigh functional. Again, this can only be transferred directly to the $k = 1$ case, but we expect it to work for larger k as well.

Let $\mathcal{V}$ be the search space associated to the left eigenvector approximations, and let V be the associated basis. We impose Petrov–Galerkin conditions on the right and left vector residuals, i.e.,

$$r_u := T(\lambda)u \perp \mathcal{V}, \qquad r_v := T(\lambda)^H v \perp \mathcal{U}, \tag{6.6}$$

which, with $u = Uc$, $v = Vd$, yield

$$V^H T(\lambda)Uc = 0, \qquad U^H T(\lambda)^H Vd = 0. \tag{6.7}$$

Solving these in general tiny nonlinear eigenvalue problems takes hardly time nor extra storage. However, solving (6.7) is the crucial point for all nonlinear Jacobi–Davidson-type methods, since the further steps are based on the approximations (λ, u, v), which are obtained from the solutions (λ, c, d) of (6.7). And since the problems are small, we can afford methods with high costs—the main issue is that we end up with proper approximations. Numerical examples are discussed in §6.6.

Compared to Algorithm 15, the left eigenvector approximations v need additional memory. In order to obtain corrections t for v, a second correction equation has to be solved. Based on the correction equation (6.3), the naive correction equation for the left eigenvector would be

$$\left(I - \frac{\dot{T}(\lambda)^H v v^H}{v^H \dot{T}(\lambda)^H v}\right) T(\lambda)^H (I - vv^H)t = -T(\lambda)^H v. \tag{6.8}$$

Obviously, there is an additional effort that comes with solving the linear system (6.8). In some cases the gained convergence enhancement makes this generalized version of JD not only faster in terms of iteration numbers but also in computational time. If the left eigenvector is sought as well, then this seems to be a good choice. Moreover, if one has computed a preconditioner for $T(\lambda)$, then its Hermitian transposed can be used for the second system as well. Note that for this version we have to require the additional assumptions $u^H \dot{T}(\lambda) u \neq 0$ and $v^H \dot{T}(\lambda)^H v \neq 0$. Therefore, it would be more natural for generalized versions to use the term $v^H \dot{T}(\lambda) u$ in the denominators of the projectors, which is nonzero close to an eigentriplet with simple eigenvalue. We will derive two methods in this style.

A different linear Jacobi–Davidson variant, namely the *two-sided Jacobi–Davidson method*, is proposed in [37]. It computes left and right eigenvectors for linear eigenproblems. There, a note on the generalization of the method for polynomial eigenvalue problems is given, which is suited for nonlinear problems as well. Petrov vectors are obtained as in equations (6.6), (6.7), the correction equations will be as follows

$$\begin{aligned} \left(I - \frac{\dot{T}(\lambda) u v^H}{v^H \dot{T}(\lambda) u}\right) T(\lambda) \left(I - \frac{\dot{T}(\lambda) u v^H}{v^H \dot{T}(\lambda) u}\right) s &= -T(\lambda) u, \\ \left(I - \frac{\dot{T}(\lambda)^H v u^H}{u^H \dot{T}(\lambda)^H v}\right) T(\lambda)^H \left(I - \frac{\dot{T}(\lambda)^H v u^H}{u^H \dot{T}(\lambda)^H v}\right) t &= -T(\lambda)^H v. \end{aligned} \tag{6.9}$$

where $s \perp v$ and $t \perp u$, this means that $[u, s]$ and $[v, t]$ are bi-orthogonal. The systems can be derived from the Newton systems with normalization conditions as follows

$$\begin{bmatrix} T(\lambda) & \dot{T}(\lambda) u \\ v^H & 0 \end{bmatrix} \begin{bmatrix} s \\ \mu \end{bmatrix} = \begin{bmatrix} -T(\lambda) u \\ 0 \end{bmatrix}, \qquad \begin{bmatrix} T(\lambda)^H & \dot{T}(\lambda)^H v \\ u^H & 0 \end{bmatrix} \begin{bmatrix} t \\ \nu \end{bmatrix} = \begin{bmatrix} -T(\lambda)^H v \\ 0 \end{bmatrix}.$$

As we have discussed in Chapter 4 when comparing the Rayleigh quotient iteration and its simplified variant, the bi-orthogonality condition makes not too much sense in the nonlinear case. Therefore one either changes (6.9) to

$$\left(I - \frac{\dot{T}(\lambda)uv^H}{v^H\dot{T}(\lambda)u}\right) T(\lambda)\left(I - uu^H\right) s \;=\; -T(\lambda)u, \tag{6.10}$$

with the standard orthogonalization $s \perp u$, or one takes the correction equations that correspond to the systems

$$\begin{bmatrix} T(\lambda) & \dot{T}(\lambda)u \\ v^H\dot{T}(\lambda) & 0 \end{bmatrix}\begin{bmatrix} s \\ \mu \end{bmatrix} = \begin{bmatrix} -T(\lambda)u \\ 0 \end{bmatrix}, \quad \begin{bmatrix} T(\lambda)^H & \dot{T}(\lambda)^H v \\ u^H\dot{T}(\lambda)^H & 0 \end{bmatrix}\begin{bmatrix} t \\ \nu \end{bmatrix} = \begin{bmatrix} -T(\lambda)^H v \\ 0 \end{bmatrix}, \tag{6.11}$$

which we have already used for the bordered version of the Rayleigh functional iteration in Algorithm 5. The correction equation in terms of projectors for the left system in (6.11) is given by

$$\left(I - \frac{\dot{T}(\lambda)uv^H}{v^H\dot{T}(\lambda)u}\right) T(\lambda)\left(I - \frac{uv^H\dot{T}(\lambda)}{v^H\dot{T}(\lambda)u}\right) s \;=\; -T(\lambda)u, \tag{6.12}$$

with the condition $s \perp \dot{T}(\lambda)^H v$. All versions are variants of Newton's method, and we expect locally cubic convergence, cf. Theorem 4.13 for the basic two-vector method. We implemented equation (6.10) and its dual equation for the two-sided Jacobi–Davidson method, which is given in Algorithm 16, and seems to be new for nonlinear eigenvalue problems.

Another generalization of the Jacobi–Davidson method suitable for nonlinear eigenproblems was introduced in [78]. Its origin lies in the singularity system (4.41) and the corresponding two-vector Algorithm 10, the generalized Rayleigh functional iteration (GRFI). GJD—as the Jacobi–Davidson formulation of the two-vector method GRFI is called—was shown to be a locally quadratically convergent method for nonlinear eigenvalue problems [78]. The motivation for the special structure of the matrix $C(\lambda, v, u)$, cf. (4.10), that defines the GRFI and hence the GJD, is the same as in the case of the two-vector method, i.e., we want to develop a method which is less sensitive with respect to the condition number of the approximated eigenvalue, than standard Newton-type methods, where the norm of the inverse Jacobian depends on the condition number of the eigenvalue, see (4.43). For linear problems, an ill-conditioned eigenvalue has left and right eigenvectors with an almost perpetual angle. For nonlinear problems, the term $y_*^H\dot{T}(\lambda_*)x_*$ is close to zero, when λ_* is ill-conditioned.

Algorithm 16 *Two-sided Jacobi–Davidson method*

Input: *approximations* (u, v) *such that* $u^H u = v^H v = 1$, *tolerance* $\epsilon > 0$

S0: *Set* $U = [\,u\,],\ V = [\,v\,]$

for $k = 1, 2, \ldots$

S1: *Solve* $V^H T(\lambda) U c = 0$ *and* $U^H T(\lambda)^H V d = 0$ *for* (λ, c, d)

S2: $u = Uc,\ v = Vd,\ r_u = T(\lambda)u,\ r_v = T(\lambda)^H v$

S3: **if** $\min\{\|r_u\|/\|u\|, \|r_v\|/\|v\|\} < \epsilon$ **stop**

S4: *Compute* $s \perp u,\ t \perp v$ *from*

$$\left(I - \frac{\dot{T}(\lambda) u v^H}{v^H \dot{T}(\lambda) u}\right) T(\lambda) \left(I - uu^H\right) s \ = \ -r_u$$

$$\left(I - \frac{\dot{T}(\lambda)^H v u^H}{u^H \dot{T}(\lambda)^H v}\right) T(\lambda)^H \left(I - vv^H\right) t \ = \ -r_v$$

S5: *MGS*(U, s), *MGS*(V, t), *normalize* s, t, *set* $U = [\,U, s\,],\ V = [\,V, t\,]$

We will briefly deduce the method. To determine the new approximations $u \in \mathcal{U}$ and $v \in \mathcal{V}$, we consider the Petrov–Galerkin conditions (6.6), and take the solutions of (6.7) to get $u = Uc$, $v = Vd$ and λ. Dependent on the type and the size of the projected problem, one gets a number of eigentriplets (λ, c, d) and has in this sense the opportunity to select an appropriate triplet (λ, u, v) with $v^H T(\lambda) u = 0$, and force convergence in a special direction in contrast to the single- or two-vector algorithms. This depends strongly on the method that is used to solve the projected problem.

The approximations u, v are improved by solving correction equations which are equivalent to the primal and dual system (4.41) of the GRFI. We deduce the first correction equation for GJD from the first equation of (4.41), i.e.,

$$\begin{bmatrix} T(\lambda) & v \\ u^H & 0 \end{bmatrix} \begin{bmatrix} s \\ \mu \end{bmatrix} = \begin{bmatrix} -r_u \\ 0 \end{bmatrix} \quad \Longleftrightarrow \quad \begin{matrix} T(\lambda)s + v\mu = -r_u \\ u^H s = 0. \end{matrix}$$

Multiplying the upper equation from the left by v^H gives $v^H T(\lambda)s + v^H v\mu = -v^H r_u = 0$, since $v^H r_u = v^H T(\lambda)u = 0$. Hence, we have $\mu = -v^H T(\lambda)s$. Substituting μ back in the upper equation yields $T(\lambda)s - vv^H T(\lambda)s = -r_u$. Since $(I - uu^H)s = s$, we end up with

$$(I - vv^H)T(\lambda)(I - uu^H)s = -r_u, \quad u^H s = 0. \tag{6.13}$$

The same procedure for the adjoint system produces the adjoint correction equation

$$(I - uu^H)T(\lambda)^H(I - vv^H)t = -r_v, \quad v^H t = 0. \tag{6.14}$$

The solutions s, t can be obtained by any preconditioned Krylov solver, as for the standard JD. They are used for expanding the subspaces after orthogonalization with respect to U, V, resp., using modified Gram–Schmidt (MGS) and subsequent normalization. The algorithm is presented in Algorithm 17. We terminate if at least one of the residuals is sufficiently small, in order to have a fair comparison with the standard nonlinear Jacobi–Davidson method, which determines only right eigenvectors. Though, if both eigenvectors are of interest, in general, one further step can be enough to compute the second one to satisfying accuracy. However, as we will see, there are examples for which it takes longer.

The major difference between the correction formulae (6.13), (6.14), the nonlinear Jacobi–Davidson correction formula (6.3) and the two-sided variants (6.9), (6.10) and (6.12) is the absence of the derivative $\dot{T}(\lambda)$ in the projectors and the fact that only orthogonal and no skew projectors occur.

Algorithm 17 *Generalized Jacobi–Davidson method*

Input: *approximations* (u, v) *such that* $u^H u = v^H v = 1$, *tolerance* $\epsilon > 0$

S0: *Set* $U = [\,u\,],\ V = [\,v\,]$

for $k = 1, 2, \ldots$

S1: *Solve* $V^H T(\lambda) U c = 0$ *and* $U^H T(\lambda)^H V d = 0$ *for* (λ, c, d)

S2: $u = Uc,\ v = Vd,\ r_u = T(\lambda)u,\ r_v = T(\lambda)^H v$

S3: **if** $\min\{\|r_u\|/\|u\|, \|r_v\|/\|v\|\} < \epsilon$ **stop**

S4: *Compute* $s \perp u,\ t \perp v$ *from*

$$(I - vv^H)T(\lambda)(I - uu^H)s = -r_u$$
$$(I - uu^H)T(\lambda)^H(I - vv^H)t = -r_v$$

S5: *MGS*(U, s), *MGS*(V, t), *normalize* s, t, *set* $U = [\,U, s\,],\ V = [\,V, t\,]$

In case of a Hermitian problem with real eigenvalues we have $x_* = y_*$. Hence, we can choose $v = u$, cf. [83], and only one projected system remains

$$(I - uu^H)T(\lambda)(I - uu^H)s = -T(\lambda)u = r_u, \quad s \perp u. \tag{6.15}$$

Here, the matrix is Hermitian whereas the matrix of the standard nonlinear Jacobi–Davidson correction formula (6.3) is not Hermitian, in general. Notice that in case of linear problems $T(\lambda) = A - \lambda I$, equation (6.15) gives the correction formula of the standard linear Jacobi–Davidson method [81]. So, for Hermitian linear problems the generalized Jacobi–Davidson method reduces to the standard method.

Up to this point we have seen two different versions of the Jacobi–Davidson method, that compute left and right eigenvectors for nonlinear eigenvalue problems. Both are based on the computation of the generalized Rayleigh functional, which yields a quadratic approximation in the neighborhood of left and right eigenvectors. Next, we want to mention another method, which is a slight variation of the standard Jacobi–Davidson method, and has to our knowledge not been formulated for nonlinear eigenproblems, yet. The original idea is embedded in the alternating Rayleigh quotient iteration [67], which was developed in order to have a globally convergent Rayleigh quotient iteration. The idea is to put the corrections for left and right eigenvectors in one subspace, resp. the search space. The alternating Jacobi–Davidson for linear problems was proposed in [37]. The nonlinear version has to be adapted likewise to the nonlinear JD and is presented in Algorithm 18.

Algorithm 18 *Alternating Jacobi–Davidson method*

Input: *eigenvector approximation* u *such that* $u^H u = 1$*, tolerance* $\epsilon > 0$

S0: *Set* $U = [\,u\,]$

for $k = 1, 2, \ldots$

S1: **[k even]** *Solve* $U^H T(\lambda) U c = 0$ *for* (λ, c)

[k odd] *Solve* $U^H T(\lambda)^H U c = 0$ *for* (λ, c)

S2: $u = Uc$, **[k even]** $r = T(\lambda)u$, **[k odd]** $r = T(\lambda)^H u$

S3: **if** $\|r\|/\|u\| < \epsilon$ **stop**

S4: *Compute approximately* $s \perp u$ *from*

[k even] $$\left(I - \frac{\dot{T}(\lambda) u u^H}{u^H \dot{T}(\lambda) u}\right) T(\lambda)(I - uu^H)s \;= -T(\lambda)u$$

[k odd] $$\left(I - \frac{\dot{T}(\lambda)^H u u^H}{u^H \dot{T}(\lambda)^H u}\right) T(\lambda)^H(I - uu^H)s \;= -T(\lambda)^H u$$

S5: *MGS*(U, s)*, normalize* s*, set* $U = [\,U, s\,]$

This method is the subspace iteration analogon to the alternating Rayleigh functional iteration of §4.2.3, which was shown to converge linearly. Although we have a Newton method included, we cannot get quadratic convergence here, since we change the direction every second step, and the Ritz values will be only linear in the angle.

We will use the following abbreviations in the remaining part of this chapter:

JD	Jacobi–Davidson	Algorithm 15
TJD	two-sided Jacobi–Davidson	Algorithm 16
GJD	generalized Jacobi–Davidson	Algorithm 17
AJD	alternating Jacobi–Davidson	Algorithm 18.

6.4 Solving the Preconditioned Correction Equation

Preconditioning plays an essential role for all Jacobi–Davidson methods. We explain the way for applying preconditioners here for Algorithm 15, along the lines of the linear Jacobi–Davidson approach, see [7], cf. also [13].

Suppose $P \approx T(\lambda)$ is some preconditioner for $T(\lambda)$. It is suggestive to project P in the same way as $T(\lambda)$ is projected, i.e., define

$$\tilde{P} = P_1 P P_2 = \left(I - \frac{\dot{T}(\lambda) u u^H}{u^H \dot{T}(\lambda) u} \right) P \left(I - u u^H \right),$$

with $P_1 = I - \frac{\dot{T}(\lambda) u u^H}{u^H \dot{T}(\lambda) u}$ and $P_2 = I - u u^H$, provided that $u^H \dot{T}(\lambda) u \neq 0$. The residual of the preconditioned correction equation according to equation (6.3), which is computed by a Krylov solver, is given by

$$r_c := \tilde{P}^{-1}(P_1 T(\lambda) P_2 s + r).$$

We use the term $\tilde{P}^{-1}$, although we know that $\tilde{P}$ is singular with rank at most $n-1$. The resulting terms will be based on the inverse of $T(\lambda)$, which is nonsingular for $\lambda \notin \lambda(T(\,\cdot\,))$.

