Algorithms and theory for polynomial eigenproblems

Taslaman, Leo

2014

MIMS EPrint: **2015.4**

Manchester Institute for Mathematical Sciences
School of Mathematics

The University of Manchester

Reports available from: http://eprints.maths.manchester.ac.uk/
And by contacting: The MIMS Secretary
School of Mathematics
The University of Manchester
Manchester, M13 9PL, UK

ISSN 1749-9097

ALGORITHMS AND THEORY FOR POLYNOMIAL EIGENPROBLEMS

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Engineering and Physical Sciences

2014

Leo Taslaman
School of Mathematics

Contents

The University of Manchester

Leo Taslaman
Doctor of Philosophy
Algorithms and theory for polynomial eigenproblems
December 22, 2014

In this thesis we develop new theoretical and numerical results for matrix polynomials and polynomial eigenproblems. This includes the cases of standard and generalized eigenproblems.

Two chapters concern quadratic eigenproblems $(M\lambda^2 + D\lambda + K)x = 0$, where M, D and K enjoy special properties that are commonly encountered in modal analysis. We discuss this application in some detail, in particular the mathematics behind discrete dampers. We show how the physical intuition of a damper that gets stronger and stronger can be mathematically proved using matrix analysis. We then develop an algorithm for quadratic eigenvalue problems with low rank damping, which outperforms existing algorithm both in terms of speed and accuracy. The first part of our algorithm requires the solution of a generalized eigenproblem with semidefinite coefficient matrices. To solve this problem we develop a new algorithm based on an algorithm proposed by Wang and Zhao [SIAM J. Matrix Anal. Appl. 12-4 (1991), pp. 654–660]. The new algorithm computes all eigenvalues in a backward stable and symmetry preserving manner.

The next two chapters are about equivalences of matrix polynomials. We show, for an algebraically closed field $\mathbb{F}$, that any matrix polynomial $P(\lambda) \in \mathbb{F}[\lambda]^{n\times m}$, $n \leq m$, can be reduced to triangular form, while preserving the degree and the finite and infinite elementary divisors. We then show that the same result holds for real matrix polynomials if we replace "triangular" with "quasi-triangular," that is, block-triangular with diagonal blocks of size 1×1 and 2×2. The proofs are constructive in the sense that we build up triangular and quasi-triangular matrix polynomials starting the Smith form. In this sense we are solving structured inverse problems. In particular, our results imply that the necessary constraints that make list of elementary divisors admissible for a real square matrix polynomial of degree ℓ are also sufficient conditions. For the case of matrix polynomials with invertible leading coefficients, we show how triangular/quasi-triangular forms, as well as diagonal and Hessenberg forms, can be computed numerically.

Finally, we present a backward error analysis of the shift-and-invert Arnoldi algorithm for matrices. This algorithm is also of interest to polynomial eigenproblems with easily constructible monic linearizations. The analysis shows how errors from the linear system solves and orthonormalization process affect the Arnoldi recurrence. Residual bounds for linear systems and columnwise backward error bounds for QR factorizations come to play, so we discuss these in some detail. The main result is a set of backward error bounds that can be estimated cheaply. We also use our error analysis to define a sensible condition for "breakdown," that is, a condition for when the Arnoldi iteration should be stopped.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Publications

Four chapters of this thesis are based on work that have either been published or been submitted for publication:

- Chapter 3 is based on the manuscript "Strongly damped quadratic matrix polynomials" [78], MIMS EPrint 2014.10, Manchester Institute for Mathematical Sciences, March 2014. This work has been submitted for publication.
- Chapter 4 is based on the manuscript "An algorithm for quadratic eigenproblems with low rank damping" [77], MIMS EPrint 2014.21, Manchester Institute for Mathematical Sciences, May 2014. This work has been submitted for publication.
- Chapter 5 is based on the paper "Triangularizing matrix polynomials" [79] (with Françoise Tisseur and Ion Zaballa), Linear Algebra and its Applications, 439(7):1679–1699, 2013.
- Chapter 7 is based on the manuscript "Backward error analysis of the shift-and-invert Arnoldi algorithm" [68] (with Christian Schröder), MIMS EPrint 2014.53, Manchester Institute for Mathematical Sciences, October 2014. This work has been submitted for publication.

Copyright Statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made **only** in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=487), in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University's Policy on Presentation of Theses.

Acknowledgements

I sincerely thank my supervisor Françoise Tisseur for all her guidance and advice, and for everything else. I am also grateful to Nick Higham for his engagement in the numerical linear algebra group at the University of Manchester. It has been my privilege to be a part of such a large and active group. I thank my collaborators Yuji Nakatsukasa, Christian Schröder and Ion Zaballa, and also my colleagues Edvin Deadman, Ramaseshan Kannan, Vanni Noferini and Sam Relton for many helpful discussions. I also thank my family and friends for their support, and last, but definitely not least, Lijing, both for all the discussions about mathematics we have had, and for the moral support she has given me.

Chapter

1 Introduction

In this thesis we study matrix polynomials and polynomial eigenproblems, both from a theoretical and an algorithmic point of view. This includes the special cases of standard and generalized eigenproblems. With a matrix polynomial we mean a matrix with entries in a polynomial ring $\mathbb{F}[\lambda]$. Here $\mathbb{F}$ can be any field, but we shall only be concerned with algebraically closed fields and the field of real numbers. We may, equivalently, consider a matrix polynomial as an expression

$$P(\lambda) = A_\ell \lambda^\ell + A_{\ell-1}\lambda^{\ell-1} + \cdots + A_0,$$

where the A_j are matrices over $\mathbb{F}$. We assume that matrix polynomials written in this notation always have nonzero leading coefficients A_ℓ. Here ℓ is the *degree* of the matrix polynomial, and if $A_\ell = I$ we say that $P(\lambda)$ is *monic*.

Much of the theory that exists for matrices over fields has been generalized to matrix polynomials. From a practical point of view, most important is probably the theory of eigenvalues and eigenspaces. In particular the generalizations to *regular* matrix polynomials, that is, square matrix polynomials with determinants that do not vanish identically.[1] If $P(\lambda)$ is regular, its *finite eigenvalues* are the points λ_0 such that $P(\lambda_0)$ is singular. We further say that $P(\lambda)$ has an eigenvalue at infinity if its leading coefficient A_ℓ is singular . The set of all eigenvalues, including infinity if A_ℓ is singular, is called the *spectrum* of $P(\lambda)$. If λ_0 is a finite eigenvalue, the

[1] In some of the early works on matrix polynomials, e.g., [50], the term "regular" refers to matrix polynomials with invertible leading coefficient.

null spaces of $P(\lambda_0)^T$ and $P(\lambda_0)$ are the left and right *eigenspaces*, respectively.[2] Similarly, the left and right eigenspaces associated with infinity are defined as the null spaces of A_ℓ^T and A_ℓ, respectively. As usual, a *left eigenvector* associated with an eigenvalue is a nonzero vector in the corresponding left eigenspace, and similarly a *right eigenvector* is one in the corresponding right eigenspace. When there is no confusion, we use the convention to omit the prefix "right" when referring to right eigenvectors and eigenspaces.

We remark that unlike the case of constant matrices, eigenvectors associated with different eigenvalues are not necessarily linearly independent. For example, $[1, 0]^T$ is an eigenvector associated with all finite eigenvalues of the matrix polynomial

$$\begin{bmatrix} p(\lambda) & 1 \\ 0 & 1 \end{bmatrix}, \quad \deg p(\lambda) > 1,$$

regardless of what the roots of $p(\lambda)$ are.

A polynomial eigenvalue problem (PEP), or simply a polynomial eigenproblem, is to find some or all eigenvalues, and possibly also the associated eigenvectors, of a given matrix polynomial. We note that the standard and generalized eigenvalue problems, $Ax = \lambda x$ and $Ax = \lambda Bx$, are the special cases obtained by setting $P(\lambda) = I\lambda - A$ and $P(\lambda) = B\lambda - A$, respectively. Inverse polynomial eigenvalue problems are also of interest. These are problems where we are given a set of eigenvalues and possible other data (such as eigenvectors, degree, etc.) and the question is whether the given data is admissible by some matrix polynomial and in that case, can we build one?

1.1 Differential equations

The polynomial eigenvalue problem has received a lot of attention due to its close connection to systems of differential equations with constant coefficients. Such equations appear frequently in various engineering applications; in particular equations of second order, which corresponds to quadratic eigenproblems (QEPs). We discuss one of these applications, namely *modal analysis*, in more detail in chapters 3 and 4. For more applications where polynomial eigenproblems arise see [10, 83] and the references therein. For now, we discuss the connection between differential equations and PEPs in more general terms. Consider the following

[2] It is important to note the transpose in this definition. There different conventions in the literature and sometimes, when the underlying field is $\mathbb{C}$, the conjugate transpose is used instead.

differential equation:

$$P\left(\frac{d}{dt}\right)u(t) = \left(A_\ell \frac{d^\ell}{dt^\ell} + A_{\ell-1}\frac{d^{\ell-1}}{dt^{\ell-1}} + \cdots + A_0\right)u(t) = f(t), \tag{1.1}$$

where the A_i are complex matrices and $t \in [0, T]$. If we assume that $P(\lambda)$ is regular with a finite eigenvalue λ_j and associated eigenvector x_j, then it is easy to see that $u(t) = e^{t\lambda_j}x_j$ is a solution of the homogeneous problem

$$P\left(\frac{d}{dt}\right)u(t) = 0. \tag{1.2}$$

Using a canonical form for matrix polynomials called the Smith form (introduced in Chapter 2), it can be shown that the solution space of (1.2) has dimension $\deg\det P(\lambda)$. This is known as Chrystal's theorem after George Chrystal [18]. See also [32, Theorem S1.6] or [51, p. 276] for a proof. It follows that when all finite eigenvalues of $P(\lambda)$ are *semisimple*, that is, their multiplicities as roots of $\det P(\lambda)$ coincide with the dimension of the associated eigenspaces, it holds that the general solution of (1.2) can be written as

$$u_h(t) = \sum_{j=1}^{\deg\det P(\lambda)} \alpha_j e^{\lambda_j t}x_j, \tag{1.3}$$

where the α_j are arbitrary constants that depend on the initial conditions, the λ_j are the finite (not necessarily distinct) eigenvalues and the x_j are associated eigenvectors such that those corresponding to the same eigenvalue are linearly independent.

We now restrict our attention to differential equations of the form (1.1) with the additional assumptions that the leading coefficient is invertible and that all eigenvalues of $P(\lambda)$ are semisimple. Since the general homogeneous solution is given by (1.3), we only discuss how to find a particular solution. We follow the approach taken by Lancaster [50]. It can be shown [50, pp. 60–65] that

$$\lambda^k P(\lambda)^{-1} = \sum_{j=1}^{n\ell} \frac{\lambda_j^k x_j y_j^T}{\lambda - \lambda_j} + \delta_{k\ell}A_\ell^{-1}, \quad k = 0\colon\ell, \tag{1.4}$$

where $\delta_{k\ell}$ is the Kronecker delta, the λ_j are the eigenvalues of $P(\lambda)$ and the x_j and y_j are associated right and left eigenvectors, respectively, chosen such that

$$y_i^T P'(\lambda_j)x_j = \delta_{ij} \quad \text{if} \quad \lambda_i = \lambda_j. \tag{1.5}$$

To find a particular solution $u_p(t)$ of (1.1), we use (1.4) and write

$$u_p(t) = P\left(\frac{d}{dt}\right)^{-1} f(t) = \sum_{j=1}^{n\ell} x_j y_j^T \left(\frac{d}{dt} - \lambda_j\right)^{-1} f(t). \tag{1.6}$$

To give this expression a meaning we define $(d/dt - \lambda_j)^{-1} f(t)$ to be any function $g(t)$ such that

$$\left(\frac{d}{dt} - \lambda_j\right) g(t) = f(t), \quad t \in [0, T].$$

Since we may simply verify that $u_p(t)$ is a solution in the end, no justification of this definition is necessary. By the method of integrating factor, we get that

$$g(t) = \int_0^t e^{\lambda_j (t-s)} f(s)\, ds$$

is one such function, and one function is enough since we are only looking for a particular solution. Substituting this expression into (1.6), yields

$$u_p(t) = \sum_{j=1}^{n\ell} x_j y_j^T \int_0^t e^{\lambda_j (t-s)} f(s)\, ds\,. \tag{1.7}$$

To see that (1.7) is a solution of (1.1), we note that (1.4) yields

$$\sum_{j=1}^{n\ell} \frac{\lambda^\ell - \lambda_j^\ell}{\lambda - \lambda_j} x_j y_j^T = A_\ell^{-1} \quad \Longleftrightarrow \quad \sum_{j=1}^{n\ell} (\lambda^{\ell-1} + \lambda^{\ell-2}\lambda_j + \cdots + \lambda_j^{\ell-1}) x_j y_j^T = A_\ell^{-1},$$

which in turn, by comparing the coefficients, implies that

$$\sum_{j=1}^{n\ell} \lambda_j^k x_j y_j^T = \delta_{k(\ell-1)} A_\ell^{-1}, \quad k = 0 \colon \ell - 1.$$

If we bear this in mind while differentiating (1.7), it is easy to see that

$$\frac{d^k}{dt^k} u_p(t) = \sum_{j=1}^{n\ell} \lambda_j^k x_j y_j^T \int_0^t e^{\lambda_j (t-s)} f(s)\, ds + \delta_{k\ell} A_\ell^{-1} f(t).$$

With these expressions for the derivatives, it is straightforward to verify that (1.7) is indeed a solution of (1.1).

Now, suppose the right hand side of (1.1) is $f(t) = f_0 e^{\theta t}$ for some θ outside the spectrum of $P(\lambda)$. This choice is interesting in view of the superposition

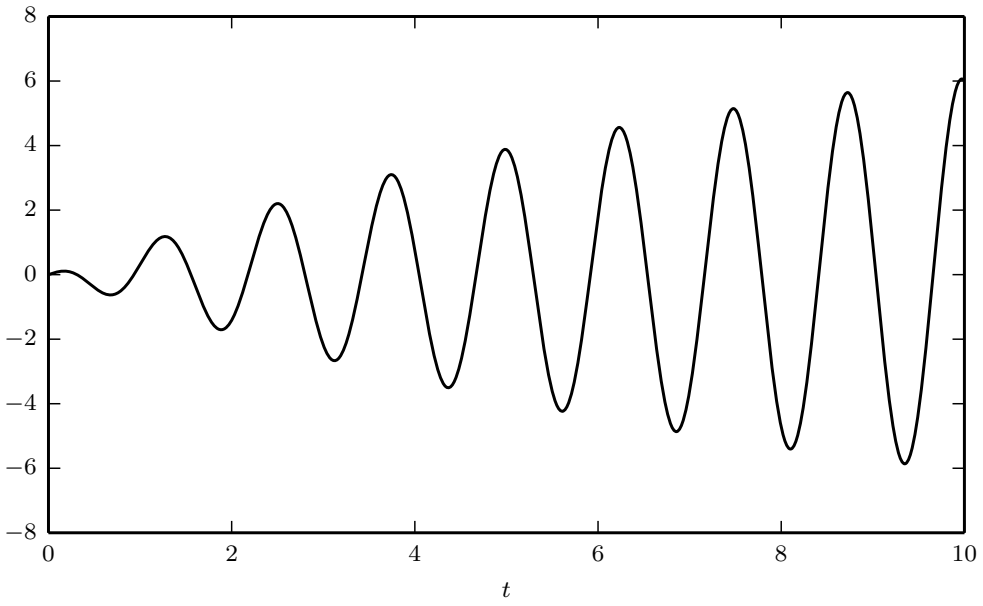

Figure 1.1: Plot of real part of the function $\psi_j(t)$ in (1.9) for $\theta = 5i$ and $\lambda_j = -0.1 + 5.1i$.

principle, since many practical right hand sides can be expanded in Fourier series and hence are sums of such functions. Substituting $f(t) = f_0 e^{\theta t}$ into (1.7) yields

$$u_p(t) = \sum_{j=1}^{n\ell} x_j y_j^T f_0 \frac{e^{\theta t} - e^{\lambda_j t}}{\theta - \lambda_j}. \tag{1.8}$$

The quotients in this expression are important: when θ is sufficiently close to λ_j and t is not too large, we get

$$\psi_j(t) := \frac{e^{\theta t} - e^{\lambda_j t}}{\theta - \lambda_j} = (t + O((\theta - \lambda_j)t^2))e^{\lambda_j t} \approx te^{\lambda_j t}. \tag{1.9}$$

If λ_j has small real part and a moderate to large imaginary part, we see that $\mathrm{Re}(\psi_j(t))$ oscillates with increasing amplitude (Figure 1.1). This is the reason behind the phenomenon called *resonance*, but it is not the full story. It may happen that $y_j^T f_0 = 0$, in which case $\psi_j(t)$ has no influence on the solution. We remark that the left and right eigenvectors y_j and x_j are normalized according to (1.5), so their norms depend on λ_j. Thus $\|x_j\|\|y_j^T f_0\||\psi_j(t)|$ may be small even if $\psi_j(t)$ is huge. It is also possible to have arbitrarily large terms $x_j y_j^T \psi_j(t)$ that cancel each other out when forming the sum (1.8). We will come back to this in Chapter 3.

Finally, we mention that more general differential equations of the form (1.1)

are also strongly connected to the underlying matrix polynomials. In Chapter 2 we will briefly discuss the case when the leading coefficient is invertible and there is no semisimplicity constraint on the eigenvalues. For other types of matrix polynomials it gets more complicated. We refer the interested reader to [32, Chapter 8] for the case of regular matrix polynomials with singular leading coefficients, and to [27, Chapter VII, § 7],[16] and [48], for the case of non-regular matrix polynomials.

1.2 Thesis outline

In Chapter 2 we introduce definitions and results, most of which are not new, that are of relevance to more than one chapter.

Chapter 3 concerns modal analysis of vibrating systems with discrete dampers. We discuss this application in some detail and connect the physics with the underlying mathematical quantities. We then show that the physical intuition of a damper that gets stronger and stronger can be proved using matrix analysis. In essence, it is shown that a structure with only very strong dampers appears to be practically undamped. We also prove a real symmetric version of (1.4), and use this formula to study particular solutions of the forced response problem.

In Chapter 4 we develop an algorithm for quadratic eigenproblems with damping matrices of low rank. Such QEPs appear in modal analysis of structures with discrete dampers. To this end, we first develop a new algorithm, based on an algorithm proposed by Wang and Zhao [89], that solves the associated undamped problem. The new algorithm is, to the author's knowledge, the first one that can find all eigenvalues of a regular pencil $A - B\lambda$, where both A and B are Hermitian and semidefinite, in a backward stable and symmetry preserving manner. That is, the algorithm computes two diagonal nonnegative matrices D_1 and D_2, such that $D_1 - D_2\lambda$ is congruent to a pencil $(A + \Delta A) - (B + \Delta B)\lambda$, where ΔA and ΔB are Hermitian and small in norm with respect to A and B, respectively. We then use our new algorithm in combination with an efficient Ehrlich-Aberth iteration. Finally, if the eigenvectors are desired, we compute these using an inverse iteration that is based on the Takagi factorization for complex symmetric matrices. Both the Ehrlich-Aberth iteration and the inverse iteration work exclusively with vectors and tall skinny matrices and contribute only with lower order terms to the total flop count. We show with numerical experiments that the proposed algorithm is both fast and accurate.

In Chapter 5 we discuss triangularization of matrix polynomials. Any square matrix over an algebraically closed field is similar to a triangular matrix (its Jordan

form, for example). If the field is the set of real numbers, then the result still holds if we replace triangular by quasi-triangular. Here the prefix *quasi* means that the diagonal blocks are of size 1×1 and 2×2. We generalize these results to matrix polynomials. More precisely, we show that for any matrix polynomial of degree ℓ in $\mathbb{F}[\lambda]^{n\times m}$, where $n \leq m$ and $\mathbb{F}$ is algebraically closed, there is a triangular/trapezoidal matrix polynomial of the same size and degree, and with the same eigenstructure (defined in Section 2.1). If $\mathbb{F} = \mathbb{R}$ we show that the same result holds if we replace triangular/trapezoidal by quasi-triangular/quasi-trapezoidal. Our proofs are constructive in the sense that we build up matrix polynomials starting from a list of invariants called elementary divisors (defined in Section 2.1). This means that we solve structured inverse eigenvalue problems. In particular, our results imply that the necessary conditions for a list of elementary divisors to be admissible for a real square matrix polynomial of degree ℓ, are also sufficient conditions. Finally, we show that any regular Hermitian matrix polynomial (that is, one with Hermitian coefficient matrices) has the same eigenstructure as some real matrix polynomial of the same size and degree. We conjecture that the other direction is true too. That is, that each regular real matrix polynomial has the same eigenstructure as some Hermitian matrix polynomial of the same size and degree.

In Chapter 6 we derive algorithms for reducing a complex matrix polynomial with nonsingular leading coefficient to triangular, diagonal and Hessenberg form, while preserving the eigenstructure, the size and the degree. If the matrix polynomial is real, we describe how to obtain quasi-diagonal and quasi-triangular forms, using similar techniques in real arithmetic. For almost all square matrix polynomials, it is shown that this is not much harder than computing the corresponding reduced from of any monic linearization (defined in Section 2.4). As part of the theory, we give a rigorous algebraic argument, based on the celebrated Abel-Ruffini theorem, for why we cannot, in general, find the eigenvalues of a matrix by applying a finite number of Givens rotations or Householder reflectors that eliminates matrix entries in the usual manner. Even though this end result is well-known in the numerical linear algebra community, it is often presented with reference to a weaker statement of the Abel-Ruffini theorem which is not strong enough to draw the desired conclusion.

Chapter 7 is about the shift-and-invert Arnoldi algorithm for constant matrices. This is also of interest to PEPs with easily constructible monic linearizations. We study the algorithm via a backward error analysis and show how errors from the linear system solves and the orthonormalization process affect the Arnoldi

recurrence. It turns out that residual bounds for linear systems and columnwise backward error bounds for QR factorizations are important, so we discuss these in some detail. The main result is a collection of backward error bounds for various versions of the algorithm, including the Hermitian shift-and-invert Lanczos algorithm with full orthogonalization. Finally, we use our error analysis to define a sensible condition for "breakdown," that is, when the small Hessenberg matrix that is computed in the algorithm should be considered to be reduced and the iteration should stop.

Chapter

2 Background material

In this chapter we collect definitions and results that are of relevance to several chapters. With the exception of Corollary 2.3.3, all results are well-known and basic to the theory of matrix polynomials. We remark that some of the terms that were defined for regular matrix polynomials in the introduction are redefined in this chapter, but for more general matrix polynomials. It is left to the reader to verify that these new definitions are valid generalizations of the old ones.

2.1 Invariants of matrix polynomials

Let $\mathbb{F}[\lambda]$ be a polynomial ring. A square matrix $U(\lambda)$ with elements in $\mathbb{F}[\lambda]$ is called *unimodular* if $\det U(\lambda) \in \mathbb{F} \setminus \{0\}$. From the formula

$$U(\lambda)^{-1} = \frac{\operatorname{adj} U(\lambda)}{\det U(\lambda)}$$

we see that any unimodular matrix has an inverse that itself is a matrix polynomial. Further, from $\det U(\lambda)^{-1} \det U(\lambda) = 1$ we see that also $U(\lambda)^{-1}$ must be unimodular. It can be shown that any $n \times n$ unimodular matrix is a product of *elementary matrix polynomials* (this follows from the proof of the Smith form, introduced below), where an elementary matrix polynomial is defined as one that satisfies (a), (b) or (c) below, for some $i, j \in \{1, 2, \ldots, n\}$.

(a) $I - e_i e_i^T - e_j e_j^T + e_i e_j^T + e_j e_i^T$.

(b) $I + (\alpha - 1)e_i e_i^T$ for some $\alpha \in \mathbb{F} \setminus \{0\}$.

(c) $I + p(\lambda)e_i e_j^T$ for some $i \neq j$.

Here e_i denotes the ith column of the identity matrix. An application, from left or right, of an elementary matrix polynomial is called an *elementary transformation*. Note that if we apply the elementary transformation above from the left, then (a) swaps rows i and j, (b) multiplies row i by α, and (c) adds row j times $p(\lambda)$ to row i. Obviously, the same holds for the columns instead of rows if the transformations are applied from the right.

If $P(\lambda) \in \mathbb{F}[\lambda]^{m\times n}$ is any matrix polynomial, and $U(\lambda) \in \mathbb{F}[\lambda]^{m\times m}$ and $V(\lambda) \in \mathbb{F}[\lambda]^{n\times n}$ are unimodular, then $U(\lambda)P(\lambda)V(\lambda)$ is said to be a *unimodular transformation* of $P(\lambda)$. Unimodular transformations form an equivalence relation $\sim$ over the set $\mathbb{F}[\lambda]^{m\times n}$ and we say that $P(\lambda)$ and $Q(\lambda)$ are *equivalent*, and write $P(\lambda) \sim Q(\lambda)$, if there exists a unimodular transformation that maps $P(\lambda)$ to $Q(\lambda)$. If $P(\lambda)$ can be transformed into $Q(\lambda)$ using constant nonsingular matrices, then we say that $P(\lambda)$ is *strictly equivalent* to $Q(\lambda)$. Since unimodular matrices have nonzero determinants independent of λ it follows immediately that equivalent regular matrix polynomials have the same finite eigenvalues, but this is not the full story. Unimodular matrices preserve much more information. This leads us to the *Smith form*, a canonical form for matrix polynomials with respect to the equivalence relation defined by unimodular transformations. Any $P(\lambda) \in \mathbb{F}[\lambda]^{n\times m}$ is equivalent to a unique diagonal matrix polynomial

$$D(\lambda) = \left[\begin{array}{ccc|c} d_1(\lambda) & & & \\ & \ddots & & 0_{r,m-r} \\ & & d_r(\lambda) & \\ \hline & 0_{n-r,r} & & 0_{n-r,m-r} \end{array}\right] \sim P(\lambda), \tag{2.1}$$

where $r =: \operatorname{rank} P(\lambda)$ and $d_1(\lambda)|\cdots|d_r(\lambda)$ are monic scalar polynomials [26, Chapter VI, §3]. Here, "|" stands for divisibility, thus $d_j(\lambda)|d_{j+1}(\lambda)$ means that $d_j(\lambda)$ is a divisor of $d_{j+1}(\lambda)$. The diagonal matrix $D(\lambda)$ in (2.1) is called the Smith form of $P(\lambda)$ and its nonzero diagonal entries $d_j(\lambda)$ are called the *invariant factors* of $P(\lambda)$. Write

$$\begin{array}{rcl} d_1(\lambda) &=& \phi_1(\lambda)^{m_{11}} \cdots \phi_s(\lambda)^{m_{1s}}, \\ d_2(\lambda) &=& \phi_1(\lambda)^{m_{21}} \cdots \phi_s(\lambda)^{m_{2s}}, \\ \vdots & & \vdots \\ d_r(\lambda) &=& \phi_1(\lambda)^{m_{r1}} \cdots \phi_s(\lambda)^{m_{rs}}, \end{array} \tag{2.2}$$

where the $\phi_i(\lambda)$ are distinct monic polynomials irreducible over $\mathbb{F}[\lambda]$, and

$$0 \leq m_{1j} \leq m_{2j} \leq \cdots \leq m_{rj}, \quad j = 1\colon s, \tag{2.3}$$

are nonnegative integers. The factors $\phi_j(\lambda)^{m_{ij}}$ with $m_{ij} > 0$ are the *finite elementary divisors* of $P(\lambda)$ with *partial multiplicity* m_{ij}. Notice that when $\mathbb{F}$ is algebraically closed, $\phi_j(\lambda)$ is linear and when $\mathbb{F} = \mathbb{R}$, $\phi_j(\lambda)$ is either linear or quadratic.

We denote by $\overline{\mathbb{F}}$ the algebraic closure of $\mathbb{F}$ and define a finite eigenvalue of a matrix polynomial $P(\lambda)$ with rank r as any scalar $\lambda_0 \in \overline{\mathbb{F}}$ such that $\operatorname{rank} P(\lambda_0) < r$, or equivalently, a root of some finite elementary divisors $\phi_j(\lambda)$ in (2.2). The *geometric multiplicity* of λ_0 is defined as the number of nonzero m_{ij} and the *algebraic multiplicity* of λ_0 as $\sum_{i=1}^{r} m_{ij}$. Note that the geometric multiplicity is bounded from above by r. In particular, all eigenvalues of an $n \times n$ matrix polynomials have geometric multiplicity at most n.

An $n \times n$ matrix polynomial $P(\lambda)$ is said to be *regular* if it is of full rank, that is, if $\operatorname{rank} P(\lambda) = n$. All other matrix polynomials are said to be *singular*. This includes all cases of non-square matrix polynomials.

The *elementary divisors at infinity* of the matrix polynomial

$$P(\lambda) = A_\ell \lambda^\ell + A_{\ell-1}\lambda^{\ell-1} + \cdots + A_0 \tag{2.4}$$

are defined as the elementary divisors of $\operatorname{rev}(P)$ at 0, where

$$\operatorname{rev}(P) := \lambda^\ell P\left(\lambda^{-1}\right) = A_0\lambda^\ell + A_1\lambda^{\ell-1} + \cdots + A_\ell$$

is the reversal of $P(\lambda)$. We omit the notation "(λ)" when referring to the reversal of $P(\lambda)$—unless we evaluate it at some point. If $P(\lambda)$ has elementary divisors at infinity, then we say that $P(\lambda)$ has eigenvalues at infinity. The associated geometric and algebraic multiplicities are defined as those of zero as an eigenvalue of $\operatorname{rev}(P)$. We refer to the set of all elementary divisors of $P(\lambda)$, that is both the finite ones and those at infinity, as the *eigenstructure* of $P(\lambda)$.

For a regular $P(\lambda) \in \mathbb{F}[\lambda]^{n\times n}$ of degree ℓ, the Smith form of $P(\lambda)$ provides the algebraic multiplicity of the eigenvalues at infinity via the degree deficiency in $\det P(\lambda)$, that is, $\ell n - \sum_{j=1}^{r} \deg d_j(\lambda)$. For singular polynomials, the Smith form does not detect the presence of elementary divisors at infinity but if $\operatorname{rank} P(\lambda) = r > \operatorname{rank} \operatorname{rev}(P)(0)$ then $P(\lambda)$ has elementary divisors at infinity.

We remark that for regular matrix polynomials $P(\lambda)$, the geometric multiplicity of an eigenvalue λ_0 coincides with the nullity of $P(\lambda_0)$ if λ_0 is finite, and the nullity of the leading coefficient otherwise. In other words, the geometric multiplicity of an eigenvalue is the dimension of the corresponding eigenspace.

For any $\lambda_0 \in \overline{\mathbb{F}}$, the invariant factors d_i of $P(\lambda)$ can be factored over $\overline{\mathbb{F}}[\lambda]$ as

$$d_i(\lambda) = (\lambda - \lambda_0)^{\alpha_i} p_i, \quad \alpha_i \geq 0, \quad p_i(\lambda_0) \neq 0.$$

The sequence of exponents $\alpha_1, \alpha_2, \ldots, \alpha_r$ with $0 \leq \alpha_1 \leq \cdots \leq \alpha_r$ is called the *partial multiplicity sequence of* $P(\lambda)$ *at* λ_0 and is denoted by

$$\mathcal{J}(P, \lambda_0) = (\alpha_1, \alpha_2, \ldots, \alpha_r).$$

This sequence is usually all zeros unless λ_0 is an eigenvalue of $P(\lambda)$. The partial multiplicity sequence for $\lambda_0 = \infty$ is defined to be

$$\mathcal{J}(P, \infty) = \mathcal{J}(\mathrm{rev}(P), 0).$$

Let $(\alpha_1, \alpha_2, \ldots, \alpha_r)$ be the partial multiplicity sequence associated with an eigenvalue λ_0. We say that λ_0 is *simple* if $\sum_{i=1}^{r} \alpha_i = 1$, *semisimple* if $\max \alpha_i = 1$ and *defective* if $\max \alpha_i > 1$.

Since equivalent matrix polynomials have the same Smith form, it follows that a matrix polynomial $P_1(\lambda) \in \mathbb{F}[\lambda]^{n \times m}$ is equivalent to $P_2(\lambda) \in \mathbb{F}[\lambda]^{n \times m}$ if and only $\mathcal{J}(P_1, \lambda_0) = \mathcal{J}(P_2, \lambda_0)$ for any $\lambda_0 \in \overline{\mathbb{F}}$. Moreover, $P_1(\lambda)$ is said to be *strongly equivalent* to $P_2(\lambda)$ if it is equivalent to $P_2(\lambda)$ and $\mathcal{J}(P_1, \infty) = \mathcal{J}(P_2, \infty)$.

The following example shows that unimodular transformations do not necessarily preserve the partial multiplicities of infinite eigenvalues, or, in other words, that strong equivalence is indeed a stronger property than "plain" equivalence.

Example 2.1.1. The regular matrix polynomial $\mathrm{diag}(1, 1, \lambda)$ has one finite elementary divisor at zero and two linear elementary divisors at infinity. Multiplying $P(\lambda)$ from the left by the elementary matrix polynomial $I_3 + \lambda e_1 e_2^T$, where e_i denotes the ith column of the identity matrix, yields

$$\widetilde{P}(\lambda) = \begin{bmatrix} 1 & \lambda & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} P(\lambda) = \begin{bmatrix} 1 & \lambda & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \lambda \end{bmatrix},$$

which is easily seen to have one finite elementary divisor at zero and one quadratic

elementary divisor at infinity with partial multiplicity two. Hence $P(\lambda)$ and $\widetilde{P}(\lambda)$ are equivalent but not strongly equivalent. ▲

Finally, we note that strict equivalence implies strong equivalence. The reverse, however, is not true, so strict equivalence is strictly stronger than strong equivalence. For example, a quadratic matrix polynomial in $\mathbb{C}[\lambda]^{n\times n}$ is in general not strictly equivalent to a triangular matrix polynomial. An argument for this is given in the introduction of Chapter 6. However, we will show in Chapter 5 that any matrix polynomial in $\mathbb{C}[\lambda]^{n\times n}$ is strongly equivalent to a triangular matrix polynomial.

2.2 Möbius transformations

The Möbius transformation is a powerful tool when dealing with infinite eigenvalues. To any nonsingular matrix $A = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right] \in \mathbb{F}^{2\times 2}$ is associated a *Möbius function* $m_A : \mathbb{F} \cup \{\infty\} \to \mathbb{F} \cup \{\infty\}$ of the form

$$m_A(z) = \frac{az+b}{cz+d}, \qquad ad - bc \neq 0,$$

where

$$m_A(\infty) = \begin{cases} a/c & \text{if } c \neq 0, \\ \infty & \text{if } c = 0, \end{cases} \qquad m_A(-d/c) = \infty \text{ if } c \neq 0.$$

Let $P(\lambda) = \sum_{j=0}^{\ell} \lambda^j A_j \in \mathbb{F}[\lambda]^{n\times m}$ with $A_\ell \neq 0$, and let $A = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right] \in \mathbb{F}^{2\times 2}$ be nonsingular. Then the *Möbius transform of $P(\lambda)$ with respect to A* is the $n \times m$ matrix polynomial $\mathcal{M}_A(P)$ defined by

$$\mathcal{M}_A(P) = (c\lambda + d)^\ell P\big(m_A(\lambda)\big) = \sum_{j=0}^{\ell} A_j (a\lambda + b)^j (c\lambda + d)^{\ell - j}. \tag{2.5}$$

As with the reversal, we omit "(λ)" when denoting a non-evaluated Möbius transform of a matrix polynomial. It follows from (2.5) that $\mathcal{M}_I(P) = P(\lambda)$ and that the reversal of $P(\lambda)$ is the Möbius transform of $P(\lambda)$ with respect to $\left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$.

We are interested in Möbius transformations due to the property stated in the next theorem. A proof can be found in [57] and [91].

Theorem 2.2.1. *Let $P(\lambda) \in \mathbb{F}[\lambda]^{n\times m}$ and let $A \in \mathbb{F}^{2\times 2}$ be nonsingular such that $c \neq 0$ and $P(a/c) \neq 0$. Then $\mathcal{J}\big(\mathcal{M}_A(P), m_A^{-1}(\lambda_0)\big) = \mathcal{J}(P, \lambda_0)$ for any $\lambda_0 \in \overline{\mathbb{F}} \cup \{\infty\}$.*

In particular, Theorem 2.2.1 allows us to carry over many results for matrix

polynomials with invertible leading coefficient to the larger class of regular matrix polynomials. Corollary 2.3.3 is an example of this usage.