Let $y = T(\lambda)s$, define $z := \tilde{P}^{-1} P_1 y$ and $\tilde{r} := \tilde{P}^{-1} r$. Then, we have $r_c = z + \tilde{r}$, since $s \perp u$. How do we compute z? We know that $\tilde{P} z = P_1 y$, i.e.,

$$\left(I - \frac{\dot{T}(\lambda) u u^H}{u^H \dot{T}(\lambda) u} \right) P (I - u u^H) z = y - \frac{\dot{T}(\lambda) u u^H}{u^H \dot{T}(\lambda) u} y,$$

and since $z \perp u$ we obtain the following expression for z

$$z = P^{-1} y + P^{-1} \dot{T}(\lambda) u \gamma, \quad \text{where} \quad \gamma = \frac{u^H P z - u^H y}{u^H \dot{T}(\lambda) u}.$$

Again, with $u^H z = 0$, we have $\gamma = -u^H P^{-1} y / u^H P^{-1} \dot{T}(\lambda) u$. Hence, we end up with

$$z = P^{-1} y - P^{-1} \dot{T}(\lambda) u \frac{u^H P^{-1} y}{u^H P^{-1} \dot{T}(\lambda) u}.$$

The term $\tilde{r}$ is obtained in the same way, which yields

$$\tilde{r} = P^{-1}y - P^{-1}\dot{T}(\lambda)u\frac{u^H P^{-1}r}{u^H P^{-1}\dot{T}(\lambda)u}.$$

An algorithm for the approximate solution of the correction equation (6.3) with left preconditioner is given in Algorithm 19, similar to the one given in [7, p. 94]. The according algorithms for the generalized variants can be derived in the same way.

Algorithm 19 *Approximate solution of JD correction equation*

Solve with left preconditioner $\tilde{P} \equiv (I - \frac{qu^H}{u^H q})P(I - uu^H)$,
for $\tilde{T} = (I - \frac{qu^H}{u^H q})T(\lambda)(I - uu^H)$, *with* $q := \dot{T}(\lambda)u$

S1: *Solve* $\hat{q}$ *from* $P\hat{q} = q$

S2: *Compute* $\mu = u^H\hat{q}$

S3: *Compute* $\tilde{r} \equiv \tilde{P}^{-1}r$ *as*

(*i*) *Solve* $\hat{r}$ *from* $P\hat{r} = r$

(*ii*) $\tilde{r} = \hat{r} - \frac{u^H\hat{r}}{\mu}\hat{q}$

S4: *Apply Krylov subspace method with starting vector* $t_0 = 0$, *operator* $\tilde{P}^{-1}\tilde{T}$, *and righthand side* $-\tilde{r}$. *For given* v *the term* $z = \tilde{P}^{-1}\tilde{T}v$ *is computed as*

(*i*) $y = T(\lambda)v$

(*ii*) *Solve* $\hat{y}$ *from* $P\hat{y} = y$

(*v*) $z = \hat{y} - \frac{u^H\hat{y}}{\mu}\hat{q}$

6.5 Asymptotic Condition Numbers

Investigating the norm of the operators and their inverses at the solution is helpful in the process of trying to understand special properties and behavior of different methods. We want to compare the standard nonlinear Jacobi–Davidson, cf. Algorithm 15, the proposed two-sided Jacobi–Davidson in Algorithm 16, and the generalized Jacobi–Davidson, cf. Algorithm 17, with respect to asymptotic condition numbers in the solution, cf. [78].

Corresponding to the methods in this order we define the projected matrices

$$\begin{aligned} A_{JD}(\lambda,u) &= \left(I-\frac{\dot T(\lambda)uu^H}{u^H\dot T(\lambda)u}\right)T(\lambda)(I-uu^H),\\ A_{TJD}(\lambda,u,v) &= \left(I-\frac{\dot T(\lambda)uv^H}{v^H\dot T(\lambda)u}\right)T(\lambda)(I-uu^H),\\ A_{GJD}(\lambda,u,v) &= \left(I-vv^H\right)T(\lambda)(I-uu^H), \end{aligned}$$

used in the (primal) correction equations of JD, TJD and GJD. Note, that TJD and GJD solve a second system for improving v. The projected matrices define nonsingular $(n-1)$-dimensional mappings

$$\mathcal{A}_{JD}:\ s\perp u\ \mapsto\ A_{JD}s\perp u,\quad \{\mathcal{A}_{TJD},\mathcal{A}_{GJD}\}:\ s\perp u\ \mapsto\ \{A_{TJD}s,A_{GJD}s\}\perp v, \tag{6.16}$$

provided that λ,u,v are sufficiently good. For the Jacobi–Davidson operator we have to assume that $x_*^Hq_*\neq 0$ where $q_*=\dot T(\lambda_*)x_*$. This condition guarantees that $u^Hq\neq 0$, where $q=\dot T(\lambda)u$, close to the solution, and does, for nonnormal T, not follow from the simplicity condition $y_*^Hq_*\neq 0$, in general. For A_{GJD} no additional condition is required.

Consider the mappings (6.16) at the solution $(\lambda,u,v)=(\lambda_*,x_*,y_*)$. Using the SVD (2.12) of $T(\lambda_*)$, the identities $I-x_*x_*^H=X_1X_1^H$, $I-y_*y_*^H=Y_1Y_1^H$ and the normalized vector $\hat q_*=q_*/\|q_*\|$, we obtain

$$\begin{array}{rcccl} A_{JD} &=& A_{JD}(\lambda_*,x_*) &=& P_1^*Y_1\Sigma_1X_1^H,\\ A_{TJD} &=& A_{TJD}(\lambda_*,x_*,y_*) &=& P_3^*Y_1\Sigma_1X_1^H,\\ A_{GJD} &=& A_{GJD}(\lambda_*,x_*,y_*) &=& Y_1\Sigma_1X_1^H, \end{array} \tag{6.17}$$

with the skew projectors $P_1^*=I-\frac{q_*x_*^H}{x_*^Hq_*}$ and $P_3^*=I-\frac{q_*y_*^H}{y_*^Hq_*}$. We have $\operatorname{rank}P_1^*=\operatorname{rank}P_3^*=n-1$ and $P_1^*\hat q_*=P_1^{*H}x_*=0$, so

$$P_1^*=X\Pi Q^H=[\,X_1\mid x_*\,]\left[\begin{array}{c|c}\Pi_1&0\\\hline 0^T&0\end{array}\right][\,Q_1\mid q_*\,]^H=X_1\Pi_1Q_1^H \tag{6.18}$$

is a unitary transformation of P_1^* with $\Pi_1=(Q_1^HX_1)^{-1}$ and

$$\|P_1^*\|=\|\Pi_1\|=\|(Q_1^HX_1)^{-1}\|=\frac{1}{|x_*^H\hat q_*|}=\frac{1}{\cos\psi_*},\quad \|(\Pi_1)^{-1}\|=1, \tag{6.19}$$

where

$$\psi_*=\measuredangle\left(\operatorname{span}\{q_*\},\operatorname{span}\{x_*\}\right)=\measuredangle\left(\operatorname{span}\{\hat q_*\},\operatorname{span}\{x_*\}\right)<\pi/2.$$

Because of $P_3^* \hat{q}_* = P_3^{*H} y_* = 0$, the same procedure for P_3^* yields

$$P_3^* = Y \hat{\Pi} Q^H = [\, Y_1 \mid y_* \,] \left[\begin{array}{c|c} \Pi_3 & 0 \\ \hline 0^T & 0 \end{array}\right] [\, Q_1 \mid q_* \,]^H = Y_1 \Pi_3 Q_1^H, \tag{6.20}$$

which is a unitary projection of P_3^* with $\Pi_3 = (Q_1^H Y_1)^{-1}$ and

$$\|P_3^*\| = \|\Pi_3\| = \|(Q_1^H Y_1)^{-1}\| = \frac{1}{|y_*^H \hat{q}_*|} = \frac{1}{\cos\phi_*}, \quad \|(\Pi_3)^{-1}\| = 1, \tag{6.21}$$

where

$$\phi_* = \measuredangle\,(\mathrm{span}\{q_*\}, \mathrm{span}\{y_*\}) = \measuredangle\,(\mathrm{span}\{\hat{q}_*\}, \mathrm{span}\{y_*\}) = \measuredangle\,(\mathrm{im}\, Q_1, \mathrm{im}\, Y_1) < \pi/2.$$

Replacing P_1^* and P_3^* in (6.17) by the representations (6.18) and (6.20), we end up with

$$A_{JD} = X_1 \Pi_1 Q_1^H Y_1 \Sigma_1 X_1^H, \quad A_{TJD} = Y_1 \Pi_3 Q_1^H Y_1 \Sigma_1 X_1^H, \quad A_{GJD} = Y_1 \Sigma_1 X_1^H.$$

Hence, the matrix representation M_{JD} of $\mathcal{A}_{JD}$ with respect to the orthogonal basis X_1 of the orthogonal complement $\{s : \ s \perp x_*\}$ and the representations M_{TJD} of $\mathcal{A}_{TJD}$ and M_{GJD} of $\mathcal{A}_{GJD}$ with respect to the orthogonal bases X_1 of $\{s : \ s \perp x_*\}$ and Y_1 of $\{t : \ t \perp y_*\}$, cf. (6.16), are given by the nonsingular $(n-1) \times (n-1)$ matrices

$$M_{JD} = \Pi_1 (Q_1^H Y_1) \Sigma_1, \quad M_{TJD} = \Pi_3 (Q_1^H Y_1) \Sigma_1, \quad M_{GJD} = \Sigma_1,$$

resp. So, for the generalized Jacobi–Davidson method we obtain immediately

$$\|\mathcal{A}_{GJD}^{-1}\| = \|M_{GJD}^{-1}\| = \|(\Sigma_1)^{-1}\| = \|T(\lambda_*)^\dagger\|,$$

where $T^\dagger$ denotes the Moore–Penrose pseudo-inverse of T.

For the Jacobi–Davidson method we have $M_{JD}^{-1} = (\Sigma_1)^{-1}(Q_1^H Y_1)^{-1}(\Pi_1)^{-1}$ which, considering (6.19) and (6.21), leads to

$$\frac{1}{\cos\phi_*} \cdot \frac{\cos\psi_*}{\|\Sigma_1\|} \le \|\mathcal{A}_{JD}^{-1}\| = \|M_{JD}^{-1}\| \le \frac{1}{\cos\phi_*} \cdot \|T(\lambda_*)^\dagger\|,$$

or in explicit terms

$$\frac{1}{|y_*^H \dot{T}(\lambda_*) x_*|} \cdot \frac{|x_*^H \dot{T}(\lambda_*) x_*|}{\|\Sigma_1\|} \le \|\mathcal{A}_{JD}^{-1}\| = \|M_{JD}^{-1}\| \le \frac{\|\dot{T}(\lambda_*)\|}{|y_*^H \dot{T}(\lambda_*) x_*|} \cdot \|T(\lambda_*)^\dagger\|. \tag{6.22}$$

The bounds follow with the inequalities $1/\|A^{-1}\| \le \|A\|$ and $1/\|A\| \le \|A^{-1}\|$ for any non-singular matrix A, which hold since $1 = \|AA^{-1}\| \le \|A\|\|A^{-1}\|$. Note that we use the spectral norm, which satisfies $\|AB\| \le \|A\|\|B\|$.

If ϕ_* is very close to $\pi/2$, i.e., if λ_* is extremely sensitive, then $\|\mathcal{A}_{JD}^{-1}\|$ will be very large which may cause problems when solving the correction equation using Krylov methods. These terms are directly connected with the condition number of λ_*, cf. §2.3.

In the linear case $T(\lambda) = A - \lambda I$ we have $\dot{T}(\lambda_*) = -I$, hence, $q_* = -x_*$, $\psi_* = 0$ and (6.22) reduces to

$$\frac{1}{|y_*^H x_*|} \cdot \frac{1}{\|\Sigma_1\|} \leq \|\mathcal{A}_{JD}^{-1}\| = \|M_{JD}^{-1}\| \leq \frac{1}{|y_*^H x_*|} \cdot \|T(\lambda_*)^\dagger\|,$$

where $|y_*^H x_*|$ can be arbitrarily small when A is strongly nonnormal.

For the two-sided Jacobi–Davidson method we have

$$M_{TJD}^{-1} = (\Sigma_1)^{-1}(Q_1^H Y_1)^{-1}(\Pi_3)^{-1},$$

which, considering (6.21), leads to

$$\frac{1}{\|\Sigma_1\|} \leq \|\mathcal{A}_{TJD}^{-1}\| = \|M_{TJD}^{-1}\| \leq \frac{1}{\cos\phi_*} \cdot \|T(\lambda_*)^\dagger\|.$$

This means that not only the generalized Jacobi–Davidson method (GJD) but also the two-sided Jacobi–Davidson method (TJD) can be expected to perform well even in badly conditioned cases, because for both methods the lower bound for the inverse operator norm does not depend on the condition number of the eigenvalue. However, for the two-sided Jacobi–Davidson method the norm of the inverse operator can be large, since the upper bound is the same as for the Jacobi–Davidson method, and that depends on the condition number of λ_*.

6.6 Numerical Examples

We compare the subspace iterating methods that have been discussed in this chapter for different examples. We change the way of counting iterations in that the term *it* in the header of tables, which denoted the number of outer iterations, is replaced by the term *ls*, which refers to the number of solved linear systems. Likewise, in figures, residual norms are plotted versus the number of linear systems that were solved. This is done in order to have a fair comparison of the standard Jacobi–Davidson method and the other variants, which—except for the alternating Jacobi–Davidson in Algorithm 18—solve two linear systems per iteration. Since we plot linear system solves in the figures, we merged the residuals for left and right eigenvectors in case of the two-sided and the generalized JD, i.e., the corresponding plotted residual in norms looks as follows

$$\|res\| := (\|res_u(0)\|, \|res_u(1)\|, \|res_v(1)\|, \|res_u(2)\|, \|res_v(2)\|, \ldots),$$

where res_u, res_v are the vectors of right and left residuals over the number of iterations. This explains potential peaks that occur when the quality of the right eigenvector and of the left eigenvector differs.

6.6.1 A Linear Problem

The convection-diffusion operator of Example 6 was tried first. The nonnormality parameters were set to $\nu = 1$ and $\mu_1 = 10$, i.e., the matrix is strongly nonnormal. This implies that the (one-sided) Rayleigh functionals used in the Jacobi–Davidson method and in the alternating Jacobi–Davidson method are not stationary and give only order one approximations, since the problem is not Hermitian, which may be a reason for the slow convergence of both methods. Both have to build large search spaces until the residual norm is finally starting to decrease, see Table 6.1 and Figure 6.1. The non-sensitivity of the generalized Jacobi–Davidson method is reflected in the fast convergence in contrast to the other methods. The two-sided Jacobi–Davidson lies in between, it is faster than the

	$\lvert\lambda_{it} - \lambda_*\rvert$	ls	$\lVert res_u\rVert$	$\lVert res_v\rVert$	gm_1	gm_2	gm_1/ls	gm_2/ls	cpu
JD	6.5e-11	94	5.7e-11		523		5.6		28.74
AJD	3.8e-11	168	2.4e-11		440	445	2.6	2.6	87.24
TJD	4.5e-11	68	1.2e-10	8.9e-11	174	170	2.6	2.5	10.91
GJD	4.9e-11	38	1.0e-08	9.3e-12	55	62	1.4	1.6	1.92

Table 6.1: *Convection-diffusion operator,* $n = 1024$.

one-sided methods, but slower than the generalized JD. Note, that in terms of iteration numbers the generalized Jacobi–Davidson method converges in only 19 iterations. It is by far the fastest method for this problem. The results reflect the analysis in Section 6.5.

6.6.2 Two Exceptional Cases

Next, we tackle the quadratic eigenvalue problems in Examples 3 and 4, which reflect exceptional cases. Example 3 was constructed such that we are able to control the parameter $\alpha = y_*^H \dot{T}(\lambda_*) x_* \ll 1$. Decreasing α causes an increase in the condition number of the eigenvalue, cf. §2.3.

One of the main problems with the one-vector methods is the unpredictable outcome. We observed in Chapter 4 that even for the moderate value $\alpha = 1e-6$ and a relatively close starting value, the inverse iteration converges to another than the sought eigenvalue.

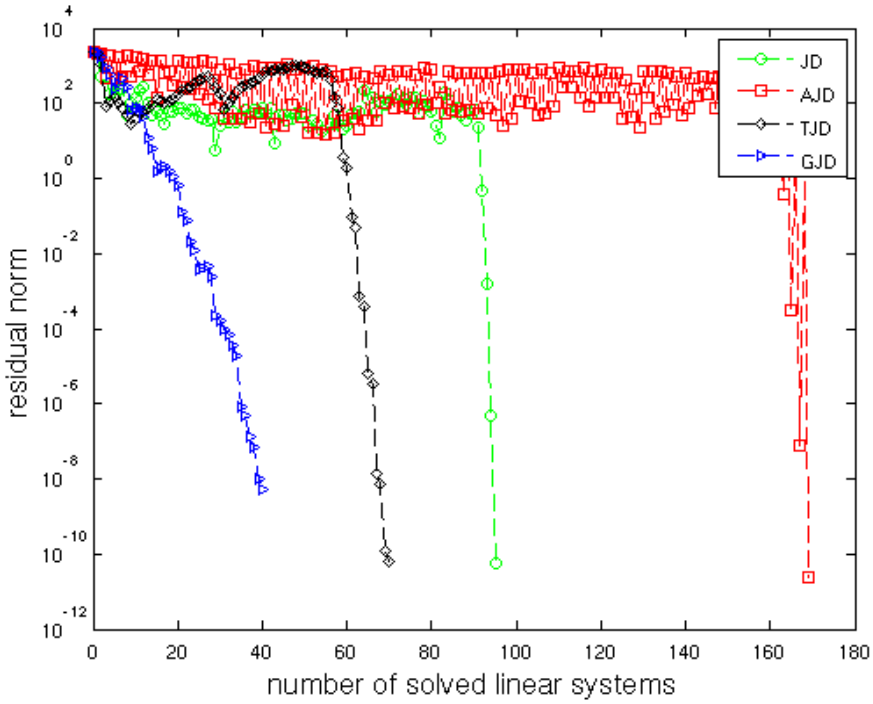

Figure 6.1: *Convection-diffusion operator,* $n = 1024$.

This should be cured by the Jacobi–Davidson algorithm through the extension of the search space. For $\alpha = 1e-6$, Table 6.2 shows that all methods converge to the sought eigenvalue, but we see that with respect to the amount of time, the generalized Jacobi–Davidson method is the fastest. All methods need less iterations (but solve more linear systems) than the standard Jacobi–Davidson method.

	$\lvert\lambda_{ls} - \lambda_*\rvert$	ls	$\lVert res_u \rVert$	$\lVert res_v \rVert$	gm_1	gm_2	gm_1/ls	gm_2/ls	cpu
JD	6.0e-08	7	1.5e-13		11		1.6		0.25
AJD	4.5e-08	10	2.6e-13		6	17	0.6	1.7	0.24
TJD	6.5e-09	12	2.2e-13	5.0e-07	9	9	0.8	0.8	0.25
GJD	4.9e-08	10	7.7e-14	7.8e-06	11	6	1.1	0.6	0.19

Table 6.2: QEP_2, $n = 100$, $\alpha = 1e-6$.

Next, we run the methods for the problem corresponding to $\alpha = 1e-12$. Results are presented in Table 6.3. The programs were terminated if one residual norm was smaller than $1e-12$. The right eigenvector iterates for all methods corresponding to the actual eigenvalue approximation satisfied this criterion after a certain number of iterations. Contrarily, the left eigenvector iterates are poor approximations at the time of termination, see

column $\|res_v\|$. We chose the matrices such that $\lambda_* = 1$ is the eigenvalue closest to zero.

	$\|\lambda_{ls} - \lambda_*\|$	ls	$\|res_u\|$	$\|res_v\|$	gm_1	gm_2	gm_1/ls	gm_2/ls	cpu
JD	2.3e-02	6	2.9e-13		13		2.2		0.08
AJD	2.1e-02	10	1.5e-13		10	28	1.0	2.8	0.21
TJD	1.7e-02	10	3.9e-13	1.4e+00	6	6	0.6	0.6	0.12
GJD	2.8e-03	10	6.9e-14	8.7e-01	10	6	1.0	0.6	0.13

Table 6.3: QEP_2, $n = 100$, $\alpha = 1e - 12$.

Therefore, we know exactly how close the approximations given by the methods are. The best accuracy of the eigenvalue approximation λ_{ls} was a deviation of $0.0028 = |\lambda_{ls} - 1|$, obtained by the generalized Jacobi–Davidson method. In case of the two-sided methods, the bad left eigenvector approximations are directly connected to the bad eigenvalue approximations, since they compute the generalized Rayleigh functional, which depends on the quality of left and right eigenvector. So, unless the two-sided methods improve the left eigenvector, the eigenvalue approximations will not get better. Several test runs, where we changed the termination criterion in that left and right residual norms are required to be small, showed that even over a large number of iterations the left eigenvector could not be improved. On the other hand, the one-sided methods do not give better results, which seems to reflect the connection to the bad condition number of the eigenvalue.

Example 4 was constructed such that $\tilde{\alpha} = x_*^H \dot{T}(\lambda_*) x_* = 0$, which was the condition that we had to exclude when considering one-sided Rayleigh functionals. Since the problem is (real) unsymmetric, this is a different condition than $\alpha \neq 0$, which is equivalent to the algebraic simplicity of the corresponding eigenvalue, provided its geometric simplicity. For the nonlinear Jacobi–Davidson method it should be supposed that $\tilde{\alpha} \neq 0$ in order to have a well-defined correction equation. In Chapter 3, we have also seen that we cannot derive the bounds nor the existence of the standard Rayleigh functional without this assumption. The generalized Jacobi–Davidson method, however, computes the generalized Rayleigh functional and its correction equations do not depend on $\tilde{\alpha}$, hence there is no need for this additional assumption. Table 6.4 and Figure 6.2 present the results for Example 4, i.e., QEP_3 with order $n = 50$. The selection of a Ritz-pair was subject to the minimal absolute value, that means we wanted to approximate $\lambda_* = 0$ in this case. All methods converge to the eigenvalue, but quite differently. The maximum number of iterations is 100, which is hit by the Jacobi–Davidson method and the two-sided Jacobi–Davidson method without converging to the desired accuracy $\|res\| < 1e - 12$. However, both converge as fast as the alternating and the generalized variant within the region of $\|res\| \approx 1e - 10$, but cannot get more accurate, cf. Figure 6.2. It is not clear why the

	$\lvert\lambda_{ls} - \lambda_*\rvert$	ls	$\lVert res_u \rVert$	$\lVert res_v \rVert$	gm_1	gm_2	gm_1/ls	gm_2/ls	cpu
JD	0.0e-00	100	5.2e-10		200		2.0		52.69
AJD	0.0e-00	7		3.7e-16	8	8	1.1	1.1	0.06
TJD	0.0e+00	100	6.4e-04	9.4e-01	100	50	1.0	0.5	8.19
GJD	0.0e-00	10	4.2e-13	2.2e-10	7	7	0.7	0.7	0.08

Table 6.4: *QEP_3, Ritz-pairs sorted for the absolute smallest eigenvalue.*

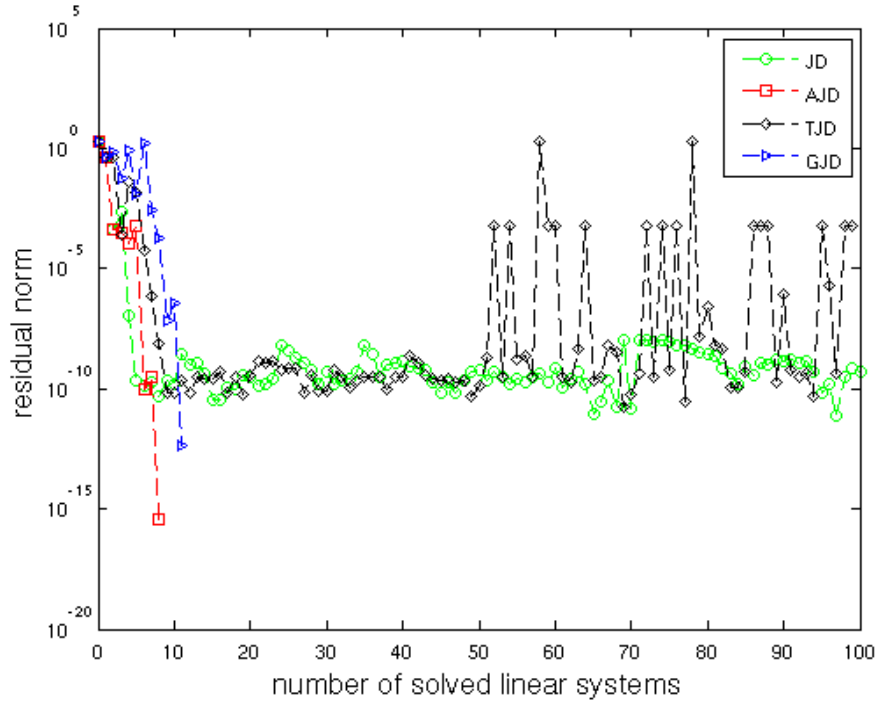

Figure 6.2: *Results for QEP_3.*

two-sided version behaves like this, since it computes the generalized Rayleigh functional, which is well-defined in the neighborhood of the solution, but the Jacobi–Davidson method behaves as anticipated. Due to $u^H\dot{T}(\lambda)u \approx 0$ close to the solution, the one-sided Rayleigh functional is not well-defined—it is not stationary anyway—and the approximations for the eigenvalue cannot get better. The same holds for the alternating version, but only at every second step. At the other steps, the well-defined functional $p(v)$ is computed, which is solution of $v^H T(p(v))^H v = 0$. Table 6.4 tells us that the alternating Jacobi–Davidson method terminates with a pretty good left eigenvector approximation. Looking at Figure 6.2 we observe that only one step before we had $\lVert res_u \rVert \approx 1e-10$, which is approximately the region where the Jacobi–Davidson method stagnates. Nevertheless, the good approximation of the eigenvalue makes it possible for the alternating JD to

converge with respect to the left eigenvector up to accuracy $1e-16$.

If we change the mechanism of selecting a Ritz-pair so that the largest positive is chosen, then only two methods converge at all, but to the eigenvalue on the other side of the spectrum, that is $\lambda = -2$. The other methods cause the eig-function which is called inside of the polyeig-function to fail, since NaN or Inf prevent convergence. This happens, when no new vector corrections are gained from the correction equation and the null-vector is regarded as solution, cf. Table 6.5.

The converging alternating JD and generalized JD present similar results. They spend the same number of iterations, but the generalized JD solves twice the number of linear systems than the alternating JD, only that the right eigenvector approximations are not quite useful in this case.

	$\lvert\lambda_{ls}-\lambda_*\rvert$	ls	$\lVert res_u\rVert$	$\lVert res_v\rVert$	gm_1	gm_2	gm_1/ls	gm_2/ls	cpu
JD	Inf	52	NaN		66		1.3		3.55
AJD	3.0e+00	49		2.1e-14	31	87	0.6	1.8	2.87
GJD	3.0e+00	98	1.2e-04	2.2e-14	49	49	0.5	0.5	6.91

Table 6.5: *QEP_3, Ritz-pairs sorted in descending order.*

6.6.3 More Examples

We want to compute an eigenvalue of the partial delay differential equation example from Example 5. Now we are faced with the problem of solving a nonlinear projected system of the form

$$\left(-\lambda V^H U + V^H A_0 U + V^H (A_1 + A_2) U \exp(-\lambda)\right) c = 0,$$

where U and V are bases of the corresponding search spaces, with $V = U$ in the Jacobi–Davidson method. Since we are interested in complex eigenvalues and cannot assume that there is a subset with real eigenvalues, we cannot use the *method of safe-guarded iteration*, see, e.g., [96], which is based on variational principles, cf. §1.3. But we can revert to the methods of Chapter 4. Since the method of successive linear problems and the residual inverse iteration did not work before, cf. Tables 4.7, 5.3, we choose the inverse iteration as inner eigensolver for the standard and the alternating Jacobi–Davidson method, and the Rayleigh functional iteration as inner eigensolver for the two-sided and generalized Jacobi–Davidson methods. However, even the inverse iteration, resp. Rayleigh functional iteration, failed in case $n = 50$. Searching for the reason of this, we started the one- and two-vector methods of Chapter 4 with the eigenvalue approximation $\lambda_i = -117 + 0.18\mathrm{i}$

that was obtained after a few steps by Jacobi–Davidson. Figure 6.3 shows the residual norms for the methods of Chapter 4 in this case. Hence, the correction equations in the

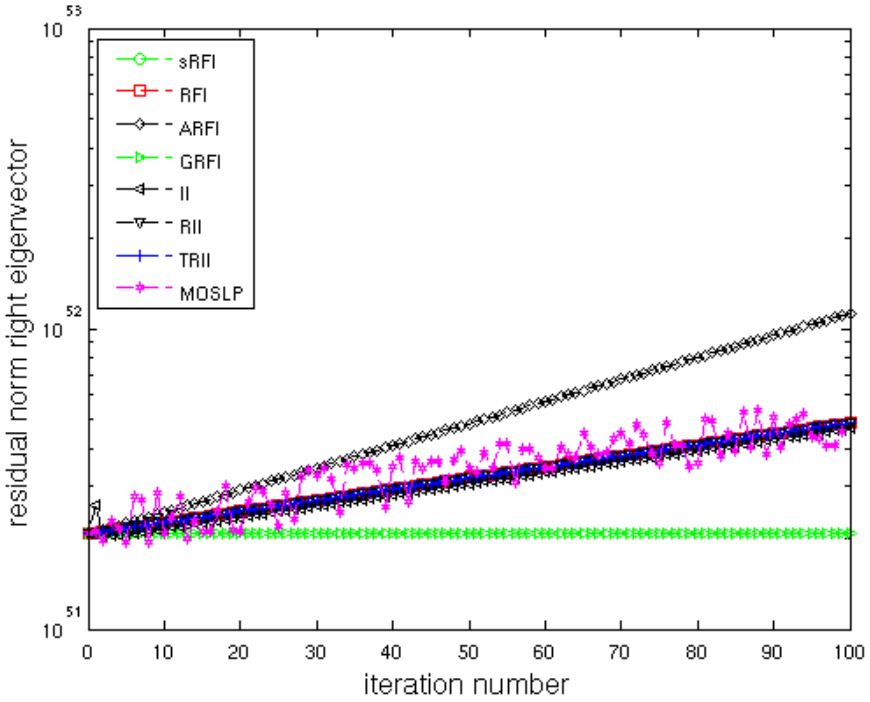

Figure 6.3: *Example PDDE*, $n = 50$, $\lambda_0 = -117 + 0.18i$.

JD methods will never have a proper approximation to start with. This is one of the main problems for algorithms for nonlinear eigenvalue problems. In the case of polynomial problems we can rely on the theory and methods for linear problems after linearization. Since this is not possible for genuine nonlinear problems, each problem needs special care and treatment. We need to be particularly careful to ensure that initial approximations are inside the basins of attraction, since we solve them by Newton methods.

Notice that in this coherence, one of the key differences between solving polynomial and non-polynomial eigenvalue problems with Jacobi–Davidson methods is the handling of starting values. In the case of non-polynomial problems the initial set is more important than in the case of polynomial problems. The reason is that in case of non-polynomial problems the current approximations (λ_k, u_k) should be used as starting values for the Rayleigh–Ritz procedure with $(\lambda_k, c_k = U^H u_k)$ as initial Ritz pair, resp. with $(\lambda_k, c_k = U^H u_k, d_k = V^H v_k)$ in case of the two-sided methods. Contrarily, when we solve a polynomial problem, we use the MATLAB polyeig-function where starting values are useless in the QZ algorithm. Therefore, when using the QZ algorithm to solve the projected

eigenproblem, an initial eigenvalue approximation is not needed, it is only helpful as a target.

Setting the order of the example delay differential operator to $n = 10$ produces a better result. Table 6.6 shows that all methods converge, although to different eigenvalues. Figure 6.4 shows that the residual norms are almost monotonically decreasing. The standard JD converges fast, followed by the generalized JD.