Finally, we remark that Möbius transformations leave eigenvectors invariant. To see this, let (λ_0, x) be an eigenpair of $P(\lambda)$. From (2.5), it follows

$$\mathcal{M}_A(P)(m_A^{-1}(\lambda_0))x = (c\lambda + d)^\ell P(\lambda_0)x = 0,$$

so $(m_A^{-1}(\lambda_0), x)$ is an eigenpair of $\mathcal{M}_A(P)$.

2.3 Defective eigenvalues

If $P(\lambda)$ is a regular matrix polynomial, its left and right eigenvectors can be used to determine whether or not an eigenvalue is defective. For constant matrices, this follows from the Jordan canonical form: if x and y are right and left eigenvectors, respectively, corresponding to the same Jordan block, then $y^T x = 0$ if and only if that Jordan block is nontrivial. Furthermore, if x and y are right and left eigenvectors corresponding to different Jordan blocks, then $y^T x = 0$. Hence, an eigenvalue is defective if and only if there exists an associated right eigenvector x such that $y^T x = 0$ for all left eigenvectors y. This result can be generalized to matrix polynomials with invertible leading coefficient.

Theorem 2.3.1 (Lancaster [50, p. 65])**.** *Let $P(\lambda)$ be a matrix polynomial with invertible leading coefficient. If λ_0 is an eigenvalue of $P(\lambda)$, then λ_0 is defective if and only if there exists an associated right eigenvector x such that $y^T P'(\lambda_0)x = 0$ for all left eigenvectors y.*

Remark 2.3.2. From the proof of Theorem 2.3.1, it follows that x comes from an arbitrary Jordan decomposition of an associated *real* linearization (defined in Section 2.4). For real eigenvalues of real matrix polynomials, we may choose a real associated Jordan chain and hence assume that x is real.

The assumption that the leading coefficient is invertible may be too strong. A way to get around this is to employ a Möbius transformation. Let λ_0 denote an arbitrary (possibly infinite) eigenvalue of a regular $n \times n$ matrix polynomial $P(\lambda)$, and let $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ be its partial multiplicity sequence. If

$$m(\lambda) = \frac{a\lambda + b}{c\lambda + d}.$$

is an invertible Möbius function, then $Q(\lambda) := (c\lambda + d)^{\deg P} P(m(\lambda))$ has the same

eigenvectors as $P(\lambda)$ and $m^{-1}(\lambda_0)$ is an eigenvalue of $Q(\lambda)$ with partial multiplicity sequence $(\alpha_1, \alpha_2, \ldots, \alpha_n)$. Suppose now that σ is not an eigenvalue of $P(\lambda)$. Then the choice $m(\lambda) = 1/\lambda + \sigma$ implies that $Q(\lambda)$ has invertible leading coefficient. Since $m^{-1}(\lambda) = 1/(\lambda - \sigma)$ we arrive at the following corollary.

Corollary 2.3.3. *Let $P(\lambda)$ be a regular matrix polynomial and assume that σ is not an eigenvalue of $P(\lambda)$. Define $Q(\lambda) = \lambda^{\deg P} P(1/\lambda + \sigma)$. If λ_0 is an eigenvalue of $P(\lambda)$ then λ_0 is defective if and only if there exists an associated eigenvector x such that $y^T Q'(1/(\lambda_0 - \sigma))x = 0$ for all left eigenvectors y.*

From Remark 2.3.2, it follows that x in Corollary 2.3.3 may be chosen to be real if $P(\lambda)$, λ_0 and σ are real.

2.4 Linearizations

Let $P(\lambda)$ be an $n \times n$ matrix polynomial of degree ℓ. A pencil $A\lambda + B$ of size $n\ell \times n\ell$ is said to be a *linearization* of $P(\lambda)$ if [32, Section 7.2]

$$A\lambda + B \sim \begin{bmatrix} P(\lambda) & \\ & I_{(n-1)\ell} \end{bmatrix}.$$

From the Smith form it follows that any linearization of $P(\lambda)$ has same finite elementary divisors as $P(\lambda)$, counting the partial multiplicities. The elementary divisors at infinity, however, are not necessarily preserved (if they are, the linearization is said to be *strong*). However, if $P(\lambda)$ has an nonsingular leading coefficient, so there are no infinite eigenvalues, then any linearization also has nonsingular leading coefficient. This follows from the equalities

$$\deg \det (A\lambda + B) = \deg \det P(\lambda) = n\ell.$$

In this case, the Jordan form J of $A^{-1}B$ is said to be the Jordan form of the matrix polynomial $P(\lambda)$. When the Jordan form exists, it carries exactly the same information as the Smith form. In fact, it is easy to construct the Smith form from the Jordan form and vice versa. This follows from the following two equivalences.

- Each Jordan block yields an elementary divisor:

$$I_\ell \lambda - \begin{bmatrix} \lambda_0 & 1 & & \\ & \lambda_0 & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_0 \end{bmatrix} \sim \begin{bmatrix} (\lambda - \lambda_0)^\ell & \\ & I_{\ell-1} \end{bmatrix}. \tag{2.6}$$

This is easily shown using elementary transformation.

- If $\gcd(p(\lambda), q(\lambda)) = 1$ then, by Bézout's identity [25, Theorem 46.9], there exist polynomials $s(\lambda)$ and $t(\lambda)$ such that $p(\lambda)s(\lambda) + q(\lambda)t(\lambda) = 1$. We have

$$\begin{bmatrix} p(\lambda) & 0 \\ 0 & q(\lambda) \end{bmatrix} \sim \begin{bmatrix} p(\lambda) & p(\lambda)s(\lambda) + q(\lambda)t(\lambda) \\ 0 & q(\lambda) \end{bmatrix} \sim \begin{bmatrix} p(\lambda)q(\lambda) & 0 \\ 0 & 1 \end{bmatrix}.$$

Note that the partial multiplicities are the sizes of the Jordan blocks. If $P(\lambda)$ is regular, but has eigenvalues at infinity, then $\mathcal{M}_A(P)$ has only finite eigenvalues for any nonsingular $A = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ such that $-d/c$ is not an eigenvalue of $P(\lambda)$. Thus the partial multiplicities of $P(\lambda)$ can be identified with the sizes of the associated Jordan blocks of $\mathcal{M}_A(P)$.

Now, suppose $P(\lambda)$ in (2.4) is regular. Since any linearization of $P(\lambda)$ shares at least its finite eigenstructure with $P(\lambda)$, one way to compute the finite eigenvalues of $P(\lambda)$ is to pick a linearization of $P(\lambda)$ and solve the associated generalized eigenvalue problem. If we chose the linearization wisely, we can also obtain the eigenvectors in this manner. Fortunately, good linearizations are easily constructed using the coefficient matrices of $P(\lambda)$ as building blocks. It is, for instance, easy to see that

$$C(\lambda) = \begin{bmatrix} I & & & \\ & \ddots & & \\ & & I & \\ & & & A_\ell \end{bmatrix} \lambda + \begin{bmatrix} & -I & & \\ & & \ddots & \\ & & & -I \\ A_0 & A_1 & \cdots & A_{\ell-1} \end{bmatrix} \tag{2.7}$$

is a linearization of $P(\lambda)$. Further, (λ_0, x_0) is an eigenpair of $P(\lambda)$ if and only if $[x_0^T \ \lambda_0 x_0^T \ \cdots \ \lambda_0^{\ell-1} x_0^T]^T$ is an eigenvector of $C(\lambda)$. Linearizations like (2.7) are not only valuable for numerical computation, but are also of theoretical interest. For example, as we will see in Chapter 3, it allows us to carry over eigenvalue perturbations result from standard matrix theory to the case of matrix polynomials.

If we take the transpose of (2.7) on the block level, we get the *left companion linearization*

$$\begin{bmatrix} I & & & \\ & \ddots & & \\ & & I & \\ & & & A_\ell \end{bmatrix} \lambda + \begin{bmatrix} & & & A_0 \\ -I & & & A_1 \\ & \ddots & & \vdots \\ & & -I & A_{\ell-1} \end{bmatrix}, \tag{2.8}$$

which plays an important role in Chapter 6. The left companion linearization can also be used to write down the solution to the differential equation (1.1) when the leading coefficient A_ℓ is invertible. In this case we can rewrite the equation as

$$A_\ell^{-1} P\left(\frac{d}{dt}\right) u(t) = A_\ell^{-1} f(t), \quad t \in [0, T],$$

where $A_\ell^{-1}P(\lambda)$ is monic. If C_L is the *left companion matrix* of $A_\ell^{-1}P(\lambda)$, that is, the constant part of the associated left companion linearization, then it can be shown that the general solution of (1.1) is given by

$$u(t) = Pe^{tC_L}x_0 + P\int_0^t e^{(t-s)C_L} R^T A_\ell^{-1} f(s)ds, \tag{2.9}$$

where $P = [I\ 0\ 0\ \cdots\ 0]$ and $R = [0\ 0\ \cdots\ 0\ I]$, and x_0 is arbitrary [32, Theorem 1.5]. In (2.9), the first and second term correspond to the homogeneous and particular solution, respectively.

2.5 Floating point arithmetic

The set $\mathbb{F}$ of floating point numbers with base $\beta \in \mathbb{N}$, precision $t \in \mathbb{N}$ and exponent range $[e_{\min}, e_{\max}]$ is defined as the set of all real numbers $\pm\beta^e \times .d_1 d_2 \ldots d_t$, where the d_i are integers such that $0 \le d_i \le \beta - 1$, and e is an integer in the interval $[e_{\min}, e_{\max}]$. We define the associated *machine precision* or *unit roundoff* to be the constant $u = \frac{1}{2}\beta^{1-t}$. Define $\mathbb{F}_\infty$ to be the floating point set with same base and precision as $\mathbb{F}$, but having $(-\infty, \infty)$ as its exponent range, and consider a rounding function $\mathrm{fl} : \mathbb{R} \to \mathbb{F}_\infty$ that maps any real number to a closest number in $\mathbb{F}_\infty$. For $x \in \mathbb{R}$, we say that $\mathrm{fl}(x)$ overflows if $|\mathrm{fl}(x)| > \max_{f\in\mathbb{F}} |f|$ and underflows if $|\mathrm{fl}(x)| < \min_{f\in\mathbb{F}} |f|$. In the absence of overflow and underflow, the standard model for arithmetic with two floating point numbers $a, b \in \mathbb{F}$ is that

$$|\mathrm{fl}(a \circ b) - (a \circ b)| \le u|a \circ b|, \tag{2.10}$$

where $\circ$ can be any of the four operations $+$, $-$, $\times$ and $\div$. Each of these floating point operations is called a *flop*.

In addition to (2.10), it is common to assume the following error bound for the square root of a nonnegative number $a \in \mathbb{F}$

$$|\mathrm{fl}(\sqrt{a}) - \sqrt{a}| \le u\sqrt{a}. \tag{2.11}$$

Unless stated otherwise, all numerical experiments shown in this thesis have been carried out in IEEE 754 double precision [42]. This floating point standard is defined with base $\beta = 2$, precision $t = 53$ and range $[e_{\min}, e_{\max}] = [-1022, 1023]$, and guarantees that (2.10) is satisfied, as long as overflow or underflow does not occur, as well as (2.11). Note that the associated unit roundoff is $2^{-53} \approx 10^{-16}$.

Chapter

3 Strongly damped quadratic matrix polynomials

3.1 Introduction

A way to prevent a structure from vibrating violently is to incorporate viscous dampers into the design. A viscous damper is a device that resists motion by producing a force proportional to the relative velocity of its ends raised to a power α. In this chapter we consider linear damping, which corresponds to dampers with $\alpha = 1$. This value of α is the default for certain product lines of seismic dampers [1]. The resisting force produced by a viscous damper arises when fluid, trapped in a cylinder, is forced through small holes (see Figure 3.1). By adjusting the size of these holes, we can make the damper stronger. But stronger is not necessarily better: if a damper is too strong, it resembles a rigid component and hence has little purpose. This suggests that a structure with only very strong dampers should be quite similar to a structure without dampers. The goal of

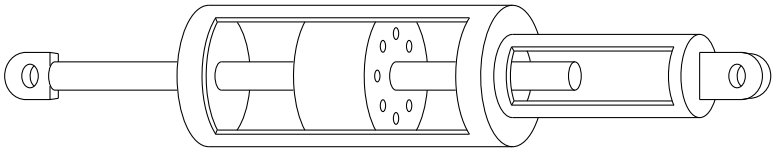

Figure 3.1: A model of a viscous damper. The larger cylinder is filled with a fluid which is forced through holes in the piston head as the piston rod moves horizontally. This causes friction and energy is dissipated and released as heat.

this chapter is to investigate this phenomenon more rigorously for discretized structures. We will do this by studying the eigenvalues and eigenspaces of a related quadratic matrix polynomial.

Consider a finite element model of a structure with r viscous dampers. If the model vibrates freely (that is, only due to initial conditions), the displacements of its nodes are given by the solutions to the equations of motion:

$$\left(M\frac{d^2}{dt^2} + sD\frac{d}{dt} + K\right)u(t) = 0. \tag{3.1}$$

Here M, sD and K are the mass matrix, damping matrix and stiffness matrix, respectively. We assume these matrices are $n \times n$, real and symmetric positive semidefinite, and further that M and K are strictly positive definite. We also assume that each damper contributes to the damping matrix with a rank one term, so $\operatorname{rank} D = r$ and $D = RR^T$ for some real $n \times r$ matrix R. If $\|D\| = 1$, say, the parameter s determines the strength of the dampers, so larger s corresponds to viscous dampers with smaller holes, and $s = 0$ yields an undamped system.

We find the solutions to (3.1) by solving the quadratic eigenproblem

$$P_s(\lambda)x = 0, \quad s \geq 0, \tag{3.2}$$

where

$$P_s(\lambda) := M\lambda^2 + sD\lambda + K. \tag{3.3}$$

The spectrum of $P_s(\lambda)$ lies in the left half plane and is symmetric with respect to the real axis. Further, if $(-\gamma + i\omega, x)$ is an eigenpair of $P_s(\lambda)$, where $\gamma, \omega \in \mathbb{R}$, and x is real if $\omega = 0$, then

$$u(t) = e^{-\gamma t}(\cos(t\omega)\mathrm{Re}(x) - \sin(t\omega)\mathrm{Im}(x)) \tag{3.4}$$

is a real solution to (3.1) and is called a *mode*.[1] We see that γ and ω correspond to damping and frequency, respectively. The solution (3.4) describes how the model switches between two configurations, given by $\mathrm{Re}(x)$ and $\mathrm{Im}(x)$, as it vibrates. If $x = ve^{i\theta}$ for some $v \in \mathbb{R}^n$ and $\theta \in \mathbb{R}$, then

$$u(t) = e^{-\gamma t}(\cos(t\omega)\cos(\theta) - \sin(t\omega)\sin(\theta))v = e^{-\gamma t}\cos(t\omega + \theta)v,$$

and we see that these two configurations must coincide and that all nodes in the

[1] The term "mode" is ambiguous and is sometimes, although not in this thesis, used to refer to an eigenvector.

model vibrate in phase. Now, if $s = 0$, it is well known that all eigenvalues are nonzero and purely imaginary, and that all eigenspaces have real bases and are pairwise M-orthogonal [50, Section 7.3]. In particular, all modes of an undamped model are undamped and those modes that correspond to simple eigenvalues are such that all nodes in the model vibrate in phase. We will see in Lemma 3.3.4 that this cannot be the case when damping is present.

To see the similarities between strongly damped structures and undamped ones, we will prove that the eigenvalues of $P_s(\lambda)$ approach nonzero points on the imaginary axis as $s \to \infty$, with the exception of $2r$ real eigenvalues which correspond to overdamped modes (that is, non-oscillating modes). This implies that the considered model has $n-r$ practically undamped modes for large enough s. For the eigenspaces of $P_s(\lambda)$ as $s \to \infty$, we will show the following. If two eigenvalues converge to distinct points on the imaginary axis, that are not complex conjugates, then the corresponding eigenspaces become more and more M-orthogonal in terms of the principal angles (defined in Section 3.2). Further, we will prove that the span of all eigenvectors associated with eigenvalues converging to a given point has an M-orthonormal basis that becomes more and more real in the sense that the norms of the imaginary parts go to zero. In particular, eigenvalues converging to points to which no other eigenvalue converges, are, for large enough s, associated with almost real eigenvectors. This corresponds to the case of simple eigenvalues for the undamped problem, and from (3.4) we see that the associated modes are such that all nodes in the model vibrate essentially in phase.

The outline of the chapter is as follows. In Section 3.2 we introduce the notion of principal angles and establish two results, Proposition 3.2.4 and Proposition 3.2.5, which are needed for Section 3.4. In Section 3.3, we study the eigenvalues of $P_s(\lambda)$ as $s \to \infty$, and prove an eigenvalue location result which extends some early work by Lancaster [50]. In Section 3.4 we study the eigenspaces of $P_s(\lambda)$ as $s \to \infty$. Finally, in Section 3.5, we discuss the forced response problem. That is, when we add a nonzero right hand side to (3.1). We also briefly discuss the case when $P_s(\lambda)$ has a nearly defective eigenvalue. Here "nearly defective" is with respect to the damping parameter s and means that $P_{s+\Delta s}(\lambda)$ has a defective eigenvalue for some small Δs.

3.2 Preliminaries

In what follows, $\langle \cdot, \cdot \rangle$ denotes an arbitrary positive definite inner product on $\mathbb{C}^n$ and $\| \cdot \|$ denotes the induced norm. Further, for a subspace $\mathcal{X}$ we define

$S(\mathcal{X}) = \{x : x \in \mathcal{X}, \|x\| = 1\}$.

The angle between two nonzero vectors u and v is defined as

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

To generalize the concept of angles to subspaces the *principal angles* (or *canonical angles*) are introduced. Given two subspaces $\mathcal{U}$ and $\mathcal{V}$, such that $p = \dim\mathcal{U} \leq \dim\mathcal{V} = q$, there are p principal angles

$$\theta_1(\mathcal{U}, \mathcal{V}) \;\leq\; \theta_2(\mathcal{U}, \mathcal{V}) \;\leq\; \cdots \;\leq\; \theta_p(\mathcal{U}, \mathcal{V}),$$

which all lie in $[0, \pi/2]$. For convenience, we shall with $\theta_{\max}(\mathcal{U}, \mathcal{V})$ refer to $\theta_p(\mathcal{U}, \mathcal{V})$. The first principal angle is defined as

$$\theta_1(\mathcal{U}, \mathcal{V}) = \min\{\, \measuredangle(u, v) \;:\; u \in S(\mathcal{U}),\; v \in S(\mathcal{V})\} = \measuredangle(u_1, v_1),$$

where u_1 and v_1 are some minimizing vectors. The remaining angles are then defined recursively by

$$\begin{aligned}\theta_i(\mathcal{U}, \mathcal{V}) &= \min\{\, \measuredangle(u, v) \;:\; u \in S(\mathcal{U}),\; v \in S(\mathcal{V}),\\ &\qquad \langle u, u_j\rangle = \langle v, v_j\rangle = 0,\; j = 1\!:\!i-1\}\\ &= \measuredangle(u_i, v_i),\end{aligned}$$

where u_i and v_i are minimizing vectors. It is clear from the definition that $\theta_i(\mathcal{U}, \mathcal{V}) = \theta_i(\mathcal{V},\mathcal{U})$ for $i = 1\!:\!p$. The vectors $u_1, u_2, \ldots, u_p$ and $v_1, v_2, \ldots, v_p$ are obviously not unique but the principal angles are. This is easily seen from the next theorem, which is due to Björck and Golub [13]. The proof in [13] is for the standard inner product, but it can easily be generalized to an arbitrary inner product. A proof is provided in Appendix A.

Theorem 3.2.1. *Suppose the inner product $\langle\cdot,\cdot\rangle$ corresponds to a Hermitian positive definite matrix A, so $\langle x, y\rangle = x^H Ay$ for any vectors x and y. If the columns of U and V form A-orthonormal bases for $\mathcal{U}$ and $\mathcal{V}$, respectively, then*

$$\theta_i(\mathcal{U}, \mathcal{V}) = \arccos(\sigma_i),$$

where σ_i is the ith largest singular value of $U^H AV$.

When $\dim\mathcal{U} = \dim\mathcal{V}$, it is well-known (see e.g., [73, p. 249] or [76]) that the

largest principal angle is given by

$$\theta_{\max}(\mathcal{U},\mathcal{V}) = \max_{u\in S(\mathcal{U})} \min_{v\in S(\mathcal{V})} \measuredangle(u,v). \tag{3.5}$$

See Appendix A for a proof.

We note that $\mathcal{U}$ and $\mathcal{V}$ are orthogonal if and only if $\theta_1(\mathcal{U},\mathcal{V}) = \pi/2$, and that $\mathcal{U} = \mathcal{V}$ if and only if $\theta_{\max}(\mathcal{U},\mathcal{V}) = 0$ and $\dim\mathcal{U} = \dim\mathcal{V}$.

In the following lemmas and propositions the calligraphic notation $\mathcal{X}(s)$ refers to a subspace which depends on the parameter $s \geq 0$. Our first lemma shows that if $u_1 + u_2$ is a unit vector, where u_1 and u_2 are almost orthogonal, then $\|u_1\| + \|u_2\| \approx 1$.

Lemma 3.2.2. *Let $\varepsilon \in (0,1)$ and assume that*

$$\lim_{s\to\infty} \theta_1(\mathcal{U}_1(s),\mathcal{U}_2(s)) = \pi/2.$$

For sufficiently large s, $\|u_1 + u_2\| = 1$, where $u_1 \in \mathcal{U}_1(s)$ and $u_2 \in \mathcal{U}_2(s)$, implies that

$$1/(1+\varepsilon) < \|u_1\|^2 + \|u_2\|^2 < 1/(1-\varepsilon).$$

Proof. The limit condition of the lemma implies that $|\langle u_1, u_2\rangle| < \|u_1\|\|u_2\|\varepsilon$ for large enough s. We have

$$|\|u_1\|^2 + \|u_2\|^2 - 1| \leq 2|\langle u_1, u_2\rangle| < 2\varepsilon\|u_1\|\|u_2\| \leq (\|u_1\|^2 + \|u_2\|^2)\varepsilon,$$

from which the lemma follows. ∎

We now use Lemma 3.2.2 to show that if a subspace $\mathcal{V}(s)$ is almost orthogonal to two subspaces $\mathcal{U}_1(s)$ and $\mathcal{U}_2(s)$, which themselves are almost orthogonal to each other, then $\mathcal{V}(s)$ is almost orthogonal to their span. More formally, we have the following lemma.

Lemma 3.2.3. *Let $\mathcal{U}(s) = \operatorname{span}\{\mathcal{U}_1(s),\mathcal{U}_2(s)\}$. If*

$$\lim_{s\to\infty} \theta_1(\mathcal{U}_1(s),\mathcal{U}_2(s)) = \lim_{s\to\infty} \theta_1(\mathcal{U}_1(s),\mathcal{V}(s)) = \lim_{s\to\infty} \theta_1(\mathcal{U}_2(s),\mathcal{V}(s)) = \pi/2,$$

then

$$\lim_{s\to\infty} \theta_1(\mathcal{U}(s),\mathcal{V}(s)) = \pi/2.$$

Proof. Let $u \in \mathcal{U}(s)$ and $v \in \mathcal{V}(s)$ be any vectors such that $\|u\| = \|v\| = 1$. We have $u = u_1 + u_2$ where $u_1 \in \mathcal{U}_1(s)$ and $u_2 \in \mathcal{U}_2(s)$ and

$$|\langle u, v\rangle| = |\langle u_1, v\rangle + \langle u_2, v\rangle| \leq |\langle u_1, v\rangle| + |\langle u_2, v\rangle|. \tag{3.6}$$

By Lemma 3.2.2 the norms of u_1 and u_2 are bounded when s is sufficiently large. Hence the right hand side of (3.6) can be forced to be arbitrarily small by taking s large enough. ∎

We can now state our first proposition of this chapter. The proposition implies that if $\mathbb{C}^n$ is decomposed into the span of p almost orthogonal subspaces, then any subspace that is almost orthogonal to all but one of these subspaces must be close to the remaining subspace in terms of the principal angles.

Proposition 3.2.4. *Suppose* $\mathcal{V}(s) \subseteq \operatorname{span}\{\mathcal{U}_1(s), \mathcal{U}_2(s) \ldots, \mathcal{U}_p(s)\}$ *and* $\dim \mathcal{U}_k(s) = \dim \mathcal{V}(s)$ *for a fixed* $k \in \{1, 2, \ldots, p\}$. *If for any* $i \neq k$ *and any* $j \neq \ell$, *it holds that*

$$\lim_{s\to\infty} \theta_1(\mathcal{V}(s), \mathcal{U}_i(s)) = \lim_{s\to\infty} \theta_1(\mathcal{U}_j(s), \mathcal{U}_\ell(s)) = \pi/2,$$

then

$$\lim_{s\to\infty} \theta_{\max}(\mathcal{V}(s), \mathcal{U}_k(s)) = 0.$$

Proof. Let $\mathcal{W}(s) = \operatorname{span}\{\mathcal{U}_i(s) \ : \ i \neq k\}$. Lemma 3.2.3 implies

$$\lim_{s\to\infty} \theta_1(\mathcal{V}(s), \mathcal{W}(s)) = \lim_{s\to\infty} \theta_1(\mathcal{U}_k(s), \mathcal{W}(s)) = \pi/2.$$

Pick N and $\varepsilon \in (0, 1)$ such that for any $s > N$ it holds that

$$\max_{\substack{u_k \in \mathcal{U}_k(s) \\ w \in \mathcal{W}(s)}} \frac{|\langle u_k, w\rangle|}{\|u_k\|\|w\|} < \varepsilon/4 \tag{3.7}$$

and

$$\max_{\substack{v \in S(\mathcal{V}(s)) \\ w \in S(\mathcal{W}(s))}} |\langle v, w\rangle| < \varepsilon/2. \tag{3.8}$$

Let $v \in S(\mathcal{V}(s))$ and write $v = u_k + w$, where $u_k \in \mathcal{U}_k(s)$ and $w \in \mathcal{W}(s)$. Due to Lemma 3.2.2 we may, by possibly choosing a larger N, assume that $\|u_k\| < 2$ (for any choice of v), so (3.7) yields $|\langle u_k, w\rangle|/\|w\| < \varepsilon/2$. We get

$$\left| \frac{|\langle u_k, w\rangle|}{\|w\|} - \|w\| \right| \leq \max_{\widetilde{w} \in S(\mathcal{W}(s))} |\langle u_k, \widetilde{w}\rangle + \langle w, \widetilde{w}\rangle| = \max_{\widetilde{w} \in S(\mathcal{W}(s))} |\langle v, \widetilde{w}\rangle| < \varepsilon/2,$$

where (3.8) is used for the last inequality, and hence $\|w\| < \varepsilon$. Further,

$$\|u_k\|^2 \geq \|v\| - \|w\|^2 - 2|\langle u_k, w\rangle| = 1 - \|w\|^2 - 2\|w\|\frac{|\langle u_k, w\rangle|}{\|w\|} > 1 - 2\varepsilon^2.$$

Note that this holds for any choice of $v = u_k + w \in \mathcal{V}(s)$ for $s > N$. Now, by (3.5), we have $\theta_{\max}(\mathcal{V}(s), \mathcal{U}_k(s)) = \arccos(x)$, where

$$\begin{aligned} x &= \min_{v \in S(\mathcal{V}(s))} \max_{\widetilde{u}_k \in S(\mathcal{U}_k(s))} |\langle v, \widetilde{u}_k\rangle| \geq \min_{\substack{u_k + w \in S(\mathcal{V}(s)) \\ u_k \in \mathcal{U}_k(s) \\ w \in \mathcal{W}(s)}} \left| \|u_k\| + \frac{\langle w, u_k\rangle}{\|u_k\|} \right| \\ &\geq \min_{\substack{u_k + w \in S(\mathcal{V}(s)) \\ u_k \in \mathcal{U}_k(s) \\ w \in \mathcal{W}(s)}} \|u_k\| - \|w\| > \sqrt{1 - 2\varepsilon^2} - \varepsilon. \end{aligned}$$

Since ε can be chosen to be arbitrarily small, the proposition follows. ∎

Finally, we show that if a subspace is close to its complex conjugate subspace, then there is an orthonormal basis for this subspace that is almost real.

Proposition 3.2.5. *Suppose the inner product $\langle \cdot, \cdot \rangle$ corresponds to a real symmetric positive definite matrix A, so $\langle x, y\rangle = x^H A y$ for any vectors x and y. If $\dim \mathcal{U}(s) = p$ for $s > N$ and*

$$\lim_{s \to \infty} \theta_{\max}(\overline{\mathcal{U}(s)}, \mathcal{U}(s)) = 0,$$

then for any $\varepsilon > 0$, for large enough s the subspace $\mathcal{U}(s)$ has an A-orthonormal basis $\{u_1, u_2, \ldots, u_p\}$ with $\|\mathrm{Im}(u_i)\| < \varepsilon$, $i = 1\!:\!p$.

Proof. Let the columns of $U = [u_1, u_2, \ldots, u_p]$ be any A-orthonormal basis of $\mathcal{U}(s)$ and note that the columns of $\overline{U}$ form an A-orthonormal basis of $\overline{\mathcal{U}(s)}$. Since A is real, $U^T A U$ is complex symmetric and hence enjoys a singular value decomposition on the form $Q \Sigma Q^T$ (also known as a Takagi factorization) [40, Corollary 4.4.4]. By the limit assumption and Theorem 3.2.1, all singular values are in $(1 - \varepsilon, 1]$ for large enough s. Define $Y = [y_1, y_2, \ldots, y_p] = U\overline{Q}$, and note that

$$Y^T A Y = Q^H U^T A U \overline{Q} = Q^H Q \Sigma Q^T \overline{Q} = \Sigma.$$

The columns of Y form an A-orthonormal basis of $\mathcal{U}(s)$, and

$$2\|\mathrm{Im}(y_i)\|^2 = \mathrm{Re}(y_i^H A y_i - y_i^T A y_i) < \varepsilon \tag{3.9}$$

for $i = 1\!:\!p$. ∎

Remark 3.2.6. For subspaces $\mathcal{U}(s)$ and $\mathcal{V}(s)$ with $\dim\mathcal{U}(s) = \dim\mathcal{V}(s) = k$, a limit condition like $\lim_{s\to\infty} \theta_{\max}(\mathcal{U}(s), \mathcal{V}(s)) = 0$ may be interpreted as the distance between $\mathcal{U}(s)$ and $\mathcal{V}(s)$ goes to zero. Here the distance is measured with the *gap metric* associated with $\langle\cdot,\cdot\rangle$. Using that gap metrics associated with different inner products are equivalent in the same sense that all norms on $\mathbb{C}^n$ are equivalent, it follows that the condition $\lim_{s\to\infty} \theta_{\max}(\mathcal{U}(s), \mathcal{V}(s)) = 0$ is independent of which positive definite inner product $\theta_{\max}$ corresponds to. See Appendix A for more details.

3.3 Eigenvalues

In this section we study the eigenvalues of $P_s(\lambda)$ defined in (3.3). Let $(\cdot)^{1/2}$ denote the principal square root, and introduce $A = M^{-1/2}DM^{-1/2}$ and $B = M^{-1/2}KM^{-1/2}$. Clearly, $P_s(\lambda)$ is equivalent to $I\lambda^2 + sA\lambda + B$, so they have the same Jordan structure. We will repetitively make use of the linearization $I\lambda - L(s)$, where

$$L(s) := \underbrace{\begin{bmatrix} sB^{-1/2}A & B^{-1/2} \\ -I & \end{bmatrix}}_{T} \underbrace{\begin{bmatrix} & I \\ -B & -sA \end{bmatrix}}_{\substack{\text{Companion} \\ \text{matrix}}} \underbrace{\begin{bmatrix} & -I \\ B^{1/2} & sA \end{bmatrix}}_{T^{-1}} = \begin{bmatrix} & B^{1/2} \\ -B^{1/2} & -sA \end{bmatrix}. \tag{3.10}$$

Lemma 3.3.1. *If $\lambda_1, \lambda_2, \ldots, \lambda_r$ are the nonzero eigenvalues of sA in* (3.10) *(not necessarily distinct), then we have Gerschgorin-like discs*

$$\mathcal{G}_0 = \left\{ z : |z| \le \|B\|_2^{1/2} \right\} \quad \textit{and} \quad \mathcal{G}_i = \left\{ z : |z - \lambda_i| \le \|B\|_2^{1/2} \right\}$$

for $i = 1\!:\!r$, such that the eigenvalues of $P_s(\lambda)$ in (3.3) *are contained in the union $\mathcal{G}_0 \cup \mathcal{G}_1 \cup \cdots \cup \mathcal{G}_r$. Furthermore, k Gerschgorin-like discs contain exactly k eigenvalues (counting multiplicities) if they are disjoint from the remaining discs.*

Proof. Apply a real orthogonal similarity transformation to $I\lambda^2 + A\lambda + B$, to obtain $I\lambda^2 + \widetilde{A}\lambda + \widetilde{B}$, where $\widetilde{A}$ is diagonal. Note that all but r of the diagonal entries of $\widetilde{A}$ must be zero. Let $\widetilde{L}(s)$ be the matrix (3.10), with $\widetilde{A}$ and $\widetilde{B}$ in place of A and B, respectively. Clearly, $\widetilde{L}(s)$ and $P_s(\lambda)$ have the same eigenstructure. Since $\widetilde{A}$ is normal and $\|\widetilde{B}\|_2 = \|B\|_2$ the lemma follows from [62, Theorem 2.1 and Corollary 2.5]. ∎

Let $\omega_{\max} = \|B\|_2^{1/2}$ and $\omega_{\min} = \sigma_{\min}(B)^{1/2}$ (where $\sigma_{\min}$ refers to the smallest

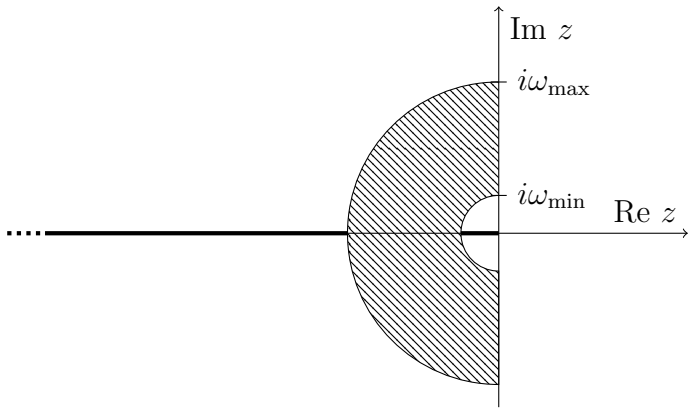

Figure 3.2: The shaded area is $\mathcal{S}$, and the left and right thick lines are $\mathcal{S}_{out}$ and $\mathcal{S}_{in}$, respectively.

singular value) and define the following sets:

$$\mathcal{S}_{in} = \{z \ : \ -\omega_{\min} < z < 0\}, \quad \mathcal{S}_{out} = \{z \ : \ z < -\omega_{\max}\} \tag{3.11}$$

and

$$\mathcal{S} = \{z \ : \ \omega_{\min} \leq |z| \leq \omega_{\max}, \ \mathrm{Re}(z) \leq 0\}. \tag{3.12}$$

See Figure 3.2 for an illustration. Lancaster showed that all nonreal eigenvalues of $P_s(\lambda)$ lie in the half annulus $\mathcal{S}$ [50, Chapter 9]. Hence any eigenvalue that is not in $\mathcal{S}$ must be in either $\mathcal{S}_{in}$ or $\mathcal{S}_{out}$. Our first goal in this section is to bound the number of such eigenvalues. To do so, we need the following lemma.

Lemma 3.3.2. *All eigenvalues of $P_s(\lambda)$ in (3.3) that lie in $\mathcal{S}_{in}$ or $\mathcal{S}_{out}$ are semisimple.*

Proof. Let λ be a real defective eigenvalue of $I\lambda^2 + sA\lambda + B$, and note that any corresponding real right eigenvector is also a left eigenvector. By Theorem 2.3.1 and Remark 2.3.2 there exists a real eigenpair (λ, v), where $\|v\|_2 = 1$, such that

$$v^T \left(\frac{d}{d\lambda}(I\lambda^2 + sA\lambda + B) \right) v = v^T(2I\lambda + sA)v = 0.$$

If we define $a = sv^TAv$ and $b = v^TBv$, we have $\lambda = -a/2$. Further,

$$v^T(I\lambda^2 + sA\lambda + B)v = \lambda^2 + a\lambda + b = \frac{a^2}{4} - \frac{a^2}{2} + b = 0,$$

which implies $a = 2\sqrt{b}$, and hence $\lambda = -\sqrt{b} \in \mathcal{S}$. ∎

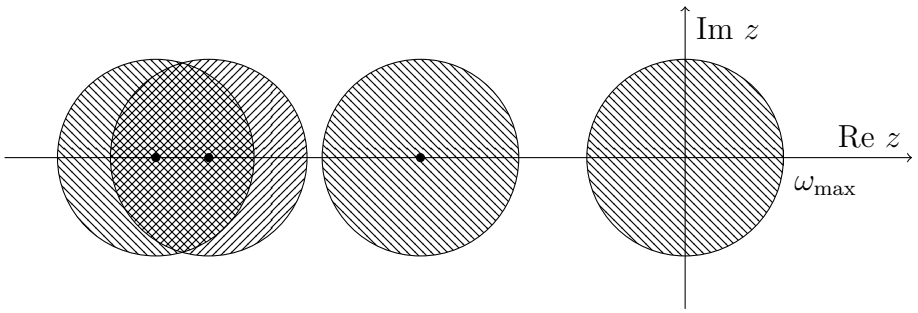

Figure 3.3: An illustration of what the Gerschgorin-like discs may look like for $r = 3$ and large enough s. The disc (from left to right) are $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3, \mathcal{G}_0$ and the dots are the nonzero eigenvalues of $-sA$. Lemma 3.3.1 implies that $\mathcal{G}_1 \cup \mathcal{G}_2 \cup \mathcal{G}_3$ contains three eigenvalues of $P_s(\lambda)$, counting multiplicities.

We now use Lemma 3.3.2 to prove the following eigenvalue location result.

Theorem 3.3.3. *The sets $\mathcal{S}_{in}$ and $\mathcal{S}_{out}$, defined in* (3.11)*, each contains at most r eigenvalues (counting multiplicities) of $P_s(\lambda)$ defined in* (3.3)*. Furthermore, r eigenvalues go to $-\infty$, and r eigenvalues go to* 0*, as $s \to \infty$.*

Proof. By Lemma 3.3.1, r eigenvalues of $P_s(\lambda)$ approach $-\infty$ as $s \to \infty$, and all remaining eigenvalues lie in $\mathcal{S} \cup \mathcal{S}_{in}$ (see Figure 3.3 for an illustration when $r = 3$). We now show that $\mathcal{S}_{out}$ cannot contain more than r eigenvalues for intermediate values of s. Let $L(s)$ denote the matrix in (3.10). By Lemma 3.3.2 all eigenvalues in $\mathcal{S}_{out}$ are semisimple and hence differentiable with respect to s [49, Theorem 6]. Consider an eigenvalue $\lambda \in \mathcal{S}_{out}$ for an arbitrary s. Since λ is real, the corresponding eigenspace of $L(s)$ has a real basis. If the columns of W form such a basis, it is easy to see that $W = [V^T B^{1/2},\ \lambda V^T]^T$, where the columns of V are real eigenvectors with respect to λ of the corresponding quadratic matrix polynomial. Furthermore, the columns of $[-V^T B^{1/2},\ \lambda V^T]^T$ form a basis of the corresponding left eigenspace, and

$$\begin{bmatrix} -V^T B^{1/2} & \lambda V^T \end{bmatrix} \begin{bmatrix} B^{1/2} V \\ \lambda V \end{bmatrix} = \lambda^2 V^T V - V^T B V$$

is positive definite since $\lambda^2 > \omega_{\max}^2 = \|B\|_2$. Suppose, without loss of generality, that V satisfies

$$\lambda^2 V^T V - V^T B V = I,$$

(otherwise we can replace V by VZ for an appropriate Z). Then, the derivatives of the eigenvalues that equal λ for the considered value of s are given by the

eigenvalues of the following matrix [49, Theorem 7],

$$[\,-V^TB^{1/2} \quad \lambda V^T\,]\,\frac{d}{ds}L(s)\begin{bmatrix} B^{1/2}V \\ \lambda V \end{bmatrix} = -\lambda^2 V^TAV,$$

which is negative semidefinite. Hence all eigenvalues that enter $\mathcal{S}_{out}$ will stay in $\mathcal{S}_{out}$ as $s \to \infty$.

Let $\sim$ denote the equivalence relation for matrix polynomials. To compute the number of eigenvalues in $\mathcal{S}_{in}$, we consider

$$\operatorname{rev} P_s(\lambda) = K\lambda^2 + sD\lambda + M \sim I\lambda^2 + s\underbrace{K^{-1/2}DK^{-1/2}}_{\widehat{A}}\lambda + \underbrace{K^{-1/2}MK^{-1/2}}_{\widehat{B}},$$

and note that $\widehat{B}$ is similar to B^{-1}. It is easy to see that λ is an eigenvalue of $P_s(\lambda)$ if and only if $1/\lambda$ is an eigenvalue of $\operatorname{rev} P_s(\lambda)$ with the same algebraic multiplicity. The proved part of the theorem implies that r eigenvalues of $\operatorname{rev} P_s(\lambda)$ go to $-\infty$ as $s \to \infty$, so r eigenvalues of $P_s(\lambda)$ must go to zero (along the negative real axis). Furthermore, $\operatorname{rev} P_s(\lambda)$ has at most r eigenvalues in $\left\{z : z < -\|\widehat{B}\|_2^{1/2}\right\}$ so $P_s(\lambda)$ has at most r eigenvalues in

$$\left\{z : -\|\widehat{B}\|_2^{-1/2} < z < 0\right\} = \left\{z : -\sigma_{\min}(B)^{1/2} < z < 0\right\} = \mathcal{S}_{in}. \qquad \blacksquare$$

Theorem 3.3.3 gives us rather large regions in which the eigenvalues lie. We are now interested in how the eigenvalues move within these regions as $s \to \infty$. To this end, we classify the eigenvalues of $P_s(\lambda)$ based on whether or not they depend on s. Eigenvalues are said to be *affected* (by damping) if they depend on s, and *unaffected* otherwise. To be more precise, let the columns of $V = [V_1,\ V_2]$ be a real M-orthogonal basis of eigenvectors of $P_0(\lambda)$, such that $\operatorname{range}(V_1)$ is spanned by all eigenvectors that are in the null space of D. If we apply the congruence transformation defined by V to $P_s(\lambda)$, the resulting matrix polynomial decomposes into the direct sum of an "undamped part" and a "damped part":

$$V^TP_s(\lambda)V = (I_k\lambda^2 + K_1) \oplus (I_{n-k}\lambda^2 + sD_2\lambda + K_2). \tag{3.13}$$

Here $k = \operatorname{rank} V_1$, $V_1^TKV_1 = K_1$, $V_2^TKV_2 = K_2$ and $V_2^TDV_2 = D_2$. The eigenvalues of $I_k\lambda^2 + K_1$ are clearly independent of s and hence unaffected. In particular they must be purely imaginary. The next lemma shows that the eigenvalues of $I_{n-k}\lambda^2 + sD_2\lambda + K_2$ are the affected eigenvalues, and furthermore that they are only purely imaginary for $s = 0$.

Lemma 3.3.4. *If λ is an eigenvalue of $I_{n-k}\lambda^2 + sD_2\lambda + K_2$, defined in* (3.13)*, for some $s > 0$, then* $\mathrm{Re}(\lambda) < 0$.

Proof. Assume the contrary and let $(\lambda = \omega i, x)$ be an eigenpair such that $\omega \in \mathbb{R} \setminus \{0\}$. If $v = V_2 x$, then

$$v^H(K - \omega^2 M + i\omega s D)v = 0,$$

and since $v^H M v$, $v^H D v$ and $v^H K v$ are real, we must have $v^H D v = 0$. Because D is symmetric positive semidefinite, this implies that $Dv = 0$. But then (λ, v) is an eigenvalue of $P_0(\lambda)$ for which $Dv = 0$. This contradicts that $v \in \mathrm{range}(V_2)$. ∎

As $s \to \infty$ we know from Theorem 3.3.3 that r eigenvalues enter $\mathcal{S}_{out}$ and go to $-\infty$, and r other eigenvalues enter $\mathcal{S}_{in}$ and go 0. We now focus on the remaining affected eigenvalues, that is, the ones that stay in $\mathcal{S}$ as $s \to \infty$.

Theorem 3.3.5. *Consider $P_s(\lambda)$ in* (3.3) *and the associated set $\mathcal{S}$ defined in* (3.12)*. For large enough s, the affected eigenvalues of $P_s(\lambda)$ in $\mathcal{S}$ are continuous functions of s that converge to purely imaginary points.*

Proof. Recall that $D = RR^T$ with $R \in \mathbb{R}^{n\times r}$, and define $t = 1/s$, $p(\lambda) = \det(M\lambda^2 + K)$, $Q(\lambda) = \lambda R^T(M\lambda^2 + K)^{-1}R$ and

$$q_t(\lambda) = \det(tp(\lambda)I_r + p(\lambda)Q(\lambda)).$$

By Lemma 3.3.4, no affected eigenvalue is a root of $p(\lambda)$ for $s > 0$. Thus, for $t > 0$, any root λ_i of $q_t(\lambda)$ that is not a root of $p(\lambda)$ is an affected eigenvalue. To see this, we simply note that

$$0 = q_t(\lambda_i) = p(\lambda_i)^r \det\left(tI_r + Q(\lambda_i)\right) = p(\lambda_i)\det\left(tI_r + Q(\lambda_i)\right) = \det P_s(\lambda_i),$$

where the matrix determinant lemma (related to the Sherman-Morrison formula) has been used for the last equality. On noting that $(M\lambda^2 + K)^{-1} = \mathrm{Adj}(M\lambda^2 + K)/p(\lambda)$, it is easy to see that $p(\lambda)Q(\lambda)$ is a matrix polynomial, so $q_t(\lambda)$ is a polynomial, and further

$$\deg q_t(\lambda) = \begin{cases} 2nr & \text{if } t \neq 0, \\ 2nr - r & \text{if } t = 0. \end{cases}$$

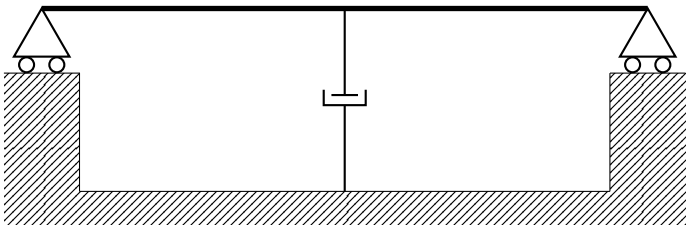

Figure 3.4: Damped beam simply supported at its ends.

From the context, it is natural to consider $q_0(\lambda)$ as a polynomial of *grade* $2nr$ (see [57]), in which case we say that $q_0(\lambda)$ has r infinite roots. Consider a finite root λ_i of $q_0(\lambda)$ of multiplicity α. In a neighborhood of $t = 0$, the solutions λ of $q_t(\lambda) = 0$ can be expanded in Puiseux series in t, and α of these series (counting multiplicities) equal λ_i at $t = 0$ [47, Chapter 5]. Thus, α roots of $q_t(\lambda)$, seen as functions of t, converge to λ_i as $t \to 0$. This shows that $2nr - r$ roots of $q_t(\lambda)$ converge to the finite roots of $q_0(\lambda)$ as $t \to 0$. Furthermore, from Theorem 3.3.3 we know that the remaining r roots go to $-\infty$ as $t \to 0$.

To show that the finite roots of $q_0(\lambda)$ are purely imaginary, we note that they all are eigenvalues of $p(\lambda)Q(\lambda)$. Since $p(-\lambda)Q(-\lambda) = -p(\lambda)Q(\lambda)^T$, $p(\lambda)Q(\lambda)$ is T-odd (by definition), so λ_i is an eigenvalue if and only if $-\lambda_i$ also is an eigenvalue [56, Theorem 4.2]. Therefore, if $q_0(\lambda)$ has a root with negative real part, it also has a root with positive real part. But this is impossible. Indeed, $q_t(\lambda)$ has a root with positive real part for some $t > 0$ only if there is an affected eigenvalue with positive real part, a contradiction. ∎

We illustrate Theorem 3.3.5 by a numerical example.

Example 3.3.6. Consider the damped beam problem from the collection NLEVP [10]. The modes of the QEP describe the vertical displacements of a beam that is supported at its ends and has a viscous damper attached to it in the middle (see Figure 3.4). We used the MATLAB function `nlevp` to create the QEP such that the coefficient matrices are of size 100×100. We then created a strongly damped version of the same problem by multiplying the damping matrix by 10^{10}. We used the algorithm described in Chapter 4 to compute all eigenvalues of both problems. The computed spectra are shown in Figure 3.5, with the exception of one large negative eigenvalue for the strongly damped problem. The experiment does indeed confirm Theorem 3.3.5. ▲

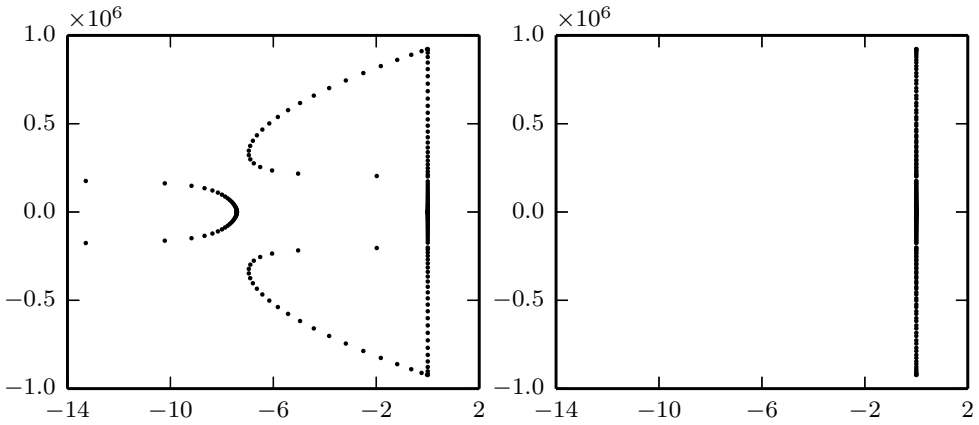

Figure 3.5: Left: The computed spectrum of the standard damped beam problem. Right: The computed spectrum of the modified strongly damped problem, with the exception of one eigenvalue around -6.3×10^{12} which is far outside the plotted window.

We end this section by discussing the simplicity of the eigenvalues of strongly damped matrix polynomials. The matrix $L(s)$ in (3.10), and hence the matrix polynomial $P_s(\lambda)$, has a constant number, k, of distinct eigenvalues for all but a finite number of values of s, known as *exceptional points* [45, p. 64]. Exceptional points are in general nonreal, so for the sake of the argument, we temporarily expand the scope and allow nonreal values of s. Now, in any simple domain not containing any exceptional points, we have a Jordan decomposition (albeit not in its usual likeness)

$$L(s) = \sum_{i=1}^{k} E_i(s)\lambda_i(s) + F_i(s),$$

where $E_i(s)$, $\lambda_i(s)$ and $F_i(s)$ denote the *eigenprojections*, (distinct) eigenvalues and *eigennilpotents*, respectively, and all are analytic in the considered domain [45, p. 68]. The following observation was made in [53, Theorem 3.3]: for purely imaginary s, $L(s)$ is skew-Hermitian so the eigennilpotents must vanish. Because any simple domain, free of exceptional points, can be expanded to contain an interval of the imaginary axis, in a manner that avoids including exceptional points, the eigennilpotents must vanish identically for any non-exceptional s. Put simply, defective eigenvalues can only exist for a finite number of exceptional values of s. This implies the following theorem.

Theorem 3.3.7. *For large enough s, all eigenvalues of $P_s(\lambda)$ (defined in (3.3)) are semisimple.*

3.4 Eigenspaces

Let $\langle \cdot, \cdot \rangle$ be the M-inner product and $\| \cdot \|$ the induced norm. In the undamped case, $s = 0$, the eigenspaces of $P_s(\lambda)$ corresponding to complex conjugate eigenvalues are identical, and any other two eigenspaces are orthogonal with respect to $\langle \cdot, \cdot \rangle$. Furthermore, all eigenspaces have real bases. The same is, in general, not true when $s > 0$. Our goal in this section is to show that for large enough s, the same properties are almost true if we restrict ourselves to the eigenspaces corresponding to eigenvalues in $\mathcal{S}$ and group together eigenspaces corresponding to "close" eigenvalues; more precise statements will be made in Theorem 3.4.1 and Corollary 3.4.3.

Suppose N is large enough so there are no exceptional points in (N, ∞); such N exists due to Theorem 3.3.7. Then there is a constant k such that $P_s(\lambda)$ has exactly k distinct eigenvalues $\lambda_1(s), \lambda_2(s), \ldots, \lambda_k(s)$ for $s > N$, which are analytic functions of s. Let $\mathcal{V}_i(s)$ denote the eigenspace corresponding to $\lambda_i(s)$ and define

$$\mathcal{U}_z(s) \;=\; \operatorname{span}\Big\{ \mathcal{V}_i(s) \;:\; \lim_{s\to\infty} \lambda_i(s) = z \Big\} \quad \text{for} \quad s > N. \tag{3.14}$$

We shall prove the following result.

Theorem 3.4.1. *Let $\langle \cdot, \cdot \rangle$ be the M-inner product and $z_1, z_2, \ldots, z_{2p}$ the distinct nonzero points on the imaginary axis to which some eigenvalue of $P_s(\lambda)$ converges as $s \to \infty$. If $\theta_1(\cdot, \cdot)$ and $\theta_{\max}(\cdot, \cdot)$ refer to the smallest and largest principal angles with respect to $\langle \cdot, \cdot \rangle$, respectively, then the following hold:*

(a) *If $z_i \neq \overline{z}_j$ and $z_i \neq z_j$ then $\lim_{s\to\infty} \theta_1\big(\mathcal{U}_{z_i}(s), \mathcal{U}_{z_j}(s)\big) = \pi/2$.*

(b) $\lim_{s\to\infty} \theta_1\big(\mathcal{U}_{z_i}(s), \mathcal{U}_{-\infty}(s)\big) = \pi/2$.

(c) $\lim_{s\to\infty} \theta_1\big(\mathcal{U}_{z_i}(s), \mathcal{U}_{0}(s)\big) = \pi/2$.

(d) $\lim_{s\to\infty} \theta_{\max}\big(\mathcal{U}_{z_i}(s), \mathcal{U}_{\overline{z}_i}(s)\big) = 0$.

(e) $\lim_{s\to\infty} \theta_{\max}\big(\mathcal{U}_{-\infty}(s), \mathcal{U}_{0}(s)\big) = 0$.

To prove Theorem 3.4.1, we need the following lemma.

Lemma 3.4.2. *Let $\mathcal{S}$ be defined by* (3.12) *and consider eigenpairs (λ_i, v_i) and (λ_j, v_j) of $P_s(\lambda)$, for some $s > 0$, for which $\|v_i\| = \|v_j\| = 1$ and $\lambda_i \in \mathcal{S}$. There are constants c_1 and c_2 which are independent of v_i, v_j and s, such that the following bounds hold:*

(a) *If* $\lambda_j \in \mathcal{S}$ *then* $|v_i^H D v_j| \leq (c_1/s)^2$.

(b) $|v_i^H D v_j| \leq c_2/s$.

Proof. Recall that $D = RR^T$ with $R \in \mathbb{R}^{n\times r}$. We have

$$RR^T v_i = -(\lambda_i s)^{-1}(M\lambda_i^2 + K)v_i.$$

Left multiplication with $M^{-1/2}$ times the Moore-Penrose pseudo-inverse $R^\dagger$, and taking norms, yield

$$\|M^{-1/2}R^T v_i\| = \frac{\|M^{-1/2}R^\dagger(M\lambda_i^2 + K)v_i\|}{|\lambda_i s|}.$$

Since $|\lambda_i|$ is bounded from below and above there is a constant c_1 such that $\|M^{-1/2}R^T v_i\| \leq c_1/s$, for any s and any choice of v_i. For case (a), an analogous argument gives $\|M^{-1/2}R^T v_j\| \leq c_1/s$ and the Cauchy-Schwartz inequality yields

$$\begin{aligned} |v_i^H D v_j| &= |\langle M^{-1/2}R^T v_i, M^{-1/2}R^T v_j\rangle| \\ &\leq \|M^{-1/2}R^T v_i\|\|M^{-1/2}R^T v_j\| \\ &\leq (c_1/s)^2. \end{aligned}$$

Similarly, for part (b), we have

$$|v_i^H D v_j| \leq \|M^{-1/2}R^T v_i\|\|M^{-1/2}R^T v_j\| \leq c_2/s.$$

for $c_2 = \|M^{-1/2}R^T\|c_1$. ∎

Proof of Theorem 3.4.1. For $i = 1\!:\!2p$, define $\mathcal{B}_i = \{z : |z - z_i| \leq \delta\}$ where $\delta > 0$ is small enough so $\mathcal{B}_i \cap \mathcal{B}_j = \emptyset$ for $i \neq j$, and let

$$\gamma = \min_{i\neq j} \operatorname{dist}(\mathcal{B}_i, \mathcal{B}_j). \tag{3.15}$$

Choose $N > 0$ such that for $s > N$ it holds that $\mathcal{B}_i$ contains all eigenvalues of $P_s(\lambda)$ that converge to z_i. The $\mathcal{B}_i$ will hereafter be referred to as "limit balls."

Now, pick $\varepsilon > 0$ and let c_1 and c_2 be the constants from Lemma 3.4.2. By possibly choosing an even larger N, we may assume that $N > c_1^2/(\varepsilon\gamma)$ and that all eigenvalues in $\mathcal{S}_{out}$, defined in (3.11), have modulus greater than c_2/ε. Consider two eigenpairs (λ_i, v_i) and (λ_j, v_j) for which $\|v_i\| = \|v_j\| = 1$. If λ_i and λ_j belong to different limit balls that are not complex conjugate sets, then we have for any

real s that

$$\begin{aligned}
\overline{\lambda}_i^2 v_i^H M v_j &= (\lambda_i^2 M v_i)^H v_j \\
&= (-(sD\lambda_i + K)v_i)^H v_j \\
&= v_i^H(-(sD\overline{\lambda}_i + K))v_j \\
&= v_i^H(sD(\lambda_j - \overline{\lambda}_i) - (sD\lambda_j + K))v_j \\
&= v_i^H(sD(\lambda_j - \overline{\lambda}_i)v_j - (sD\lambda_j + K)v_j) \\
&= s(\lambda_j - \overline{\lambda}_i)v_i^H D v_j + \lambda_j^2 v_i^H M v_j,
\end{aligned}$$

and further

$$-v_i^H M v_j = \frac{s v_i^H D v_j}{\overline{\lambda}_i + \lambda_j}. \tag{3.16}$$

From (3.15) we have $|\overline{\lambda}_i + \lambda_j| \geq \gamma$, and by part (a) of Lemma 3.4.2 $|s v_i^H D v_j| \leq c_1^2/s$. Thus, $|v_i^H M v_j| < \varepsilon$ for $s > N$. Since this bound is independent of which normalized eigenvectors v_i and v_j we picked, and $\varepsilon > 0$ is arbitrary, we have proved part (a) of the theorem.

If λ_i is in a limit ball and $\lambda_j \in \mathcal{S}_{out}$, then v_i and v_j also satisfy (3.16). By part (b) of Lemma 3.4.2 $|s v_i^H D v_j| \leq c_2$ and $|\overline{\lambda}_i + \lambda_j| \geq |\lambda_j| > c_2/\varepsilon$. Hence, $|v_i^H M v_j| < \varepsilon$ for $s > N$, and we have shown part (b) of the theorem.

Since $\operatorname{rev} P_s(\lambda)$ has the same eigenspaces as $P_s(\lambda)$, part (c) follows immediately if we consider part (b) for the reversed matrix polynomial.

For $s > N$, the eigenvectors corresponding to the eigenvalues of $P_s(\lambda)$ in $\mathcal{B}_i$ and $\mathcal{S}_{out}$, span the subspaces $\mathcal{U}_{z_i}(s)$ and $\mathcal{U}_{-\infty}(s)$, respectively. Furthermore, due to Theorem 3.3.7, $\dim \mathcal{U}_{z_i}(s)$ and $\dim \mathcal{U}_{-\infty}(s)$ equal the sums of the algebraic multiplicities of all eigenvalues in $\mathcal{B}_i$ and $\mathcal{S}_{out}$, respectively. Since $z_1, z_2, \ldots, z_{2p}$ can be paired into complex conjugates, we may assume, without loss of generality, that $\operatorname{Im}(z_i) > 0$ for $i = 1\!:\!p$. We have

$$\dim \mathcal{U}_{z_1} + \dim \mathcal{U}_{z_2} + \cdots + \dim \mathcal{U}_{z_p} + \dim \mathcal{U}_{-\infty} = n.$$

Part (a) and part (b), which we just proved, imply that

$$\lim_{s\to\infty} \theta_1\big(\mathcal{U}_{\overline{z}_i}(s), \mathcal{U}_z(s)\big) = \pi/2$$

for $z \in \{z_1, z_2, \ldots, z_p, -\infty\} \setminus \{z_i\}$. Hence, part (d) follows from Proposition 3.2.4. Similarly, we have

$$\lim_{s\to\infty} \theta_1\big(\mathcal{U}_0(s), \mathcal{U}_z(s)\big) = \pi/2$$

for $z \in \{z_1, z_2, \ldots, z_p\}$, so also part (e) follows from Proposition 3.2.4. ∎

The next corollary is an immediate consequence of part (d) of Theorem 3.4.1.

Corollary 3.4.3. *For any $\varepsilon > 0$, $\mathcal{U}_{z_i}(s)$ in (3.14) has an M-orthonormal basis $\{u_1, u_2, \ldots, u_k\}$, where $\|\mathrm{Im}(u_j)\| < \varepsilon$, $j = 1\!:\!k$, for large enough s.*

Proof. The corollary follows immediately from Proposition 3.2.5. ∎

We end this section by a numerical experiment that illustrates the implications of Corollary 3.4.3 to the damped beam problem that was discussed in Example 3.3.6.

Example 3.4.4. As in Example 3.3.6, we created several versions of the 100×100 damped beam problem by multiplying the original damping matrix by a parameter s. We then used the algorithm described in Chapter 4 to solve these eigenproblems for a sequence of increasing values of s. We then used the technique explained in the proof of Proposition 3.2.5 to compute the normalized eigenvectors $y_1, y_2, \ldots, y_{2n}$ with *smallest* imaginary part. For eigenspaces of dimension one, it follows from (3.9) that this technique indeed gives us the normalized eigenvectors with smallest imaginary parts. Since the Takagi factorization is trivial to compute for scalars, such eigenvectors are easily determined once any nonzero vectors in the associated eigenspaces are known. We then computed the norm of the imaginary parts of all eigenvectors $y_1, y_2, \ldots, y_{2n}$ and saved the largest one for each value of s. In Figure 3.6 this norm is plotted as a function of s. The figure confirms the theory: when the damper gets strong enough the eigenvectors become almost real, meaning that all nodes in the discretized model vibrate essentially in phase. ▲

3.5 Forced response

The solution of the homogeneous differential equation (3.1) is known as the *free response*. The *forced response* is the corresponding particular solution when a nonzero right hand side $f(t)$ is added to the equation. As we saw in Section 1.1, the forced response is also strongly connected to the eigenvalues and eigenspaces. We now discuss this in more detail. Consider the differential equation

$$P_s\left(\frac{d}{dt}\right) u(t) = f_0 e^{i\omega t}, \quad t \in [0, T], \tag{3.17}$$

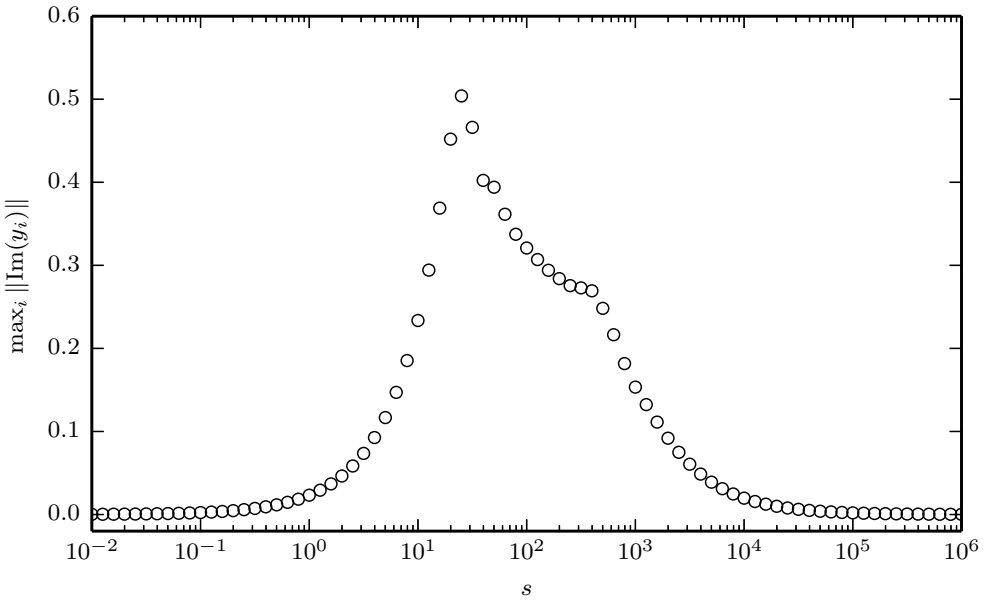

Figure 3.6: The maximum of the norms of the imaginary parts of all the computed eigenvectors $y_1, y_2, \ldots, y_{2n}$ of the modified damped beam problems as a function of the damping parameter s.

where $P_s(\lambda)$ is as in (3.3), and assume $i\omega$, $\omega \in \mathbb{R}$, is not an eigenvalue $P_s(\lambda)$. Suppose further that s is large enough so all eigenvalues are semisimple. By (1.8)

$$u_p(t) = \sum_{j=1}^{n\ell} x_j y_j^T f_0 \psi_j(t), \quad \psi_j(t) := \frac{e^{i\omega t} - e^{\lambda_j t}}{i\omega - \lambda_j},$$

is a particular solution of (3.17), where the λ_j are the eigenvalues of $P_s(\lambda)$ and the y_j and x_j are the associated left and right eigenvectors, respectively, normalized such that

$$y_i^T \left(\frac{d}{d\lambda} P_s(\lambda) \Big|_{\lambda=\lambda_j} \right) x_j = y_i^T (2\lambda_j M + sD) x_j = \delta_{ij} \quad \text{if} \quad \lambda_i = \lambda_j. \tag{3.18}$$

Since our matrix polynomial is real and symmetric, we can derive a nicer "symmetric" formula for the response. Because (1.8) followed from (1.4), we aim to find a "symmetric version" of (1.4). We have the following theorem, which shows that the assumptions under which Lancaster [50, pp. 127–129] derived formulas for the forced response are always true.

Theorem 3.5.1. *Let $P(\lambda) \in \mathbb{R}[\lambda]^{n\times n}$ be symmetric, of degree ℓ, and with nonsingular leading coefficient A_ℓ. If all eigenvalues of $P(\lambda)$ are semisimple, then it*

holds that

$$\lambda^k P(\lambda)^{-1} = \sum_{j=1}^{n\ell} \frac{\lambda_j^k x_j x_j^T}{\lambda - \lambda_j} + \delta_{k\ell} A_\ell^{-1}, \quad k = 1\!:\!\ell,$$

where the (λ_j, x_j) *are eigenpairs such that* $x_i^T P'(\lambda_j) x_j = \delta_{ij}$ *if* $\lambda_i = \lambda_j$*. Furthermore, if* λ_j *is real, then there is an* x_j *that is real or purely imaginary.*

Proof. By [50, p. 65], equation (1.4) holds for left and right eigenvectors y_j and x_j that satisfy (1.5). Lancaster showed that such eigenvectors existed as long as all eigenvalues are semisimple. Recall that, for symmetric problems, each right eigenvector is a left eigenvector and vice versa. Suppose λ_j has multiplicity s and let the columns of $Y_j = [y_{j_1}\ y_{j_2}\ \cdots\ y_{j_s}]$ and $X_j = [x_{j_1}\ x_{j_2}\ \cdots\ x_{j_s}]$ be associated eigenvectors such that (1.5) hold, or equivalently, such that

$$Y_j^T P'(\lambda_0) X_j = I_s. \tag{3.19}$$

We need to show that there exists an eigenvector matrix Z_j such that

$$Z_j^T P'(\lambda_0) Z_j = I_s. \tag{3.20}$$

Since, $P(\lambda)$ is symmetric, X_j and Y_j span the same subspace, so $Y_j = X_j S$. We have

$$X_j^T P'(\lambda_0) X_j = S^{-T}, \tag{3.21}$$

where S^{-T} is complex symmetric. If $U\Sigma U^T$ is a Takagi factorization of S^{-T} then $Z_j = X_j \overline{U} \Sigma^{-1/2}$ satisfies (3.20).

For the second part, we assume that X_j is real. We may do this, since (3.19) was deduced in [50] by first considering an arbitrary matrix X_j such that range(X_j) is the right eigenspace associated with λ_j, and then explicitly construct the corresponding Y_j. Taking the real part of (3.19), yields $\mathrm{Re}(Y_j)^T P'(\lambda_j) X_j = I_s$, which shows that $\mathrm{Re}(Y_j)$ is of full rank. Further, taking the real part of $Y_j^T P(\lambda_j) = 0$, shows that all columns of $\mathrm{Re}(Y_j)$ are (left) eigenvectors associated with λ_0. Hence, we may assume that both X_j and Y_j in (3.19) are real. In this case, S^{-T} in (3.21) becomes real symmetric and thus enjoys a real spectral decomposition $S^{-T} = Q\Lambda Q^T$. If we define $Z_j = X_j Q_j \Lambda^{-1/2}$, then each column is either real or purely imaginary, and Z_j satisfies (3.20). ∎

Using Theorem 3.5.1 and the discussion in Section 1.1, we can now write down

"symmetric" formulas for particular solutions of differential equations of the form

$$P\left(\frac{d}{dt}\right)u(t) = f(t), \quad t \in [0,T],$$

where $P(\lambda)$ is as in Theorem 3.5.1 and $f(t)$ is piecewise continuous. In particular, it follows that

$$u_p(t) = \sum_{j=1}^{n\ell} x_j x_j^T f_0 \psi_j(t), \quad \psi_j(t) := \frac{e^{i\omega t} - e^{\lambda_j t}}{i\omega - \lambda_j}, \tag{3.22}$$

is a solution to (3.17), where the (λ_j, x_j) are eigenpairs such that

$$x_j^T(2\lambda_j M + sD)x_j = 1, \tag{3.23}$$

and x_j is real or purely imaginary if λ_j is real.

We saw in Section 3.3 that strongly damped matrix polynomials can have negative eigenvalues arbitrarily close to the origin. This raises the question: what happens when ω is small and λ_j is small and negative? In such situations we certainly have that $i\omega$ is close to λ_j, in which case (1.9) shows that $\psi_j(t)$ can be quite large. To answer this, we investigate the M-norm of the terms $x_j x_j^T f_0 \psi_j(t)$ in (3.22). Define m, d and k through

$$\|x_j\|_M^2(m\lambda^2 + d\lambda + k) = x_j^T(M\lambda^2 + sD\lambda + K)x_j \tag{3.24}$$

and note that (3.23) and $m\lambda_j^2 + d\lambda_j + k = 0$ imply

$$\|x_j\|_M^2 = \frac{1}{|2m\lambda_j + d|} = \frac{1}{|2m\lambda_j - m\lambda_j - k/\lambda_j|} = \frac{|\lambda_j|}{|m\lambda_j^2 - k|}.$$

If $v_j := x_j/\|x_j\|_M$, then we can bound the term under investigation as follows:

$$\|x_j\|_M |x_j^T f_0| \left|\frac{e^{i\omega t} - e^{\lambda_j t}}{i\omega - \lambda_j}\right| = |v_j^T f_0| \left|\frac{e^{i\omega t} - e^{\lambda_j t}}{i\omega - \lambda_j}\right| \left|\frac{\lambda_j}{m\lambda_j^2 - k}\right|. \tag{3.25}$$

Furthermore, if λ_j is negative, then the right hand side of (3.25) is bounded by $2|v_j^T f_0|/|m\lambda_j^2 - k|$. Now, if λ_j is a small negative eigenvalue such that

$$\lambda_j^2 \ll \omega_{\min} := \min_{\|u\|_M = 1} u^H K u,$$

then $|m\lambda_j^2 - k| = |k| - |\lambda_j|^2 \gtrsim \omega_{\min}$. Thus the term $x_j x_j^T f_0 \psi_j(t)$ is well behaved

provided that the undamped problem does not have a tiny eigenvalue.