	$\lvert\lambda_{ls}\rvert$	ls	$\lVert res_u \rVert$	$\lVert res_v \rVert$	gm_1	gm_2	gm_1/ls	gm_2/ls	cpu
JD	1.9e+01	4	7.5e-15		4		1.0		0.01
AJD	1.2e+01	9		5.1e-13	5	13	0.6	1.4	0.04
TJD	1.2e+01	10	1.2e-11	1.2e-11	5	5	0.5	0.5	0.01
GJD	1.9e+01	6	3.1e-14	9.6e-14	3	3	0.5	0.5	0.02

Table 6.6: *Example PDDE, $n = 10$. The projected nonlinear system is solved by the inverse iteration and the Rayleigh functional iteration, resp.*

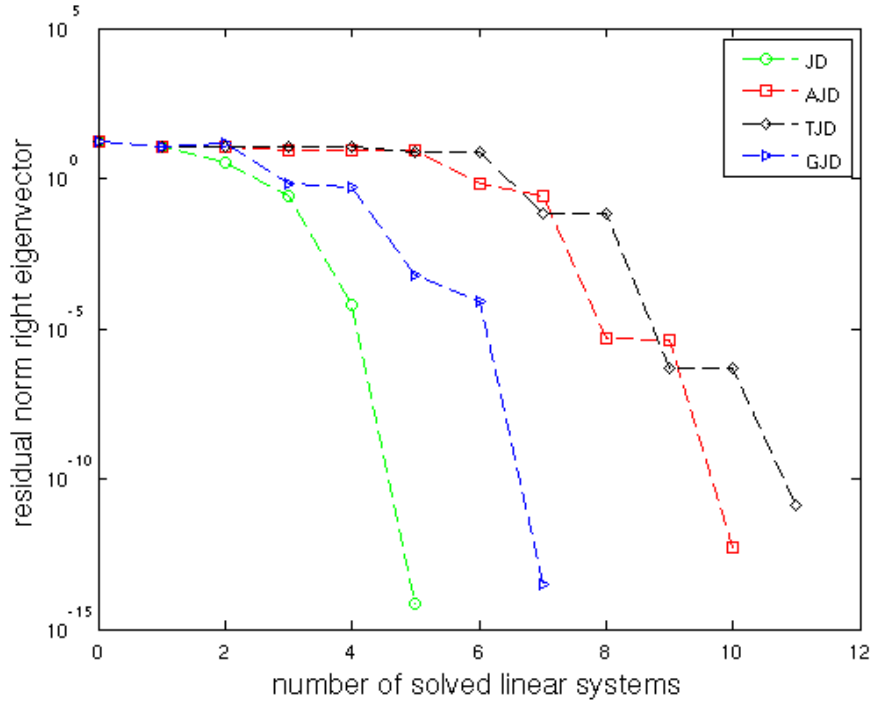

Figure 6.4: *Example PDDE, $n = 10$. The projected nonlinear system is solved by the inverse iteration and the Rayleigh functional iteration, resp.*

6.7 Conclusion

Different versions of the nonlinear Jacobi–Davidson algorithm have been analyzed. We presented nonlinear versions of the two-sided and the alternating Jacobi–Davidson methods, and recalled the generalized Jacobi–Davidson method. Cubic convergence for the underlying two-sided Rayleigh functional iteration, which corresponds to the two-sided Jacobi–Davidson method, has already been shown in Chapter 4. The application of preconditioners was explained in analogy to the linear Jacobi–Davidson method.

The theoretical analysis of asymptotic condition numbers shows that the generalized Jacobi–Davidson method is insensitive with respect to the condition number of the eigenvalue, whilst the upper bound for the condition number of the operator for the two-sided Jacobi–Davidson method depends on the condition number of the eigenvalue, and for the standard Jacobi–Davidson method, upper and lower bound for the condition number of the operator depend on the condition number of the eigenvalue, i.e., theoretically, the Jacobi–Davidson method will not give accurate results in case of ill-conditioned eigenvalues.

The validity of the theoretic bounds is confirmed by numerical examples as for the case of the linear ill-conditioned convection-diffusion operator, for which the generalized Jacobi–Davidson is the fastest method by far. There, the Jacobi–Davidson method needs a huge search space before it finally converges. For a quadratic ill-conditioned eigenproblem, however, all methods were not able to compute the strongly ill-conditioned eigenvalue accurately. In this particular example this may be connected to the bad left eigenvalue approximations that are gathered by the two-sided and the generalized methods. As we have discussed, Example 4, which was constructed such that $x_*^H \dot{T}(\lambda_*) x_* = 0$, violates an initial condition of the standard Jacobi–Davidson, since then the left projector is not well-defined close to the solution. This is reflected by the behavior of JD, which converges only up to a residual norm about 10^{-10} where it stagnates. The alternating JD converges corresponding to the left eigenvector, the generalized JD converges with respect to both eigenvectors.

Moreover, we have seen that computing the inner nonlinear (projected) eigenproblem for truly nonlinear eigenproblems is problematic, since the available methods (discussed in Chapter 4) are locally convergent methods that may fail for bad starting values. For linear problems, the Jacobi–Davidson method is relatively robust with respect to the initial approximations. The main reason for this advantage is that for the projected linear problem all eigenvalues can be determined easily, in general. The proper selection guards the way to the wanted part of the spectrum. For nonlinear eigenvalue problems this is the

crucial point. There is no straightforward way to determine all eigenvalues, since there is no Schur form or eigendecomposition, in general. Hence, when we cannot ensure that we have computed, say, the smallest eigenvalue for the projected system in every single iteration, then we compute corrections for different eigenvector approximations all the time and convergence happens more or less by chance.

Chapter 7

Nonlinear Complex Symmetric Jacobi–Davidson

7.1 Introduction

Applications often yield problems that contain real symmetric matrices. When parts of the spectrum are complex or when imaginary numbers appear explicitly, then the operator $T(\lambda)$ is no longer real for all λ, but it stays symmetric. Consider for instance

$$\left(K_0 - \sum_{l=1}^{m} (\mathrm{i}\lambda^{2l-1} C_l + \lambda^{2l} M_l) \right) x = 0,$$

with real symmetric matrices K_0, C_l, $M_l \in \mathbb{R}^{n\times n}$, which are associated with the dynamic governing equations of a structure submitted to viscous damping [20]. This problem is not Hermitian, as problems with real symmetric coefficient matrices are in general, but it is complex symmetric, cf. §1.2. Even in case of a polynomial eigenvalue problem with solely real symmetric matrices, one cannot suppose a global characterization like $T(\lambda) = T(\bar{\lambda})^H$ for all λ in the spectrum, unless special definiteness criterions are satisfied. Nevertheless, complex symmetric problems have special properties that should be utilized. In this chapter we will see what they are and how this can be done.

We consider nonlinear eigenvalue problems of the form

$$T(\lambda)x = 0, \quad \text{where} \quad T(\lambda) = T(\lambda)^T \quad \text{for all } \lambda \in \mathbb{C}, \tag{7.1}$$

where we assume that $T(\,.\,) : \mathbb{C} \to \mathbb{C}^{n\times n}$ is holomorphic on an open disk around the simple eigenvalue λ_*.

Another example for a complex symmetric problem is given by

$$T(\lambda)x = \left(K - \lambda M + \mathrm{i} \sum_{l=1}^{k} (\lambda - \sigma_l^2)^{\frac{1}{2}} W_l \right) x = 0, \tag{7.2}$$

where K, M, W_l are $n \times n$ real symmetric matrices, K is positive semi-definite, M is positive definite and σ_l are given nonnegative scalars [52], cf. Chapter 1. Similar rational problems govern free vibration of plates with elastically attached masses [95], and vibrations of fluid-solid models [93]. A nonoverdamped mass-spring system yields the quadratic problem in Section 7.6.

Complex symmetric problems in general do not have special spectral properties like Hermitian or real symmetric problems have. The reason for this follows already for the special linear problem from the fact that *any* matrix is similar to a complex symmetric matrix [39, p. 209]. The eigenvalues do not appear in pairs or quadruples, like for Hamiltonian problems, they are not restricted to one half plane. Let us give a small example of a quadratic problem to illustrate this

$$T(\lambda)x = (\lambda^2 M + \lambda C + K)x, \quad M = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} \mathrm{i} & 0 \\ 0 & 1 - \mathrm{i} \end{bmatrix}, \quad K = \begin{bmatrix} 2 & 1 \\ 1 & -2 \end{bmatrix}.$$

The eigenvalues are approximately $-0.01 - 2.04\mathrm{i}$, $-2 + 0.64\mathrm{i}$, $-0.06 + 1.12\mathrm{i}$, $1.02 + 0.28\mathrm{i}$, i.e., they are not symmetric with respect to one of the axes.

Often, complex symmetric problems are treated like general problems, i.e., as if there is no structure. Thus, algorithms for general problems are applied and one expects and gets the order of convergence and the costs as one has for general eigenvalue problems. For instance, if we have a polynomial problem where we want to deflate eigenvectors by a nonequivalence deflation, as is done in [41] for a quadratic eigenproblem, then we require left and right eigenvectors. If the problem is Hermitian with real eigenvalues, then left and right eigenvectors coincide. If it is not, then one expects almost double costs, in general, since the left eigenvectors have to be computed as well. In case of complex symmetric problems this can be avoided due to the extra structure, because the following holds:

$$T(\lambda_*)x_* = 0 \quad \Longleftrightarrow \quad x_*^T T(\lambda_*) = 0 \quad \Longleftrightarrow \quad \bar{x}_*^H T(\lambda_*) = 0,$$

i.e., if (λ_*, x_*) is an eigenpair for T, then $\bar{x}_*$ represents the corresponding left eigenvector. Hence, for complex symmetric problems we have

$$\bar{x}_* = y_*.$$

If we change the usual inner product defined by $(x, y) := y^H x$ to the bilinear form

$$(x, y)_T := y^T x, \tag{7.3}$$

then we can apply methods especially suited for Hermitian problems with real eigenvalues where $T(\lambda_*) = T(\lambda_*)^H$, and expect to get the same order of convergence as for Hermitian problems. However, the bilinear form (7.3) does not define an inner product, since $(x, x)_T = 0$ does not imply $x = 0$. Hence, we have to require the additional assumption that *quasi-null vectors*

$$(x, x)_T = 0, \qquad x \neq 0 \tag{7.4}$$

do not appear in computations and as solution. Suppose that the right eigenvector x_* satisfies the quasi-null condition (7.4), i.e., $0 = x_*^T x_* = y_*^H x_*$, where y_* is the corresponding left eigenvector. For linear problems this is a contradiction to the assumption that the corresponding eigenvalue λ_* is simple, hence quasi-null vectors cannot be eigenvectors for simple eigenvalues in case of linear problems. We will see that quasi-null vectors do not influence our algorithms, since we design them appropriately, see the discussion in §7.3.

Simple eigenvalues of complex symmetric eigenvalue problems are characterized by the following corollary, which follows immediately from Proposition 2.2.

Corollary 7.1 *Let λ_* be a geometrically simple eigenvalue for the complex symmetric problem* (7.1)*, and let x_* be the corresponding unit norm eigenvector. Then, the eigenvalue λ_* is algebraically simple if and only if*

$$x_*^T \dot{T}(\lambda_*) x_* \neq 0.$$

Proof. Since the left eigenvector satisfies $y_* = \bar{x}_*$, we have $\alpha = x_*^T \dot{T}(\lambda_*) x_* = y_*^H \dot{T}(\lambda_*) x_*$, which is nonzero if and only if λ_* is algebraically simple, provided that it is geometrically simple, cf. the proof of Proposition 2.2. □

In the remaining parts of this chapter the characteristic properties of complex symmetric eigenvalue problems are exploited. We make use of many results and algorithms of the previous chapters, starting with the adapted Rayleigh functional, the local existence and stationarity of which follows immediately from the generalized Rayleigh functional. As before, the Rayleigh functional is the essential ingredient of the complex symmetric Rayleigh functional iteration, which will be shown to converge locally cubically, the complex symmetric residual inverse iteration and the complex symmetric Jacobi–Davidson method, which will be introduced. Numerical experiments show the advantages of the new methods compared to the original methods if applied to complex symmetric nonlinear eigenvalue problems.

7.2 The Rayleigh Functional

In analogy to the Rayleigh functional and the generalized Rayleigh functional definitions and analysis of Chapter 3, we define the Rayleigh functional $p_T(u)$ for complex symmetric T as follows

$$\begin{aligned} & p_T : u \in \mathcal{K}_\varepsilon(x_*) \longrightarrow p_T(u) \in D \in \mathbb{C}, \\ & p_T(cu) = p_T(u) \quad \text{for all} \quad c \neq 0, && (7.5) \\ & (T(p_T(u))u, u)_T = 0, && (7.6) \\ & (\dot{T}(p_T(u))u, u)_T \neq 0, && (7.7) \end{aligned}$$

with $\mathcal{K}_\varepsilon(x_*)$ as defined in (3.9). In contrast to the case where the Rayleigh functional is defined with respect to the standard inner product, i.e., as solution of $u^H T(p(u))u = 0$, we are able to differentiate

$$g(p_T(u), u) := u^T T(p_T(u))u = 0 \tag{7.8}$$

with respect to u even if $T(\lambda)$, u and p_T are complex-valued, if p_T is differentiable.

Theorem 7.2 *Let $D \subset \mathbb{C}$ be an open set, and let $T : D \to \mathbb{C}^{n\times n}$, where $T(\lambda) = T(\lambda)^T$, be holomorphic in D. Let $\lambda_* \in D$ be an algebraically simple eigenvalue of T, and let $\tau_* > 0$ be such that $S_* := \bar{S}(\lambda_*, \tau_*) = \{\lambda \in \mathbb{C} : |\lambda - \lambda_*| \leq \tau_*\} \subset D$. Then, there exist constants $0 < \tau_0 < \tau_*, 0 < \varepsilon_0 < \pi/2$, such that for all $u \in \mathcal{K}_{\varepsilon_0}(x_*)$ the Rayleigh functional $p_T \equiv p_T(u) \in S_0 := \bar{S}(\lambda_*, \tau_0)$ is uniquely defined as solution of equation (7.8) and is holomorphic in S_0. The derivative is given by*

$$\frac{\partial p_T(u)}{\partial u} \quad = \quad -\frac{2u^T T(p_T(u))}{u^T \dot{T}(p_T(u))u}. \tag{7.9}$$

Proof. Using the indefinite bilinear form enables us to differentiate $g(p_T, u)$ with respect to both arguments, since the Cauchy–Riemann differential equations hold and g is real-differentiable. Thus, the differentiability of T is carried over to g. We start with considering $g(\lambda, u) = u^T T(\lambda)u$. Then, we obtain $\frac{\partial g}{\partial u}[h] = 2h^T T(\lambda)u$ and $\frac{\partial g}{\partial \lambda} = u^T \dot{T}(\lambda)u$. Since $\frac{\partial g}{\partial \lambda}(\lambda_*, x_*) = x_*^T \dot{T}(\lambda_*)x_* \neq 0$ for the simple eigenvalue λ_*, the implicit function theorem guarantees the existence of a unique $p_T = p_T(u) \in \bar{S}(\lambda_*, \tau_0)$ for some constant τ_0, $0 < \tau_0 < \tau_*$, and vectors $u \in \mathcal{K}_{\varepsilon_0}(x_*)$ with constant ε_0, $0 < \varepsilon_0 < \pi/2$, which satisfies $g(u) = g(p_T(u), u) = u^T T(p_T(u))u = 0$, and which is holomorphic there. Differentiating g with respect to u gives

$$\begin{aligned} \frac{\partial g}{\partial u}[h] \quad &= \quad 2h^T T(p_T(u))u + u^T \dot{T}(p_T(u)) \left(\frac{\partial p_T}{\partial u} h\right) u \\ &= \quad \left\{ 2u^T T(p_T(u)) + (u^T \dot{T}(p_T(u))u) \frac{\partial p_T}{\partial u} \right\} h = 0, \end{aligned} \tag{7.10}$$

for all $h \neq 0$, which yields (7.9) immediately. □

This result implies stationarity without extra work.

Lemma 7.3 *Let $D \subset \mathbb{C}$ be an open set, and let $T : D \to \mathbb{C}^{n\times n}$, where $T(\lambda) = T(\lambda)^T$, be holomorphic in D. Let $\lambda_* \in D$ be an algebraically simple eigenvalue of T, and let $\tau_* > 0$ be such that $S_* := \bar{S}(\lambda_*, \tau_*) = \{\lambda \in \mathbb{C} : |\lambda - \lambda_*| \leq \tau_*\} \subset D$. Then, there exist constants $0 < \tau_0 < \tau_*$, $0 < \varepsilon_0 < \pi/2$, such that for all $u \in \mathcal{K}_{\varepsilon_0}(x_*)$ the Rayleigh functional $p_T(u)$ exists in $\bar{S}(\lambda_*, \tau_0)$, and it is stationary at the eigenvector $u = x_*$.*

Proof. The first part was shown in Theorem 7.2 and implies equation (7.9). Stationarity is given if the first order derivative vanishes. But the right hand side of equation (7.9) equals zero only if $u = x_*$ is an eigenvector. □

Therefore, the Taylor series for $p_T(u)$ with respect to the eigenvector x_* looks as follows

$$p_T(u) = \lambda_* + (u - x_*)^T \nabla^2 p_T(x_*)(u - x_*) + R_T, \tag{7.11}$$

where $R_T = o(\|u - x_*\|^2)$. Hence, we have

$$|p_T(u) - \lambda_*| \leq C\|u - x_*\|^2, \tag{7.12}$$

with $C := \|\nabla^2 p_T(x_*)\| + \frac{|R_T|}{\|u-x_*\|^2}$ and $\lim_{\|u-x_*\|\to 0} |R_T|/\|u - x_*\|^2 = 0$.