3.5.1 A remark on nearly defective systems

We end this chapter by discussing the application of the formula (3.22) to QEPs that are nearly defective with respect to small perturbations in the damping parameter s. In view of Theorem 3.3.7, such QEPs are not strongly damped. From (3.25), we see that the "response in the jth mode," that is, $x_j x_j^T f_0 \psi_j(t)$, can be large when $m\lambda_j^2 \approx k$ (where m and k are defined in (3.24)). We now show that this can happen when a small perturbation of the damping parameter s yields a defective eigenvalue. Suppose $P_s(\lambda)$ has a defective eigenvalue λ_j for $s = a$. We are interested in the behavior of the eigenvalues and eigenspaces in the neighborhood of $s = a$. Since the left companion matrix associated with $M^{-1}P_s(\lambda)$ depends continuously (in fact linearly) on the *real* parameter s, we may parametrize the eigenvalue λ_j as a continuous function of s [45, pp. 106–110]. In other words, there exists a continuous function $\lambda_j(s)$ such that $\det P_s(\lambda_j(s)) = 0$ and $\lambda_j(a) = \lambda_j$, the defective eigenvalue of $P_a(\lambda)$ referred to above. Now, suppose $\dim \operatorname{null} P_s(\lambda_j(s)) = 1$ near $s = a$. Then it can be shown that the associated eigenspace is (a) continuous in s, where the continuity is understood in the gap metric (see Appendix A), and (b) has a continuous spanning vector $v_j(s)$ [31, pp. 408–411]. Define

$$f(s) := \frac{v_j(s)^T(2\lambda_j(s)M + sD)v_j(s)}{\|v_j(s)\|_M^2} = \frac{v_j(s)^T(M\lambda_j(s)^2 - K)v_j(s)}{\lambda_j(s)\|v_j(s)\|_M^2},$$

and note that $f(s)$ is continuous. By Theorem 2.3.1, we have $f(a) = 0$, so $f(s)$ takes arbitrarily small values in a neighborhood of a. For s_0 such that $\lambda_j(s_0) \in \mathcal{S}$ (defined in (3.12)) and $0 < f(s_0) < \epsilon/\omega_{\max}$ we have

$$|m(s_0)\lambda_j(s_0)^2 - k(s_0)| < \epsilon,$$

where $m(s_0) = v_j(s_0)^T M v_j(s_0)/\|v_j(s_0)\|_M^2$ and $k(s_0) = v_j(s_0)^T K v_j(s_0)/\|v_j(s_0)\|_M^2$. In view of (3.25), this shows that if $P_s(\lambda)$ has a defective eigenvalue for $s = a$, then (3.22) can have arbitrarily large terms in a neighborhood of $s = a$, at, say, $t = t_0 > 0$, even when $i\omega$ is bounded away from the spectrum of $P_s(\lambda)$ for s close to a. However, from (2.9), it is clear that the solution is bounded at $t = t_0$ for all s in a small interval $(a - \epsilon, a + \epsilon)$, $\epsilon > 0$. This shows that when the problem is nearly defective (in the discussed sense), severe cancellation can take place when summing up the terms in (3.22).

Chapter

4 A quadratic eigensolver for problems with low rank damping

4.1 Introduction

In this chapter we consider quadratic eigenproblems with damping matrices of low rank. More precisely, we consider problems of the form

$$\left(M\lambda^2 + D\lambda + K\right) x = 0, \tag{4.1}$$

where M, D and K are real, $n \times n$ and positive semidefinite matrices, (M, K) is regular (that is, $\det(M\lambda + K) \not\equiv 0$) and $r := \operatorname{rank} D \ll n$. Without the low rank term $D\lambda$, a simple substitution, $\omega = -\lambda^2$, turns (4.1) into a definite generalized eigenproblem (GEP):

$$Kx = \omega Mx, \tag{4.2}$$

which is much easier to solve. Our goal is to first develop an algorithm that solves (4.2) such that all eigenvalues are computed in a symmetry preserving and backward stable manner. We then design a fast Ehrlich-Aberth iteration that modifies the solution of (4.2) until we have found the eigenvalues of the damped problem (4.1). Finally, if the eigenvectors are desired, we compute these using a special inverse iteration that is based on the Takagi factorization for complex symmetric matrices.

4.1.1 Motivation

QEPs of the form (4.1) appear naturally in modal analysis of physical structures. Modal analysis is the study of the vibrational properties of structures. We now discuss this application briefly. The discussion serves not only as a motivation, but also allows us to use our intuition of mechanical systems to understand the choice of starting points used in our algorithm later on.

As we saw in Chapter 3, the homogeneous differential equation

$$\left(M\frac{d^2}{dt^2} + D\frac{d}{dt} + K\right)u(t) = 0 \tag{4.3}$$

play an important role in structural engineering. Before we assumed that mass and stiffness matrices M and K were both symmetric positive definite. In this chapter we relax these conditions and allow both to be symmetric positive semidefinite, as long as the pencil (M, K) is regular. The damping matrix D is symmetric positive semidefinite as before, which implies that $M\lambda^2 + D\lambda + K$ is regular (since it evaluates to a positive definite matrix for large positive values of λ). In theory, M should be strictly positive definite, but singular M is common in practice due to further simplifications by the engineers. Singular M may, for instance, arise when the lumped mass model is used. In this case there is no inertia with respect to the rotational degrees of freedom of a beam, say, so the corresponding mass matrix becomes singular [46, 69]. The stiffness matrix K is positive definite or positive semidefinite, depending on the boundary conditions. The damping matrix depends on what kind of damping is modeled. We consider the case of discrete (linear) damping, which refers to the physical objects called viscous dampers that were discussed in Section 3.1. When a viscous damper is modeled with finite elements, it appears as a positive semidefinite rank one term of the damping matrix. Hence, if the structure only has a few viscous dampers, their contribution to the damping matrix is of low rank. This is indeed the case in several applications. It is not uncommon that less than ten dampers are used, and in some cases as few as one or two [1].

If only a few viscous dampers are used, and (M, K) is regular, then the corresponding QEP we need to solve has the structure of (4.3). In particular, if (λ, x) is an eigenpair of (4.1), then $u(t) = e^{\lambda t}x$ is a solution to (4.3). We recall from Chapter 3 that the spectrum is symmetric with respect to the real axis and lies in the left half plane. Thus, if $(-d + i\omega, x)$ with $\omega > 0$ is an eigenpair of (4.1),

then so is $(-d - i\omega, \overline{x})$. It follows that the real function

$$u(t) = e^{-d}(\cos(t\omega)\mathrm{Re}(x) + \sin(t\omega)\mathrm{Im}(x)).$$

is a solution to (4.3). Notice how the real and imaginary parts of the eigenvalue correspond to damping and frequency respectively.

4.1.2 Existing algorithms and our objective

The conventional way of solving (4.1) is through linearization, which means that the problem is rewritten as GEP of twice the size. This approach does not respect the special structure of problem (4.1). There do exist symmetric linearizations, but no stable algorithm that can preserve this symmetry is currently available.

Recently, Hammarling, Munro and Tisseur proposed a linearization based algorithms for finding all eigenpairs of general regular quadratic eigenproblems [34]. Their algorithm, called `quadeig`, is backward stable in the unstructured sense described in Section 2.3, as long as the damping is not too strong. The bulk of the computation lies in solving the linearized problem, for which the QZ algorithm is used. The QZ algorithm is estimated to use $50m^3$ flops ($30m^3$ flops if we only want the eigenvalues), where m is the size of the matrices [33, p. 413]. Since `quadeig` works on a linearization we have $m = 2n$, where n is size of the coefficient matrices M, D and K. We get an estimated complexity of $400n^3$ flops ($240n^3$ flops for eigenvalues only).

We shall develop an algorithm that exploits the structure of problem (4.1) and whose main complexity lies in finding all eigenpairs of the definite GEP (4.2). There are several methods for solving (4.2), but no existing algorithms are both backward stable and symmetry preserving. We develop an algorithm, based on an algorithm proposed by Wang and Zhao [89]. Given the pencil (M, K), where M and K are as in (4.2), the new algorithm computes a nonnegative diagonal pencil (M_D, K_D) that is congruent to $(M + \Delta M, K + \Delta K)$, where ΔM and ΔK are symmetric and small in norm with respect to M and K, respectively. Thus, it computes all eigenvalues of (4.2) in a backward stable and symmetry preserving manner. This new algorithm is interesting on its own, since GEPs of the form (4.2) are not uncommon in applications. Further, we prove a result that gives expressions for the number of zero and infinite eigenvalues. These expressions can be evaluated using quantities that are conveniently computed as byproducts in our algorithm for (4.2). This allows us to deflate all such eigenvalues as soon as (4.2) has been solved. Finally, we mention that our algorithm for (4.2) is estimated to need only

$26n^3$ flops if M and K are nonsingular, and up to $43n^3$ flops otherwise. This means that the estimated flop count for our quadratic eigensolver is significantly lower than QZ-based solvers like `quadeig` and MATLAB's `polyeig`.

The outline of the chapter is as follows. In Section 4.2 necessary background material is discussed. This includes definitions of backward errors for QEPs and the Ehrlich-Aberth method. In Section 4.3 we review Wang and Zhao's algorithm for definite GEPs and develop a new algorithm based on it that can solve (4.2) in a backward stable and symmetry preserving manner. In Section 4.4 our algorithm for solving (4.1) is described and in Section 4.5 we present the results from numerical experiments. In Section 4.6 we discuss the application to the large scale case, the possibility of generalizations and related work.

4.2 Preliminaries

4.2.1 Backward errors for polynomial eigenproblems

Let $(\widetilde{x}, \widetilde{\lambda})$ denote a computed eigenpair of a matrix polynomial

$$P(\lambda) = \sum_{k=0}^{\ell} A_k \lambda^k \quad \text{and let} \quad \Delta P(\lambda) = \sum_{k=0}^{\ell} \Delta A_k \lambda^k$$

denote a perturbation of $P(\lambda)$. If $\widetilde{\lambda}$ is finite, we follow [81] and define the relative backward error of the computed eigenpair $(\tilde{\lambda}, \widetilde{x})$ as

$$\eta_P(\widetilde{\lambda}, \widetilde{x}) = \min\{\epsilon : (P + \Delta P)(\widetilde{\lambda})\tilde{x} = 0,\ \|\Delta A_i\| \le \epsilon \|A_i\|,\ i = 0\colon \ell\}, \tag{4.4}$$

and the relative backward error of the computed eigenvalue $\tilde{\lambda}$ as

$$\eta_P(\widetilde{\lambda}) = \min_{\widetilde{x} \neq 0} \eta_P(\widetilde{\lambda}, \widetilde{x}). \tag{4.5}$$

In general $\|\cdot\|$ can be any matrix norm; in this chapter, however, we will only use the spectral norm, so $\|\cdot\| = \|\cdot\|_2$. For the spectral norm, it was proved in [81] that

$$\eta_P(\widetilde{\lambda}, \widetilde{x}) = \|P(\widetilde{\lambda})\widetilde{x}\| \Big(\|\widetilde{x}\| \sum_{k=0}^{\ell} \|A_k\| |\widetilde{\lambda}|^k \Big)^{-1} \tag{4.6}$$

and

$$\eta_P(\widetilde{\lambda}) = \Big(\|P(\widetilde{\lambda})^{-1}\| \sum_{k=0}^{\ell} \|A_k\| |\widetilde{\lambda}|^k \Big)^{-1}. \tag{4.7}$$

Notice that

$$\eta_P(\widetilde{\lambda}, \widetilde{x}) = \eta_{\operatorname{rev} P}(1/\widetilde{\lambda}, \widetilde{x}) \quad \text{and} \quad \eta_P(\widetilde{\lambda}) = \eta_{\operatorname{rev} P}(1/\widetilde{\lambda})$$

for $\widetilde{\lambda} \neq 0$, where

$$\operatorname{rev} P(\lambda) := \sum_{k=0}^{\ell} A_{\ell-k}\lambda^k.$$

Since infinite eigenvalues of $P(\lambda)$ are defined as the zero eigenvalues of $\operatorname{rev} P(\lambda)$, it is natural to define

$$\eta_P(\infty, \widetilde{x}) = \eta_{\operatorname{rev} P}(0, \widetilde{x}) \quad \text{and} \quad \eta_P(\infty) = \eta_{\operatorname{rev} P}(0).$$

We also note that if $Q(\lambda)$ is related to $P(\lambda)$ via a simple parameter scaling, so

$$Q(\lambda) = \sum_{k=0}^{\ell} (s^k A_k)\lambda^k,$$

then

$$\eta_P(s\widetilde{\lambda}, \widetilde{x}) = \eta_Q(\widetilde{x}, \widetilde{\lambda}) \quad \text{and} \quad \eta_P(s\widetilde{\lambda}) = \eta_Q(\widetilde{\lambda}). \tag{4.8}$$

4.2.2 Ehrlich-Aberth iteration

The Ehrlich-Aberth method [3, 23] is an algorithm for simultaneously finding all roots of a scalar polynomial. If $p(\lambda) = 0$ is the scalar polynomial equation we want to solve, then the algorithm takes starting points $\lambda_1^{(0)}, \ldots, \lambda_\ell^{(0)}$, where $\ell = \deg(p)$, and then update these points via

$$\lambda_k^{(i+1)} = \lambda_k^{(i)} - \frac{q(\lambda_k^{(i)})}{1 - q(\lambda_k^{(i)}) \sum_{j \neq k} \dfrac{1}{\lambda_k^{(i)} - \lambda_j^{(i)}}}, \tag{4.9}$$

where $q(\lambda) := p(\lambda)/p'(\lambda)$. Clearly these updates can be done in parallel, which is nice, but if we insist to update in sequential order we might as well use the latest approximations available. This leads to the slightly faster Gauss-Seidel version:

$$\lambda_k^{(i+1)} = \lambda_k^{(i)} - \frac{q(\lambda_k^{(i)})}{1 - q(\lambda_k^{(i)}) \left(\sum_{j<k} \dfrac{1}{\lambda_k^{(i)} - \lambda_j^{(i+1)}} - \sum_{j>k} \dfrac{1}{\lambda_k^{(i)} - \lambda_j^{(i)}} \right)}. \tag{4.10}$$

In practice, the Ehrlich-Aberth method exhibits rapid convergence to isolated simple eigenvalues if good starting points are provided. The algorithm also converges for multiple and tightly clustered eigenvalues, but more iterations are generally required in these cases.

Recently, Bini and Noferini [11] used the Ehrlich-Aberth method for finding the eigenvalues of regular matrix polynomials. If $P(\lambda)$ is such a matrix polynomial, their algorithm applies the Ehrlich-Aberth iteration to the equation $\det P(\lambda) = 0$, and for the selection of starting points, it makes use of Newton polygons. For the evaluation of $p(\lambda)/p'(\lambda)$, which is the most expensive part of the updating process, they used Jacobi's formula

$$\frac{d}{d\lambda} \det P(\lambda) = \operatorname{trace}\left(P(\lambda)^{-1} P'(\lambda)\right) \det P(\lambda)$$

to obtain

$$p'(\lambda)/p(\lambda) = \operatorname{trace}\left(P(\lambda)^{-1} P'(\lambda)\right). \tag{4.11}$$

By using (4.11), each update costs $O(n^3)$ flops.

Since the method is iterative, some stopping criterion is needed. Bini and Noferini gave two suggestions: either stop updating λ_i when the condition number of $P(\lambda_i)$ is large enough, or when the associated backward error (4.7) is small enough. Both criteria require $O(n^3)$ flops to check.

The Ehrlich-Aberth method can only be used to find the eigenvalues. If also the eigenvectors are sought, these can be found afterwards using inverse iteration or the SVD—both techniques requires $O(n^3)$ flops per eigenvector.

The algorithm in [11] demonstrated superb accuracy in numerical tests, but is unfortunately an expensive alternative for solving QEPs. Applied to an $n \times n$ QEP the complexity is $O(n^4)$—assuming the starting points are good enough so the number of iterations is independent of n.

4.3 GEPs with semidefinite matrices

Wang and Zhao [89] proposed an algorithm for solving

$$Ax = \lambda Bx, \tag{4.12}$$

where $A, B \in \mathbb{R}^{n \times n}$ are symmetric positive definite. Their method is outlined in Algorithm 4.1.

Algorithm 4.1: Wang and Zhao's algorithm.

Description: Solves $(A - \lambda B)x = 0$ where $A, B \in \mathbb{R}^{n\times n}$ are positive definite.

1 Compute Cholesky decompositions $A = L_A L_A^T$ and $B = L_B L_B^T$.
2 Compute the QR factorization $[L_A\ L_B]^T = QR$.
3 Define $Q_1 = [I_n\ 0_{n\times n}]Q$ and $Q_2 = [0_{n\times n}\ I_n]Q$.
4 Compute the singular values $\sigma_1(Q_1) \geq \sigma_2(Q_1) \geq \cdots \geq \sigma_n(Q_1)$ of Q_1.
5 Compute the singular values $\sigma_1(Q_2) \geq \sigma_2(Q_2) \geq \cdots \geq \sigma_n(Q_2)$ of Q_2 and a corresponding matrix V of right singular vectors.
6 Compute eigenvalues: $\lambda_i = \sigma_i(Q_1)/\sigma_{n-i+1}(Q_2)$ for $i = 1\colon n$.
7 Compute eigenvectors: $x_i = R^{-1}(Ve_i)$ for $i = 1\colon n$.

To see why Algorithm 4.1 works, we note that Q on line 2 has a CS decomposition

$$Q = \begin{bmatrix} U_1 & \\ & U_2 \end{bmatrix} \begin{bmatrix} C \\ S \end{bmatrix} V^T,$$

where $Q_1 = U_1 C V^T$ and $Q_2 = U_2 S V^T$ are singular value decompositions (SVDs). Since each column of Q has unit norm, so must be the case for each column of $[C\ S]^T$. In other words, it must hold that $c_{ii}^2 + s_{ii}^2 = 1$ for $i = 1\colon n$. If we define $X = R^{-1}V$, then we have

$$X^T A X = V^T R^{-T} L_A L_A^T R^{-1} V = V^T Q_1^T Q_1 V = C^2 \tag{4.13}$$

and similarly

$$X^T B X = S^2. \tag{4.14}$$

Now consider the case when A or B (possibly both) are singular but still positive semidefinite, and the pencil $A - \lambda B$ is regular. For such problems, Algorithm 4.1 still works after a small modification: instead of computing Cholesky factorizations on line 1, we compute any other factorizations such that

$$A = L_A L_A^T \quad \text{and} \quad B = L_B L_B^T,$$

where L_A and L_B need not be triangular. If A, say, is singular we can, for example, use the factorization given by $L_A = U\Lambda^{1/2}$ where $A = U\Lambda U^T$ is a spectral decomposition. Regarding the eigenvectors, we have $Rx = 0$ only if $(A - \lambda B)x = 0$ independent of λ. Hence, the assumption that $A - \lambda B$ is regular implies that R is invertible.

Wang and Zhao showed that if no rounding errors occur on line 6 of Algorithm 4.1, then the algorithm finds the exact eigenvalues of a perturbed problem

$$(A + \Delta A)x = \lambda(B + \Delta B)x, \tag{4.15}$$

where ΔA and ΔB are *symmetric* and $\|\Delta A\|/(\|A\| + \|B\|)$ and $\|\Delta B\|/(\|A\| + \|B\|)$ are both small. Here (and below) "small" refers to a modest multiple of machine precision that depends on n. Put differently, they showed that the pairs $(\sigma_{n-i+1}(Q_2), \sigma_i(Q_1))$, $i = 1\!:\!n$, are the exact eigenvalues in *homogeneous form* of the homogeneous pencil $\alpha(A + \Delta A) - \beta(B + \Delta B)$. Note that the same backward errors ΔA and ΔB apply for all homogeneous eigenvalues. Now, let (λ_i, x_i) be an exact eigenpair of the perturbed GEP (4.15). If we take into account the rounding errors on line 6 of Algorithm 4.1, then we have $\widetilde{\lambda}_i = \lambda_i(1 + \delta)$ where δ is real and less than the unit roundoff (in modulus). We get

$$(A + \Delta A)x_i = \lambda_i(B + \Delta B)x_i = \widetilde{\lambda}_i(B + \Delta B)(1 + \delta)^{-1}x_i = \widetilde{\lambda}_i(B + \Delta B_i)x_i,$$

where ΔB_i is symmetric and $\|\Delta B_i\|/(\|A\|+\|B\|)$ is small. Note, that the backward error ΔB_i depends on λ_i, so in contrast to the homogeneous case, we no longer have one pair of backward errors, $(\Delta A, \Delta B)$, for all computed eigenvalues.

The error analysis in [89] is oblivious to which factorizations are being done on line 1 of Algorithm 4.1 as long as $L_A L_A^T = A + \Delta\widetilde{A}$ and $L_B L_B^T = B + \Delta\widetilde{B}$, where $\|\Delta\widetilde{A}\|/\|A\|$ and $\|\Delta\widetilde{B}\|/\|B\|$ are both small. Therefore the same backward error result holds if we compute these factorizations from the spectral decomposition, which can be computed stably using e.g., the QR algorithm [80], [33, p. 464].

The above backward error result does not say anything about the magnitude of $\|\Delta A\|/\|A\|$ and $\|\Delta B\|/\|B\|$, and is hence not satisfactory with respect to the backward error defined in (4.5). Fortunately, this can be fixed by an eigenvalue parameter scaling. If we use Algorithm 4.1 (possibly with our modification to handle singular matrices) to solve $Ax = \lambda(sB)x$, with $s = \|A\|/\|B\|$, rather than (4.12), then we get computed eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ such that the backward errors $\eta_{A-(sB)\lambda}(\lambda_i)$, $i = 1\!:\!n$ are small. From (4.8), we see that $s\lambda_1, s\lambda_2, \ldots, s\lambda_n$ have small backward errors as eigenvalue approximations to (4.12). Taken together, our modifications leads to a new algorithm which is summarized in Algorithm 4.2; the corresponding flop count is shown in Table 17.[1]

[1]Wang and Zhao pointed out that the cost of the QR factorization can be reduced if we can take advantage of the triangular structure of L_A and L_B (assuming A and B are nonsingular). For simplicity, this will not be exploited in this thesis.

Algorithm 4.2: Modified Wang-Zhao algorithm.

Description: Solves $(A - \lambda B)x = 0$ where $A, B \in \mathbb{R}^{n\times n}$ are positive semidefinite and $A - \lambda B$ is regular.

if A is nonsingular **then**
    Compute Cholesky factorizations $A = L_A L_A^T$.
else
    Compute a spectral decomposition $A = U_A \Lambda_A U_A^T$ and set $L_A = U_A \Lambda_A^{1/2}$.
end
if B is nonsingular **then**
    Compute Cholesky factorizations $B = L_B L_B^T$.
else
    Compute a spectral decomposition $B = U_B \Lambda_B U_B^T$ and set $L_B = U_B \Lambda_B^{1/2}$.
end
Let $s = \|A\|/\|B\|$ (If $\|A\|$ or $\|B\|$ are unknown, estimations suffice).
Compute the QR factorization $[L_A \ \sqrt{s} L_B]^T = QR$.
Define $Q_1 = [I_n \ 0_{n\times n}]Q$ and $Q_2 = [0_{n\times n} \ I_n]Q$.
Compute the singular values $\sigma_1(Q_1) \geq \sigma_2(Q_1) \geq \cdots \geq \sigma_n(Q_1)$ of Q_1.
Compute the singular values of $\sigma_1(Q_2) \geq \sigma_2(Q_2) \geq \cdots \geq \sigma_n(Q_2)$ of Q_2 and a corresponding matrix V of right singular vectors.
Compute eigenvalues $\lambda_i = s\sigma_i(Q_1)/\sigma_{n-i+1}(Q_2)$ for $i = 1\colon n$.
Compute eigenvectors $x_i = R^{-1}(Ve_i)$ for $i = 1\colon n$.

Table 4.1: Flop count estimation for Algorithm 4.2.

Cholesky factorization	$(1/3)n^3$	[33, p. 164]
Symmetric QR Algorithm	$9n^3$	[33, p. 463]
Householder QR factorization ($2n \times n$)	$(12 + 2/3)n^3$	[33, p. 249]
Singular values	$(2 + 2/3)n^3$	[33, p. 493]
Singular values + right singular vectors	$12n^3$	[33, p. 493]
Triangular linear system	n^2	[33, p. 107]
Alg. 4.2: A and B nonsingular	$29n^3$	
Alg. 4.2: A xor B singular	$(37 + 2/3)n^3$	
Alg. 4.2: A and B singular	$(46 + 1/3)n^3$	

We remark that the two if-then-else statements in Algorithm 4.2 can be executed in parallel. Similarly, the computation of the SVD quantities of Q_1 and Q_2 (line 14 and 15) can be done in parallel.

The backward error analysis in [89] only concerns the eigenvalues. Since the eigenvectors are given by $R^{-1}V$ the quality of the computed eigenvectors depend on the triangular matrix R. As mentioned above, this matrix is always invertible, but it may be ill-conditioned. In exact arithmetic we have $R^TR = A + sB$, with $s = \|A\|/\|B\|$, so R is ill-conditioned exactly when there exists a vector v such that both $v^TAv/\|A\|$ and $v^TBv/\|B\|$ are small.

4.4 Main algorithm

The proposed algorithm for solving (4.1) is outlined briefly in Algorithm 4.3.

Algorithm 4.3: Main algorithm

Description: Computes all eigenvalues/eigenpairs of (4.1).

1 Compute an $S \in \mathbb{R}^{n\times r}$ such that $D = SS^T$.

2 Compute the *undamped eigenvalues* (that is, the eigenvalues of $M\lambda^2 + K$) and a nonsingular $X \in \mathbb{R}^{n\times n}$ such that

$$X^T(M\lambda^2 + SS^T\lambda + K)X = M_d\lambda^2 + \widehat{S}\widehat{S}^T\lambda + K_d =: P(\lambda), \tag{4.16}$$

where M_d and K_d are diagonal.

3 Lock undamped eigenvalues that are also eigenvalues of (4.1).

4 Compute the eigenvalues of (4.16) by the Ehrlich-Aberth iteration. Return the computed eigenvalues if the eigenvectors are not requested.

5 Compute the eigenvectors of (4.16) by inverse iteration.

6 Return (λ_i, Xv_i), $i = 1\!:\!2n$, where (λ_i, v_i) is a computed eigenpair of (4.16).

For the first step of Algorithm 4.3, we can find S by, for instance, computing the spectral decomposition of D.

The second step of Algorithm 4.3 essentially reduces to solving a definite GEP. It is easy to see that X must be an eigenvector matrix corresponding to $K - M\omega$. Furthermore, if ω_k, $k = 1\!:\!n$, are the eigenvalues of $K - M\omega$, then the undamped eigenvalues are given by $\pm i\sqrt{\omega_k}$ if ω_k is finite, and ∞ otherwise. We use Algorithm 4.2 to find all eigenpairs of $K - M\omega$. Note that there is no need to

form the matrices X^TMX and X^TKX explicitly: from (4.13) and (4.14) we see that M_d and K_d are given by the singular values computed in Algorithm 4.2.

The description of step 3, 4 and 5 are more involved, so we discuss these in separate subsections.

4.4.1 Step 3: Initial locking

It may happen that some undamped eigenvalues are also eigenvalues to (4.1). Since there is no need to do any further work on such eigenvalues, we declare them "locked." When deciding which eigenvalues to lock, we treat zero and infinite eigenvalues separately from nonzero finite eigenvalues. The reason for this is that it is not unlikely that zero and infinite eigenvalues of (4.1) are defective. Matrix polynomials (and constant matrices for that matter) with defective eigenvalues are often regarded as degenerate cases. Indeed, if we randomly generate the coefficient matrices of a matrix polynomial, it will almost surely have no defective eigenvalues. However, we will see that this is not necessarily the case if we impose rank constraints on the coefficient matrices and force them to be positive semidefinite.

Suppose (λ_k, x_k), where $\lambda_k \neq 0, \infty$, is a computed eigenpair of $M\lambda^2 + K$ and let $\eta(\lambda_k, x_k)$ denote the corresponding backward error with respect to $M\lambda^2 + D\lambda + K$. We declare λ_k "locked" if $\eta(\lambda_k, x_k)$ is small enough. In our code, "small enough" is defined as less than nu where u is machine precision.

The next proposition provides a method to determine how many of the zero and infinite eigenvalues to lock.

Proposition 4.4.1. *Let $Q(\lambda) = M\lambda^2 + D\lambda + K$ be the matrix polynomial in* (4.1)*, with (M, K) regular. The number of zero eigenvalues is given by*

$$\dim \operatorname{null}(K) + \dim(\operatorname{null}(D) \cap \operatorname{null}(K)),$$

and the number of infinite eigenvalues is given by

$$\dim \operatorname{null}(M) + \dim(\operatorname{null}(D) \cap \operatorname{null}(M)).$$

Proof. For readability, we introduce the following variables

$$k := \dim \operatorname{null}(K) \quad \text{and} \quad \ell := \dim(\operatorname{null}(D) \cap \operatorname{null}(K)).$$

Pick a real invertible X_1 such that $X_1^TMX_1$ and $X_1^TKX_1$ are diagonal and the first k diagonal elements of $X_1^TKX_1$ are zero. This can be done since (M, K)

is a definite pencil. Further, pick an invertible $X_2 \in \mathbb{R}^{k\times k}$ such that the first $\ell \leq k$ columns $X_2 \oplus I_{n-k}$ is a basis for $\operatorname{null}(X_1^T D X_1) \cap \operatorname{null}(X_1^T K X_1)$ and the first k columns is a basis for $\operatorname{null}(X_1^T K X_1)$. Let $X = X_1(X_2 \oplus I_{n-k})$ and note that $X^T Q(\lambda) X$ decomposes into a direct sum

$$X^T Q(\lambda) X = M_1\lambda^2 \oplus (M_2\lambda^2 + D_2\lambda + K_2),$$

where M_1 is $\ell \times \ell$ and $\operatorname{null}(D_2) \cap \operatorname{null}(K_2) = \{0\}$. Note that $M_1\lambda^2$ has exactly 2ℓ zero eigenvalues and $Q_2(\lambda) := M_2\lambda^2 + D_2\lambda + K_2$ has at least $k-\ell$ zero eigenvalues. We need to show that $Q_2(\lambda)$ has exactly $k-\ell$ zero eigenvalues, or equivalently, that all its zero eigenvalues are semisimple. To this end, we observe that $Q_2(\lambda)$ is real and symmetric, so all right eigenvectors associated with zero are also left eigenvectors. Next, we pick $\sigma > 0$ such that $\det Q_2(\sigma) \neq 0$ and define

$$\widehat{Q}(\lambda) = \lambda^2 Q_2(1/\lambda + \sigma).$$

From Corollary 2.3.3 it follows that zero is a defective eigenvalue of $Q_2(\lambda)$ only if there exists a real right eigenvector x such that

$$x^T \widehat{Q}'(-1/\sigma)x = x^T(D_2 + K_2/\sigma)x = 0.$$

Since D_2 and K_2 are both positive semidefinite, such x must lie in $\operatorname{null}(D_2) \cap \operatorname{null}(K_2)$ and hence cannot exist.

The number of infinite eigenvalues equals the number of zero eigenvalues of $\operatorname{rev} Q(\lambda) := K\lambda^2 + D\lambda + M$. Thus, the other half of the proposition can be shown analogously if we consider $\operatorname{rev} Q(\lambda)$ instead of $Q(\lambda)$. ∎

Remark 4.4.2. The number of "missing eigenvectors" corresponding to the zero and infinite eigenvalues are given by $\dim(\operatorname{null}(D) \cap \operatorname{null}(K))$ and $\dim(\operatorname{null}(D) \cap \operatorname{null}(M))$, respectively. Hence, defective eigenvalues are always present if, for example, $\operatorname{rank} D = 1$ and $\dim \operatorname{null}(K) = 2$.

By Proposition 4.4.1, the number of zero and infinite eigenvalues depends on $\operatorname{null}(K)$ and $\operatorname{null}(M)$, respectively. These spaces are available from the corresponding spectral decompositions, which are computed in Algorithm 4.2. If the columns of $N_1 \in \mathbb{R}^{n\times k_1}$ and $N_2 \in \mathbb{R}^{n\times k_2}$ constitute bases for $\operatorname{null}(K)$ and $\operatorname{null}(M)$, respectively, then there are $2k_1 - \operatorname{rank}(DN_1)$ zero eigenvalues and $2k_2 - \operatorname{rank}(DN_2)$ infinite eigenvalues. These quantities can be computed numerically using the SVD.

4.4.2 Step 4: Computing eigenvalues

We now discuss how to use the Ehrlich-Aberth method to exploit the structure of (4.16) in order to find all eigenvalues. We focus on the following three questions.

1. How do we pick the starting points?

2. How do we compute (4.11) efficiently?

3. Which stopping criterion should we use?

For starting points we use the undamped eigenvalues with small (in a relative sense) random perturbations added to the unlocked eigenvalues. These perturbations are added to break symmetries, since it is well-known that the Ehrlich-Aberth method may fail to converge due to certain symmetries [3]. Suppose, for example, that (4.1) has two real simple eigenvalues and all undamped eigenvalues are finite and nonzero. Assume further that we want to use the update rule (4.9). If we do not add the perturbations, then starting points can be paired into complex conjugates, and the update rule (4.9) preserves this symmetry. Hence, convergence to real simple eigenvalues is impossible. Another problem occurs if two starting points are the same. In this case we get division by zero in the Ehrlich-Aberth updates. Also this problem disappears when we add random perturbations to the starting points.

The rationale behind using the undamped eigenvalues as starting points becomes more clear if we think about the application discussed in section 4.1.1. In this case the eigenvalues correspond to vibrational properties (frequency and damping) of a physical structure, and the undamped eigenvalues correspond to vibrational properties of the *same structure*, but with the strength of the dampers set to zero. If the damping is small or moderate, our choice of starting points seems reasonable. But what if the damping is strong? In this case we note that a strong viscous damper (that is, one with small holes in its piston head, see Figure 3.1) is in some sense similar to a rigid component. We expect the spectrum to respect this similarity. In the case when M and K are positive definite, we saw in Chapter 3, that all eigenvalues of $M\lambda^2 + sD\lambda + K$, converge to points on the imaginary axis as $s \to \infty$, with the exception of $\operatorname{rank} D$ eigenvalues which go to $-\infty$. This means that $\operatorname{rank} D$ eigenvalues can be arbitrarily far from all the staring points. Fortunately, as will be seen in section 4.5.3, Ehrlich-Aberth works quite well in practice when only a few starting points are "way off."

The computation of trace $(P(\lambda)^{-1}P'(\lambda))$ for fix values of λ is the core of our Ehrlich-Aberth iteration. We compute this trace using the Sherman-Morrison-Woodbury formula in combination with standard trace laws. The precise procedure is outlined in Algorithm 4.4; we leave out the tedius algebra that justifies it. If M_d and K_d are stored as vectors and Algorithm 4.4 is implemented accordingly, then total flop count is only $4n + 2rn + 4r^2n$ (counting only terms with a factor n). Since there are $2n$ eigenvalues, and we expect each eigenvalue to converge in a few steps, the complexity in n of our Ehrlich-Aberth iteration is quadratic.

Algorithm 4.4: Computation of trace $(P(\lambda)^{-1}P'(\lambda))$.

Description: Computes $t = \text{trace}\,(P(\lambda)^{-1}P'(\lambda))$ where
$$P(\lambda) = M_d\lambda^2 + \widehat{S}\widehat{S}^T\lambda + K_d.$$

1 $A := M_d\lambda^2 + K_d$
2 $B := A^{-1}\widehat{S}$
3 $C := \widehat{S}^T B$
4 $D := I_r + \lambda C$
5 $E := M_d B$
6 $F := CD^{-1}C$
7 $G := (B^T E)D^{-1}$
8 $t := 2\lambda\text{trace}(M_d A^{-1}) + \text{trace}(C) - 2\lambda^2\text{trace}(G) - \lambda\text{trace}(F)$

When an eigenvalue has converged, we mark it as "locked" and do not update it in subsequent iterations. We are done when all eigenvalues are locked. The obvious question is "When do we declare an eigenvalue 'converged'?" One approach is to estimate the backward error (4.7) with respect to (4.16), and lock computed eigenvalues if their backward errors are smaller than some tolerance, say machine precision. If we use the `normest1` algorithm [38] in combination with the Sherman-Morrison-Woodbury formula, such estimation requires only $O(n)$ flops if we count r as a small constant. We found, however, that we often get better results (both in terms of accuracy and speed) with the following heuristic strategy: lock $\lambda_k^{(i)}$ when

$$\left|\lambda_k^{(i)} - \lambda_k^{(i+1)}\right| < \text{tol} \times \left|\lambda_k^{(i)}\right|,$$

where tol is initially set to be machine precision, and is then relaxed by a factor 10 each 50th iteration. This is the stopping condition used in our numerical experiments. Here the number 50 is somewhat arbitrary. From experience, convergence of most eigenvalues is usually obtained within 10 iterations. Some eigenvalues

requires more iterations, but the idea is that if 50 is not enough then the problem is most likely not the number of iterations, but rather that the tolerance is too stringent. We stress that the argument is based purely on experience, so there may very well exist examples where this strategy fails. We remark, however, that some kind of relaxation strategy for the tolerance is necessary also when the eigenvalue backward error is used as a stopping condition—otherwise the iteration may go on forever. We comment more on this at the end section 4.5.3.

4.4.3 Step 5: Computing eigenvectors

When all eigenvalues have been found we compute the corresponding eigenvectors. Since the computation of eigenvectors corresponding to different eigenvalues are completely decoupled, this phase of the algorithm is embarrassingly parallel. We now discuss how to determine an eigenvector v_i of $P(\lambda)$ corresponding to a computed eigenvalue λ_i. If λ_i is an undamped eigenvalue, then v_i has already been found; otherwise, more work is required. The next proposition provides one method for computing v_i.

Proposition 4.4.3. *Let λ_i be an eigenvalue of $P(\lambda)$ but not of $Q(\lambda) := P(\lambda) - SS^T\lambda$. Then all eigenvectors associated with λ_i lie in the range of $Q(\lambda_i)^{-1}S$.*

Proof. Suppose (v_i, λ_i) is an eigenpair of $P(\lambda)$ and write $v_i = Q(\lambda_i)^{-1}Sx + y$ where $y \perp \text{range}(Q(\lambda_i)^{-1}S)$. We need to show that $y = 0$. We have

$$\begin{aligned} 0 = P(\lambda_i)v_i &= P(\lambda_i)(Q(\lambda_i)^{-1}Sx + y) \\ &= (Q(\lambda_i) + \lambda_i SS^T)(Q(\lambda_i)^{-1}Sx + y) \\ &= S(I_r + \lambda_i S^T Q(\lambda_i)^{-1}S)x + Q(\lambda_i)y + \lambda_i SS^T y, \end{aligned}$$

which implies that $Q(\lambda_i)y \in \text{range}(S)$, or equivalently, that $y \in \text{range}(Q(\lambda_i)^{-1}S)$. The result now follows from the definition of y. ∎

Remark 4.4.4. A consequence of Proposition 4.4.3 is that the geometric multiplicity of λ_i cannot exceed the rank of S.

Proposition 4.4.3 implies that if λ_i is computed exactly, then it is enough to look for eigenvectors in the r dimensional subspace $\text{range}(Q(\lambda_i)^{-1}S)$. Furthermore, we see from the proof that $Q(\lambda_i)^{-1}Sx$ is an eigenvector of $P(\lambda)$ for any $x \in \text{null}(I_r + \lambda_i S^T Q(\lambda_i)^{-1}S)$. Since x can be computed cheaply from the SVD of $I_r + \lambda_i S^T Q(\lambda_i)S$, this yields a fast method for computing v_i. In practice, however,

the computed eigenvalues contain errors so Proposition 4.4.3 is strictly speaking not applicable, and the discussed method may lead to inaccurate eigenvectors. The computed eigenvectors are, however, often very good (frequently with perfect backward errors of order 10^{-16}) and serve as excellent starting vectors for inverse iteration.

There are several approaches to inverse iteration for polynomial eigenproblems, see [64, 65]. The approach we take is (to the author's knowledge) new. It is designed for real symmetric matrix polynomials and is slightly cheaper than the alternatives—although it may be argued that the savings are negligible in our context. The idea is to iterate according to

$$v_i^{(k+1)} = P(\lambda_i)^{-1}\overline{v}_i^{(k)}/\|v_i^{(k)}\|. \tag{4.17}$$

So, why does this work? To answer this, we note that $P(\lambda_i)$ is complex symmetric and hence enjoys an SVD on the form $\overline{U}\Sigma U^H$ (also known as the Takagi factorization). If $U = [u_1 \ u_2 \ \cdots \ u_n]$, $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$ and $v_i^{(k)} = \alpha_1 u_1 + \alpha_2 u_2 + \cdots + \alpha_n u_n$, then

$$P(\lambda_i)^{-1}\overline{v}_i^{(k)} = U\Sigma^{-1}\overline{U}^H\overline{v}_i^{(k)} = \sum_{j=1}^{n}\frac{\alpha_j}{\sigma_j}u_j. \tag{4.18}$$

Since σ_n is tiny when λ_i is close to an eigenvalue, we expect (4.18) to be huge in the direction of u_n. This is delightful, since the vector u_n is the best possible eigenvector approximation we can hope for in the sense that $\eta_P(\lambda_i, u_n) = \eta_P(\lambda_i)$. As usual with inverse iteration, the ill-conditioning of $P(\lambda_i)$ is benign since the matrix magnifies errors in the direction of the desired vector.

To compute (4.17) we use the Sherman-Morrison-Woodbury formula with the starting vector described above. Since the starting vector already is a good approximation, we only take one step of inverse iteration in our code. The complexity for computing one eigenvector of (4.16) with this technique is linear in n.

Another way to solve the linear systems from (4.17) is to use QR factorization and back substitution. This is an attractive option from a stability point of view, albeit more expensive. If the technique used in [82] is employed, each QR factorization can be computed with $O(rn^2)$ flops. The idea is to compute, in a bottom-up fashion, a sequence of (real) Givens rotations $U_1U_2\ldots U_{n-1} =: U$ such that US is trapezoidal and UM_d and UK_d are r-Hessenberg. This implies that $UP(\lambda_i)$ is r-Hessenberg for any λ_i, so its QR factorization can be computed efficiently using Givens rotations. We did not use this approach for our numerical

experiments. It may, however, be the method of choice when only a few eigenvectors are sought, or when the eigenvectors are computed in parallel.

We end this section with a negative remark: the approach to first find the eigenvalues, and then the eigenvectors via inverse iteration, is flawed when multiple eigenvalues are present. In this case we may approximate the same eigenvector several times. An obvious "solution" is to compute an invariant subspace rather than individual eigenvectors; inverse iteration and our choice of starting vector can indeed be generalized to subspaces. The problem is that it is hard to a priori decide what the dimension of the subspaces should be.

4.5 Numerical experiments

We implemented Algorithm 4.3 in MATLAB 2012b. Our code is written in serial, so it does not, for instance, exploit that the workload in steps 4 and 5 of the algorithm is embarrassingly parallel. Individual MATLAB functions that are being called, may, however, be multithreaded. For the Ehrlich-Aberth iteration, we used the Gauss-Seidel updates shown in (4.10). The first part of our algorithm (step 1–3) make use of MATLAB's core routines `svd` and `qr`. The second part of our algorithm (step 4–6) is written in "pure" MATLAB code (except for the computation of small $r \times r$ SVDs) and is sometimes slower than the first part even though the flop count is much lower. Since we expect this speed difference to wane if the entire algorithm is implemented in a low-level language, we sometimes state explicitly how much time is spend on the second part.

We compared our algorithm to `quadeig`, the MATLAB implementation of the eigensolver in [34] for unstructured QEPs. In the core of this implementation we find MATLAB's `eig` routine, which performs the real QZ algorithm in this case.

The numerical experiments were carried out in MATLAB R2012b in IEEE double precision arithmetic on a machine with the following specifications.

Memory	16GB (4X4GB) 1333 MHz DDR3 Non-ECC
Processor	Intel® Core™ i7-2600 (8M Cache, 3.40 GHz)
Operating System	Windows® 7 Professional (64Bit)

4.5.1 The damped beam

The damped beam from the collection of nonlinear eigenvalue problems, NLEVP [10], was studied earlier in Example 3.3.6 and Example 3.4.4. The construction of the coefficient matrices is explained in [37], where it is also shown that half of the

eigenvalues are undamped. This makes it an ideal problem for our algorithm. We modeled the problem such that the coefficient matrices were of size 1000×1000. Algorithm 4.3 computed all eigenpairs in 2 seconds while `quadeig` needed 112 seconds. The backward errors for the computed eigenpairs are shown in Figure 4.1. We remark that there is no guarantee that two backward errors plotted with the same x-coordinate correspond to the same eigenvalue. We see that both algorithm performed well in terms of backward stability (all backward errors are less than n times the machine precision). The spectrum, as it was computed by the two algorithms, are shown in Figure 4.2.

Let us pause a for a while and discuss Figure 4.2. We know that all eigenvalues must lie in the left half plane, and half of them on the imaginary axis. Hence the real parts of some of the computed eigenvalues from `quadeig` must be inaccurate, even though Figure 4.1 shows all backward errors are about 10^{-14}. In terms of relative errors, this is consistent with the "approximate bound"

$$\text{forward error} \lesssim \text{backward error} \times \text{condition number},$$

for the *unstructured* forward error, if we define the condition number conformably with the backward error introduced in section 4.2.1. This condition number is

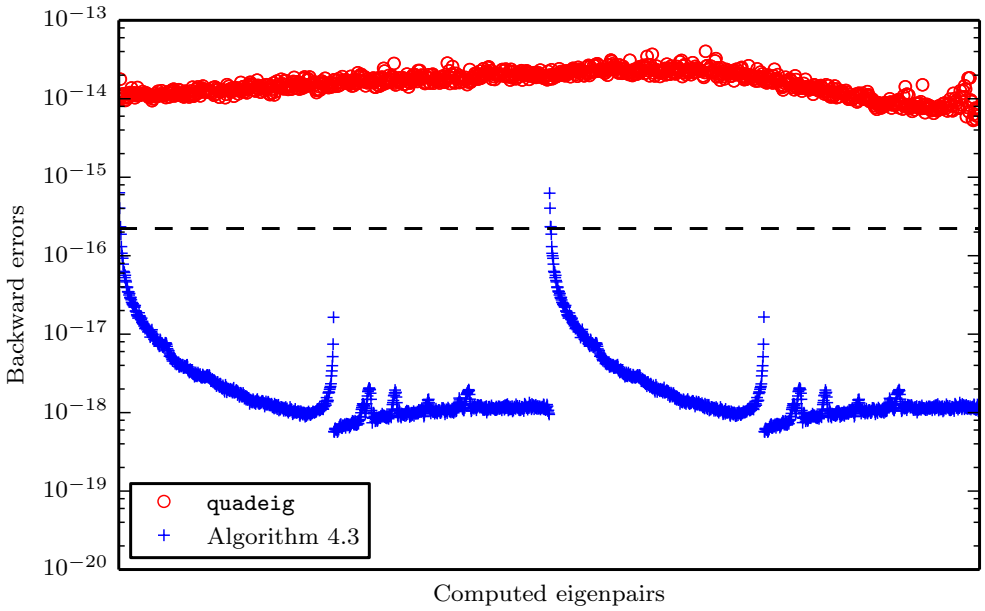

Figure 4.1: Backward errors of computed eigenpairs for the damped beam. The dashed line indicates the machine precision.

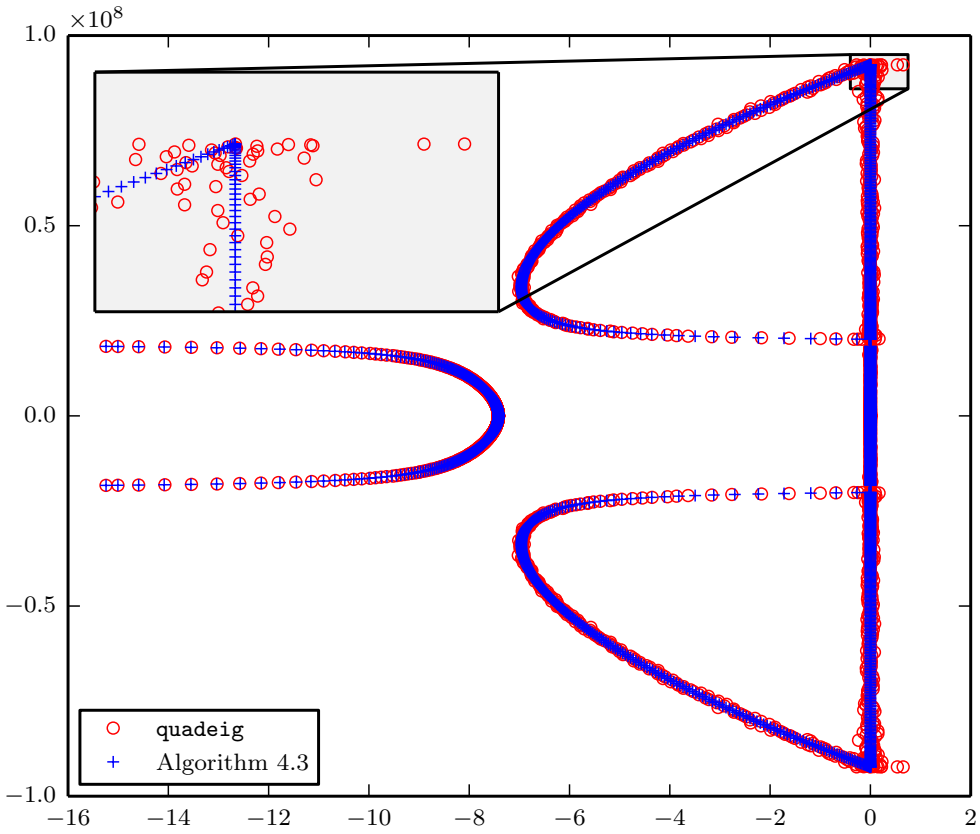

Figure 4.2: Computed spectra of the damped beam.

given by

$$\kappa(\lambda) = \frac{\|M\|\|\lambda|^2 + \|D\|\|\lambda| + \|K\|}{|\lambda||v^T(2M\lambda + D)v|},$$

if λ is a simple nonzero eigenvalue of (4.1) and v is an associated normalized eigenvector [81]. If we, for example, evaluate the condition number of the upper-right-most eigenvalue using the computed quantities from Algorithm 4.3, we find that the condition number is of order 10^7. Assuming this answer is of the correct order of magnitude, the *relative* forward error is at most of order $10^{-14} \times 10^7 = 10^{-7}$. For the *absolute* forward error, we note that the modulus of the eigenvalue in question is about 10^8, so the absolute forward error is at most of order $10^{-14} \times 10^7 \times 10^8 = 10$. This explains why we see some red circles in the right half plane. The unstructured forward error bound does not, however, explain the nice pattern in the spectrum produced by Algorithm 4.3. One explanation is the problem has a lot of structure that Algorithm 4.3 respects.

Finally, recall that each undamped eigenvalue is computed as $\pm i\omega_k$, for some real eigenvalue ω_k of $K - M\omega$, and is then locked if it is also an eigenvalue of the damped problem. This explains the straight line of blue crosses on the imaginary

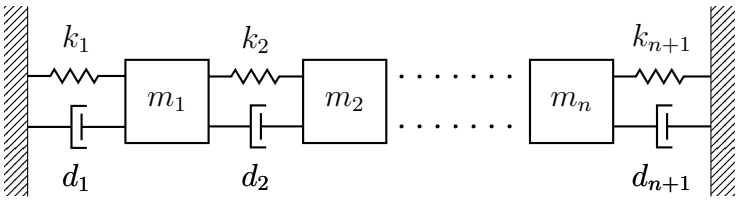

Figure 4.3: A simple mass-spring-damper system.

axis in Figure 4.2. However, even if we bypass the initial locking phase and add relative perturbations of order 10^{-8} to *all* eigenvalues, our Ehrlich-Aberth iteration returns half the eigenvalues in a strip centered at the imaginary axis of width about 10^{-13}.

4.5.2 A mass-spring-damper system

Our next QEP comes from a simple mass-spring-damper system; the particular setup is shown in Figure 4.3. To make the problem more interesting, we introduced defective infinite eigenvalues by setting some of the masses, as well as most damping coefficients, to zero. We defined $n = 1000$,

$$m_i = \begin{cases} 0 & \text{if } i \in \{1, n\} \\ 1 & \text{otherwise,} \end{cases} \qquad d_i = \begin{cases} 1/100 & \text{if } i \in \mathcal{J} := \{12, n/2+1, n-10\} \\ 0 & \text{otherwise} \end{cases}$$

and $k_i = 1$ for $i = 1\!:\!n$. Notice that there only are three effective dampers, that is, with nonzero damping coefficients d_i. The corresponding mass, damping and stiffness matrices are given by

$$M = I - e_1 e_1^T - e_n e_n^T, \quad D = \frac{1}{100} \sum_{i \in \mathcal{J}} (e_{i-1} - e_i)(e_{i-1} - e_i)^T$$

and

$$K = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{bmatrix},$$

respectively [88, p. 2]. Further, Proposition 4.4.1 implies that the associated QEP has four defective infinite eigenvalues. We solved the eigenproblem using the new algorithm and `quadeig`. Although this QEP has the same size as the damped

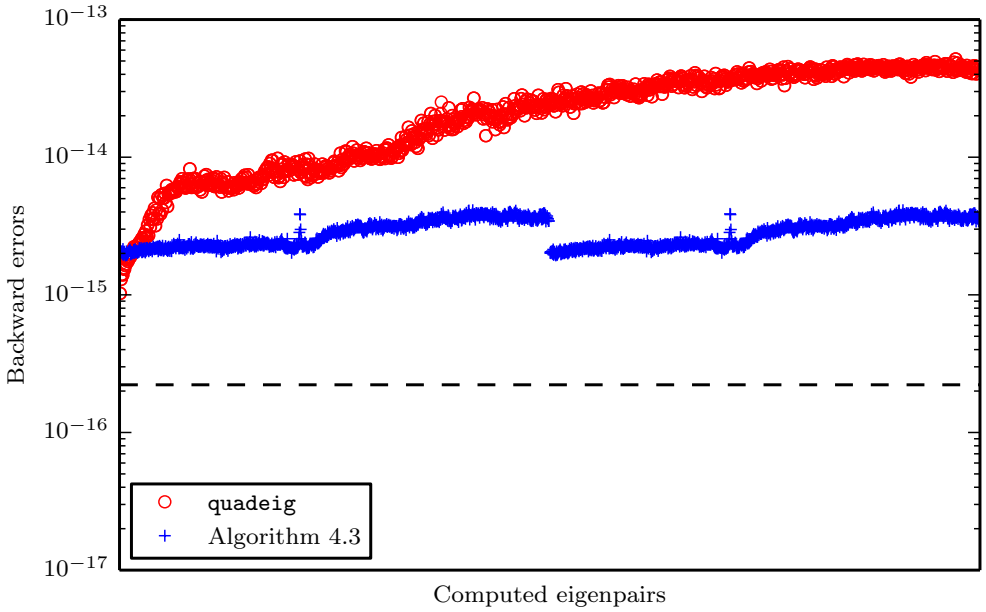

Figure 4.4: Backward errors of computed eigenpairs for the mass-spring-damper system described in section 4.5.2. The dashed line indicates the machine precision.

beam problem in Section 4.5.1, we expect the computation time for Algorithm 4.3 to be longer for this problem. There are two reasons for this. First, only four eigenvalues are locked initially (that is, the infinite eigenvalues), in contrast to *half* the eigenvalues of the damped beam problem. Second, the damping matrix of this problem has larger rank. The computation time for Algorithm 4.3 was 5 seconds, where more than half the time was spend on the second part (step 4–6) of the algorithm; the computation time for `quadeig` was 110 seconds. The backward errors for the computed eigenpairs are shown in Figure 4.4. As in Figure 4.1, two errors plotted with the same x-coordinate may correspond to different eigenvalues.

Once the eigenvalues have been computed, Algorithm 4.3 computes the eigenvectors at a negligible cost. As mentioned in the introduction of this chapter, this is not the case for `quadeig`, which is significantly faster if only the eigenvalues are sought. Therefore, we reran this experiment with the modification that only the eigenvalues were computed. This time `quadeig` required 74 seconds, while Algorithm 4.3 still used 5 seconds.

4.5.3 QEPs with random coefficient matrices

We created random coefficient matrices using the MATLAB commands

```
M = randn(n);  D = randn(n,5);  K = randn(n);
M = M*M'; D = D*D'; K = K*K';
```

Table 4.2: Backward errors and execution times for the tested algorithms. The last columns shows how many times, on average, each eigenvalue approximation was updated in the Ehrlich-Aberth iteration.

n	`quadeig`		Algorithm 4.3		
	max $\eta(\lambda, v)$	time	max $\eta(\lambda, v)$	time (step 4–6)	Av. #upd
200	6.3e−15	0.4	1.7e−15	0.6 (0.6)	8.2
600	1.6e−14	20.4	1.3e−15	3.9 (3.4)	7.9
1000	2.9e−14	110.3	1.3e−15	8.7 (7.2)	7.9
1400	3.7e−14	313.5	2.1e−15	15.8 (12.2)	7.9
1800	5.5e−14	702.2	1.7e−15	43.2 (34.4)	7.7
2200	6.3e−14	1296.6	1.7e−15	42.0 (26.1)	7.7
2600	7.0e−14	2143.2	1.8e−15	58.7 (34.5)	7.7
3000	7.8e−14	3299.6	1.9e−15	84.1 (44.3)	7.7

and solved the corresponding problem for different values of n. Note that the rank of the damping matrix is 5 for each test problem. The results are shown in Table 4.2. As expected, Algorithm 4.3 scales much better with n than `quadeig`.

Our next experiment concerns strongly damped problems, or more precisely, problems for which $\|D\|$ is much larger than $\|M\|$ and $\|K\|$. Such problems have badly scaled linearizations, even if parameter scalings are employed [24]. This implies that linearization based algorithms, such as `quadeig`, cannot compute all eigenpairs backward stably, unless the same problem is solved twice using two different linearizations [92]. The proposed algorithm is "linearization free" and does not share this drawback. However, it is still worth investigating the performance on strongly damped problems. There are two reasons for this:

1. There are rank D eigenvalues that are far away from all starting points.

2. As seen in Chapter 3, eigenvalues may cluster around the origin.

We generated test problems using the following MATLAB code:

```
M = randn(250); D = randn(250,r); K = randn(250);
M = M*M'; D = s*(D*D'); K = K*K';
```

The results for different values of s and r are shown in Table 4.3. We see that the norm of D does affect the accuracy. However, the increase in the worst case

Table 4.3: Backward errors and execution time for Algorithm 4.3. As in Table 4.2, Av. #upd denotes the number of average Ehrlich-Aberth updates per eigenvalue.

s	$r = 5$		$r = 25$	
	max $\eta(\lambda, v)$	Av. #upd	max $\eta(\lambda, v)$	Av. #upd
1e+00	1.1e−15	7.9	1.5e−15	17.0
1e+02	3.0e−15	6.9	1.4e−14	12.8
1e+04	1.8e−14	7.3	8.4e−13	17.0
1e+06	4.6e−14	7.2	8.2e−13	20.9
1e+08	3.7e−14	7.0	6.2e−13	26.0
1e+10	4.3e−14	7.8	1.4e−12	29.8
1e+12	2.3e−14	8.3	9.0e−13	36.1
1e+14	1.3e−14	9.0	3.7e−13	42.2

backward error is modest (a factor 10 or 100) and appears to stagnate as s grows. The results in Table 4.3 can be explained as follows. When s is large, there are $2r$ real eigenvalues; half of them cluster around the origin and the other half are large and negative. As s grows, our algorithm fails to satisfy the initial stopping condition for some eigenvalues, in particular the real ones, and therefore relaxes the tolerance (after 50 iterations). This is the reason for the growth in the worst case backward errors. It also explains the increase in average number of Ehrlich-Aberth steps taken per eigenvalue. We remark that taking more steps before relaxing the tolerance does not *necessarily* improve the accuracy. The problem is not always that the iterates are "lost" and far away from the eigenvalues they should approximate, but rather that the Ehrlich-Aberth corrections, which are computed using the Sherman-Morrison-Woodbury formula, are not accurate enough. The author has not found any example where the updates computed using the Sherman-Morrison-Woodbury formula was seriously inaccurate, but has encountered cases where the associated eigenvalue backward error stagnates before it reaches nu, while a naive update using MATLAB's backslash yields backward errors at machine precision. This is why a relaxed tolerance sometimes is needed also if the eigenvalue backward error is used as stopping condition, unless the initial tolerance is known, a priori, to be attainable. Such knowledge, however, would require a rigorous error analysis of the Sherman-Morrison-Woodbury formula as well as error bounds for the computed eigenvalue backward error estimates. This

is outside the scope of this thesis.

4.6 Discussion

4.6.1 The large scale case

Todays models of vibrating structures are often so large that it becomes unfeasible to find all eigenpairs. Even in cases when it is possible, all eigenpairs are rarely of interest. Instead subspace based methods are used to target the most important eigenvalues. When a good subspace has been found, a smaller "projected" eigenproblem needs to be formed. There are several ways of forming this smaller problem. A good approach is to project directly onto the coefficient matrices, in style of an orthogonal Rayleigh-Ritz procedure [9]. This leads to a smaller system that shares the essential structure that all coefficient matrices are positive semidefinite. More precisely, if the columns of U span a computed subspace of dimension k, then we form

$$Q(\lambda) := U^T \left(M\lambda^2 + D\lambda + K \right) U,$$

which is a matrix polynomial of size $k \times k$. The next step is to find *all* eigenpairs of $Q(\lambda)$. If k is significantly larger than the number of discrete dampers (recall that less than 10 dampers is not uncommon in practice) then we have a problem on same form as (4.1) to which the proposed algorithm can be applied.

4.6.2 Generalizations

The main idea behind this work was that the structure "diagonal plus low rank" can be exploited to quickly compute eigenvalues and eigenvectors. We considered a rather special QEP, but it is also possible to apply this idea to other types of eigenvalue problems. A major obstacle, however, is the choice of starting points for the Ehrlich-Aberth iteration. Consider, for example, a rank one modification of a standard eigenvalue problem $Ax = \lambda x$. If we already have a spectral decomposition $A = S\Lambda S^{-1}$, then for any u and v, $S^{-1}(A + vu^T)S$ is the sum of a diagonal matrix and a rank one matrix, so we can apply the techniques discussed in this chapter to quickly compute all its eigenpairs—*assuming good starting points are available.* The analogue of the starting points used in our algorithm, would be the eigenvalues of A. Unfortunately, we cannot, without further insight into the problem, argue that this choice is any good. In fact, it can be arbitrarily bad: Ackermann's

formula (for pole placement) [4] states that a rank one modification—albeit an extreme one— is enough to change the spectrum of any given nonderogatory matrix arbitrarily.

One situation where good starting points are available appears in homotopy methods. Consider, for example, the following problem: Given a vector u and spectral decomposition $A = S\Lambda S^{-1}$ such that all eigenvalues of A have negative real part, find the smallest $t \geq 0$ such that $A + tuu^H$ has a purely imaginary eigenvalue. If we define $x = S^{-1}u$ and $y^H = u^H S$, then $\Lambda + txy^H$ is similar to $A + tuu^H$ and on the form "diagonal plus rank one." Hence one way to attack the problem is to solve a sequence of eigenvalue problems

$$\Lambda + t_i xy^H, \quad i = 0, 1, 2, \ldots, \quad \text{where} \quad 0 = t_0 < t_1 < t_2 < \cdots,$$

by an Ehrlich-Aberth iteration that exploits the structure and uses the eigenvalues of the previous step as starting points.

4.6.3 Related work

A completely different approach, which applies to a larger family of QEPs with low rank damping, was recently studied by Lu, Huang, Bai and Su [55]. Their algorithm is based on a trimmed linearization of a rational eigenvalue problem that approximates the QEP of interest. This means that the bulk of the computation comes from solving a non-definite generalized eigenvalue problem of size $m \times m$, where m is larger than the matrix size n, but significantly smaller than $2n$. An advantage with this approach is that the algorithm is directly applicable to large scale problems.

Chapter

5 Triangularizing matrix polynomials

5.1 Introduction

In this chapter we consider the following problem: given a matrix polynomial $P(\lambda) \in \mathbb{F}[\lambda]^{n\times m}$, construct, when possible, a triangular or trapezoidal matrix polynomial $T(\lambda) \in \mathbb{F}[\lambda]^{n\times m}$, of the same degree and eigenstructure as $P(\lambda)$. If this is possible, then $P(\lambda)$ is said to be *triangularizable*. Recall that the eigenstructure of a matrix polynomial refers to the eigenvalues and their partial multiplicities or, equivalently, to the elementary divisors of the matrix polynomial, including those at infinity.

To make the text in this chapter more readable, we say that a matrix is triangular if all entries (i, j) such that $i > j$ are zero; this includes trapezoidal matrices. Similarly, we say that an $n \times m$ matrix is quasi-triangular if the leading $\min(n, m) \times \min(n, m)$ submatrix is quasi-triangular.

We will show that when $\mathbb{F}$ is algebraically closed, any $P(\lambda) \in \mathbb{F}[\lambda]^{n\times m}$ with $n \leq m$ is triangularizable, thereby extending an earlier but little-known result by Gohberg, Lancaster and Rodman for square monic matrix polynomials with complex coefficient matrices [32, proof of Theorem 1.7]. Over the real numbers, not all matrix polynomials are triangularizable. We will prove, however, that as long as $n \leq m$, all matrix polynomials in $\mathbb{R}[\lambda]^{n\times m}$ are *quasi-triangularizable*, that is, strongly equivalent to some quasi-triangular matrix polynomial. Our results extend in a non-trivial way some recent results by Tisseur and Zaballa for square regular quadratic matrix polynomials [84]. Our proofs concerning the

reduction to triangular and quasi-triangular forms are constructive provided that the elementary divisors (finite and at infinity) of the original matrix polynomial $P(\lambda)$ are available. Since this is the only information that is used, we are solving the following inverse problem: given a list of elementary divisors (finite and at infinity) over a field $\mathbb{F}$, determine under what conditions they are admissible by a triangular/quasi-triangular matrix polynomial in $\mathbb{F}[\lambda]^{n\times m}$ of a fixed degree, and, in case they are admissible, design a constructive procedure to obtain it. We end this chapter by discussing such inverse problems and state a conjecture for the inverse Hermitian polynomial eigenvalue problem.

The chapter is organized as follows. In Section 5.2 we discuss how the Möbius transformation will be employed. Section 5.3 concerns triangularization of matrix polynomials over algebraically closed fields and Section 5.4 treats the analogous quasi-triangularization over the real numbers. In Section 5.5 the inverse polynomial eigenvalue problems solved in the previous sections are identified and the case of Hermitian matrix polynomials is considered.

Since all derivations in this chapter are carried out over polynomial rings, the notation "(λ)" is omitted for scalar polynomials.

5.2 Application of the Möbius transformation

Recall the Möbius transformation for matrix polynomials introduced in Section 2.2. For a given matrix polynomial $P(\lambda) \in \mathbb{F}[\lambda]^{n\times m}$, $n \leq m$, we will use the following technique to prove that it is strongly equivalent to a triangular or quasi-triangular matrix polynomial of the same degree:

(i) If $P(\lambda)$ has elementary divisors at infinity, we apply a Möbius transformation to $P(\lambda)$ with Möbius function m_A such that $\mathcal{M}_A(P)$ only has finite elementary divisors and $(\mathcal{M}_{A^{-1}} \circ \mathcal{M}_A)(P) = P(\lambda)$ up to a product by a nonzero scalar. If $P(\lambda)$ has no eigenvalues at infinity then we take $A = I_2$ so that $\mathcal{M}_A(P) = P(\lambda)$.

(ii) We then show that $\mathcal{M}_A(P)$ is equivalent to a triangular or quasi-triangular matrix polynomial $T(\lambda)$.

Notice that $\mathcal{M}_{A^{-1}}(T)$ is triangular or quasi-triangular as $T(\lambda)$ is triangular or quasi-triangular, respectively. We claim that, provided that $\mathbb{F}$ is an infinite field, a Möbius function always exists that satisfies the two conditions of item (i) and $\mathcal{M}_{A^{-1}}(T)$ is strongly equivalent to $P(\lambda)$. In fact, let $a, c \in \mathbb{F}$, $c \neq 0$, such

that a/c is not an eigenvalue of $P(\lambda)$ and take $b, d \in \mathbb{F}$ such that $A = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ is nonsingular. Write $P(\lambda) = \widetilde{A}_\ell(c\lambda - a)^\ell + \widetilde{A}_{\ell-1}(c\lambda - a)^{\ell-1} + \cdots + \widetilde{A}_0$, with suitable $\widetilde{A}_0, \ldots, \widetilde{A}_\ell$. Given that a/c is not an eigenvalue of $P(\lambda)$ it follows that $\operatorname{rank} \widetilde{A}_0 = \operatorname{rank} P(a/c) = \operatorname{rank} P(\lambda)$. Also,

$$\mathcal{M}_A(P) = \widetilde{A}_0(c\lambda + d)^\ell + (bc - ad)\widetilde{A}_1(c\lambda + d)^{\ell-1} + \cdots + (bc - ad)^\ell \widetilde{A}_\ell$$

so that the leading coefficient of $\mathcal{M}_A(P)$ is $c^\ell \widetilde{A}_0$. Hence $\operatorname{rank} \operatorname{rev}(\mathcal{M}_A(P))(0) = \operatorname{rank} \widetilde{A}_0 = \operatorname{rank} P(\lambda)$ and $\mathcal{M}_A(P)$ has no eigenvalues at infinity. Now, $m_{A^{-1}}(z) = (-dz + b)/(cz - a)$ and

$$\mathcal{M}_{A^{-1}}\big(\mathcal{M}_A(P)\big) = (c\lambda - a)^\ell \mathcal{M}_A(P)(m_{A^{-1}}(\lambda)) = (bc - ad)^\ell P(\lambda).$$

This proves our claim about the existence of a Möbius function that satisfies the two conditions of item (i) for each matrix polynomial $P(\lambda)$. The second part of the claim (that $P(\lambda)$ and $\mathcal{M}_{A^{-1}}(T)$ are strongly equivalent) is an immediate consequence of Theorem 2.2.1.

5.3 Triangularization over algebraically closed fields

We start with a deflation procedure which will be used to construct upper triangular matrix polynomials with diagonal entries of a specified degree. The techniques used to prove the following two results appear for $\mathbb{F} = \mathbb{C}$ in the proof of a theorem by Gohberg, Lancaster and Rodman on the inverse problem for linearizations [32, proof of Theorem 1.7].