Inequality (7.12) states the key fact why it is reasonable to use the complex symmetric Rayleigh functional instead of the standard Rayleigh functional. For the standard functional, Corollary 3.22 gives the corresponding bound, i.e.,

$$|p - \lambda_*| \quad \leq \quad \frac{10}{3} \frac{\|T(\lambda_*)\|}{|x_*^H \dot{T}(\lambda_*) x_*|} \tan \xi,$$

with the additional assumption $x_*^H \dot{T}(\lambda_*) x_* \neq 0$. Thus, for $\varepsilon_0 < \pi/3$, Lemma 2.10 (i) increases the upper bound to

$$|p - \lambda_*| \quad \leq \quad \frac{20}{3} \frac{\|T(\lambda_*)\|}{|x_*^H \dot{T}(\lambda_*) x_*|} \|u - x_*\|.$$

This means that the standard functional results in a linear bound in $\|u - x_*\|$, whereas the complex symmetric functional yields a quadratic one. Moreover, if x_* is a complex vector, then the denominator $\tilde{\alpha} = x_*^H \dot{T}(\lambda_*) x_*$ in the bound is again a different term than $\alpha = y_*^H \dot{T}(\lambda_*) x_* = x_*^T \dot{T}(\lambda_*) x_*$, which is nonzero in case λ_* is simple.

With inequality (7.12) we already have a bound which is quadratic in the distance of the approximation u and the exact eigenvector x_*. For the further analysis, however, we would

like to have an expression in terms of the angle $\xi = \measuredangle(\mathrm{span}\{u\}, \mathrm{span}\{x_*\})$, since this is the proper measure as we have discussed in §2.5. Lemma 2.10 states that $\sin\xi \leq \|u - x_*\|$ without equality in general. Hence, we cannot obtain a bound in the angles from (7.12). But, Theorem 3.5 applied to this context gives the bound immediately.

Corollary 7.4 *Let $D \subset \mathbb{C}$ be an open set, and let $T : D \to \mathbb{C}^{n\times n}$, where $T(\lambda) = T(\lambda)^T$, be holomorphic in D. Let $\lambda_* \in D$ be an algebraically simple eigenvalue of T, and let $\tau_* > 0$ be such that $S_* := \bar{S}(\lambda_*, \tau_*) = \{\lambda \in \mathbb{C} : |\lambda - \lambda_*| \leq \tau_*\} \subset D$. Then there exist constants $0 < \tau_0 < \tau_*$, $0 < \varepsilon_0 < \pi/2$ such that for all $u \in \mathcal{K}_{\varepsilon_0}(x_*)$, there exists a unique $p_T = p_T(u) \in S_0 := \bar{S}(\lambda_*, \tau_0)$ with $g(p_T, u) = (T(p_T(u))u, u)_T = 0$. Moreover, one has*

$$|p_T(u) - \lambda_*| \quad \leq \quad \frac{8}{3}\,\frac{\|T(\lambda_*)\|}{|x_*^T \dot{T}(\lambda_*) x_*|}\tan^2\xi. \tag{7.13}$$

Proof. We have already shown the existence of $p_T \in S_0$. The bound is shown as in Theorem 3.5 with $y_* = \bar{x}_*$ and $Y_1 = \bar{X}_1$. □

In practice it is often too complicated to determine the Rayleigh functional explicitly for nonlinear problems, as is for instance possible for the quadratic eigenproblem. For linear problems, equation (7.8) yields the complex symmetric Rayleigh quotient

$$p(u) = \frac{u^T A u}{u^T u},$$

uniquely, cf. [6], for which our bounds hold as well. In the nonlinear case, the issue is far more complicated, since the solutions of the nonlinear equation (7.8) have to be determined in the first place and next, one of the roots has to be selected, which, in the ideal case, approximates the sought eigenvalue. The easier way is to compute the Newton step for equation (7.8), i.e.,

$$p_{LT}(\lambda, u) = \lambda - \frac{u^T T(\lambda) u}{u^T \dot{T}(\lambda) u}, \tag{7.14}$$

with $p_{LT} :\ (\lambda, u) \in\ S_0 \times \mathcal{K}_{\varepsilon_0}(x_*) \mapsto p_{LT} \in \mathbb{C}$. The restriction to this domain guarantees that p_{LT} is well-defined. This is the appropriate analogon to the generalized Rayleigh quotient of Lancaster [49], cf. §3.4. It follows from [48] that this quotient is stationary with the argument that left eigenvector and its approximations are replaced by the conjugated right eigenvector and its approximations.

We want to give a bound for the distance to the exact eigenvalue in order to compare it with bound (7.13).

Theorem 7.5 *Under the assumptions and with the constants of Corollary 7.4, the iterate*

(7.14) *is well-defined for all* $\lambda \in S_0 = \bar{S}(\lambda_*, \tau_0)$ *and all* $u \in \mathcal{K}_{\varepsilon_0}(x_*)$*, and we have*

$$|p_{LT}(\lambda, u) - \lambda_*| \leq \frac{4\|T(\lambda_*)\|}{|x_*^T \dot{T}(\lambda_*) x_*|} \tan^2 \xi + \frac{2L|\lambda - \lambda_*|^2}{|x_*^T T(\lambda_*) x_*| \cos^2 \xi}, \tag{7.15}$$

with L *as defined in* (2.11)*.*

Proof. The proof follows from the proof of Theorem 3.32, with $v = \bar{u}$ and $y_* = \bar{x}_*$. □

Now we can determine the distance between the two approximations given by the complex symmetric Rayleigh quotient $p_{LT}(\lambda, u)$ and the complex symmetric Rayleigh functional $p_T(u)$.

Corollary 7.6 *Under the assumptions and with the constants of Corollary 7.4, we have*

$$|p_T(u) - p_{LT}(\lambda, u)| \leq \frac{20\|T(\lambda_*)\|}{3|x_*^T \dot{T}(\lambda_*) x_*|} \tan^2 \xi + \frac{2L|\lambda - \lambda_*|^2}{|x_*^T \dot{T}(\lambda_*) x_*| \cos^2 \xi},$$

for all $\lambda \in S_0$ *and all* $u \in \mathcal{K}_{\varepsilon_0}(x_*)$*.*

Proof. Since $|p_T(u) - p_{LT}| \leq |p_T(u) - \lambda_*| + |p_{LT} - \lambda_*|$, the estimate follows with (7.13) and (7.15). □

7.3 Complex Symmetric Rayleigh Functional Iteration

The Rayleigh functional gives an eigenvalue approximation for a given eigenvector approximation u. But how can u be improved? We first consider the Newton update.

Let u be an eigenvector approximation corresponding to the eigenvalue approximation λ. One step of the Newton method for (7.1), extended by the normalizing condition $u^H u = 1$, is equivalent to the system

$$\begin{bmatrix} T(\lambda) & \dot{T}(\lambda)u \\ u^H & 0 \end{bmatrix} \begin{bmatrix} x \\ \mu \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{7.16}$$

Note that we have to use the inner product for the normalizing condition if we do not want to check the extra condition $u^T u \neq 0$. We have already observed in §4.2 that it makes no difference for the proof of convergence, whether we impose $w^H u = 1$, with $w = u$, or $\tilde{w}^H u = 1$, with $\tilde{w} = \bar{u}$, as long as $|w^H x_*| > 0$, respectively $|\tilde{w}^H x_*| > 0$, holds uniformly. Though, the last condition excludes quasi-null vectors.

The Rayleigh functional iteration in the complex symmetric case can be written as follows:

Algorithm 20 *Complex symmetric Rayleigh functional iteration*

Input: $(\lambda_0, u_0),\ \|u_0\| = 1$

for $k = 0, 1, 2, ...$

S1: *Solve* $T(\lambda_k)x_{k+1} = \dot{T}(\lambda_k)u_k$, *set* $u_{k+1} = x_{k+1}/\|x_{k+1}\|$

S2: *Solve* $u_{k+1}^T T(\lambda_{k+1})u_{k+1} = 0$ *for* λ_{k+1}

Theorem 7.7 *Let $D \subset \mathbb{C}$ be an open set and let $T : D \to \mathbb{C}^{n\times n}$, where $T(\lambda) = T(\lambda)^T$, be twice continuously differentiable (holomorphic) in D. Let $\lambda_* \in D$ be an algebraically simple eigenvalue of T with corresponding eigenvector x_* with $\|x_*\| = 1$, and let $\tau_* > 0$ be such that $S_* := \{\lambda \in \mathbb{C} : |\lambda - \lambda_*| \le \tau_*\} \subset D$. Then, there exist constants $0 < \varepsilon_0 < \pi/3$, $0 < \tau_0 \le \tau_*$ and $K_0^a > 0$, $C_0^a > 0$, $K_0^d > 0$, $C_0^d > 0$, $K > 0$, $Q_0 > 0$ with $\tau_0 Q_0 \le 1/2$, such that the complex symmetric Rayleigh functional iteration in Algorithm 20 is well-defined for every initial pair (λ_0, u_0), $u_0 \in \mathcal{K}_{\varepsilon_0}(x_*)$, $\|u_0\| = 1$, and $\lambda_0 \in \bar{S}(\lambda_*, \tau_0)$. Moreover, it converges in the following sense*

$$\lim_{k\to\infty} \lambda_k = \lambda_*, \quad \lim_{k\to\infty} \xi_k = 0,$$

where $\xi_k = \measuredangle(\mathrm{span}\{u_k\}, \mathrm{span}\{x_*\})$, and the rate of convergence is characterized by

$$\sin\xi_{k+1} \;\le\; \|u_{k+1} - x_*^{u_k}\| \le \begin{cases} \frac{1}{2}\sin\xi_k, \\ K_0^a \sin^3\xi_k, \\ K_0^d \|u_k - x_*^{u_k}\|^3, \end{cases} \tag{7.17}$$

and

$$|\lambda_{k+1} - \lambda_*| \le \begin{cases} 4K\left(\frac{1}{2}\right)^{2^{k+1}}, \\ C_0^a |\lambda_k - \lambda_*|^2 \sin^2\xi_k, \\ C_0^d |\lambda_k - \lambda_*|^2 \|u_k - x_*^{u_k}\|^2, \end{cases} \tag{7.18}$$

where $x_*^{u_k} = x_*/u_k^H x_*$. The R-order is at least 3.

Proof. All preparations for the proof have already been done in Lemma 4.5, Lemma 4.11, Theorem 4.7 and Corollary 4.12. We can use the proof for the two-sided Rayleigh functional iteration in Theorem 4.13, and in particular the inequality (4.28), i.e.,

$$\sin\xi_{k+1} \le |\lambda_k - \lambda_*| \left\{2K_1 \sin\xi_k + K_2|\lambda_k - \lambda_*|\right\},$$

with K_1, K_2 from Theorem 4.7. Since the eigenvalue update is computed as Rayleigh functional, inequality (7.13) gives

$$|\lambda_{k+1} - \lambda_*| \leq K \tan^2 \xi_{k+1},$$

with $K = 8\|T(\lambda_*)\|/3|\alpha|$ as before. Hence, the rest follows with Theorem 4.13, with $\eta_k = \xi_k$. □

If we want to work directly on the corrections s_k, where $u_{k+1} = u_k + s_k$, we can use the bordered matrix formulation (4.7). In order to use the symmetry of $T(\lambda)$ it is, however, more reasonable to redefine the normalizing condition such that we work with the bordered complex symmetric matrix

$$C(\lambda, u) := \begin{bmatrix} T(\lambda) & \dot{T}(\lambda)u \\ u^T\dot{T}(\lambda)^T & 0 \end{bmatrix}. \tag{7.19}$$

The normalizing condition is reasonable, since we know that $u^T\dot{T}(\lambda)u \neq 0$ close to an eigenpair. Note, that the derivative of $T(\lambda)$ also satisfies $\dot{T}(\lambda) = \dot{T}(\lambda)^T$. A version of Algorithm 20 with the symmetrically bordered matrix is presented in Algorithm 21.

Algorithm 21 *Complex symmetric Rayleigh functional iteration (bordered version)*

Input: (λ_0, u_0) *where* $\|u_0\| = 1$

for $k = 0, 1, 2, ...$

S1: *Solve* $\begin{bmatrix} T(\lambda_k) & \dot{T}(\lambda_k)u_k \\ u_k^T\dot{T}(\lambda_k)^T & 0 \end{bmatrix} \begin{bmatrix} s_k \\ \mu_k \end{bmatrix} = -\begin{bmatrix} T(\lambda_k)u_k \\ 0 \end{bmatrix}$, *set* $u_{k+1} = \dfrac{u_k + s_k}{\|u_k + s_k\|}$

S2: *Solve* ${u_{k+1}}^T T(\lambda_{k+1})u_{k+1} = 0$ *for* λ_{k+1}

The linear systems in both, Algorithm 20 and 21, use a complex symmetric matrix. Hence, it is appropriate to use a complex symmetric solver. Several methods solving complex symmetric problems are available, see, e.g., the complex symmetric adaption of the quasi-minimal residual method (QMR) [24], the complex conjugate gradient method (COCG) [91], which is an adaptation of the bi-conjugate gradient algorithm, and the conjugate gradient type iterative method CSYM [16], which is, in contrast to the first two, no Krylov subspace method.

Let us come back to the question, whether quasi-null vectors play a role in the Rayleigh functional iteration. We know that quasi-null vectors can be eigenvectors for nonlinear problems where $\dot{T}(\lambda) \neq \pm I$, see, e.g., the example in §2.1, even if the eigenvalue is simple.

But, for the proof of Theorem 7.7 we did not need to exclude quasi-null vectors. The reason is that in the nonlinear case the term $x_*^T\dot{T}(\lambda_*)x_*$ is the important term, which must not be zero. The assumption that the approximated eigenvalue is simple is necessary for the nonsingularity of the bordered matrix. It is also necessary for the existence of the Rayleigh functional. So, if the eigenvalue is simple, then we only need to be careful with the normalizing condition, to either use the Euclidean norm to normalize the vectors or to normalize as in Algorithm 21, i.e., by using the term $x_*^T\dot{T}(\lambda_*)x_*$. Then, quasi-null vectors will not influence the algorithm.

The following Lemma reveals an interesting property of the Rayleigh functional, although we did not use it in the proof.

Lemma 7.8 *Under the assumptions of Theorem 7.2, the complex symmetric Rayleigh functional satisfies*

$$\frac{\partial^2 p_T(x_*)}{\partial u^2} = -\frac{2T(\lambda_*)}{x_*^T\dot{T}(\lambda_*)x_*}. \tag{7.20}$$

Proof. Again, existence and the holomorphic property have been shown in Theorem 7.2. Due to (7.1) we also have $\dot{T}(\lambda) = \dot{T}(\lambda)^T$. Since the derivative given in (7.9) applied to a vector is an element in $\mathbb{C}$, we can write

$$\left(\frac{\partial p_T}{\partial u}h\right)\left(\frac{\partial p_T}{\partial u}k\right) = k^T\left(\frac{\partial p_T}{\partial u}\right)^T\frac{\partial p_T}{\partial u}h.$$

Thus, differentiating equation (7.8) twice, i.e., differentiating (7.10) yields

$$\begin{aligned}\frac{\partial^2 g}{\partial u^2}[h,k] &= 2(k^T\dot{T}(p_T(u))u)\frac{\partial p_T}{\partial u}[h] + \left(u^T\ddot{T}(p_T(u))\left(\frac{\partial p_T}{\partial u}[k]\right)u\right)\frac{\partial p_T}{\partial u}[h] + \\ &\quad (u^T\dot{T}(p_T(u))u)\frac{\partial^2 p_T}{\partial u^2}[k,h] + 2k^TT(p_T(u))h + 2u^T\dot{T}(p_T(u))\left(\frac{\partial p_T}{\partial u}[k]\right)h \\ &= k^T\Big\{2T(p_T(u)) + 2\dot{T}(p_T(u))u\frac{\partial p_T}{\partial u} + 2\left(\dot{T}(p_T(u))u\frac{\partial p_T}{\partial u}\right)^T + \\ &\quad (u^T\ddot{T}(p_T(u))u)\left[\left(\frac{\partial p_T}{\partial u}\right)^T\left(\frac{\partial p_T}{\partial u}\right)\right] + u^T\dot{T}(p_T(u))u\frac{\partial^2 p_T}{\partial u^2}\Big\}h \;=\; 0,\end{aligned}$$

for all vectors h, k, where $\ddot{T}(p_T) = \frac{\partial^2 T(p_T)}{\partial p_T^2}$. Hence, we have

$$\frac{\partial^2 p_T(u)}{\partial u^2} = -\frac{1}{u^T\dot{T}(p_T(u))u}\Big\{2T(p_T(u)) + 2\dot{T}(p_T(u))u\frac{\partial p_T}{\partial u} + 2\left(\dot{T}(p_T(u))u\frac{\partial p_T}{\partial u}\right)^T + (u^T\ddot{T}(p_T(u))u)\left[\left(\frac{\partial p_T}{\partial u}\right)^T\left(\frac{\partial p_T}{\partial u}\right)\right]\Big\},$$

where the denominator is nonzero for vectors u in $\hat{\mathcal{K}}_{\varepsilon_0}$, since λ_* is simple. Applying $u = x_*$ yields (7.20), since p_T is stationary at the eigenvector x_*. □

7.4 Complex Symmetric Residual Inverse Iteration

We have already analyzed the different convergence rates of the residual inverse iteration methods with appropriate Rayleigh functionals given in Algorithms 2 and 7. We formulate the complex symmetric version in Algorithm 22, where the complex symmetric Rayleigh functional is used.