Lemma 5.3.1. *Let $d_1 | \cdots | d_n$ be monic polynomials with coefficients in $\mathbb{F}$ and define $\ell_j := \deg d_j$. Assume that for a given positive integer q and a pair of indices (i, j) such that $\ell_i \leq q < \ell_j$, there is a polynomial s with $\deg s < \ell_j$ such that $d_{k-1} | s d_i | d_k$ for some index $k \leq j$. Then $D(\lambda) = \operatorname{diag}(d_1, \ldots, d_n)$ is equivalent to $\widetilde{D}(\lambda) + d_i e_{k-1} e_j^T$, where e_i denotes the ith column of the $n \times n$ identity matrix and*

$$\widetilde{D}(\lambda) = \operatorname{diag}(\underbrace{d_1, \ldots, d_{i-1}, d_{i+1}, \ldots d_{k-1}}_{k-2 \text{ terms}}, s d_i, \underbrace{d_k, \ldots, d_{j-1}}_{j-k \text{ terms}}, -d_j/s, \underbrace{d_{j+1}, \ldots, d_n}_{n-j \text{ terms}}).$$

Proof. We obtain $\widetilde{D}(\lambda) + d_i e_{k-1} e_j^T$ by performing the following elementary transformations on $D(\lambda)$:

(i) add to column j column i multiplied by s,

(ii) add to row j row i multiplied by $-d_j/(sd_i)$,

(iii) permute columns i and j,

(iv) successively interchange rows t and $t+1$ for $t = i\,{:}\,k-1$, so that rows i, $i+1$, …, $k-2$, $k-1$ of the new matrix are rows $i+1$, $i+2$, …, $k-1$ and i, respectively, of the former one,

(v) permute columns i to $k-1$ in the same way as the rows in (iv). ∎

Theorem 5.3.2. *Let $d_1|\cdots|d_n$ be monic polynomials with coefficients in an algebraically closed field $\mathbb{F}$. Then there exists a monic triangular matrix polynomial $P(\lambda) \in \mathbb{F}[\lambda]^{n\times n}$ of degree ℓ and with $d_1, \ldots, d_n$ as invariant factors if and only if $\sum_{j=1}^{n} \deg d_j = \ell n$.*

Proof. The "only if" part is trivial. For the "if" part, suppose that there are monic polynomials $d_1|\cdots|d_n$ such that $\sum_{j=1}^{n} \ell_j = \ell n$, where $\ell_j = \deg d_j$ and let $D(\lambda) = \operatorname{diag}(d_1, \ldots, d_n)$. If $\ell_1 = \ell$ then $\ell_i = \ell$ for $i = 2\,{:}\,n$. Hence $D(\lambda)$ is a monic triangular (in fact diagonal) matrix polynomial of degree ℓ and the construction is done. If, on the other hand, $\ell_1 < \ell$, then $\ell_n > \ell$ and so $\ell_1 < \ell_1 + \ell_n - \ell < \ell_n$, then there is a monic polynomial s of degree $\ell_n - \ell$ such that $d_{k-1}|sd_1|d_k$ for some index k, $1 < k \le n$. By Lemma 5.3.1, $D(\lambda)$ is equivalent to

$$T_1(\lambda) = \begin{bmatrix} d_2 & & & & & & & \\ & \ddots & & & & & & \\ & & d_{k-1} & & & & & \\ & & & d_1 s & \cdots & & \cdots & d_1 \\ & & & & d_k & & & \vdots \\ & & & & & \ddots & & \\ & & & & & & d_{n-1} & \vdots \\ & & & & & & & -d_n/s \end{bmatrix} = \begin{bmatrix} D_1(\lambda) & t_1 \\ & -d_n/s \end{bmatrix},$$

where $D_1(\lambda) = \operatorname{diag}(d_2, \ldots, d_{k-1}, d_1 s, \ldots, d_{n-1}) = \operatorname{diag}(d_1^{(1)}, \ldots, d_{n-1}^{(1)})$ is in Smith form. If $\deg D_1 > \ell$ then we look for a new index k and a monic polynomial s_1 of degree $\deg d_{n-1}^{(1)} - \ell$ such that $d_{k-1}^{(1)}|s_1 d_1^{(1)}|d_k^{(1)}$. Apply the elementary transformations of Lemma 5.3.1 to the whole matrix $T_1(\lambda)$ so that the obtained matrix has the degree ℓ polynomial $-d_{n-1}^{(1)}/s_1$ as the $(n-1)$st diagonal entry. Notice that these elementary transformations can modify the off-diagonal elements of the nth column

of $T_1(\lambda)$ but the degree ℓ polynomial $-d_n/s$ in position (n,n) remains unchanged. We repeat this deflation process until all diagonal entries are of degree ℓ. The resulting matrix polynomial $T(\lambda)$ is upper triangular but not necessarily monic or of degree ℓ. We can, however, use the diagonal entries to eliminate off-diagonal terms of degree larger than $\ell-1$ in the following way: if $\deg t_{ij} > \ell-1$ for $i<j$ then $t_{ij} = t_{ii}b_{ij} + c_{ij}$ for some b_{ij} and c_{ij} such that $\deg c_{ij} < \deg t_{jj} = \ell$ or $c_{ij} = 0$. Then adding to the ith column of $T(\lambda)$ the jth column multiplied by $-b_{ij}$ reduces the (i,j) entry to zero or to some polynomial with degree strictly less than ℓ. Note that to reduce the degree of all the off-diagonal entries of $T(\lambda)$ we must work bottom up. We now have all degree ℓ polynomials on the diagonal, and their leading coefficient is plus or minus 1. Thus, after multiplication by a diagonal sign matrix, we end up with a monic upper triangular matrix polynomial of degree ℓ. ∎

Notice that for a given list of invariant factors $d_1|\cdots|d_n$ such that $\sum_{j=1}^n \deg d_j = \ell n$, there may be more than one monic triangular matrix polynomial of degree ℓ having $d_1,\ldots,d_n$ as invariant factors, as illustrated by the following example.

Example 5.3.3. Let $\mathbb{F}=\mathbb{C}$ and $d_1 = 1$, $d_2 = (\lambda^2+1)\lambda^2$ and $d_3 = (\lambda^2+1)\lambda^2(\lambda-1)$ be given. Note that $d_1|d_2|d_3$ and $\sum_{j=1}^3 \deg d_j = 9$ so that by Theorem 5.3.2 there is a 3×3 monic triangular matrix polynomial $T(\lambda)$ of degree 3 with d_1,d_2,d_3 as invariant factors. To construct $T(\lambda)$ we first look for a degree $\deg d_3 - \ell = 2$ polynomial s and an index k such that $d_{k-1}|sd_1|d_k$. We have to take $k=2$ and for s we can choose either $s=\lambda^2$ or $s=\lambda^2+1$. Both choices yield a different upper triangular cubic matrix polynomial.

- If $s=\lambda^2$ then by Lemma 5.3.1 $D(\lambda) = \operatorname{diag}(d_1,d_2,d_3)$ is equivalent to $T_1(\lambda) = \operatorname{diag}\big(\lambda^2,(\lambda^2+1)\lambda^2,-(\lambda^2+1)(\lambda-1)\big) + e_2e_3^T$. Then we look for a linear polynomial s_1 such that $\lambda^2|s_1\lambda^2|\lambda^2(\lambda^2+1)$. We can take either $s_1=\lambda-i$ or $s_1=\lambda+i$. The latter choice yields

$$T_a(\lambda) = \begin{bmatrix} \lambda^2(\lambda+i) & \lambda^2 & 1 \\ 0 & -\lambda^2(\lambda-i) & -(\lambda-i) \\ 0 & 0 & -(\lambda^2+1)(\lambda-1) \end{bmatrix}.$$

- The choice $s=\lambda^2+1$ leads to the real matrix polynomial

$$T_b(\lambda) = \begin{bmatrix} (\lambda^2+1)\lambda & \lambda^2+1 & 1 \\ 0 & -(\lambda^2+1)\lambda & -\lambda \\ 0 & 0 & -\lambda^2(\lambda-1) \end{bmatrix}. \qquad \blacktriangle$$

We are now ready to state the main result of this section, which generalizes [57, Theorem 9.3].

Theorem 5.3.4. *For an algebraically closed field $\mathbb{F}$, any $P(\lambda) \in \mathbb{F}[\lambda]^{n\times m}$ with $n \le m$ is triangularizable.*

Proof. Assume $P(\lambda)$ has degree ℓ and rank r. Since $\mathbb{F}$ is algebraically closed, it is infinite (otherwise we could define the polynomial $1 + \Pi_{\lambda_i\in\mathbb{F}}(\lambda - \lambda_i)$, which does not have any roots in $\mathbb{F}$) and from the discussion in Section 5.2 it follows that there is a Möbius function m_A induced by a nonsingular matrix $A \in \mathbb{F}^{2\times 2}$ such that if $\mathcal{M}_A(P)$ is triangularizable then $P(\lambda)$ is strongly equivalent to triangular matrix polynomial of degree ℓ. We will now show that $\mathcal{M}_A(P)$ is triangularizable. By [57, Proposition 3.29], $\operatorname{rank}\mathcal{M}_A(P) = \operatorname{rank} P(\lambda) = r$. Let

$$D(\lambda) = \operatorname{diag}(d_1, \dots, d_r, 0, \dots, 0) \in \mathbb{F}[\lambda]^{n\times m}$$

be the Smith form of $\mathcal{M}_A(P)$. Because all minors of $\mathcal{M}_A(P)$ of order r are of degree at most $r\ell$, and because the greatest common divisor of all such minors is invariant under unimodular transformations (see [26, p. 140] or [32, Theorem S1.2]), it holds that $\sum_{j=1}^{r} \deg d_j \le r\ell$. We consider two cases.

Case 1 $\sum_{j=1}^{r} \deg d_j = r\ell$**.** By Theorem 5.3.2, the regular part $\operatorname{diag}(d_1, \dots, d_r) \in \mathbb{F}[\lambda]^{r\times r}$ of $D(\lambda)$ is equivalent to an $r \times r$ upper triangular matrix polynomial of degree ℓ. Hence $\mathcal{M}_A(P)$ is triangularizable.

Case 2 $\sum_{j=1}^{r} \deg d_j < r\ell$**.** If $r = m$, then $\mathcal{M}_A(P)$ is square and regular, that is, $\mathcal{M}_A(P)$ has $m\ell$ eigenvalues, a contradiction. Thus $r < m$. Starting with $\widetilde{T}_0(\lambda) = \operatorname{diag}(d_1, \dots, d_r) \in \mathbb{F}[\lambda]^{r\times r}$, we follow the construction in Theorem 5.3.2 until we reach a step, say $r - k$, such that the matrix polynomial has the form

$$\widetilde{T}_{r-k}(\lambda) = \begin{bmatrix} \widetilde{d}_1 & & & & * & \cdots & * \\ & \widetilde{d}_2 & & & \vdots & \ddots & \vdots \\ & & \ddots & & \vdots & \ddots & \vdots \\ & & & \widetilde{d}_k & * & \cdots & * \\ & & & & * & \cdots & * \\ & & & & & \ddots & \vdots \\ & & & & & & * \end{bmatrix}, \tag{5.1}$$

where $\deg \widetilde{d}_j < \ell$ for $j = 1\colon k$ and the asterisks on the diagonal denote polynomials of degree ℓ. Now, as in the proof of Theorem 5.3.2, suppose we have applied

an appropriate sequence of elementary transformations to reduce the degree of the off-diagonal entries of $\widetilde{T}_{r-k}(\lambda)$ to polynomials of degree strictly less than ℓ. Then $\mathcal{M}_A(P)$ is equivalent to the upper triangular matrix polynomial of degree ℓ, $T_{r-k}(\lambda) = \widetilde{T}_{r-k}(\lambda) \oplus 0_{n-r,m-r}$. Note that $\widetilde{T}_{r-k}(\lambda)$ has a singular leading coefficient so $\operatorname{rank}\operatorname{rev}(T_{r-k})(0) < r$. This means that $T_{r-k}(\lambda)$ has elementary divisors at infinity, and it is hence not strongly equivalent to $\mathcal{M}_A(P)$. We now show how to remove the elementary divisors at infinity while maintaining the upper triangular form. Note that since $r < m$, the last column of $T_{r-k}(\lambda)$ is a zero column. Thus permuting the columns according to $(1, 2, \ldots, n)$ (cycle notation) preserves the triangular structure. Define g_i through $\deg(\lambda^{g_i}\widetilde{d}_i) = \ell$. Using a sequence of k right elementary operations we obtain the equivalent matrix polynomial

$$T(\lambda) = \begin{bmatrix} \lambda^{g_1}\widetilde{d}_1 & \widetilde{d}_1 & & & & * & \cdots & * \\ & \lambda^{g_2}\widetilde{d}_2 & \widetilde{d}_2 & & & \vdots & \ddots & \vdots \\ & & \ddots & \ddots & & \vdots & \ddots & \vdots \\ & & & \lambda^{g_k}\widetilde{d}_k & \widetilde{d}_k & * & \cdots & * \\ & & & & & * & \cdots & * \\ & & & & & & \ddots & \vdots \\ & & & & & & & * \end{bmatrix} \oplus 0_{n-r,m-r-1}, \tag{5.2}$$

which is still upper triangular and of degree ℓ. It now remains to show that $\operatorname{rev} T(\lambda) = \lambda^\ell T(\lambda^{-1})$ has no elementary divisor at zero. For this we write $\widetilde{d}_i$ in factorized form

$$\widetilde{d}_i = \begin{cases} \prod_{j=1}^{\ell-g_i}(\lambda - \lambda_{ij}) & \text{if } \ell > g_i, \\ 1 & \text{otherwise,} \end{cases}$$

and let

$$c_i = \begin{cases} \prod_{j=1}^{\ell-g_i}(1 - \lambda\lambda_{ij}) & \text{if } \ell > g_i, \\ 1 & \text{otherwise.} \end{cases}$$

Then $\operatorname{rev}(T)$ equals

$$\begin{bmatrix} c_1 & \lambda^{g_1}c_1 & & & & * & \cdots & * \\ & c_2 & \lambda^{g_2}c_2 & & & \vdots & \ddots & \vdots \\ & & \ddots & \ddots & & \vdots & \ddots & \vdots \\ & & & c_k & \lambda^{g_k}c_k & * & \cdots & * \\ & & & & & * & \cdots & * \\ & & & & & & \ddots & \vdots \\ & & & & & & & * \end{bmatrix} \oplus 0_{n-r,m-r-1}.$$

By construction, the polynomials represented as asterisks on the diagonal of $\operatorname{rev}(T)$ do not annihilate when evaluated at zero, and similarly, $c_i(0) \neq 0$. Therefore $\operatorname{rank}\operatorname{rev}(T)(0) = r$ and so $\operatorname{rev}(T)$ has no elementary divisors at zero. Hence, the upper triangular matrix polynomial $T(\lambda)$ in (5.2) is strongly equivalent to $\mathcal{M}_A(P)$, that is, $\mathcal{M}_A(P)$ is triangularizable. ∎

Remark 5.3.5. If $n > m$, we cannot always triangularize; see Example 5.3.7. The construction fails when we can no longer guarantee that $r < m$, implying that we cannot permute the nonzero part of the matrix one step to the right. However, using similar arguments, we can in this case ensure that $r < n$. By permuting the nonzero part of the matrix one step down instead of one step to the right, we can still build a matrix polynomial with the correct elementary divisors; this matrix will have Hessenberg structure (all entries (i, j) are zero for $i + 1 > j$).

We illustrate Theorem 5.3.4 with the following example taken from [87, Example 1].

Example 5.3.6. The quadratic matrix polynomial

$$Q(\lambda) = \begin{bmatrix} \lambda^2 + \lambda & 4\lambda^2 + 3\lambda & 2\lambda^2 \\ \lambda & 4\lambda - 1 & 2\lambda - 2 \\ \lambda^2 - \lambda & 4\lambda^2 - \lambda & 2\lambda^2 - 2\lambda \end{bmatrix}$$

has Smith form

$$D(\lambda) = \begin{bmatrix} 1 & -1 & -1 \\ -\lambda & 1+\lambda & \lambda \\ 0 & -\lambda & 1 \end{bmatrix} Q(\lambda) \begin{bmatrix} 1 & -3 & 6 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \lambda - 1 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

and since $\det(Q(\lambda)) = \det(D(\lambda)) \equiv 0$, $Q(\lambda)$ is singular. This matrix polynomial has only one finite elementary divisor. Note that $\operatorname{rank}\operatorname{rev}(Q)(0) = 1 < 2 = \operatorname{rank} Q(\lambda)$, so $Q(\lambda)$ has elementary divisors at infinity. Now the Smith form of $\operatorname{rev}(Q)$, given by $\widetilde{D}(\lambda) = \operatorname{diag}(1, \lambda^2(\lambda - 1), 0)$, reveals an elementary divisors at infinity for $Q(\lambda)$ with partial multiplicity 2.

As zero is not eigenvalue of $Q(\lambda)$, bearing in mind the technique described in Section 5.2, we can take $a = 0$ and $c = 1$. In fact, if $A = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$ then $\mathcal{M}_A(Q) = \operatorname{rev}(Q)$ has no elementary divisors at infinity and we can follow the proof of Theorem 5.3.4 with this matrix. We start the triangularization process with the submatrix $\operatorname{diag}(1, \lambda^2(\lambda - 1))$. Lemma 5.3.1 with $(i, j) = (1, 2)$, $k = 2$ and $s(\lambda) = \lambda - 1$ yields

$\widetilde{T}(\lambda) = \operatorname{diag}(\lambda - 1, -\lambda^2) + e_1 e_2^T$. Hence, $\widetilde{D}(\lambda)$ is equivalent to

$$T_1(\lambda) = \begin{bmatrix} \lambda - 1 & 1 & 0 \\ 0 & -\lambda^2 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

The matrix polynomial $T_1(\lambda)$, which is quadratic and upper triangular, has a singular leading coefficient indicating that $T_1(\lambda)$ and $\mathcal{M}_A(Q)$ are not strongly equivalent. Its elementary divisors at infinity can be removed as described in the proof of Theorem 5.3.4. First we permute the columns according to (1,2,3), to obtain:

$$T_2(\lambda) = \begin{bmatrix} 0 & \lambda - 1 & 1 \\ 0 & 0 & -\lambda^2 \\ 0 & 0 & 0 \end{bmatrix}.$$

Second, multiply the second column by λ and add it to the first one. This yields:

$$T(\lambda) = \begin{bmatrix} \lambda(\lambda - 1) & \lambda - 1 & 1 \\ 0 & 0 & -\lambda^2 \\ 0 & 0 & 0 \end{bmatrix},$$

which is strongly equivalent to $\mathcal{M}_A(Q)$. Finally,

$$\mathcal{M}_{A^{-1}}(T) = \operatorname{rev}(T) = \begin{bmatrix} \lambda + 1 & -\lambda^2 + \lambda & \lambda^2 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{bmatrix}$$

is quadratic, triangular and strongly equivalent to $Q(\lambda)$. ▲

The next example shows triangularization is not always possible when $n > m$.

Example 5.3.7. The quadratic matrix polynomial $Q(\lambda) = \left[\begin{smallmatrix} \lambda \\ \lambda^2 \end{smallmatrix}\right]$ has the Smith form $D(\lambda) = \left[\begin{smallmatrix} \lambda \\ 0 \end{smallmatrix}\right]$. A triangular matrix polynomial $T(\lambda) = \left[\begin{smallmatrix} q \\ 0 \end{smallmatrix}\right]$ has the Smith form $D(\lambda)$ if and only if $q = \lambda$, but then $\deg T(\lambda) \neq \deg Q(\lambda)$. ▲

5.4 Quasi-triangularization over the real numbers

We now concentrate on the non-algebraically closed field $\mathbb{R}$. Although some real matrix polynomials are triangularizable over $\mathbb{R}[\lambda]$ (see for instance Example 5.3.3 and Example 5.3.6), it is shown in [84] that not all quadratic real matrix polynomials

are triangularizable over $\mathbb{R}[\lambda]$. A characterization of all matrix polynomials that are triangularizable over $\mathbb{R}[\lambda]$ is given [79].

We now show that any $P(\lambda) \in \mathbb{R}[\lambda]^{n\times m}$ with $n \leq m$ is quasi-triangularizable. We start with an analogue of Theorem 5.3.2 for real matrix polynomials.

Theorem 5.4.1. *Let $d_1|\cdots|d_n$ be monic polynomials with coefficients in $\mathbb{R}$. Then there exists a monic quasi-triangular matrix polynomial $T(\lambda) \in \mathbb{R}[\lambda]^{n\times n}$ of degree ℓ and with $d_1, \ldots, d_n$ as invariant factors if and only if $\sum_{j=1}^{n} \deg d_j = \ell n$.*

Proof. We only prove the "if" part as the "only if" part is trivial. Suppose that there are monic polynomials $d_1|\cdots|d_n$ such that $\sum_{j=1}^{n} \ell_j = \ell n$, where $\ell_j = \deg d_j$ and $\ell_1 < \ell$. Let $T_n(\lambda) = \operatorname{diag}(d_1, \ldots, d_n)$. We start by constructing an upper triangular matrix polynomial equivalent to $T_n(\lambda)$ whose diagonal entries have degree either ℓ or $\ell + 1$ or $\ell - 1$.

Assume that there is a pair of indices (i, j) such that $\ell_i < \ell < \ell_j$ for which there is a real polynomial s of degree $\ell_j - \ell$ such that $d_{k-1}|sd_i|d_k$ for some index $k \leq j$. It is important for us to remark that the existence of such a real polynomial s is equivalent to the existence of a real polynomial r of degree $\ell_j - \ell - \deg(d_{k-1}/d_i)$ such that $r|(d_k/d_{k-1})$. Then by Lemma 5.3.1, $T_n(\lambda)$ is equivalent to a matrix polynomial of the form

$$\begin{bmatrix} \ddots & & & & & & \\ & sd_i & & d_i & & & \\ & & \ddots & & & & \\ & & & -d_j/s & & & \\ & & & & d_{j+1} & & \\ & & & & & \ddots & \\ & & & & & & d_n \end{bmatrix}.$$

By permuting rows and columns we can move the degree ℓ polynomial $-d_j/s$ to the lower right corner while keeping the upper triangular form and the $(n-1)\times(n-1)$ leading submatrix in Smith form. Specifically,

$$\begin{bmatrix} \ddots & & & & & & \\ & sd_i & & & & d_i & \\ & & \ddots & & & & \\ & & & d_{j+1} & & & \\ & & & & \ddots & & \\ & & & & & d_n & \\ & & & & & & -d_j/s \end{bmatrix}.$$

We say that a polynomial $-d_j/s$ has been *deflated* to the (n,n) position. We repeat this deflation procedure for all possible pairs of indices (i,j) satisfying the above conditions. Also, by means of appropriate permutations of rows and columns which do not introduce nonzero entries in the lower triangular part of the matrix, we can move (deflate) all the diagonal entries of degree ℓ down to the lower right part of the matrix. We end up with a matrix polynomial of the following form

$$T_p(\lambda) = \begin{bmatrix} c_1 & & & & * & \cdots & * \\ & c_2 & & & \vdots & \ddots & \vdots \\ & & \ddots & & \vdots & \ddots & \vdots \\ & & & c_p & * & \cdots & * \\ & & & & * & \cdots & * \\ & & & & & \ddots & \vdots \\ & & & & & & * \end{bmatrix},$$

where the c_i are polynomials such that $c_1|c_2|\cdots|c_p$ (that is, the $p \times p$ leading principal submatrix of $T_p(\lambda)$ is in Smith form), $\sum_{i=1}^{p} \deg c_i = p\ell$, and the asterisks on the diagonal denote polynomials of degree ℓ. We redefine ℓ_i to be the degree of c_i. Note that if $\ell_1 = \ell$ then $T_p(\lambda)$ has all its diagonal entries of degree ℓ.

Suppose that $\ell_1 < \ell$, which implies that $p \geq 2$. We show that if we cannot deflate a degree ℓ polynomial, then we can consecutively deflate two polynomials of degree $\ell+1$ and $\ell-1$, respectively. If $p = 2$ and there is no real polynomial s of degree $\ell_2 - \ell$ such that $c_1|sc_1|c_2$ then there is no real polynomial r of degree $\ell_2 - \ell$ such that $r|(c_2/c_1)$. This implies that c_2/c_1 has no linear factor and $\ell_2 - \ell$ is odd. Thus there is a degree $\ell_2 - (\ell+1)$ polynomial s_1 such that $c_1|s_1c_1|c_2$. Then using the procedure described in Lemma 5.3.1 with $(i,j) = (1,2)$, $k = 2$ and s_1, we deflate a degree $\ell+1$ polynomial in position $(2,2)$ leaving s_1c_1 of degree $\ell-1$ in position $(1,1)$.

We now assume that $p > 2$ and that for any pair of indices (i,j) with $\ell_i < \ell < \ell_j$ we cannot find a real polynomial s of degree $\ell_j - \ell$ such that $c_{k-1}|sc_i|c_k$ for any index k, $i < k \leq j$. Then there is no real polynomial r of degree $\ell_j - \ell - \deg(c_{k-1}/c_i)$ such that $r|(c_k/c_{k-1})$. It follows then that c_k/c_{k-1} contains no linear factors and $\ell_j - \ell$ $(= \deg s)$ and $\ell_{k-1} - \ell_i$ $(= \deg(c_{k-1}/c_i))$ have different parity. We consider three cases.

Case 1 $\ell_2 < \ell < \ell_{p-1}$. Then $\ell_1 < \ell < \ell_p$ and there is a degree $\ell_p - (\ell+1)$ polynomial s_1 such that for some index $k \leq p$, $c_{k-1}|s_1c_1|c_k$. We use the procedure described in Lemma 5.3.1 with $(i,j) = (1,p)$, s_1 and the index k to deflate the

degree $\ell+1$ polynomial $-c_p/s_1$ to position (p,p). This produces a matrix $T_{p-1}(\lambda)$, whose $(p-1)\times(p-1)$ leading principal submatrix is a Smith form still having c_2 and c_{p-1} as diagonal elements. We then repeat the argument using c_2 and c_{p-1} and a polynomial s_2 of degree $\ell_{p-1}-(\ell-1)$ such that $c_{k-1}|s_2c_2|c_k$ for some index $k\leq p-1$ to deflate the degree $\ell-1$ polynomial $-c_{p-1}/s_2$ to position $(p-1,p-1)$.

Case 2 $\ell_2>\ell$. As explained above, there are no linear factors in c_2/c_1, so ℓ_1 and ℓ_2 have the same parity. Further, $\ell_3-\ell$ is odd and c_3/c_2 contains no linear factors (otherwise there would be a real polynomial r of degree $\ell_3-\ell-\deg(c_3/c_1)$ such that $r|(c_3/c_2)$ and so $c_2|sc_1|c_3$ for some polynomial of degree $\ell_3-\ell$; a contradiction). Hence ℓ_1, ℓ_2 and ℓ_3 have the same parity. Using c_1, c_2 and a polynomial s_1 of even degree $\ell_2-\ell-1$ such that $c_1|s_1c_1|c_2$, we apply the procedure described in Lemma 5.3.1 to deflate the degree $\ell+1$ polynomial $-c_2/s_1$ to the (p,p) diagonal entry. This produces a triangular matrix whose $(p-1)\times(p-1)$ leading principal submatrix is $\mathrm{diag}(s_1c_1,c_3,\ldots,c_p)$. Note that $\deg(c_3/(s_1c_1))$ is even and $\ell_3-\ell$ is odd. Note also that $\ell_2>\ell$ implies that $2\ell\geq\ell_1+\ell_2$ and so $\ell_3-\deg(s_1c_1)=\ell_3-\ell_2+\ell+1-\ell_1\geq\ell_3-\ell+1$. Hence, we can always find a polynomial s_2 of degree $\ell_3-\ell+1$ such that $s_1c_1|s_2s_1c_1|c_3$ and deflate the degree $\ell-1$ polynomial $-c_3/s_2$ to position $(p-1,p-1)$.

Case 3 $\ell_{p-1}<\ell$. The condition $\sum_{i=1}^{p}\deg c_i=p\ell$ implies $\ell_p-\ell>0$. Furthermore, $\ell_p-\ell$ is odd and $\ell_p-\ell_{p-1}$ is even because otherwise there would be a real polynomial s of degree $\ell_p-\ell$ such that $s|(c_p/c_{p-1})$ and so $c_{p-1}|sc_{p-1}|c_p$. This would imply that the degree ℓ polynomial $-c_p/s$ could be deflated by using Lemma 5.3.1. Now we use c_{p-1}, c_p and a polynomial s_1 of even degree $\ell_p-\ell-1$ satisfying $c_{p-1}|s_1c_{p-1}|c_p$ to deflate the degree $\ell+1$ polynomial $-c_p/s_1$ to position (p,p). We are left with $\mathrm{diag}(c_1,\ldots,c_{p-2},s_1c_{p-1})$. Notice that $\ell_p+\ell_{p-1}\geq 2\ell$. If $\deg(s_1c_{p-1})=\ell-1$ (that is, $\ell_p+\ell_{p-1}=2\ell$) we have already deflated two polynomials of degrees $\ell+1$ and $\ell-1$ to positions (p,p) and $(p-1,p-1)$. Otherwise, we look for a real polynomial s_2 of degree $\deg(s_1c_{p-1})-\ell+1=\ell_{p-1}+\ell_p-2\ell$ such that $c_{p-2}|s_2c_{p-2}|s_1c_{p-1}$. Note that $\deg s_2$ is even so we can always construct it. Using the procedure described in Lemma 5.3.1 with $(i,j)=(p-2,p-1)$, $k=p-1$ and s_2, we deflate the degree $\ell-1$ polynomial $-s_1c_{p-1}/s_2$ to the $(p-1,p-1)$ position.

We repeat these processes until all diagonal entries of the matrix polynomials are of degree either ℓ, $\ell+1$ or $\ell-1$, and each diagonal entry of degree $\ell+1$ is directly preceded by a diagonal element of degree $\ell-1$. It now remains to transform the resulting upper triangular matrix polynomial to quasi-triangular

form with entries of degree at most ℓ. We assume that the diagonal entries have been scaled to become monic. Now suppose that all the entries below row i are of degree ℓ or less. If the ith diagonal entry is of degree ℓ then we use the procedure described at the end of Theorem 5.3.2 to reduce the entries in row i except the (i,i) entry to polynomials of degree strictly less than ℓ. If the ith diagonal entry is of degree $\ell+1$, then the $(i-1)$th diagonal entry is of degree $\ell-1$. We use the procedure described at the end of Theorem 5.3.2 to reduce the entries in rows i and $i-1$ except those on the diagonal to polynomials of degree at most ℓ for row i and polynomials of degree at most $\ell-2$ for row $i-1$. Hence rows $i-1$ and i look like

$$\begin{bmatrix} 0 & \cdots & 0 & \tilde{d}_{i-1} & \lozenge & \lozenge & \cdots & \lozenge \\ 0 & \cdots & 0 & 0 & \tilde{d}_i & \times & \cdots & \times \end{bmatrix}, \quad \begin{matrix} \deg \tilde{d}_{i-1} = \ell - 1, & \deg \lozenge \leq \ell - 2, \\ \deg \tilde{d}_i = \ell + 1, & \deg \times \leq \ell. \end{matrix}$$

Next, we add λ times row $i-1$ to row i, and then $-\lambda$ times column $i-1$ to column i leading to

$$\begin{bmatrix} 0 & \cdots & 0 & \tilde{d}_{i-1} & * & * & \cdots & * \\ 0 & \cdots & 0 & \lambda\tilde{d}_{i-1} & e_i & * & \cdots & * \end{bmatrix},$$

where $\deg e_i \leq \ell$ and no entry hiding behind the asterisks is of degree larger than ℓ. By moving upwards through the matrix in this way we end up with real quasi-triangular matrix polynomial $T(\lambda) = \sum_{j=0}^{\ell} \lambda^j T_j$. Since $\sum_{j=1}^{n} \deg d_j = \ell n$, $T(\lambda)$ has ℓn finite eigenvalues, implying that the leading coefficient T_ℓ is nonsingular. The matrix polynomial $T_\ell^{-1} T(\lambda)$ is of degree ℓ, is monic, real and quasi-triangular, and has $d_1, \ldots, d_n$ as invariant factors. ∎

Example 5.4.2. Let

$$D(\lambda) = \operatorname{diag}\big(1, (\lambda^2+1)^2, (\lambda^2+1)^2, (\lambda^2+1)^2\big) = \operatorname{diag}(d_1, d_2, d_3, d_4)$$

be the Smith form of a 4×4 cubic matrix polynomial. We follow the proof of Theorem 5.4.1 to construct a quasi-triangular polynomial of degree $\ell = 3$ with Smith form $D(\lambda)$. Notice that $\ell_1 < \ell < \ell_2 = \ell_3 = \ell_4$, where $\ell_i = \deg d_i$ and that there is no real polynomial s of degree $\ell_2 - \ell = 1$ such that $1|s|d_2$. This corresponds to Case 2 in the proof of Theorem 5.4.1. Following the instructions yields $s_1 = 1$, so the first part of Case 2 is simply a permutation of d_2 to the lower right corner. Because $d_2 = d_3 = d_4$, this does not modify $D(\lambda)$ but to follow Case 2 in detail, we now consider the matrix $\operatorname{diag}(d_1, d_3, d_4, d_2)$. Next, we look for a degree $\ell_3 - \ell + 1 = 2$ real polynomial s_2 such that $1|s_2|d_3$. We have

to take $s_2 = \lambda^2 + 1$. Then, by Lemma 5.3.1, $D(\lambda)$ is equivalent to $T_1(\lambda) = \mathrm{diag}(\lambda^2+1, (\lambda^2+1)^2, -(\lambda^2+1), (\lambda^2+1)^2) + e_1 e_3^T$. It remains to apply the last step of the proof of Theorem 5.4.1 to block triangularize the polynomial. This leads to

$$T(\lambda) = \begin{bmatrix} \lambda^2+1 & -\lambda(\lambda^2+1) & 1 & -\lambda \\ \lambda(\lambda^2+1) & \lambda^2+1 & \lambda & -\lambda^2 \\ 0 & 0 & \lambda^2+1 & -\lambda(\lambda^2+1) \\ 0 & 0 & \lambda(\lambda^2+1) & \lambda^2+1 \end{bmatrix}. \qquad \blacktriangle$$

We can now state the analogue of Theorem 5.3.4 for real polynomials.

Theorem 5.4.3. *Any $P(\lambda) \in \mathbb{R}[\lambda]^{n\times m}$ with $n \le m$ is quasi-triangularizable.*

Proof. The proof is along the same line as that presented for Theorem 5.3.4. We only sketch it and point out the differences.

We apply a Möbius transform $\mathcal{M}_A$ to $P(\lambda)$ induced by a real 2×2 nonsingular matrix A such that $\mathcal{M}_A(P)$ has no elementary divisors at infinity. We compute the Smith form $D(\lambda)$ of $\mathcal{M}_A(P)$, and let $\mathrm{diag}(d_1, \ldots, d_r)$ denote the regular part of $D(\lambda)$, where $r = \operatorname{rank} P(\lambda)$. Starting with $\mathrm{diag}(d_1, \ldots, d_r)$, we follow the triangularization procedure in the proof of Theorem 5.4.1 with two small modifications if $\sum_{j=1}^{r} \deg d_j < \ell r$:

(i) We stop the induction procedure when the remaining (non-deflated) diagonal elements are of degrees strictly less than ℓ.

(ii) If the induction procedure reaches case 3, then item (i) assures that $\ell_p > \ell$. We might, however, have $\ell_p + \ell_{p-1} < 2\ell - 1$ ($\ell_p + \ell_{p-1}$ is even so $\ell_p + \ell_{p-1} = 2\ell - 1$ is not possible), in which case we deflate a polynomial of degree $\ell - 1$ to position (p, p). The remaining diagonal elements are of degrees strictly less than ℓ so we stop the induction.

Now, all diagonal elements of degree $\ell + 1$ are preceded by a diagonal element of degree $\ell - 1$. Hence, we can perform the block-triangularization as in Theorem 5.4.1. Finally, we remove unwanted elementary divisors at infinity using the procedure described in Theorem 5.3.4. ∎

Remark 5.4.4. In the singular case $n \le m$, $r < m$, the procedure for removing elementary divisors at infinity moves the nonzero quasi-triangular part of the matrix polynomial one column to the right. This means that the resulting matrix polynomial is in fact triangular.

Remark 5.4.5. If $n > m$, then we can, similar to Remark 5.3.5, build a strongly equivalent matrix polynomial that would be quasi-triangular if the first row was deleted.

5.5 Inverse problems

The main objective of this chapter was the characterization of the real and complex matrix polynomials that can be reduced to triangular or trapezoidal form while preserving the degree and the finite and infinite elementary divisors. However, as a by-product, we solved a structured inverse polynomial eigenvalue problem. Recall that problems concerning the construction of matrix polynomials having certain eigenvalues or elementary divisors are called inverse polynomial eigenvalue problems. In [32, Theorem 1.7] a monic inverse polynomial eigenvalue problem is solved over $\mathbb{C}$ (in fact over any algebraically closed field). Since monic matrix polynomials have no elementary divisors at infinity, the Smith form contains all the information about elementary divisors. It is shown in the above reference that in order to build such an $n \times n$ matrix polynomial of degree ℓ, the only constraints on the list of its elementary divisors are

(i) the geometric multiplicities are bounded by n (because any regular $n \times n$ matrix polynomial has n invariant factors), and

(ii) the sum of the partial multiplicities of all the elementary divisors is $n\ell$.

This is generalized to matrices with nonsingular leading coefficients over arbitrary fields in [60, Theorem 5.2]. From Theorem 5.3.4 and Remark 5.3.5 it follows that we can realize a list of finite and infinite elementary divisors by an $n \times m$ matrix polynomial of degree ℓ over an algebraically closed field if and only if condition (i) above and

(iii) the sum of the partial multiplicities of all elementary divisors, including those at infinity, is at most $\ell \min(m, n)$,

are satisfied, thereby extending the result in [32, Theorem 1.7] and [60, Theorem 5.2]. Furthermore, from Theorem 5.4.3 and Remark 5.4.5, we get the solution to the corresponding inverse problem over $\mathbb{R}[\lambda]$. As one could expect, the only additional constraint on a *complex* list of elementary divisors is that nonreal elementary divisors must come in complex conjugate pairs.

Constraints on the structure of matrix polynomials often impose constraints on the elementary divisors. We have described these constraints in the case of real

triangular matrix polynomials and have shown that there are no constraints for complex ones other that (i) and (iii).

Recall that a matrix polynomial is called *Hermitian* or *self-adjoint* if all the coefficient matrices are Hermitian. If the leading coefficient is nonsingular, it is well-known that all nonreal elementary divisors come in complex conjugate pairs [30, Lemma 1.2]. Given a regular Hermitian matrix polynomial $P(\lambda)$, we can always find a real Möbius transformation m_A such that $\mathcal{M}_A(P)$ is Hermitian and has nonsingular leading coefficient. Hence, it follows from Theorem 2.2.1 that also the nonreal elementary divisors of $P(\lambda)$ must come in complex conjugate pairs. This constraint on the list of elementary divisors is exactly the same constraint as in the inverse polynomial eigenvalue problem over $\mathbb{R}[\lambda]$. We have proved the following result.

Theorem 5.5.1. *Any regular Hermitian matrix polynomial is strongly equivalent to a real matrix polynomial.*

We conjecture that the theorem is true in the other direction too.

Conjecture 5.5.2. *Any regular real matrix polynomial is strongly equivalent to a Hermitian matrix polynomial, and vice versa.*

Note that the set of Hermitian matrix polynomials of degree ℓ and size $n \times n$ has exactly the number of same degrees of freedom as the set of real $n \times n$ matrix polynomial of degree ℓ. That is, both sets can be identified with $\mathbb{R}^{n^2\ell}$.

Since any regular real matrix polynomial can be mapped to a real matrix polynomial with invertible leading coefficient using a real Möbius transformation, it is enough to prove the conjecture for matrix polynomials without infinite eigenvalues.

It is easy to see that the conjecture is true for pencils. Given a real pencil with invertible leading coefficient, we may, for instance, construct a Hermitian pencil from the associated Jordan form as follows: permute the diagonal blocks so Jordan blocks with complex conjugate eigenvalues of the same size appear in pairs on the diagonal. Then multiply each such pair, $J_s(\lambda) \oplus J_s(\overline{\lambda})$ (here s denotes the size of the Jordan block), by the *sip matrix*

$$\begin{bmatrix} & & & 1 \\ & & 1 & \\ & \iddots & & \\ 1 & & & \end{bmatrix},$$

of size $2s \times 2s$, from left or right. The resulting submatrix is then Hermitian.

Another argument makes use of the fact that any real matrix is the product of two Hermitian matrices, one of which is invertible [17]. As mentioned above, we may assume that the real pencil we start with has invertible leading coefficient. Multiplying the entire pencil with the inverse of this leading coefficient yields a real monic pencil $I\lambda - A$. If $A = H_1^{-1}H_2$, where H_1 and H_2 are Hermitian, then $H_1\lambda + H_2$ is Hermitian and strongly equivalent (in fact strictly equivalent) to $I\lambda - A$.

Mackey and Tisseur [59] recently showed that the conjecture is true for quadratic matrix polynomials. Their proof is constructive and quite involved, and it is unclear if it can be generalized to higher order matrix polynomials.

We end this chapter by proving that the conjecture is true for 2×2 matrix polynomials. We need the following lemma.

Lemma 5.5.3. *Let $p(\lambda) = \sum_{i=0}^{\ell} \alpha_i \lambda^i$, $\alpha_\ell \neq 0$, be a scalar real polynomial. For any $k \in \{1, 2, \ldots, \ell - 1\}$, there exists a shift σ such that the coefficient in front of λ^k in $p(\lambda + \sigma)$ is nonzero.*

Proof. By the binomial theorem we have

$$p(\lambda + \sigma) = \sum_{i=0}^{\ell} \alpha_i (\lambda + \sigma)^i = \sum_{i=0}^{\ell} \sum_{j=0}^{i} \alpha_i \binom{i}{j} \lambda^j \sigma^{i-j},$$

where the coefficient in front of λ^k is given by

$$\sum_{i=k}^{\ell} \alpha_i \binom{i}{k} \sigma^{i-k} = \alpha_\ell \binom{\ell}{k} \sigma^{\ell-k} + \text{lower order terms in } \sigma.$$

Choosing σ large enough yields the result. ∎

Theorem 5.5.4. *Any regular real matrix polynomial of size 2×2 is strongly equivalent to a Hermitian matrix polynomial.*

Proof. As mentioned above, it is enough to prove the theorem for real matrix polynomials with invertible leading coefficients. Let $\operatorname{diag}(d_1, d_2)$ be the Smith form of such a matrix polynomial of degree ℓ. If the coefficient in d_1 in front of λ^ℓ is zero, we use Lemma 5.5.3 and consider a shifted matrix polynomial instead. Let ℓ_1 and ℓ_2 denote the degrees of d_1 and d_2 respectively. Multiply d_2 by -1 and consider the ansatz

$$\begin{bmatrix} 1 & 0 \\ \bar{x} & 1 \end{bmatrix} \begin{bmatrix} d_1 & 0 \\ 0 & -d_2 \end{bmatrix} \begin{bmatrix} 1 & x \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} d_1 & d_1 x \\ d_1 \bar{x} & d_1 x \bar{x} - d_2 \end{bmatrix}. \tag{5.3}$$

Because $\ell_1 + \ell_2 = 2\ell$ we have that $\ell_2 - \ell_1$ is even. Write

$$d_2 = \underbrace{\left(p\lambda^{\ell+1} + \alpha\lambda^{\ell_1}\right)}_{t_\alpha} + \left(s - \alpha\lambda^{\ell_1}\right),$$

where $\deg(s) = \ell$ and $\alpha \in \mathbb{R}$. Polynomial long division yields

$$t_\alpha = d_1(q + \alpha) + r_\alpha,$$

where $\deg(r_\alpha) < \ell_1$ and $\deg(q + \alpha) = \ell_2 - \ell_1$. For large α the real polynomial $q + \alpha$ has no real roots, and hence $q + \alpha = x\bar{x}$ for some nonreal polynomial x. Thus the $(2, 2)$ entry in right hand side of (5.3) is

$$d_1 x\bar{x} - d_2 = d_1 x\bar{x} - \left(d_1 x\bar{x} + r_\alpha + s - \alpha\lambda^{\ell-1}\right) =: p.$$

Note that p is real and of degree ℓ. We reuse the letter q, and do another long division to obtain $d_1 x = pq + r$. Taking the complex conjugate yields $d_1\bar{x} = p\bar{q} + \bar{r}$. We get

$$\begin{bmatrix} 1 & -q \\ 0 & 1 \end{bmatrix} \begin{bmatrix} d_1 & d_1 x \\ d_1\bar{x} & p \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\bar{q} & 1 \end{bmatrix} = \begin{bmatrix} d_1(1 - \bar{q}x - q\bar{x}) + p\bar{q}q & r \\ \bar{r} & p \end{bmatrix}.$$

Since the determinant is of degree 2ℓ, the $(1, 1)$ element on the right hand side is of degree ℓ. Finally, if we used Lemma 5.5.3, then we shift back to obtain a matrix polynomial that is equivalent to what we started with, Hermitian and of degree ℓ. ∎

Chapter

6

Reduction of matrix polynomials to simpler forms

6.1 Introduction

In many applications it is beneficial to first reduce matrices to simpler forms. For example, without the initial reduction to Hessenberg form, the complexity of the QR algorithm for computing eigenvalues would be $O(n^4)$ instead of $O(n^3)$. As a further example, the spectral decomposition $S^{-1}DS$ of an $n \times n$ matrix A can be used to decouple the system of ODEs $A\dot{u}(t) = u(t) + f(t)$ into n independent scalar ODEs, $D\dot{y}(t) = y(t) + g(t)$, where $y(t) = Su(t)$ and $g(t) = Sf(t)$. These equations can then be solved in parallel. In hopes of that reduced forms can also be useful for computations on matrix polynomials, we discuss in this chapter how to reduce matrix polynomials to triangular, diagonal and Hessenberg form, and related block forms, while preserving the degree and eigenstructure. The reductions are done by means of structure preserving similarity transformations applied to the left companion linearization. We discuss the construction and existence of these transformations and illustrate with MATLAB code how they can be performed in practice. The problem of given a matrix polynomial, finding an equivalent matrix polynomials of the same degree and of simpler form, in particular diagonal form, has been studied before [19, 20, 28, 63]. In contrast to previous work, our construction handles the reduction to the different structures mentioned above in a uniform manner. Our construction also shed some light on how equivalent matrix polynomials relate to each other, without mentioning unimodular transformations.

6.1.1 Structure preserving transformations

Consider matrix polynomials

$$P(\lambda) = P_\ell \lambda^\ell + P_{\ell-1}\lambda^{\ell-1} + \cdots + P_0 \quad \text{with} \quad \det(P_\ell) \neq 0, \tag{6.1}$$

over $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$. It is, in general, not possible to reduce such matrix polynomials to the simpler forms mentioned above using only strict equivalence. For example, if the degree $\ell > 1$, then there exist nonsingular matrices E and F such that $EP(\lambda)F$ is triangular, only if the first column f_1 of F is such that $\operatorname{rank}[P_\ell f_1 \ \cdots \ P_1 f_1 \ P_0 f_1] = 1$. But finding such vector f_1 is clearly not possible in general.

We saw in Chapter 5, that unimodular transformations provide enough freedom to reduce any square matrix polynomial to triangular form over $\mathbb{C}$ and quasi-triangular form over $\mathbb{R}$, while preserving the degree. Of course, this includes the case of Hessenberg form. Further, it is an easy exercise to show that any complex/real matrix polynomial with semisimple eigenstructure is equivalent to a diagonal/quasi-diagonal matrix polynomial of the same degree. But how can we compute these simpler forms in practice? The reductions described in Chapter 5 are based on applying unimodular transformations to the Smith form, which is not convenient from a numerical point of view. To avoid this, we work with linearizations instead. Suppose $P(\lambda)$ has the same eigenstructure as $R(\lambda) = I\lambda^\ell + \sum_{j=0}^{\ell-1} R_j \lambda^j$ and take any monic linearization $I\lambda - A$ of $P(\lambda)$. Note that $I\lambda - A$ is also a linearization of $R(\lambda)$. The Gohberg, Lancaster, Rodman theory [32] tells us that there is an $\ell n \times n$ matrix X such that (A, X) is a left standard pair for $R(\lambda)$, which means that the $\ell n \times \ell n$ matrix

$$S = [X \ AX \ \cdots \ A^{\ell-1}X] \tag{6.2}$$

is nonsingular and

$$A^\ell X + A^{\ell-1} X R_{\ell-1} + \cdots + AXR_1 + XR_0 = 0. \tag{6.3}$$

Taken together, (6.2) and (6.3) can be rewritten as

$$S^{-1}AS = \begin{bmatrix} & & -R_0 \\ I & & -R_1 \\ & \ddots & \vdots \\ & I & -R_{\ell-1} \end{bmatrix} =: C_L(R) \tag{6.4}$$

showing that A is similar to the left companion matrix associated with $R(\lambda)$. Since S preserve the left companion structure, we say that the similarity transformation defined by S is a *structure preserving transformation*. In fact, for any given monic linearization $I\lambda - A$ of $P(\lambda)$, any nonsingular matrix S of the form (6.2) will always transform A into a left companion matrix associated with some matrix polynomial, as in (6.4). Note that if $Y := (e_\ell \otimes I)S^{-1}$ and $R_\ell = I_n$, then it follows from [51, Theorem 14.2.5 and Theorem 14.7.1] that

$$R_{\ell-j} = -\sum_{i=\ell-j+1}^{\ell} YA^{i+j-1}XR_i, \quad j = 1\!:\!\ell.$$

The above discussion suggests that in order to reduce $P(\lambda)$ in (6.1) to a simpler form, it is enough to find an $n\ell \times n$ matrix X such that S in (6.2) is nonsingular and $S^{-1}AS$ has the desired zero pattern, where A can be any matrix such that $I\lambda - A$ is a linearization of $P(\lambda)$.

In the generic case, when all eigenvalues are distinct, it turns out to be surprisingly easy to find X such that $S^{-1}AS$ is the left companion matrix of a matrix polynomial in triangular, diagonal or Hessenberg form. We illustrate this with a snippet of MATLAB code. The code below generates a random monic cubic matrix polynomial, computes the left companion form of an equivalent triangular matrix polynomial and plots the (numerical) zero pattern of it.

```
n = 5; deg = 3;  % size and degree
P0 = randn(n); P1 = randn(n); P2 = randn(n);
C_P = [ zeros(n) zeros(n) -P0;
        eye(n)   zeros(n) -P1;
        zeros(n) eye(n)   -P2 ];
[U,~] = schur(C_P,'complex');
X = U*kron(eye(n), ones(deg,1));
S = [X C_P*X C_P^(deg-1)*X];
C_R = S\C_P*S;
spy(abs(C_R)>1e-12)
```

If we replace `schur(C_P,'complex')` by `eig(C_P)`, then `C_R` becomes the companion matrix of an equivalent diagonal matrix polynomial; and if we replace `schur(C_P,'complex')` by `hess(C_P)` and `ones(deg,1)` by `eye(deg,1)`, then `C_R` becomes the companion matrix of an equivalent matrix polynomial in Hessenberg form. The code can be generalized to any degree and works as long as the block Krylov matrix S is nonsingular, which it is for almost all coefficient matrices.

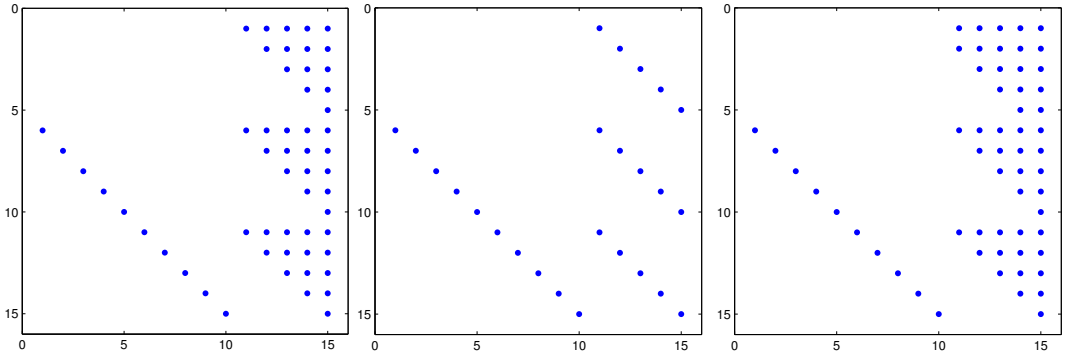

Figure 6.1: Spy plots for the reduced matrix polynomials obtained by the code shown below: triangular (left), diagonal (middle), and Hessenberg (right).

Spy plots from one execution of the above MATLAB code and the two discussed modifications of it are shown in Figure 6.1. In this chapter we discuss why and when the above code works. In the rare cases when it fails, we describe whenever possible what has to be achieved for the reductions to go through.

To be slightly more general, we also consider reduction to certain block forms. For a given $n \times n$ matrix polynomial of degree ℓ with entries in $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$, we consider the following reduced forms:

- block-diagonal form:

$$D(\lambda) = D_1(\lambda) \oplus D_2(\lambda) \oplus \cdots \oplus D_k(\lambda) \in \mathbb{F}[\lambda]^{n\times n} \tag{6.5}$$

  of degree ℓ with $D_i(\lambda) \in \mathbb{F}[\lambda]^{s_i \times s_i}$, $1 \le i \le k$ and $s_1 + \cdots + s_k = n$,

- block-triangular form:

$$T(\lambda) = \begin{bmatrix} T_{11}(\lambda) & T_{12}(\lambda) & \cdots & T_{1k}(\lambda) \\ & T_{22}(\lambda) & & \vdots \\ & & \ddots & \vdots \\ & & & T_{kk}(\lambda) \end{bmatrix} \in \mathbb{F}[\lambda]^{n\times n}, \tag{6.6}$$

  of degree ℓ with $T_{jj}(\lambda) \in \mathbb{F}[\lambda]^{s_j \times s_j}$, $1 \le j \le k$ and $s_1 + \cdots + s_k = n$, and

- Hessenberg form:

$$H(\lambda) = \lambda^\ell H_\ell + \cdots + \lambda H_1 + H_0 \in \mathbb{F}[\lambda]^{n\times n}, \tag{6.7}$$

  with coefficient matrices H_i, $i = 0\colon \ell$, in Hessenberg form.

6.2 Conditions for reduction

Let $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$. For matrices $A \in \mathbb{F}^{m\times m}$ and $V \in \mathbb{F}^{m\times k}$ we define the block Krylov matrix

$$K_\ell(A, V) = [V \; AV \; \cdots \; A^{\ell-1}V] \in \mathbb{F}^{m\times \ell k}$$

and the block Krylov subspace

$$\mathcal{K}_\ell(A, V) = \operatorname{range} K_\ell(A, V).$$

For a subspace $\mathcal{X}$ of $\mathbb{F}^m$ and a matrix A operating on that subspace we define $A\mathcal{X} = \{Ax \; : \; x \in \mathcal{X}\}$.

Assume that $P(\lambda)$ is given by (6.1) and let $\lambda I - A$ be any monic linearization of $P(\lambda)$, for example, the left companion linearization of $P_\ell^{-1}P(\lambda)$. Recall that we are looking for a matrix $X \in \mathbb{F}^{\ell n\times n}$ such that

(i) $S := [X \; AX \; \cdots \; A^{\ell-1}X]$ is nonsingular, and

(ii) $I\lambda - S^{-1}AS$ is the left companion linearization of one of the reduced forms in (6.5)–(6.7).

If (i) holds, then $S^{-1}AS$ is a left companion matrix associated with a monic matrix polynomial, say $R(\lambda) = \lambda^\ell I + \cdots + \lambda R_1 + R_0$, and

$$S^{-1}AS(e_n \otimes I_\ell) = S^{-1}A^\ell X = -\begin{bmatrix} R_0 \\ R_1 \\ \vdots \\ R_{\ell-1} \end{bmatrix}. \tag{6.8}$$

We see that the (i, j) element of $R(\lambda)$, $i \neq j$, is zero if and only if the vector $S^{-1}A^\ell x_j$ has zeros in the entries $i, i + n, \ldots, i + (\ell - 1)n$, where x_j denotes the jth column of X. From $S^{-1}[X \; AX \; \cdots \; A^{\ell-1}X] = I$ and (6.8) it follows that

$$[R(\lambda)]_{ij} \equiv 0, \; i \neq j \quad \Longleftrightarrow \quad A^\ell x_j \in \mathcal{K}_\ell(A, [x_1 \; \cdots \; x_{i-1} \; x_{i+1} \cdots \; x_n]). \tag{6.9}$$

We are now ready to state our main theorem, but before we do so we introduce some new notation. For the block reductions (6.5) and (6.6), it is useful to partition X as $X = [X_1 \; X_2 \; \cdots \; X_k]$, where $X_j \in \mathbb{F}^{\ell n\times s_j}$ and $s_1 + \cdots + s_k = n$. Finally, we let $x_{1:i}$ and $X_{1:i}$ denote the matrices $[x_1 \; x_2 \; \cdots \; x_i]$ and $[X_1 \; X_2 \; \cdots \; X_i]$, respectively.

Theorem 6.2.1. *Let $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$ and consider $P(\lambda) \in \mathbb{F}[\lambda]^{n\times n}$ of degree ℓ and with nonsingular leading matrix coefficient. Let $\lambda I - A$ be any monic linearization*

of $P(\lambda)$. Then $P(\lambda)$ is equivalent to a monic matrix polynomial $R(\lambda)$ of degree ℓ having one of the reduced forms (6.5)–(6.7) *if and only if there exists $X \in \mathbb{F}^{\ell n \times n}$ such that*

(i) *the matrix $[X \ AX \ \cdots \ A^{\ell-1}X] \in \mathbb{F}^{\ell n \times \ell n}$ is nonsingular, and*

(ii) (a) *$\mathcal{K}_\ell(A, X_i)$ is A-invariant for $i = 1\!:\!k$ for block-diagonal form as in* (6.5)*,*

(b) *$\mathcal{K}_\ell(A, X_{1:i})$ is A-invariant for $i = 1\!:\!k$ for block-triangular form as in* (6.6)*,*

(c) $\operatorname{range} A^\ell x_{1:i} \subset \mathcal{K}_\ell(A, x_{1:i+1})$ *for $i = 1\!:\!n-1$ for Hessenberg form as in* (6.7)*.*

Proof. ($\Rightarrow$) Suppose that $P(\lambda)$ is equivalent to $R(\lambda)$. Then $\lambda I - A$ is also a monic linearization of $R(\lambda)$ and as explained in the introduction, there is a matrix X such that (A, X) is a left standard pair for $R(\lambda)$, which implies (i) and $AS = SC_L(R)$ where $S = [X \ AX \ \cdots \ A^{\ell-1}X]$.

Now suppose that $R(\lambda)$ has the block-diagonal form of $D(\lambda)$ in (6.5). Let $\Pi_i \in \mathbb{F}^{\ell n \times \ell s_i}$ be the projection matrix such that $K_\ell(A, X_i) = K_\ell(A, X)\Pi_i = S\Pi_i$. Then from $AS = SC_L(D)$ we have that

$$AK_\ell(A, X_i) = AS\Pi_i = SC_L(D)\Pi_i = S\Pi_i C_L(D_i) = K_\ell(A, X_i)C_L(D_i),$$

which proves (ii)(a). The proofs for (ii)(b)–(c) are similar.

($\Leftarrow$) Suppose that there exists X such that $S = [X \ AX \ \cdots \ A^{\ell-1}X]$ is nonsingular. Then the matrix $S^{-1}AS$ is the left companion form of a monic matrix polynomial of degree ℓ, say $R(\lambda)$, equivalent to $P(\lambda)$.

Now $AS = SC_L(R)$, (ii)(a) and (6.9) imply that the $n \times n$ blocks $R_0, \ldots, R_{\ell-1}$ in the last block column of $S^{-1}AS$ (see (6.8)) are block-diagonal with k diagonal blocks, the ith diagonal block being $s_i \times s_i$, where s_i is the number of columns of X_i, $i = 1\!:\!k$. The proofs for (ii)(b)–(c) are similar. ∎

6.3 Construction of the matrix X

We discuss in this section a way to construct the matrix X in Theorem 6.2.1 such that properties (i) and (ii) hold.

We start by proving some technical results. Let $I\lambda - A$ be the left companion matrix of a monic matrix polynomial $P(\lambda)$ of size $n \times n$ and degree ℓ. Further, let

Π denote the permutation matrix

$$[e_1\ e_{n+1}\ \cdots\ e_{(\ell-1)n+1}\ e_2\ e_{n+2}\ \cdots\ e_{(\ell-1)n+2}\ \cdots\ e_n\ e_{2n}\ \cdots\ e_{\ell n}].$$

Then the permuted linearization $I\lambda - \Pi^T A \Pi$ is called the *left companion linearization of $P(\lambda)$ in controller form*. If we view this linearization as an $\ell \times \ell$ block pencil, then it has the same zero structure as $P(\lambda)$. Furthermore, the diagonal $\ell \times \ell$ blocks are the companion matrices of the corresponding scalar polynomials on the diagonal of $P(\lambda)$.

Using the controller form, it is easy to deduce the next theorem.

Theorem 6.3.1 (Complex case)**.** *Suppose $A \in \mathbb{C}^{\ell n \times \ell n}$ has no eigenvalue with geometric multiplicity greater than n. Then A has a Schur decomposition*

$$A = Q \begin{bmatrix} T_{11} & * & * & * \\ & T_{22} & * & * \\ & & \ddots & * \\ & & & T_{nn} \end{bmatrix} Q^H,$$

where the diagonal blocks $T_{ii} \in \mathbb{C}^{\ell \times \ell}$, $i = 1\colon n$, are nonderogatory.

Proof. Since A has no eigenvalue with geometric multiplicity greater than n, it follows from [32, Proof of Theorem 1.7] that $I\lambda - A$ is a linearization of an $n \times n$ upper triangular monic matrix polynomial $P(\lambda)$ of degree ℓ. This matrix polynomial has a left companion linearization in controller form, which itself must be monic. Let H denote the constant matrix of this linearization. Then $A = SHS^{-1}$ for some S. Further, H is block upper triangular, with blocks of size $\ell \times \ell$, and all diagonal blocks must be nonderogatory (since they are companion matrices). Let $U_i T_i U_i^H$ be a Schur decomposition of the ith diagonal block and set $U = U_1 \oplus U_2 \oplus \cdots \oplus U_n$. Then

$$H = UTU^H, \quad \text{with} \quad T = \begin{bmatrix} T_1 & * & * & * \\ & T_2 & * & * \\ & & \ddots & * \\ & & & T_n \end{bmatrix},$$

is a Schur decomposition. Finally, let $QR = SU$ be a QR factorization, and note that $A = Q(RTR^{-1})Q^H$ is a Schur decomposition of A. Since the ith diagonal $\ell \times \ell$ block of RTR^{-1} is similar to T_i the theorem is proved. ∎

We now prove the real analogue of Theorem 6.3.1.

Theorem 6.3.2 (Real case)**.** *Suppose $A \in \mathbb{R}^{\ell n \times \ell n}$ has no eigenvalue with geometric multiplicity greater than n. Then A has a real Schur decomposition*

$$A = Q \begin{bmatrix} T_{11} & * & * & * \\ & T_{22} & * & * \\ & & \ddots & * \\ & & & T_{ss} \end{bmatrix} Q^T, \tag{6.10}$$

where each T_{ii} is either of size $\ell \times \ell$ and nonderogatory or of size $2\ell \times 2\ell$ and such that all eigenvalues have geometric multiplicity one or two.

Proof. Since all eigenvalues of A have geometric multiplicity at most n, it follows that $I\lambda - A$ has a real Smith form $D(\lambda) \oplus I_{(\ell-1)n}$ with $\deg \det D(\lambda) = n\ell$. By Theorem 5.4.1, $D(\lambda)$ is equivalent to some real monic quasi-triangular matrix polynomial $T(\lambda)$ of degree ℓ. It follows that

$$I\lambda - A \sim \begin{bmatrix} D(\lambda) & \\ & I_{(\ell-1)n} \end{bmatrix} \sim \begin{bmatrix} T(\lambda) & \\ & I_{(\ell-1)n} \end{bmatrix},$$

where $\sim$ denotes the equivalence relation for matrix polynomials. In other words, A is a linearization of some monic quasi-triangular matrix polynomial of degree ℓ. If H denotes the constant matrix of the left companion linearization of $T(\lambda)$ in controller form, then the rest of the proof is essentially the same as last part of the proof of Theorem 6.3.1, with the only difference that we consider the real Schur decomposition instead of the complex. ∎

We now restrict ourselves to the (highly generic) case when no eigenvalue of A has algebraic multiplicity larger than n. In this case, the Schur decompositions in Theorem 6.3.1 and Theorem 6.3.2 can be computed relatively easily as follows. First compute any (real or complex) Schur decomposition. Then reorder the diagonal entries/blocks using the procedure in [8] according to rules described below. We discuss the real and complex case separately.

Complex case: Suppose there are k eigenvalues of algebraic multiplicity n, and note that $k \leq \ell$. Reorder the Schur form such that the leading $k \times k$ submatrix has one instance of each of these eigenvalues. If there are $k < \ell$ such eigenvalues, pick any $\ell - k$ distinct eigenvalues of algebraic multiplicity less than n and reorder the diagonal such that these appear after the first k eigenvalues that were deflated. The leading $\ell \times \ell$ submatrix obtained in this way has simple eigenvalues and is thus nonderogatory. By continuing inductively on the lower left $(n-1)\ell \times (n-1)\ell$ part of the matrix, we arrive at the desired Schur form.

Real case: The procedure over $\mathbb{R}$ is similar, but to keep the decomposition real, we need to move nonreal eigenvalues in complex conjugate pairs. First, reorder the Schur form such that one instance of each eigenvalue of algebraic multiplicity n appears in the leading $k \times k$ matrix (assuming there were k such eigenvalues). If $k = \ell$, continue inductively as in the complex case. If $k < \ell$ and k is even, move k_2 2×2 blocks, with pairwise different eigenvalues of algebraic multiplicity less than n, so they appear directly after the deflated $k \times k$ submatrix on the diagonal. Here $k_2 \leq (\ell - k)/2$ should be chosen to be as large as possible. If the leading $\ell \times \ell$ block is nonderogatory, continue inductively as in the complex case. Otherwise move $\ell - k - 2k_2$ distinct *real* eigenvalues of algebraic multiplicity less than n so they appear after the k_2 2×2 blocks we just deflated. If $k < \ell$ and k is odd, divide into two cases. If there is a real eigenvalue of algebraic multiplicity less than n, deflate one instance of that eigenvalue and continue by deflating 2×2 blocks as above, so the leading $\ell \times \ell$ matrix becomes nonderogatory. If there is no such real eigenvalue available we aim to form a leading $2\ell \times 2\ell$ block where all eigenvalues have algebraic multiplicity at most two. Reorder the Schur form so the leading $2k \times 2k$ submatrix contains two of each eigenvalues of algebraic multiplicity n. Since $2k < 2\ell$ is always even and all real eigenvalues have algebraic multiplicity n, there are $\ell - k$ available 2×2 blocks with pairwise distinct eigenvalues of algebraic multiplicity less than n. Move these so they appear after the leading $2k \times 2k$ submatrix. The leading $2\ell \times 2\ell$ submatrix now has the desired property. Continue inductively as in the complex case, with $n - 2$ instead of n.

Theorem 6.3.1 and Theorem 6.3.2 will be used in combination with the following lemmas.

Lemma 6.3.3 (Complex case)**.** *If $B \in \mathbb{C}^{\ell \times \ell}$ is nonderogatory, then there exists $x \in \mathbb{C}^{\ell}$ such that the Krylov matrix $K_\ell(B, x)$ is nonsingular.*

Proof. Since B is nonderogatory it is similar to the companion matrix C of its characteristic polynomial [40, Theorem 3.3.15], that is, $B = S^{-1}CS$ for some nonsingular matrix S. It is now easy to see that $K_\ell(C, e_1) = I$. Hence letting $x = S^{-1}e_1$ yields the desired result. ∎

The next lemma is the real counterpart of Lemma 6.3.3.

Lemma 6.3.4 (Real case)**.** *Let $B \in \mathbb{R}^{2\ell \times 2\ell}$ have eigenvalues with geometric multiplicity at most 2. Then there exist two real vectors x and y such that $K_\ell(B, [x\ y])$ is nonsingular.*

Proof. We can rearrange the real Jordan decomposition of B so that

$$S^{-1}BS = \begin{array}{c} \\ m_1 \\ m_2 \end{array} \begin{array}{c} \begin{array}{cc} m_1 & m_2 \end{array} \\ \begin{bmatrix} J_1 & \\ & J_2 \end{bmatrix} \end{array} \in \mathbb{R}^{2\ell \times 2\ell}, \quad m_1 \geq m_2 > 0$$

with J_1 and J_2 nonderogatory. This latter property implies that J_1 and J_2 are similar (via real arithmetic) to left companion matrices $C_1 \in \mathbb{R}^{m_1 \times m_1}$ and $C_2 \in \mathbb{R}^{m_2 \times m_2}$, respectively. Hence there exists a nonsingular $W \in \mathbb{R}^{2\ell \times 2\ell}$ such that

$$W^{-1}BW = C_1 \oplus C_2 =: C.$$

It suffices to prove that there exist $u, v \in \mathbb{R}^{2\ell}$ such that $M = [K_\ell(C, u)\ K_\ell(C, v)]$ is nonsingular because taking $x = W^{-1}u$ and $y = W^{-1}v$ then yields the desired result.

If $m_1 = m_2$ then $u = e_1$ and $v = e_{\ell+1}$ yield $M = I_{2\ell}$ and we are done. If $m_1 > m_2$, we let $u = e_1$ and $v = e_{\ell-m_2+1} + e_{m_1+1}$. Then direct calculations show that

$$M = \begin{array}{c} \begin{array}{cccc} \ell-m_2 & m_2 & m_2 & \ell-m_2 \end{array} \\ \begin{bmatrix} I & & & \\ & I & I & \\ & & & I \\ & & I & * \end{bmatrix} \end{array},$$

where $*$ denotes some irrelevant $m_2 \times (\ell - m_2)$ matrix. It is now easy to see that M has full column rank, and thus is nonsingular. ∎

Finally we have a lemma that can be seen as a block generalization of Lemma 6.3.3 and Lemma 6.3.4.

Lemma 6.3.5. *Let $\mathbb{F}$ denote $\mathbb{C}$ or $\mathbb{R}$. If all eigenvalues of $A \in \mathbb{F}^{k\ell \times k\ell}$ have geometric multiplicity at most k, then there exists $X \in \mathbb{F}^{k\ell \times k}$ such that $K_\ell(A, X)$ is nonsingular.*

Proof. We will handle the real and complex case simultaneously. Let $A = ZTZ^{-1}$ be the decomposition from Theorem 6.3.1 or Theorem 6.3.2 and denote the diagonal blocks by T_{ii}, $i = 1\colon s$. For each T_{ii} we define W_i in the following way. If T_{ii} is of size $\ell \times \ell$ take W_i to be the vector in Lemma 6.3.3 such that $K_\ell(T_{ii}, W_i)$ is nonsingular, and if T_{ii} is of size $2\ell \times 2\ell$ take W_i to be the $2\ell \times 2$ matrix whose columns are the two real vectors in Lemma 6.3.4. Letting $W = W_1 \oplus W_2 \oplus \cdots \oplus W_s$ and $X = ZW$ yields $K_\ell(A, X) = ZK_\ell(T, W)$, which is of full rank. ∎

6.3.1 Reduced forms

We now have all the necessary results to construct the matrix X in Theorem 6.2.1 such that properties (i) and (ii) therein hold.

Proposition 6.3.6 (Block-triangular form)**.** *Let $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$. If*

$$A = Z \begin{bmatrix} T_{11} & * & * & * \\ & T_{22} & * & * \\ & & \ddots & * \\ & & & T_{kk} \end{bmatrix} Z^{-1} \in \mathbb{F}^{\ell n \times \ell n}, \tag{6.11}$$

where T_{ii} is of size $s_i\ell \times s_i\ell$ and each eigenvalue of T_{ii} has multiplicity at most s_i, for $i = 1\!:\!k$ with $s_1 + \cdots + s_k = n$ then there exists $X = [X_1\ X_2\ \cdots\ X_k]$ with $X_i \in \mathbb{F}^{n\ell \times s_i}$ such that S in (6.2) is nonsingular and $\mathcal{K}_\ell(A, X_{1:i})$ is A-invariant for $i = 1\!:\!k$.

Proof. By Lemma 6.3.5, we can for each T_{ii} pick a V_i such that $K_\ell(T_{ii}, V_i)$ is nonsingular. Thus, if we form $X = Z(V_1 \oplus V_2 \oplus \cdots \oplus V_k)$, we have that $K_\ell(A, X)$ is nonsingular. Further, if we let $Z_{1:i}$ denote the first $s_1 + s_2 + \cdots + s_i$ columns of Z, then we have

$$\begin{aligned} A\mathcal{K}_\ell(A, X_{1:i}) &= \operatorname{range} AZ_{1:i} \begin{bmatrix} K_\ell(T_{11}, V_1) & * & * & * \\ & K_\ell(T_{22}, V_2) & * & * \\ & & \ddots & * \\ & & & K_\ell(T_{ii}, V_i) \end{bmatrix} \\ &= \operatorname{range} Z_{1:i} \begin{bmatrix} K_\ell(T_{11}, T_{11}V_1) & * & * & * \\ & K_\ell(T_{22}, T_{22}V_2) & * & * \\ & & \ddots & * \\ & & & K_\ell(T_{ii}, T_{ii}V_i) \end{bmatrix} \\ &\subset \mathcal{K}_\ell(A, X_{1:i}), \end{aligned}$$

for $i = 1\!:\!k$. ∎

Remark 6.3.7. The proof of Proposition 6.3.6 provides a means to construct X. From the proof we see that the columns of $X_{1:i}$ must be a basis for the invariant subspace of A corresponding to the eigenvalues of $T_{11}, T_{22}, \ldots, T_{ii}$.