Algorithm 22 *Complex symmetric residual inverse iteration*

Input: (λ_0, u_0), *normalization vector* w, *such that* $w^H u_0 = 1$
for $k = 0, 1, 2, ...$
S1: *Solve* $u_k^T T(\lambda_{k+1}) u_k = 0$ *for* λ_{k+1}
S2: *Compute the residual* $r_k = T(\lambda_{k+1}) u_k$
S3: *Solve* $T(\lambda_0) s_k = r_k$ *for* s_k
S4: *Set* $x_{k+1} = u_k - s_k, \quad u_{k+1} = x_{k+1}/w^H x_{k+1}$

It is advisable to update the shift λ_0 now and then to accelerate convergence. The only difference to the original method (Algorithm 2) is given in step S1, where the complex symmetric Rayleigh functional is computed instead of the standard Rayleigh functional. Since this will give a quadratic approximation of the eigenvalue, the complex symmetric version of the residual inverse iteration converges as follows:

Theorem 7.9 *Let* λ_* *be a simple eigenvalue of* $T(\lambda) = T(\lambda)^T$. *Suppose that* $T(\lambda)$ *is twice continuously differentiable and that* x_* *is the corresponding right eigenvector normalized by* $w^H x_* = 1$. *Then, the complex symmetric residual inverse iteration in Algorithm 22 converges for all* λ_0 *sufficiently close to* λ_*, *if* $\xi_0 = \measuredangle(\mathrm{span}\{u_0\}, \mathrm{span}\{x_*\}) \leq \pi/3$, *and we have*

$$\frac{\|u_{k+1} - x_*\|}{\|u_k - x_*\|} = \mathcal{O}(|\lambda_0 - \lambda_*|), \qquad |\lambda_{k+1} - \lambda_*| = \mathcal{O}(\|u_k - x_*\|^2).$$

Moreover, if the shift λ_0 *is updated by* λ_k *in every iteration, then the R-order of the two-sided residual inverse iteration method is at least 2, and we have*

$$\|u_{k+1} - x_*\| = \mathcal{O}(\|u_k - x_*\| \|u_{k-1} - x_*\|^2).$$

Proof. The proof follows immediately from the proof of Theorem 4.17 with $y_* = \bar{x}_*$ and $v_k = \bar{u}_k$, which implies that $\|v_k - y_*\| = \|u_k - x_*\|$. □

7.5 Nonlinear Complex Symmetric Jacobi–Davidson

We have already discussed nonlinear versions of the Jacobi–Davidson method in Chapter 6 and want to give a complex symmetric (nonlinear) Jacobi–Davidson method here. In [6] a Jacobi-Davidson method for solving complex symmetric linear problems has been presented and cubic convergence of the underlying complex symmetric Rayleigh functional iteration has been shown. We have already shown the locally cubic convergence of the complex symmetric Rayleigh functional iteration for nonlinear eigenproblems, and want to bring it in a subspace expanding context, which yields the Jacobi–Davidson method presented in Algorithm 23.

Suppose that we have a search space $\mathcal{U}$ with orthonormal basis U of size $k \times n$. Then, new approximations are obtained from the nonlinear Rayleigh–Ritz step with respect to the bilinear form (7.3), i.e., we impose the Ritz–Galerkin condition on the actual residual

$$r := T(\lambda)u \perp_T \mathcal{U}. \tag{7.21}$$

Writing $u = Uc$, $c \in \mathbb{C}^k$ gives the equivalent formulation

$$U^T T(\lambda) U c = 0.$$

This is a nonlinear eigenvalue problem of size $k \times k$, which can be tackled with a dense solver like the method of successive linear problems [72], residual inverse iteration [61], or the nonlinear inverse iteration, cf. Chapter 4. It is important to note that this is also a complex symmetric problem, since $(U^T T(\lambda) U)^T = U^T T(\lambda) U$, independent of U, hence we should use a method with the particular adaptions, as for instance the complex symmetric Rayleigh functional iteration in Algorithm 21, or the complex symmetric residual inverse iteration in Algorithm 22.

After the Ritz pair (λ, c) is computed, the new vector $u = Uc$ is improved by a vector s, which is obtained by solving a correction equation that is equivalent to a Newton step. Consider system (7.16) with $x = u + s$. Multiplying the first equation by u^T, rearranging for μ and backtransforming into the equation yields, cf. §6.2,

$$\left(I - \frac{\dot{T}(\lambda) u u^T}{u^T \dot{T}(\lambda) u}\right) T(\lambda)(I - uu^H)s = -\left(I - \frac{\dot{T}(\lambda) u u^T}{u^T \dot{T}(\lambda) u}\right) r = -r, \tag{7.22}$$

since $s \perp u$ with the second equation, and $r \perp_T u$ because of (7.21), provided that $u^T \dot{T}(\lambda) u \neq 0$. Corollary 7.1 implies that $u^T \dot{T}(\lambda) u \neq 0$ holds for (λ, u) in the neighborhood of (λ_*, x_*), if λ_* is an algebraically simple eigenvalue.

In order to get a complex symmetric projected matrix, i.e., the left projector equals the transposed right projector regarding $T(\lambda)$, one could use the matrix (7.19) to obtain

$$\left(I - \frac{\dot{T}(\lambda)uu^T}{u^T\dot{T}(\lambda)u}\right) T(\lambda) \left(I - \frac{uu^T\dot{T}(\lambda)}{u^T\dot{T}(\lambda)u}\right) s = -r, \tag{7.23}$$

with the additional condition $s \perp_T \dot{T}(\lambda)u$.

Algorithm 23 *Complex symmetric Jacobi–Davidson method (JDCS)*

Input: *starting vector* u, $u^Tu \neq 0$, *tolerance* $\epsilon > 0$

S0: *Set* $U = [\,u\,]$

for $k = 1, 2, \ldots$

S1: *Solve* $U^TT(\lambda)Uc = 0$ *for* (λ, c)

S2: $u = Uc,\ r = T(\lambda)u,\ q = \dot{T}(\lambda)u$

S3: **if** $\|r\|_2/\|u\|_2 < \epsilon$ **stop**

S4: *Compute* $s \perp u$ *from* $\left(I - \frac{qu^T}{u^Tq}\right) T(\lambda)(I - uu^H)s = -r$

S5: $MGS(U, s)$, $s = s/\|s\|_2$, $U = [\,U, s\,]$

In most cases in applications preconditioning is necessary. The preconditioner has to be projected in the same way as $T(\lambda)$ is projected from the left and the right, which can be done as described in §6.4. However, applying such a preconditioner to the correction equation with complex symmetric operator destroys the symmetry, and one has to use a solver for unsymmetric systems, anyway.

The norms used in Algorithm 23 refer to the spectral norm. Also one could orthonormalize s with respect to a complex symmetric modified Gram–Schmidt orthogonalization, as is done in [6], we use the standard inner product, hence standard modified Gram–Schmidt procedure. This is numerically more stable and $U^HU = I$ can always be achieved in contrast to $U^TU = I$ (if there is a quasi-null vector among the eigenvectors), in exact arithmetic. It makes no difference for the correction equation, as we have seen for the bordered matrix system in §4.2.

Since we have not analyzed block Rayleigh functionals, it is not clear if the 2-norm orthogonalization of U leads to worse approximations coming from the Rayleigh–Ritz step. On the other hand, we are only interested in one of the Ritz pairs, where the second order of the Rayleigh functional in the angle should still hold.

7.6 Numerical Examples

When we want to compare complex symmetric methods with the standard (one-sided) methods, we have to make sure that the underlying problem is complex, indeed. Because for symmetric problems with real coefficients, real eigenvalues and eigenvectors can occur, depending on the problem. In this case there are no differences between the Jacobi–Davidson method and its complex symmetric version. So, when we want to see differences in the performance of the methods, we have to ensure that the eigenvector is complex, because otherwise $u^H T(p)u = u^T T(p)u$ and the standard Rayleigh functional gives the same approximation of second order as the complex symmetric functional.

Example 7 Nonoverdamped mass-spring system. *Consider the connected damped spring system as described and illustrated in [87] with masses of weight 1, where all springs have the same constant κ and all dampers have the same constant τ, except for the first and the last one which have 2κ, 2τ, resp.. The vibration of this system is governed by a second-order differential equation, the corresponding quadratic eigenvalue problem is given by*

$$T(\lambda)x = (\lambda^2 M + \lambda C + K)x = 0,$$

where

$$M = I, \quad C = \tau \begin{bmatrix} 3 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 3 \end{bmatrix}, \quad K = \kappa \begin{bmatrix} 3 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 3 \end{bmatrix}.$$

For $\kappa = 10$, $\tau = 1$ the problem is nonoverdamped but stable, since all eigenvalues lie in the left half plane, see the left figure in Figure 7.1.

As we have already observed in Chapter 1, this problem is not only complex symmetric but also Hermitian, which implies that the complex eigenvalues are symmetric with respect to the real axis, i.e., for $\lambda \in \lambda(T(\,.\,))$, we have $\bar{\lambda} \in \lambda(T(\,.\,))$ as well. The addition of both properties implies that for all eigentriples $(\lambda_*, x_*, \bar{x}_*)$, the triples $(\bar{\lambda}, \bar{x}_*, x_*)$ are eigentriples, too.

We started with testing the one-vector methods, comparing original and complex symmetric methods. Results are shown in Table 7.1. We were not able to find initial values for which all methods converged—and which were so bad that differences for the methods appeared. The residual inverse iteration methods were less sensitive with respect to the starting values and converged almost always to the eigenvalue smallest in absolute value

$\lambda_* \approx -0.5 \pm 3.12\mathrm{i}$. The complex symmetric version (CSRII) is reasonably faster than the standard residual inverse iteration (RII).

	$\lvert\lambda_{ls}\rvert$	ls	$\lVert res_u \rVert$	gm	gm/ls	cpu
RFI	4.5e+00	50	2.9e-02	443	8.9	0.62
CSRFI	1.0e+01	50	1.0e+02	224	4.5	0.29
RII	3.2e+00	17	9.7e-12	40	2.4	0.10
CSRII	3.2e+00	12	3.1e-15	17	1.4	0.06

Table 7.1: *Nonoverdamped mass-spring system,* $\lambda_0 = -3\mathrm{i}$.

Since Example 7 refers to a quadratic eigenvalue problem, we can use the MATLAB polyeig-function to solve the projected quadratic eigenproblem inside of the Jacobi–Davidson methods. Table 7.2 and the right figure in Figure 7.1 show the results with the clear advantage of the complex symmetric Jacobi–Davidson algorithm, which is about one third faster than the Jacobi–Davidson algorithm.

	$\lvert\lambda_{ls}\rvert$	ls	$\lVert res_u \rVert$	gm	gm/ls	cpu
JD	3.2e+00	15	5.2e-14	36	2.4	0.78
JDCS	3.2e+00	10	9.9e-14	18	1.8	0.26

Table 7.2: *Nonoverdamped mass-spring system.*

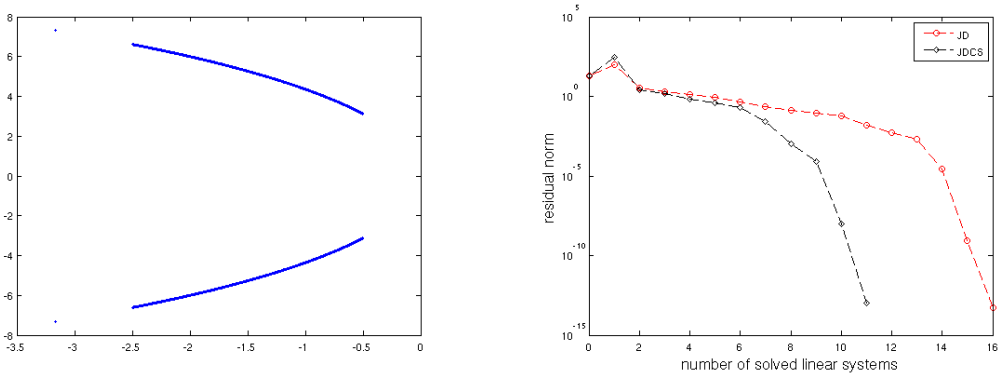

Figure 7.1: *The left figure shows the spectrum of the nonoverdamped mass-spring system, with* $n = 1000$. *The right figure shows the residual norms for the Jacobi–Davidson method and the complex symmetric Jacobi–Davidson method applied to the nonoverdamped mass-spring system.*

The partial differential delay equation example, see Example 5, is also a complex symmetric example, since the matrices A_0, A_1, A_2 and I are real symmetric, but complex eigenvalues occur. Solving this with Jacobi–Davidson-type methods is tricky, since we need to solve a genuine nonlinear problem to obtain approximations from the projected systems. We have already done so in Chapter 4, but as we have seen in Tables 4.7, 5.3 and Figure 6.3, several methods, including the method of successive linear problems, for dense small systems, fail to converge. On the other hand, we have developed two one-vector methods that are adapted to the complex symmetric case, namely the complex symmetric Rayleigh functional iteration and the complex symmetric residual inverse iteration. Both have been shown to converge faster, in theory. Numerical experiments show that when it comes to the question of convergence or divergence, then the behavior of the complex symmetric methods coincides with the behavior of the basic methods, in general. This means that if the original method converges, then the complex symmetric method will do also and the other way round, at least in the one-dimensional case. For the subspace extending methods we find the opposite quite often, in the sense that either JD converges or the complex symmetric JD, in case of non-polynomial problems.

We use the inverse iteration method as inner method for the Jacobi–Davidson algorithm and the complex symmetric Rayleigh functional iteration of Algorithm 21 as inner method for the complex symmetric Jacobi–Davidson algorithm (JDCS). The inner eigensolvers are restricted to at most 20 iterations and stop if the (inner) residual norm is smaller than $1e-10$. The stopping tolerance for the JD methods was also set to $1e-10$.

As we have discussed in the examples section of Chapter 6, it is natural to provide the projected eigenproblem solver with the initial values $(\lambda_k, c_k = U^H u_k)$. If the search space is kept orthogonal with respect to the bilinear form (7.3) in case of the complex symmetric Jacobi–Davidson, i.e., $U^T U = I$ holds approximately, then we have to use $(\lambda_k, c_k = U^T u_k)$ as initial Ritz pair.

In our case it did not make a difference whether we orthogonalized with respect to the inner product $(.,.)$ or the bilinear form $(.,.)_T$. For the partial delay differential system, JD and JDCS often deliver the same convergence history, when the problem stays real. Therefore, we started with a random vector with a small complex part, which resulted in a diverging Jacobi–Davidson method but a fast converging complex symmetric Jacobi–Davidson method, see Table 7.3 and the left figure in Figure 7.2. These versions depend as heavily on the starting value and vectors as the one-vector methods, and we either observe fast convergence or fast divergence, as the example shows. When we start the methods without handing over initial vector approximations to the inner eigenproblem solvers, and use random vectors in the Rayleigh functional iterations instead, but the

	$\|\lambda_{ls}\|$	ls	$\|res\|$	gm	gm/ls	cpu
JD	2.9e+01	49	1.4e+13	50	1.0	1.16
JDCS	1.2e+01	4	4.1e-15	4	1.0	0.01

Table 7.3: *Example PDDE with $n = 10$, current approximations are used as starting vectors for the projected eigenproblem solver.*

recent eigenvalue approximation, then directions will change often. However, in most runs, the methods converged sooner or later. Results for this test are presented in Table 7.4 and the right figure of Figure 7.2. The complex symmetric Jacobi–Davidson seems to

	$\|\lambda_{ls}\|$	ls	$\|res\|$	gm	gm/ls	cpu
JD	5.7e+00	18	4.1e-11	19	1.1	2.21
JDCS	5.7e+00	9	1.6e-11	9	1.0	0.07

Table 7.4: *Example PDDE with $n = 10$, random starting vectors are used for the projected eigenproblem solver everytime.*

be more stable and needs only half of the number of iterations compared to the standard Jacobi–Davidson.

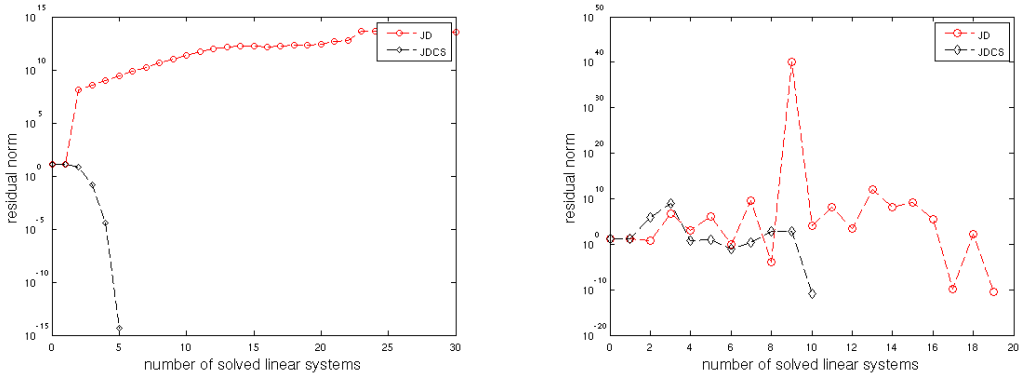

Figure 7.2: *Example PDDE with $n = 10$, the left figure corresponds to the run where current approximations were used as starting vectors for the projected eigenproblem solver. The right figure corresponds to the run where random starting vectors were used for the projected eigenproblem solver everytime.*

The genuine nonlinear eigenvalue problem (7.2) is solved in [51] for $k = 1$ and $k = 2$. Initial eigenpairs close to a shift σ_0^2 are determined by solving the corresponding linearized problem—linearized in the sense of the truncated Taylor series—with the shift-and-invert

Arnoldi method [57]. Then, a nonlinear Rayleigh–Ritz algorithm is proposed, where the projected nonlinear eigenvalue problem is solved by the inverse iteration method (not the complex symmetric version though), and new vectors are gained by the nonlinear Arnoldi update, i.e., the residual inverse iteration step. This makes sense, since the shift σ_0^2, which is already known, is close to the sought eigenvalues. The matrix $T(\sigma_0^2)$ is fixed through the whole process, i.e., new vectors are computed as $x = T(\sigma_0^2)^{-1}T(\lambda)u$, where u is the actual approximation. Instead of using the bilinear form (7.3) for the Ritz–Galerkin condition, Liao kept the search space real by using $Q = \mathrm{orth}([\mathrm{Re}(U)\ \mathrm{Im}(U)])$, i.e., real and imaginary parts of new corrections are orthogonalized. In this way, he achieves that $Q^H T(\lambda)Q$ equals $Q^T T(\lambda)Q$, which is related to the complex symmetric Rayleigh functional. However, this approach doubles the size of the search space. The numerical experiments in [51, p. 63f] show that significant savings can be made in comparison to the case where the complex search space U is used with $U^H T(\lambda)U$. Our analysis shows the reason for this behavior, and that an effective way is to solve with $U^T T(\lambda)U$ for the complex search space U, which saves the orthonormalizing costs and is only of half size compared to $Q^T T(\lambda)Q$.

7.7 Conclusion

Though sometimes not realized as such, complex symmetric eigenvalue problems occur frequently in applications. The characteristic of these problems is the relation between left and right eigenvectors, i.e., the left eigenvector equals the conjugated right eigenvector. This property makes a straightforward application of previous results and algorithms possible, that have been presented in this thesis. In this chapter, we developed the appropriate nonlinear complex symmetric Rayleigh functional, showed local existence of which, a quadratic bound in the angle of the corresponding eigenvector and its aproximation, and stationarity. This Rayleigh functional is the basis for the complex symmetric Rayleigh functional iteration and the complex symmetric residual inverse iteration, which we have introduced and showed to be locally cubically, respectively quadratically, convergent. The complex symmetric Rayleigh functional iteration, in turn, is the basis for the complex symmetric Jacobi–Davidson method, which converges cubically without extra costs compared to the quadratically convergent standard (nonlinear) Jacobi–Davidson method if applied to complex symmetric problems, which different numerical examples confirmed.

Chapter 8

Summary and Outlook

Nonlinear eigenvalue problems are as fascinating as challenging. The nonlinearity seems to put them out of reach somehow. Obtaining general results, which can be used for the actual development of algorithms, is a complicated issue. We are far away from generally working black box solvers nowadays, especially for nonpolynomial problems. However, modern applications frequently generate nonlinear eigenvalue problems, and methods become more advanced simultaneously. However, a thorough theoretical basis is often missing.

The overall topic of this thesis was the nonlinear eigenvalue problem in general. We gave basic facts holding for most of the problems that we know of, which will be helpful for further analysis. In contrast to the theses [11, 11, 51], we did not start out with a special problem, resp. class of problems, but our results apply to a wide range of problems. All kinds of linearizations were avoided, we treated the problem as it is with the minimal assumption that T is twice continuously differentiable. It turned out that assuming $\dot{T}$ to be Lipschitz continuous is not enough to prove the cubic convergence for the two-sided Rayleigh functional iteration. We presented some general facts for nonlinear eigenvalue problems, like the equivalent formulation for an algebraically simple eigenvalue: $\alpha = y_*^H \dot{T}(\lambda_*) x_* \neq 0$. The term α occurs frequently throughout the theoretical analysis in denominators, i.e., the results are restricted to simple eigenvalues. We developed a representation for the inverse operator $T(\lambda)^{-1}$ for $\lambda \notin \lambda(T(\,.\,))$, which motivates the nonlinear inverse iteration techniques, and a term for the condition number of an eigenvalue, followed by basic facts on angles and distances between vectors in the complex plane. We concluded that the angle between two vectors is the proper measure of approximation in the complex plane, rather than the 2-norm distance.

We have considered a generalization of the (two-sided) Rayleigh quotient for nonlinear eigenvalue problems, in particular a standard and a generalized Rayleigh functional, for which we have shown locally unique existence, first order perturbation bounds and er-

ror bounds in the vectors, as well as stationarity. Hence, the Rayleigh functionals are consistent generalizations of the Rayleigh quotients, i.e., applying $T(\lambda) = A - \lambda I$ yields the Rayleigh quotients, and the stationarity property is preserved. However, the results depend strictly on the algebraic simplicity of the considered eigenvalue, meaning that $y_*^H \dot{T}(\lambda_*) x_* \neq 0$ is required. In case of the standard functional, this condition is replaced by $x_*^H \dot{T}(\lambda_*) x_* \neq 0$ which is a different condition if working on non-Hermitian problems.

Chapter 4 provides a detailed Newton theory for the extended system consisting of the eigenvalue problem and a normalization condition. We have derived a bound for the Newton update which couples the eigenvalue and the eigenvector sequence. The results on Rayleigh functionals of Chapter 3 immediately led to the locally cubic convergence of the two-sided Rayleigh functional iteration. This iteration combines the inverse iteration steps for left and right eigenvector with the two-sided Rayleigh functional. Therefore it is the nonlinear equivalent to Ostrowski's two-sided Rayleigh quotient iteration. In the same way, we have developed a two-sided residual inverse iteration and proved its locally quadratic convergence. We have compared both with well-known methods and with methods, which had not been adapted to the nonlinear case before. These one- and two-vector methods are essential for the task of solving small dense nonlinear problems inside of subspace methods for nonlinear eigenvalue problems as the Jacobi–Davidson method and the nonlinear Arnoldi method. We have analyzed costs and assumptions under which the methods work and have run some examples.

We have shown that the two-sided methods can be improved by adding an additional Rayleigh functional computation in between the two correction equations. The R-order of the so-called half-step methods is reasonably larger than the R-order of the original methods, as numerical examples have illustrated. In detail, the R-order of the half-step two-sided Rayleigh functional iteration is 4. This implies that dividing the half-step two-sided Rayleigh functional iteration in two parts yields a method with R-order two that has the same costs as the inverse iteration, but computes the left eigenvector as well. In general, two-sided methods have the advantage that the condition number of the eigenvalue can be estimated during runtime.

The discussed one- and two-vector methods were used for the analysis of the subspace extending Jacobi–Davidson-type methods. Different versions of the nonlinear Jacobi–Davidson algorithm have been analyzed regarding their correction equations and appropriate nonlinear forms have been shown. The theoretical analysis of asymptotic condition numbers revealed that the generalized Jacobi–Davidson method is insensitive to the condition number of the eigenvalue, whilst the upper bound for the condition number of the operator for the two-sided Jacobi–Davidson method depends on the condition number of

the eigenvalue, and for the standard Jacobi–Davidson method upper and lower bound for the condition number of the operator depend on the condition number of the eigenvalue, i.e., theoretically, the Jacobi–Davidson method can deliver inaccurate results in case of ill-conditioned eigenvalues. The numerical experiments in Chapter 6 have shown that all methods cannot compute a strongly ill-conditioned eigenvalue accurately, however, in this particular example this may be connected to the bad left eigenvalue approximation that is gathered by the two-sided methods. The standard Jacobi–Davidson stagnates with an residual norm around 10^{-10} for Example 4, where $x_*^H \dot{T}(\lambda_*)x_* = 0$, whereas the alternating and generalized JD give accurate results. Moreover, we have seen that computing the inner nonlinear (projected) eigenproblem may be problematic, since the available methods (discussed in Chapter 4) are locally convergent methods, which can fail for bad starting values. This is one of the main problems in the computation of eigenvalues for nonlinear problems. Therefore, it is important to have good starting values and vectors—for instance obtained by solving the linearized problem as is done in [51].

One way to obtain a larger region of convergence is to use the concept of damping, i.e., new corrections are multiplied by a parameter ω, $0 < \omega \leq 1$, before they are added to the old approximation, resp. the search space. In this sense, a damped Jacobi–Davidson method could be an alternative, which needs further investigations.

Finally, we have introduced the appropriate Rayleigh functional for complex symmetric problems and have shown stationarity and first order perturbation bounds for it. We have analyzed the corresponding complex symmetric Rayleigh functional iteration and have shown locally cubic convergence. We have derived the complex symmetric residual inverse iteration in the same manner, and have proven at most quadratic convergence. Therefore, the complex symmetric Jacobi–Davidson method converges locally cubically without extra costs compared to the quadratically convergent standard Jacobi–Davidson method. Numerical experiments have illustrated the theoretical results.

The natural next step regarding Chapters 6 and 7 is to formulate and test the two-sided and the complex symmetric nonlinear Arnoldi methods, since both residual inverse iteration methods have been analyzed here, and to compare them with the corresponding Jacobi–Davidson-type method numerically.

In this thesis, we have considered the task to compute one eigenvalue (and corresponding eigenvectors). Applications request in general more than one, or even harder, a set of eigenvalues from the inner spectrum. There are several techniques to obtain more than one eigenvalue, but it is problematic to guarantee that all eigenvalues in a specified domain are found. For linear eigenproblems this can be ensured by shifting, but there is no equivalent formulation for nonlinear problems. Moreover, since there exists no Schur decomposition,

we cannot rely on a complete set of Schur vectors. In few special cases we can compute a partial Schur form, cf. [56].

A promising strategy to obtain a set of eigenvalues is to shift a computed eigenvalue from the boundary of the spectrum to zero or infinity, depending on which eigenvalues are desired, and compute the next one at the outer spectrum afterwards. Since the problem is changed everytime an eigenvalue has converged, this procedure is called nonequivalence deflation [17,29,41]. For the quadratic complex symmetric operator $T(\lambda) = \lambda^2 M + \lambda C + K$, with positive definite K, this looks as follows. Suppose we have a converged eigenpair (λ_1, x_1) available. Since K is positive definite we may scale x_1 such that $x_1^T K x_1 = 1$. Let $\theta_1 = 1/(x_1^T M x_1)$, provided that $x_1^T M x_1 \neq 0$. Then we deflate the original problem by a rank-1-transformation and derive a new complex symmetric quadratic eigenproblem $T_1(\lambda)x = (\lambda^2 M_1 + \lambda C_1 + K_1)x = 0$, where

$$\begin{aligned} M_1 &= M - \theta_1 M x_1 x_1^T M, \\ C_1 &= C + \frac{\theta_1}{\lambda_1}(M x_1 x_1^T K + K x_1 x_1^T M), \\ K_1 &= K - \frac{\theta_1}{\lambda_1^2} K x_1 x_1^T K. \end{aligned}$$

The transformation $T_1(\lambda)$ has the same spectrum as $T(\lambda)$, but λ_1 is mapped to infinity. If $T(\lambda)$ is not complex symmetric then x_1^T must be replaced by y_1^H and we need to compute the left eigenvector additionally everytime. Both eigenvectors are required with high accuracy in order to deflate the problem in a proper way. Otherwise, one iteratively obtains a different problem with different eigenvalues. We have seen several methods that can be used for this. Therefore, it seems reasonable in any case—either for general or for complex symmetric problems—to combine the generalized/complex symmetric Jacobi–Davidson, resp. nonlinear Arnoldi methods, with the nonequivalence deflation. This approach can be used for further comparisons of the different JD methods. In particular the available information on the conditioning of the eigenvalue throughout the process is valuable.

Also not clear at this point, is the quality of approximation given by the solution of the nonlinear Rayleigh–Ritz procedure in dependency on the input. So far, we have only analyzed the one-dimensional case. Obviously, if (λ, c) is a Ritz pair, i.e., $U^H T(\lambda) U c = 0$, then $u^H T(\lambda) u = 0$ with $u = Uc$, hence λ is the Rayleigh functional $p(u)$, and the corresponding bounds hold. However, the other direction, where U includes some perturbed eigenvectors needs further attention.

Bibliography

[1] P.-A. Absil, R. Sepulchre, P. Van Dooren, and R. Mahony. Cubically convergent iterations for invariant subspace computations. *SIAM J. Matrix Anal. Appl.*, 26:70–96, 2004.

[2] P.-A. Absil and P. Van Dooren. Two-sided Grassmann-Rayleigh quotient iteration. Technical Report UCL-INMA-2007.024, Universite Catholique de Louvain, 2007.

[3] A. L. Andrew, K.-W. E. Chu, and P. Lancaster. Derivatives of eigenvalues and eigenvectors of matrix functions. *SIAM J. Matrix Anal. Appl.*, 14:903–906, 1993.

[4] A. L. Andrew, K.-W. E. Chu, and P. Lancaster. On the numerical solution of nonlinear eigenvalue problems. *Computing*, 55:91–111, 1995.

[5] P. M. Anselone and L. B. Rall. The solution of characteristic value-vector problems by Newton's method. *Numer. Math.*, 11:38–45, 1968.

[6] P. Arbenz and M. E. Hochstenbach. A Jacobi-Davidson method for solving complex symmetric eigenvalue problems. *SIAM J. Sci. Comput.*, 25:1655–1673, 2004.

[7] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, editors. *Templates for the Solution of Algebraic Eigenvalue Problems. A Practical Guide*. SIAM, Philadelphia, 2000.

[8] S. Batterson and J. Smillie. The dynamics of Rayleigh quotient iteration. *SIAM J. Numer. Anal.*, 26:624–636, 1989.

[9] S. Batterson and J. Smillie. Rayleigh quotient iteration for nonsymmetric matrices. *Math. Comp.*, 55(191):169–178, 1990.

[10] H. Behnke and F. Sommer. *Theorie der analytischen Funktionen einer komplexen Veränderlichen*. Springer-Verlag, Berlin - Göttingen - Heidelberg, 2nd edition, 1962.

[11] M. Betcke. *Iterative Projection Methods for Symmetric Nonlinear Eigenvalue Problems with Applications*. PhD thesis, TU Hamburg-Harburg, 2007.

[12] M. Betcke and H. Voss. Stationary Schrödinger equations governing electronic states of quantum dots in the presence of spin-orbit splitting. *Appl. Math.*, 52(3):267–284, 2007.

[13] T. Betcke and H. Voss. A Jacobi-Davidson-type projection method for nonlinear eigenvalue problems. *Future Generation Computer Systems*, 20:363–372, 2004.

[14] T. Braconnier and F. Chaitin-Chatelin. Roundoff induces a chaotic behaviour for eigensolvers applied on highly nonnormal matrices. In J. Perriaux et al., editor, *Computational Science for the 21st Century*. John Wiley, 1997.

[15] P. Büchel, M. Lücke, D. Roth, and R. Schmitz. Pattern selection in the absolutely unstable regime as a nonlinear eigenvalue problem: Taylor vortices in axial flow. *Phys. Rev. E*, 53:4764–4777, 1996.

[16] A. Bunse-Gerstner and R. Stöver. On a conjugate gradient-type method for solving complex symmetric linear systems. *Lin. Alg. Appl.*, 287:105–123, 1998.

[17] J. Carvalho, B. N. Datta, W.-W. Lin, and C.-S. Wang. Eigenvalue embedding in a quadratic pencil using symmetric low rank updates. Technical report 2001–3, National Center for Theoretical Science, National Tsinghua University, Hsinchu 300, Taiwan, Republic of China, 2001.

[18] R. Courant. Über die Eigenwerte bei den Differentialgleichungen der mathematischen Physik. *Math. Z.*, 7:1–57, 1920.

[19] R. J. Duffin. A minmax theory for overdamped networks. *Arch. Rational Mech. Anal.*, 4:221–233, 1955.

[20] N. A. Dumont. On the solution of generalized non-linear complex-symmetric eigenvalue problems. *Int. J. Numer. Meth. Engng.*, 71:1534–1568, 2007.

[21] E. Fischer. Über quadratische Formen mit reellen Koeffizienten. *Monatshefte für Mathematik und Physik*, 16:234–249, 1905.

[22] D. R. Fokkema, G. L. G. Sleijpen, and H. A. van der Vorst. Accelerated inexact Newton schemes for large systems of nonlinear equations. *SIAM J. Sci. Comput.*, 19:657–674, 1998.

[23] M. Freitag and A. Spence. Rayleigh quotient iteration and simplified Jacobi–Davidson method with preconditioned iterative solves. Technical report, University of Bath, http://hdl.handle.net/10247/167, 2007. to appear in Lin. Alg. Appl.

[24] R. W. Freund. Conjugate gradient-type methods for linear systems with complex symmetric coefficient matrices. *SIAM Sci. Stat. Comp.*, 13:425–448, 1992.

[25] I. Gohberg, P. Lancaster, and L. Rodman. *Matrix Polynomials*. Academic Press, New York, 1982.

[26] M. A. Golberg. A generalized Rayleigh quotient for eigenvalue problems nonlinear in the parameter. *J. Optimization Theory Appl.*, 11:146–158, 1973.

[27] P. Guillaume. Nonlinear eigenproblems. *SIAM J. Matrix Anal. Appl.*, 20:575–595, 1999.

[28] C.-H. Guo, N. J. Higham, and F. Tisseur. Detecting and solving hyperbolic quadratic eigenvalue problems. MIMS EPrint 2007.117, Manchester Institute for Mathematical Sciences, The University of Manchester, 2007.

[29] J.-S. Guo, W.-W. Lin, and C.-S. Wang. Numerical solutions for large sparse quadratic eigenvalue problems. *Linear Algebra Appl.*, 225:57–89, 1995.

[30] K. P. Hadeler. Mehrparametrige und nichtlineare Eigenwertaufgaben. *Arch. Rational Mech. Anal.*, 27:306–328, 1967.

[31] K. P. Hadeler. Variationsprinzipien bei nichtlinearen Eigenwertaufgaben. *Arch. Rational Mech. Anal.*, 30:297–307, 1968.

[32] H. Heuser. *Lehrbuch der Analysis, Teil 2*. B.G. Teubner Stuttgart, Leipzig, 10. edition, 1998.

[33] V. Heuveline and C. Bertsch. On multigrid methods for the eigenvalue computation of nonselfadjoint elliptic operators. *East-West J. Numer. Math.*, 8:275–297, 2000.

[34] N. J. Higham, D. S. Mackey, N. Mackey, and F. Tisseur. Symmetric linearizations for matrix polynomials. *SIAM J. Matrix Anal. Appl.*, 29(1):143–159, 2006.

[35] N. J. Higham, D. S. Mackey, and F. Tisseur. Definite matrix polynomials and their linearization by definite pencils. MIMS EPrint 2007.97, School of Mathematics, University of Manchester, 2007.

[36] N. J. Higham, D. S. Mackey, F. Tisseur, and S. D. Garvey. Scaling, sensitivity and stability in the numerical solution of quadratic eigenvalue problems. *Int. J. Numer. Meth. Engng*, 73:344–360, 2008.

[37] M. E. Hochstenbach and G. L. G. Sleijpen. Two-sided and alternating Jacobi-Davidson. *Lin. Algebra Appl.*, 358:145–172, 2003.

[38] M. E. Hochstenbach and G. L. G. Sleijpen. Harmonic and refined Rayleigh–Ritz for the polynomial eigenvalue problem. *Num. Lin. Alg. Appl.*, 15(1):35–54, 2008.

[39] R. Horn and C. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, New York, 1985.

[40] H. Y. Hu and Z. H. Wang. *Dynamics of Controlled Mechanical Systems with Delayed Feedback*. Springer, Berlin, 2002.

[41] T.-M. Hwang, W.-W. Lin, and V. Mehrmann. Numerical solution of quadratic eigenvalue problems with structure-preserving methods. *SIAM J. Sci. Comput.*, 24(4):1283–1302, 2003.

[42] I. Ipsen. Computing an eigenvector with inverse iteration. *SIAM Rev.*, 39:254–291, 1997.

[43] E. Jarlebring. *The Spectrum of Delay-Differential Equations: Numerical Methods, Stability and Perturbation*. PhD thesis, TU Braunschweig, 2008.

[44] E. Jarlebring and H. Voss. Rational Krylov for nonlinear eigenproblems, an iterative projection method. *Appl. Math.*, (50):543–554, 2005.

[45] L. Kaufman. Eigenvalue problems in fiber optic design. *SIAM J. Matrix Anal. Appl.*, 28(1):105–117 (electronic), 2006.

[46] A Kiełbasiński and H. Schwetlick. *Numerische lineare Algebra. Eine computerorientierte Einführung*. Deutscher Verlag der Wissenschaften, Berlin, 1988. Also: Harri Deutsch Verlag, Thun-Frankfurt, 1988.

[47] D. Kressner. *Numerical Methods for General and Structured Eigenvalue Problems*. Lecture Notes in Computational Science and Engineering. Springer, Berlin - Heidelberg, 2005.

[48] P. Lancaster. A generalised Rayleigh quotient iteration for lambda-matrices. *Arch. Rat. Mech. Anal.*, 8:309–322, 1961.

[49] P. Lancaster. *Lambda-Matrices and Vibrating Systems*. Dover Publications, Mineola, New York, 2002.

[50] H. Langer. Über stark gedämpfte Scharen im Hilbertraum. *J. Math. Mech.*, 17:685–705, 1968.

[51] B.-S. Liao. *Subspace Projection Methods for Model Order Reduction and Nonlinear Eigenvalue Computation*. PhD thesis, University of California, Davis, 2007.

[52] B.-S. Liao, Z. Bai, L.-Q. Lee, and K. Ko. Solving large scale nonlinear eigenvalue problems in next-generation accelerator design. Preprint August 19, 2006, University of California, Davis, CSE-TR, 2001.

[53] D. S. Mackey, N. Mackey, C. Mehl, and V. Mehrmann. Palindromic polynomial eigenvalue problems: Good vibrations from good linearizations. *SIAM J. Matrix Anal. Appl.*, 28:1029–1051, 2006.

[54] D. S. Mackey, N. Mackey, C. Mehl, and V. Mehrmann. Vector spaces of linearizations for matrix polynomials. *SIAM J. Matrix Anal. Appl.*, 28(4):971–1004, 2006.

[55] A. S. Markus. *Introduction to the Spectral Theory of Polynomial Operator Pencils*. American Mathematical Society, Providence, RI, USA, 1988.

[56] K. Meerbergen. Locking and restarting quadratic eigenvalue solvers. *SIAM J. Sci. Comput.*, 22:1814–1839, 2001.

[57] K. Meerbergen and A. Spence. Implicitly restarted Arnoldi with purification for the shift-invert transformation. *Math. Comp.*, 66:667–689, 1997.

[58] V. Mehrmann and H. Voss. Nonlinear eigenvalue problems: A challenge for modern eigenvalue methods. *GAMM–Mitteilungen*, 27:121–152, 2004.

[59] V. Mehrmann and D. Watkins. Structure-preserving methods for computing eigenpairs of large sparse skew-Hamiltonian/Hamiltonian pencils. *SIAM J. Sci. Comput.*, 22:1905–1925, 2001.

[60] V. Mehrmann and D. Watkins. Polynomial eigenvalue problems with Hamiltonian structure. *Electr. Trans. Num. Anal.*, 13:106–113, 2002.

[61] A. Neumaier. Residual inverse iteration for the nonlinear eigenvalue problem. *SIAM J. Numer. Anal.*, 22:914–923, 1985.

[62] S.-I. Niculescu. *Delay effects on stability: A robust control approach.* Springer, London, 2001.

[63] Y. Notay. Convergence analysis of inexact Rayleigh quotient iteration. *SIAM J. Matrix Anal. Appl.*, 24:627–644, 2003.

[64] J. M. Ortega and W. C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables.* Academic Press, New York, London, 1970. Republished 2000 by SIAM, Philadelphia.

[65] M. R. Osborne. A new method for the solution of eigenvalue problems. *Comput. J.*, 7:228–232, 1964.

[66] A. Ostrowski. On the convergence of the Rayleigh quotient iteration for the computation of the characteristic roots and vectors. I–VI. *Arch. Rational Mech. Anal.*, 1–4:1:233–241, 2:423–428, 3:325–340, 3:341–347, 3:472–481, 4:153–165, 1958/59.

[67] B. N. Parlett. The Rayleigh quotient iteration and some generalizations for nonnormal matrices. *Math. Comp.*, 28:679–693, 1974.

[68] H. Poincaré. Sur les équations aux dérivées partielles de la physique mathématique. *Amer. J. Math.*, 12:211–294, 1890.

[69] Lord Rayleigh. Some general theorems relating to vibrations. *Proc. London Math. Soc.*, 4:357–368, 1873.

[70] E. H. Rogers. A minimax theory for overdamped systems. *Arch. Rational Mech. Anal.*, 16:89–96, 1964.

[71] K. Rothe. *Lösungsverfahren für nichtlineare Matrixeigenwertaufgaben mit Anwendungen auf die Ausgleichselementmethode.* Wiss. Beitr. aus europ. Hochschulen, Reihe 11, Mathematik. Verlag an der Lottbek, Ammersbek, 1989. Diss. Univ. Hamburg 1989.

[72] A. Ruhe. Algorithms for the nonlinear eigenvalue problem. *SIAM J. Numer. Anal.*, 10:674–689, 1973.

[73] A. Ruhe. Rational Krylov, a practical algorithm for large sparse nonsymmetric matrix pencils. *SIAM J. Sci. Comp.*, 19:1535–1551, 1998.

[74] J. W. Schmidt. On the R-order of coupled sequences. *Computing*, 26:333–342, 1981.

[75] C. Schröder. *Palindromic and Even Eigenvalue Problems - Analysis and Numerical Methods*. PhD thesis, TU Berlin, 2008.

[76] H. Schwetlick. *Numerische Lösung nichtlinearer Gleichungen*. Deutscher Verlag der Wissenschaften, Berlin, 1979. Auch: R. Oldenbourg Verlag, München-Wien, 1979.

[77] H. Schwetlick and R. Lösche. A generalized Rayleigh quotient iteration for computing simple eigenvalues of nonnormal matrices. *Z. Angew. Math. Mech.*, 80:9–25, 2000.

[78] H. Schwetlick and K. Schreiber. A primal-dual Jacobi-Davidson-like method for nonlinear eigenvalue problems. *ZIH Preprint, TU Dresden*, ZIH-IR-0613:1–20, 2006.

[79] V. S. Shchesnovich and S. B. Cavalcanti. Rayleigh functional for nonlinear systems. *ArXiv Nonlinear Sciences e-prints*, November 2004.

[80] V. Simoncini and D. B. Szyld. Theory of inexact Krylov subspace methods and applications to scientific computing. *SIAM J. Scient. Computing*, 25:454–477, 2003.

[81] G. L. G. Sleijpen and H. A. van der Vorst. A Jacobi-Davidson iteration method for linear eigenvalue problems. *SIAM J. Matrix Anal. Appl.*, 17:401–425, 1996.

[82] S. I. Solov′ëv. Preconditioned iterative methods for a class of nonlinear eigenvalue problems. *Linear Algebra Appl.*, 415(1):210–229, 2006.

[83] A. Spence and C. Poulton. Photonic band structure calculations using nonlinear eigenvalue techniques. *J. Comput. Phys.*, 204:65–81, 2005.

[84] G. Stepan. *Retarded Dynamical Systems: Stability and Characteristic Functions*. Longman, New York, 1989.

[85] G. W. Stewart and J. Sun. *Matrix Perturbation Theory*. Academic Press, Boston, 1990.

[86] F. Tisseur. Backward error and condition of polynomial eigenvalue problems. *Linear Algebra Appl.*, 309:339–361, 2000.

[87] F. Tisseur and K. Meerbergen. The quadratic eigenvalue problem. *SIAM Rev.*, 43:235–286, 2001.

[88] R. E. L. Turner. Some variational principles for nonlinear eigenvalue problems. *J. Math. Anal. Appl.*, 17:151–160, 1967.

[89] R. E. L. Turner. A class of nonlinear eigenvalue problems. *J. Func. Anal.*, 7:297–322, 1968.

[90] H. Unger. Nichtlineare Behandlung von Eigenwertaufgaben. *Z. Angew. Math. Mech.*, 30:281–282, 1950.

[91] H. A. van der Vorst and J. B. M. Melissen. A Petrov–Galerkin type method for solving $Ax = b$, where A is symmetric complex. *IEEE Trans. Magnetics*, 26:706–708, 1990.

[92] H. Voss. A maxmin principle for nonlinear eigenvalue problems with application to a rational spectral problem in fluid-solid vibration. *Appl. Math.*, 48:607–622, 2003.

[93] H. Voss. A rational spectral problem in fluid-solid vibration. *Elec. Trans. Num. Anal.*, 16:94–106, 2003.

[94] H. Voss. An Arnoldi method for nonlinear eigenvalue problems. *BIT*, (44):387–401, 2004.

[95] H. Voss. Eigenvibrations of a plate with elastically attached loads. In P. Neittaanmäki, T. Rossi, S. Korotov, E. Onate, J. Periaux, and D. Knörzer, editors, *Proceedings of the European congress on computational methods in applied sciences and engineering*, ISBN 951-39-1869-6, Jyväskylä, Finland, 2004.

[96] H. Voss. Numerical methods for sparse nonlinear eigenvalue problems. Report 70, Arbeitsbereich Mathematik, TU Hamburg-Harburg, 2004. Proc. XVth Summer School on Software and Algorithms of Numerical Mathematics, Hejnice, Czech Republic.

[97] H. Voss. A Jacobi–Davidson method for nonlinear and nonsymmetric eigenproblems. *Computers & Structures*, 85:1284–1292, 2007.

[98] H. Voss. A minmax principle for nonlinear eigenproblems depending continuously on the eigenparameter. Report 122, TU Hamburg–Harburg, 2008.

[99] H. Voss and B. Werner. A minimax principle for nonlinear eigenvalue problems with applications to nonoverdamped systems. *Math. Meth. Appl. Sci.*, 4:415–424, 1982.

[100] H. Voss and B. Werner. Solving sparse nonlinear eigenvalue problems. Technical report, Inst. für Angew. Mathematik, Universität Hamburg, 1982.

[101] N. Wagner and L. Gaul. Eigenpath analysis of transcendental two-parameter eigenvalue problems. European Congress on Computational Methods in Applied Sciences and Engineering ECCOMAS, Jyväskylä, 24–28 July 2004, http://www.imamod.ru/∼serge/arc/conf/ECCOMAS_2004/ECCOMAS_V2/proceedings/pdf/389.pdf, 2004.

[102] B. Werner. Das Spektrum von Operatorenscharen mit verallgemeinerten Rayleighquotienten. *Arch. Rat. Mech. Anal.*, 42:223–238, 1971.

[103] H. Weyl. Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). *Math. Ann.*, 71:441–479, 1912.

[104] Y. Zhou. Studies on Jacobi–Davidson, Rayleigh quotient iteration, inverse iteration generalized Davidson and Newton updates. *Numer. Linear Algebra Appl.*, 13(7):621–642, 2006.