Proposition 6.3.8 (Block-diagonal form)**.** *Let $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$. If*

$$A = Z \begin{bmatrix} D_{11} & & & \\ & D_{22} & & \\ & & \ddots & \\ & & & D_{kk} \end{bmatrix} Z^{-1} \in \mathbb{F}^{\ell n \times \ell n}, \tag{6.12}$$

where D_{ii} is of size $s_i\ell \times s_i\ell$ and each eigenvalue of D_{ii} has multiplicity at most s_i, for $i = 1\colon k$ with $s_1 + \cdots + s_k = n$ then there exists $X = [X_1\ X_2\ \cdots\ X_k]$ with $X_i \in \mathbb{F}^{n\ell \times s_i}$ such that S in (6.2) is nonsingular and $\mathcal{K}_\ell(A, X_i)$ is A-invariant for $i = 1\colon k$.

The proof is similar to that of Proposition 6.3.6 and is omitted. We have the following analogue of Remark 6.3.7.

Remark 6.3.9. The columns of X_i are a basis for the invariant subspace of A corresponding to the eigenvalues of D_{ii}.

Clearly the decomposition in Proposition 6.3.8 is not possible for an arbitrary number of blocks. Indeed, the linear matrix polynomial $I\lambda - J_\alpha$, where

$$J_\alpha = \begin{bmatrix} \alpha & 1 & & \\ & \alpha & \ddots & \\ & & \ddots & 1 \\ & & & \alpha \end{bmatrix}$$

is of size $\ell \times \ell$, cannot be reduced to a block diagonal structure with smaller block size.

Let $I\lambda - A$ be a linearization of $P(\lambda)$ in (6.1). From Theorem 6.2.1 and Proposition 6.3.8, we see that reduction to diagonal form is possible if we can partition the Jordan blocks associated with A into n sets, such that

(a) each set has at most one Jordan block of each eigenvalue, and

(b) the sizes of all Jordan blocks in each set sum up to ℓ.

The result also holds in the opposite direction. That is, it is possible to reduce $P(\lambda)$ to diagonal form, *only if* we can partition the Jordan blocks of A such that (a) and (b) hold. To see this, we simply note that any diagonal monic matrix polynomial $D(\lambda) = d_1(\lambda) \oplus d_2(\lambda) \oplus \cdots \oplus d_n(\lambda)$ has left companion linearization in controller form: $I\lambda - (C_L(d_1) \oplus C_L(d_2) \oplus \cdots \oplus C_L(d_n))$. The following question arises: When is it possible to partition the Jordan blocks such that (a) and (b)

are satisfied? This problem was recently solved by Lancaster and Zaballa [52] for the special case of quadratic matrix polynomials with nonsingular leading matrix coefficient, and by Zúñiga Anaya [93] for general regular quadratics. For matrix polynomial of higher degree the problem is still open.

Proposition 6.3.10 (Hessenberg form)**.** *Let* $\mathbb{F} = \mathbb{C}$ *or* $\mathbb{R}$ *and consider*

$$A = ZHZ^{-1} \in \mathbb{F}^{\ell n \times \ell n}, \tag{6.13}$$

where H *is upper Hessenberg and partitioned in* $\ell \times \ell$ *blocks. Assume that the* $\ell \times \ell$ *diagonal blocks are unreduced. If we let* $X = Z[e_1\ e_{\ell+1}\ \cdots\ e_{(n-1)\ell+1}]$ *then* $K_\ell(A, X)$ *is nonsingular and* $\operatorname{range} A^\ell x_{1:i} \subset \mathcal{K}_\ell(A, x_{1:i+1})$ *for* $i = 1\!:\!n-1$.

Proof. Let $j_{1:i} := [e_1\ e_{\ell+1}\ \cdots\ e_{(i-1)\ell+1}]$. We have $K_\ell(A, X) = ZK_\ell(H, j_{1:n})$, which is obviously nonsingular. Furthermore,

$$\operatorname{range} A^\ell x_{1:i} = \operatorname{range} ZH^\ell j_{1:i} \subset Z\mathcal{K}_\ell(H, j_{1:i+1}) = \mathcal{K}_\ell(A, x_{1:i+1}),$$

completing the proof. ∎

In practice we are interested in Hessenberg decompositions $A = UHU^H$, where U is unitary or real orthogonal, depending on whether we work over $\mathbb{C}$ or $\mathbb{R}$. By the implicit Q-theorem [33, Theorem 7.4.2], the Hessenberg matrix H is uniquely defined, up to signs, by the first column of U. Hence a random Hessenberg matrix similar to A via unitary/real orthogonal transformations, can be constructed using, e.g., the Arnoldi algorithm with a random starting vector. If a matrix has distinct eigenvalues, the resulting Hessenberg matrix will be unreduced with probability one. Since this is the generic case for matrix polynomials, Proposition 6.3.8 may be used to reduce almost all matrix polynomials to Hessenberg form, without further care.

If a matrix on the other hand has an eigenvalue of geometric multiplicity greater than one, then any similar Hessenberg matrix is necessarily reduced. Now, according to Proposition 6.3.8 the reduction to Hessenberg form is still valid if H is reduced, as long as the diagonal $\ell \times \ell$ blocks are unreduced. This means that all zeros on the subdiagonal are in the positions $(\ell+1, \ell), (2\ell+1, 2\ell), \ldots, ((n-1)\ell+1, (n-1)\ell)$. If we have a zero in any other position on the subdiagonal, $K_\ell(A, X)$ becomes singular and the reduction will fail. This raises the following question: is it possible to move zeros on the subdiagonal, from unwanted to wanted positions, using a finite number of Givens rotations or Householder reflectors that are constructed

from the entries in the matrix in the usual manner to introduce zeros? The following argument shows that the answer is in general "no." For an introduction to the algebra the argument is based upon, we refer the reader to [67, Chapter 3] or [25, Section 56]. Assume that the answer to the above question is "yes" and suppose we want to compute all eigenvalues of a Hessenberg matrix $H \in \mathbb{R}^{n\times n}$. To make the argument crystal clear we may think of H as the left companion matrix of an arbitrary monic real polynomial $p(x)$ of degree n. Consider the field $\mathbb{K}_0 := \mathbb{Q}(h_{11}, h_{12}, \ldots, h_{nn})$, that is, the smallest field containing rational numbers and all entries of H. Clearly, it holds that $H \in \mathbb{K}_0^{n\times n}$. When Givens rotations or Householder reflectors are constructed to introduce zeros upon operation on a matrix with entries in a field $\mathbb{F} \subset \mathbb{R}$, they have entries in a *pure (field) extension* $\mathbb{F}(s)$ of $\mathbb{F}$. (By definition, $\mathbb{F}(s)$ is a pure extension of $\mathbb{F}$ if $s^m \in \mathbb{F}$ for some integer $m \geq 1$.) Suppose, for example, we want to eliminate the second component of $[x, y]^T \in \mathbb{F}^2 \subset \mathbb{R}^2$ using a Givens rotation G. Then

$$G = \frac{1}{\sqrt{x^2 + y^2}} \begin{bmatrix} x & y \\ -y & x \end{bmatrix},$$

so we have $G[x, y]^T \in \mathbb{K}(\sqrt{x^2 + y^2})^2$. Thus, if we start with $H \in \mathbb{K}_0^{n\times n}$, one application of a Givens rotation or Householder reflector—that are constructed in the usual manner—moves our matrix to a larger set $\mathbb{K}_1^{n\times n} := \mathbb{K}_0(\sqrt{x})^{n\times n}$, for some $x \in \mathbb{K}_0$. Thus, if we apply t, say, Givens rotations or Householder reflectors to our initial matrix H, we extend the field, in which our matrix entries lie, t times, using only pure extensions. If $\mathbb{K}_i$ denotes the relevant pure extension of $\mathbb{K}_{i-1}$, then we have

$$\mathbb{K}_0 \subseteq \mathbb{K}_1 \subseteq \cdots \subseteq \mathbb{K}_t.$$

The final field $\mathbb{K}_t$ is therefore (by definition) a *radical extension* of $\mathbb{K}_0$.

Now, consider $H \oplus 1$ and note that $H \oplus 1$ is a reduced Hessenberg matrix with a zero on the last entry on the first subdiagonal. Suppose we can move this zero to the middle of the subdiagonal using Givens rotations or Householder reflectors that are constructed in the usual manner. We get a similar matrix of the form

$$\begin{bmatrix} H_{11} & H_{12} \\ & H_{22} \end{bmatrix},$$

where H_{11} and H_{22} are square. The spectrum of $H \oplus 1$ is the union of the spectra of H_{11} and H_{22}, so we may continue by computing the eigenvalues of these smaller matrices. If we continue recursively, by forming $H_{11}\oplus 1$ and $H_{22}\oplus 1$, and successively

split the spectra, then we eventually end up with several small eigenvalue problems of size at most 4×4. These small matrices all have entries in a radical extension $\mathbb{K}_t$ of $\mathbb{K}_0$. Furthermore, since the associated characteristic polynomials are of degree at most 4, their eigenvalues lie in a radical extension of $\mathbb{K}_t$. In other words, all roots of the characteristic polynomial of H, which is $p(x)$, lie in a radical extension of $\mathbb{K}_0$. Thus, $p(x)$ is (by definiton) *solvable by radicals* over $\mathbb{K}_0$. Since $p(x)$ is arbitrary, this contradicts the Abel-Ruffini theorem, which states that for $\ell \geq 5$ there are polynomials $a_\ell x^\ell + a_{\ell-1}x^{\ell-1} + \cdots + a_0$ that are not solvable by radicals over $\mathbb{Q}(a_0, a_1, \ldots, a_\ell)$. A classic example is $x^5 - x + 1$.

We remark that the argument does not depend on the fact that $H \oplus 1$ has the zero we want to move in the last position on the subdiagonal. Indeed, the argument still holds if we replace $H \oplus 1$ by, say, $H \oplus A$, for any $A \in \mathbb{Q}^{2\times 2}$.

We end this chapter with two remarks about the usage of the Abel-Ruffini theorem in the field of numerical linear algebra.

Remark 6.3.11. The Abel-Ruffini theorem is commonly mentioned as an argument against the existence of a general *direct* eigenvalue algorithm. If we only consider algorithms that perform additions, subtractions, multiplications, divisions, and root extractions, then the argument is indeed valid. This is clearly the case for any algorithm that only applies a finite number of Givens rotations or Householder reflectors, that are implemented in the usual manner to introduce zeros. We cannot, however, reason about an eigenvalue algorithm, that computes, say, a logarithm at some point, using the Abel-Ruffini theorem.

Remark 6.3.12. As it is stated above, the Abel-Ruffini theorem implies that we cannot write down a general (finite) formula for the roots of polynomials of degree $\ell \geq 5$ using only $+$, $-$, $\times$, $\div$, $\sqrt[k]{}$ with $k \in \mathbb{N}$, the rational numbers and the polynomial coefficients. This weaker statement is sometimes also referred to as the Abel-Ruffini theorem. However, the non-existence of such a general formula for the roots is of less interest to the field of numerical linear algebra. This weaker result does not even rule out the existence of a direct algorithm that *only* performs, say, Givens rotations to eliminate matrix entries in the usual manner. The subtlety here is that each polynomial could (hypothetically) have its own set of formulas for its roots, even though there are no *general* formulas that applies to all polynomials of a certain degree. For instance, we need the stronger form of the Abel-Ruffini theorem stated above to rule out the existence of a Givens rotation based direct eigenvalue algorithm that makes use of a (nontrivial) pivoting strategy, or contains any other conditional statements that depend on the input matrix.

Chapter

7

Error analysis of the shift-and-invert Arnoldi algorithm

7.1 Introduction

Consider an implementation of the Arnoldi method [7, 90]. Not much meaning can be given to the computed quantities if they deviate too much from the recurrence that underpins the algorithm in exact arithmetic:

$$AV_k = V_{k+1}\underline{H}_k, \quad \underline{H}_k = H(1:k+1, 1:k).$$

Luckily, good implementations, where in particular the orthogonalization is done with care, can be shown to be backward stable [6, 21, 29, 70] in the sense that the computed quantities V_{k+1} and $\underline{H}_k$ satisfy an exact recurrence with a slightly perturbed matrix:

$$(A + \Delta A)V_k = V_{k+1}\underline{H}_k. \tag{7.1}$$

This means that we can compute a basis of an exact Krylov subspace corresponding to a nearby matrix. Since the basis will in general not be perfectly orthonormal, so $V_{k+1}^H V_{k+1} \neq I$, we use the term "Krylov recurrence" instead of "Arnoldi recurrence" when referring to recurrences like (7.1). If A is Hermitian, then it can be shown that the computed basis spans a Krylov subspace associated with a perturbed *Hermitian* matrix $A + \Delta A$ [44]. There is a catch in this case, though: the small $(k+1) \times k$ matrix associated with this Krylov subspace is in general not the computed Hessenberg matrix.

In this chapter we perform a similar backward error analysis of the shift-and-invert Arnoldi algorithm. For example, we show that an implementation of the Arnoldi method applied to A^{-1}, yields computed matrices V_{k+1} and $\underline{H}_k$ such that

$$(A + \Delta A)^{-1} V_k = V_{k+1} \underline{H}_k,$$

and we give an upper bound for $\|\Delta A\|_2$. Perturbed versions of the shift-and-invert Arnoldi algorithm have been considered in the literature as a part of the theory of *inexact methods*, see [54, 61]. However, these results neglect that the orthonormalization is not performed exactly, and furthermore, assume bounds on linear system residuals that may be unattainable (more on this in Section 7.2). We consider more general linear system residuals and take the error from the orthonormalization into account. Our analysis of how the orthonormalization errors propagate into the shift-and-invert Krylov recurrence highlights the importance of *columnwise* backward error bounds for QR factorizations, and is thus of a different flavor than the corresponding analysis for standard Arnoldi, done in, for example [21].

We also use our error analysis to motivate when "breakdown" should be declared, that is, when $h_{j+1,j}$ may be considered to be "numerically zero."

The algorithm we study can be divided into two main subproblems: solving linear systems and orthonormalizing vectors. We state our backward error results in such a way that they are independent of how these subproblems are being solved, but we also discuss relevant and commonly used approaches for solving these two tasks.

7.1.1 Technical outline

We study floating point implementations of Algorithm 7.1, where A is assumed to be of size $n \times n$, σ is the shift, b is the starting vector, and k is the maximum number of steps we perform. Throughout the chapter $\|\cdot\|$ refers to the spectral norm. In exact arithmetic, the function on line 4 of Algorithm 7.1 is defined as

$$\text{orthogonalization}(w_j, V_j) := [w_j - V_j(V_j^H w_j), V_j^H w_j],$$

which corresponds to classical Gram-Schmidt if implemented as it stands. In floating point arithmetic, however, orthogonalization routines with better numerical properties, such as modified Gram-Schmidt, are usually employed.

In the jth iteration in Algorithm 7.1, a new vector w_j is computed and

Algorithm 7.1: The Shift-and-invert Arnoldi algorithm

Input: A, σ, b, k
Output: $V_{k+1} := [v_1, \ldots, v_{k+1}]$, $\underline{H}_k = [h_{ij}]_{i=1:k+1, j=1:k}$

```
1 v_1 = b/||b||
2 for j = 1, 2, ..., k
3     w_j = (A − σI)^{-1} v_j
4     [w'_j, h_{1:j}] = orthogonalization(w_j, V_j)
5     h_{j+1 j} = ||w'_j||
6     if h_{j+1 j} = 0 break
7     v_{j+1} = w'_j / h_{j+1 j}
8 end for
```

decomposed into a linear combination of $v_1, \ldots, v_j$ and a new component that will be the definition of v_{j+1}. In exact arithmetic, this can be described by the Arnoldi recurrence $(A - \sigma I)^{-1} v_j = V_k h_{1:j,j} + h_{j+1,j} v_{j+1}$. When the corresponding computation is done in practice, however, errors are present in all steps of the computation. First, we need to solve a linear system. If we use a direct solver the matrix $A - \sigma I$ needs to be formed. We consider the rounding error in this step as part of the residual from the linear system. This does not affect the norm of the residual significantly, because the rounding error is very small,

$$\|\text{float}(A - \sigma I) - (A - \sigma I)\| < \max_{1 \le i \le n} |a_{ii} - \sigma| u \le u \|A - \sigma I\|.$$

Here $\text{float}(A - \sigma I)$ refers to the computed shifted matrix and u is the unit roundoff. Let r_j be the residual from the linear system, so

$$(A - \sigma I) w_j = v_j + r_j \tag{7.2}$$

is the actual linear system that has been solved. Then we have the following equality for the computed quantities:

$$(A - \sigma I)^{-1}(v_j + r_j) = w_j = V_{j+1} h_{1:j+1,j} + g_j,$$

where g_j is an error coming from the orthonormalization process. Defining $f_j = r_j - (A - \sigma I) g_j$ and $F_k = [f_1 \; f_2 \; \cdots \; f_k]$ yields a perturbed recurrence

$$(A - \sigma I)^{-1}(V_k + F_k) = V_{j+1} \underline{H}_k.$$

We discuss the residual r_j and the error g_j in Section 7.2 and Section 7.3, respectively, and provide bounds for both quantities. In Section 7.4, we use these bounds in order to bound F_k, and subsequently the backward error for the shift-and-invert Krylov recurrence. In Section 7.5, we explain how the idea of implicit restarting can be used to gain further insight into the backward error. We also discuss in what sense we have Hermitian backward errors if the method is applied to a Hermitian matrix A. Finally, we talk about breakdown conditions: in floating point arithmetic, the test if $h_{j+1,j} = 0$ in Algorithm 7.1 is rarely done. Instead one usually checks whether $h_{j+1,j}$ is "small enough." This case is referred to as *breakdown*. A sensible definition of "small enough" is when the quantity is dominated by errors. We discuss this in more detail and derive backward error bounds for this case.

7.1.2 Notation

The scalar σ refers to a shift while $\sigma_{\min}(X)$ refers to the smallest singular value of X. The dagger notation $X^\dagger$ refers to the Moore-Penrose pseudo-inverse of X. The lower letter u is reserved to denote the unit roundoff if real arithmetic is used, and $\sqrt{5}$ times the unit roundoff if complex arithmetic is used (see Appendix B). When the matrix size is understood from the context, we denote zero matrices and identity matrices as 0 and I, respectively. Similarly, the vector e_i denotes the ith column of the identity matrix whose size is understood from the context. For a matrix X, the lower case x_i refers to the ith column of X and X_k to $[x_1\ x_2\ \cdots\ x_k]$, that is, the first k columns of X.

7.2 Errors from linear systems

In this section we discuss bounds on the residual r_j from (7.2).

7.2.1 Backward error bounds

The normwise backward error associated with a computed solution y of a linear $Ax = b$ is defined as

$$\eta_{A,b}(y) := \min\{\epsilon : (A + \Delta A)y = b + \Delta b, \|\Delta A\| \leq \epsilon\|A\|, \|\Delta b\| \leq \epsilon\|b\|\},$$

and given by the formula

$$\eta_{A,b}(y) = \|r\|/(\|A\|\|y\| + \|b\|) \tag{7.3}$$

where $r = Ay - b$ [66]. See also [36, p. 120]. This result is true for any vector norm $\|\cdot\|$ and its subordinate matrix norm. Thus, if we solve the linear systems in Algorithm 7.1, up to a backward error ϵ_{bw}, then it holds that

$$\|r_j\| \leq (\|A - \sigma I\|\|w_j\| + \|v_j\|)\epsilon_{\text{bw}}, \tag{7.4}$$

where r_j is defined in (7.2). If the linear systems are solved by a backward stable direct method, we have $\epsilon_{\text{bw}} \leq \phi(n)u$, where $\phi(n)$ is an algorithm dependent constant. If we are interested in the smallest possible ϵ_{bw} such that (7.4) holds, then we need to compute $\|r_j\|/(\|A - \sigma I\|\|w_j\| + \|v_j\|)$. However, this may not be feasible for the spectral norm, due to the term $\|A - \sigma I\|$. In these cases we can replace $\|A - \sigma I\|$ by a lower bound (the tighter the better), and thus obtain an upper bound for ϵ_{bw}. We can for instance do a few iterations of the power method applied to $(A - \sigma I)^H(A - \sigma I)$. MATLAB's `normest` function does exactly this. This would lead to a lower bound of $\|A - \sigma I\|$, since convergence is always from below. Another possibility is to use the (lower) bound in [39]. We can also bound the matrix spectral norm in terms of the corresponding infinity-norm or 1-norm. The following proposition shows that such bounds can be satisfactory for many sparse matrices, in particular those which can be permuted to banded form.

Proposition 7.2.1. *Let k_{row} and k_{col} denote the maximum number of nonzero entries in a row and column of A, respectively. Then the following two upper and lower bounds hold:*

$$\begin{aligned}
\frac{1}{\sqrt{k_{\text{col}}}}\|A\|_2 \;&\leq\; \|A\|_\infty \;\leq\; \sqrt{k_{\text{row}}}\|A\|_2, \\
\frac{1}{\sqrt{k_{\text{row}}}}\|A\|_2 \;&\leq\; \|A\|_1 \;\leq\; \sqrt{k_{\text{col}}}\|A\|_2.
\end{aligned}$$

Proof. We have $\|A\|_\infty = \|Ax\|_\infty$ for some x with $\|x\|_\infty = 1$ and at most k_{row} nonzeros. We get

$$\|A\|_\infty \leq \|Ax\|_\infty \leq \|Ax\|_2 \leq \|A\|_2\|x\|_2 \leq \sqrt{k_{\text{row}}}\|A\|_2,$$

which is the desired upper bound for $\|A\|_\infty$. Further, we have

$$\|A\|_1 = \|A^T\|_\infty \leq \sqrt{k_{\text{col}}}\|A^T\|_2 = \sqrt{k_{\text{col}}}\|A\|_2,$$

which is the desired upper bound for $\|A\|_1$.

The lower bounds follow from [86, Theorem 4.2]. ∎

The inequality (7.4) can also be used as a stopping criterion for iterative linear system solvers [5]. In this case, ϵ_{bw} denotes the desired backward error, which is given prior to execution. If we replace $\|A - \sigma I\|$ with a lower bound, then we get a more stringent stopping criterion.

7.2.2 Residual reduction bounds

An alternative to (7.4) is to use the bound

$$\|r_j\| \leq \|v_j\|\epsilon_{\text{tol}}. \tag{7.5}$$

This bound is commonly used as a stopping condition when the linear systems are solved by iterative methods. Unfortunately, as a stopping condition, (7.5) "may be very stringent, and possibly unsatisfiable" [36, p. 336]. See also [22, pp. 72–73] for a 2×2 example that illustrates the pitfall of comparing the norm of the residual with the norm of the right hand side. However, since (7.5) is de facto commonly used in computer codes it is still worth to study it under the assumption that the stopping criterion is met.

7.2.3 Auxiliary residual bounds

In order to treat both (7.4) and (7.5) in a unified way, we consider the following auxiliary bound

$$\|r_j\| \leq \|v_j\|\epsilon_1 + \|A - \sigma I\|\|w_j\|\epsilon_2. \tag{7.6}$$

Clearly, the substitutions $(\epsilon_1, \epsilon_2) \leftarrow (\epsilon_{\text{bw}}, \epsilon_{\text{bw}})$ and $(\epsilon_1, \epsilon_2) \leftarrow (\epsilon_{\text{tol}}, 0)$ give back (7.4) and (7.5), respectively. We can simplify the bound in (7.6) in cases when $A - \sigma I$ is not too ill-conditioned with respect to ϵ_2. To see this we need the following lemma.

Lemma 7.2.2. *If* $\kappa(A - \sigma I)\epsilon_2 < 1$ *and* (7.6) *hold, then*

$$\|r_j\| \leq \frac{\epsilon_1 + \kappa(A - \sigma I)\epsilon_2}{1 - \kappa(A - \sigma I)\epsilon_2}\|v_j\|.$$

Proof. We have

$$\begin{aligned}\|r_j\| &\leq \|A-\sigma I\|\|(A-\sigma I)^{-1}(v_j+r_j)\|\epsilon_2 + \|v_j\|\epsilon_1 \\ &\leq \kappa(A-\sigma I)\|v_j+r_j\|\epsilon_2 + \|v_j\|\epsilon_1 \\ &\leq \kappa(A-\sigma I)(\|v_j\|+\|r_j\|)\epsilon_2 + \|v_j\|\epsilon_1.\end{aligned}$$

Reordering gives the result. ∎

The next result yields a family of new residual bounds independent of $\|v_j\|$.

Proposition 7.2.3. *Let $(A-\sigma I)^{-1}(v_j+r_j) = w_j$ and assume (7.6) hold. If*

$$0 < \frac{\epsilon_1 + \kappa(A-\sigma I)\epsilon_2}{1-\kappa(A-\sigma I)\epsilon_2} \leq \gamma < 1, \tag{7.7}$$

then

$$\|r_j\| \leq \left(\epsilon_2 + \frac{\epsilon_1}{1-\gamma}\right)\|A-\sigma I\|\|w_j\|.$$

Proof. From (7.6) we have

$$\|r_j\| \leq \left(\epsilon_2 + \epsilon_1\frac{\|v_j\|}{\|A-\sigma I\|\|w_j\|}\right)\|A-\sigma I\|\|w_j\|.$$

Thus we need to show $\|v_j\|/(\|A-\sigma I\|\|w_j\|) \leq 1/(1-\gamma)$. We have

$$\frac{\|v_j\|}{\|A-\sigma I\|\|w_j\|} = \frac{\|v_j\|}{\|A-\sigma I\|\|(A-\sigma I)^{-1}(v_j+r_j)\|} \leq \frac{\|v_j\|}{\|v_j+r_j\|},$$

and from the reverse triangle inequality,

$$\frac{\|v_j\|}{\|v_j+r_j\|} \leq \frac{\|v_j\|}{|\|v_j\|-\|r_j\||}.$$

Now, by Lemma 7.2.2 and assumption (7.7), we have

$$\|r_j\| \leq \frac{\epsilon_1+\kappa(A-\sigma I)\epsilon_2}{1-\kappa(A-\sigma I)\epsilon_2}\|v_j\| \leq \gamma\|v_j\|.$$

Putting everything together yields

$$\frac{\|v_j\|}{\|A-\sigma I\|\|w_j\|} \leq \frac{\|v_j\|}{|\|v_j\|-\|r_j\||} \leq \frac{1}{1-\gamma}.$$ ∎

In particular if $\kappa(A-\sigma I) \leq (1-2\epsilon_1)/(3\epsilon_2)$, then we have $\kappa(A-\sigma I)\epsilon_2 < 1$

and can take $\gamma = 1/2$ in Proposition 7.2.3, to obtain

$$\|r_j\| \leq (2\epsilon_1 + \epsilon_2)\|A - \sigma I\|\|w_j\|. \tag{7.8}$$

This is the same bound as we get from (7.6) if we replace (ϵ_1, ϵ_2) with $(0, 2\epsilon_1 + \epsilon_2)$. In particular, if the linear systems are solved in a backward stable manner so that (7.4) holds, and $\kappa(A-\sigma I) \leq (1-2\epsilon_{\text{bw}})/(3\epsilon_{\text{bw}})$, then (7.8) holds with $2\epsilon_1+\epsilon_2 = 3\epsilon_{\text{bw}}$.

7.3 Errors from orthonormalization

In this section we are concerned with the orthonormalization error

$$g_j = w_j - V_{j+1}h_{1:j+1,j}.$$

Up to signs, this error can be viewed as the backward error in the $(j+1)$st column of a perturbed QR factorization

$$[v_1\ w_1\ w_2\ \cdots\ w_k] = V_{k+1}[e_1\ \underline{H}_k] + [0\ g_1\ g_2\ \cdots\ g_k]. \tag{7.9}$$

Thus, we are interested in *columnwise* backward error bounds for QR factorizations. The next theorem shows how such bounds can be obtained from normwise backward error bounds given in the spectral norm or the Frobenius norm. It applies to floating point algorithms $\texttt{qr}(\cdot)$ that are unaffected by power-of-two column scalings, in the sense that if $[Q, R] = \texttt{qr}(A)$, then $[Q, RD] = \texttt{qr}(AD)$ for any $D = \text{diag}(d_1, d_2, \ldots, d_k)$ where the d_i are powers of 2. Barring underflow and overflow, this covers commonly used algorithms such as classical and modified Gram-Schmidt with and without (possibly partial) reorthogonalization, Householder QR and Givens QR.

Theorem 7.3.1. *Let $\texttt{qr}(A)$ denote an algorithm that computes an approximate QR factorization of an $n \times k$ matrix A in floating point arithmetic. Suppose further that $[Q, RD] = \texttt{qr}(AD)$ for any $D = \text{diag}(d_1, d_2, \ldots, d_k)$ where the d_i are powers of 2. If Q and R denote the computed factors, $\Delta A = A - QR$ and $\|\Delta A\|_* \leq \gamma\|A\|_* u$, where $\|\cdot\|_*$ denotes the spectral norm or the Frobenius norm, then $\|\Delta a_i\| \leq 2\gamma\sqrt{k}\|a_i\|u$ for $i = 1\!:\!k$.*

Proof. For $i = 1\!:\!k$, we define $d_i = 2^{-\lfloor \log_2 \|a_i\| \rfloor}$, so $1 \leq \|a_i\|d_i < 2$. Since ΔAD is

the backward error from $\mathtt{qr}(AD)$ we have

$$d_i\|\Delta a_i\| = \|\Delta AD e_i\| \le \|\Delta AD\|_* \le \gamma\|AD\|_* u < 2\gamma\sqrt{k}\|AD e_i\|u = d_i 2\gamma\sqrt{k}\|a_i\|u,$$

for $i = 1:k$, from which the theorem follows. ∎

The constant γ in Theorem 7.3.1 is obviously algorithm dependent and many bounds for it exist in the literature. Some of them contain both n and k [72], and others only k [12, 2], [36, Theorem 19.13]. In [36, p. 361] a columnwise bound depending on n and k is given. For Krylov methods we usually have $n \gg k$, so bounds independent from n should certainly be favored. We shall assume that

$$\|g_j\| \le \eta(n,k)\|w_j\|u, \tag{7.10}$$

holds for some function $\eta(n,k)$.

7.3.1 Columnwise errors in modified Gram-Schmidt

Our next theorem shows that for modified Gram-Schmidt (MGS), with and without one round of reorthogonalization, η in (7.10) does not depend on n and is given by

$$\eta(n,k) = \zeta k,$$

where ζ is a modest constant. We need the following forward error result for _axpy operations.

Lemma 7.3.2. *Let α be a scalar and x and y vectors. If*

$$s = \mathrm{float}(\alpha x + y) - (\alpha x + y) \quad \textit{then} \quad \|s\| \le 2(\|\alpha x\| + \|y\|)u.$$

Proof. The ith component of $\alpha x + y$ can be viewed as the inner product $[x_i\ y_i][\alpha\ 1]^T$. Thus the componentwise forward error is bounded by $|s| \le 2u(|\alpha x| + |y|)$ [43] (see also Appendix B). We get $\|s\| \le \|2u(|\alpha x| + |y|)\| \le 2(\|\alpha x\| + \|y\|)u$. ∎

The next theorem gives columnwise backward error bounds for MGS with and without one round of reorthogonalization.

Theorem 7.3.3. *Let Q and R denote the computed factors in the QR decomposition of an $n \times k$ matrix A, which was obtained by a floating point implementation of modified Gram-Schmidt with or without one round of reorthogonalization. Assume*

(i) $\|q_j\| = 1$ *for* $j = 1\!:\!k$, *and*

(ii) $(1+(n+3)u)^k < 1+\delta$ *for some* $\delta > 0$.

Then there exists a ΔA *such that* $A + \Delta A = QR$ *with* $\|\Delta a_j\| \leq cj\|a_j\|u$, *where* $c = 4(1+\delta)$ *if no reorthogonalization was done and* $c = 10(1+\delta)^2$ *if one round of reorthogonalization was done.*

Let us pause for a while and discuss the assumptions before we proceed with the proof. Assumption (i) is imposed to keep our analysis cleaner; it does not affect our final bounds in any significant way. Assumption (ii) is needed for the following reason: if we compute $y = \text{float}(x - q_j(q_j^H x))$ for some $1 \leq j \leq k$, then, assuming (i), the quantity $1+(n+3)u = 1+\|q_j\|^2(n+3)u$ is an upper bound for $\|y\|/\|x\|$ [36, Lemma 3.9]. Thus, (ii) guarantees that we can apply a sequence of k elementary "floating point" projections of the form $I - q_i q_i^H$ to any vector x, and the resulting vector will be bounded in norm by $(1+\delta)\|x\|$.

Proof of Theorem 7.3.3. Let $R^{(1)}$ and $R^{(2)}$ denote the *strictly* upper triangular matrices containing the orthogonalization coefficients corresponding to the first and second round of orthogonalization, respectively. We define $R^{(2)} \equiv 0$, if no reorthogonalization is done. Assume for a while that $R^{(1)}$ and $R^{(2)}$ are given, and suppose we want to compute

$$a_j - \sum_{i=1}^{j-1} r_{ij}^{(1)} q_i - \sum_{i=1}^{j-1} r_{ij}^{(2)} q_i.$$

This can be viewed as $2(j-1)$ _axpy operations. We define $a_j^{(0)} = a_j$ and

$$a_j^{(i)} = \begin{cases} \text{float}(a_j^{(i-1)} - r_{ij}^{(1)} q_i) & \text{for } i = 1\!:\!j-1, \\ \text{float}(a_j^{(i-1)} - r_{(i-j+1)j}^{(2)} q_{i-j+1}) & \text{for } i = j\!:\!2(j-1). \end{cases}$$

Using Lemma 7.3.2 yields

$$a_j^{(i)} = \begin{cases} a_j^{(i-1)} - r_{ij}^{(1)} q_i + s_i & \text{for } i = 1\!:\!j-1, \\ a_j^{(i-1)} - r_{(i-j+1)j}^{(2)} q_{i-j+1} + s_i & \text{for } i = j\!:\!2(j-1), \end{cases}$$

where

$$\|s_i\| \leq \begin{cases} 2(\|r_{ij}^{(1)} q_i\| + \|a_j^{(i-1)}\|)u & \text{for } i = 1\!:\!j-1, \\ 2(\|r_{(i-j+1)j}^{(2)} q_{i-j+1}\| + \|a_j^{(i-1)}\|)u & \text{for } i = j\!:\!2(j-1). \end{cases}$$

Now, $a_j^{(i-1)}$ is also the result of applying $i-1$ elementary floating point projections

to a_j, so the discussion prior to the proof gives $\|a_j^{(i-1)}\| < (1+(n+3)u)^{i-1}\|a_j\|$. Further, from (ii) we have $(1+nu)\|a_j^{(i-1)}\| < (1+\delta)\|a_j\|$ for $i = 1\!:\!j-1$ and $(1+nu)\|a_j^{(i-1)}\| < (1+\delta)^2\|a_j\|$ for $i = j\!:\!2(j-1)$. The forward error of a computed inner product $\mathrm{float}(x^H y)$, where x and y are of length n, is bounded by $nu\|x\|\|y\|$ [43]. See also Appendix B. It follows that

$$|r_{ij}^{(1)}| \le |\mathrm{float}(q_i^H a_j^{(i-1)})| \le |q_i^H a_j^{(i-1)}| + nu\|a_j^{(i-1)}\| < (1+\delta)\|a_j\|$$

and, similarly, that $|r_{ij}^{(2)}| < (1+\delta)^2\|a_j\|$. Thus s_i is bounded by

$$\|s_i\| \le \begin{cases} 4(1+\delta)\|a_j\|u & \text{for } i = 1\!:\!j-1, \\ 4(1+\delta)^2\|a_j\|u & \text{for } i = j\!:\!2(j-1). \end{cases}$$

We have

$$a_j - \sum_{i=1}^{j-1} r_{ij}^{(1)} q_i - \sum_{i=1}^{j-1} r_{ij}^{(2)} q_i = a_j^{(2(j-1))} - \sum_{i=1}^{2(j-1)} s_i.$$

If we define $d_i = \mathrm{float}(\|a_j^{(2(j-1))}\|)$ and $q_j = \mathrm{float}(a_j^{(2(j-1))}/d_j)$ and note that

$$a_j^{(2(j-1))} = q_j d_j + f_j \quad \text{with} \quad \|f_j\| \le \|a_j^{(2(j-1))}\|u < (1+\delta)^2\|a_j\|u,$$

then we get

$$a_j - \sum_{i=1}^{j-1} (r_{ij}^{(1)} + r_{ij}^{(2)}) q_i - d_j q_j = f_j - \sum_{i=1}^{2(j-1)} s_i.$$

Finally, defining $R = \mathrm{float}(R^{(1)} + R^{(2)}) + \mathrm{diag}(d_1, d_2, \ldots, d_k)$ yields

$$\Delta a_j := a_j - \sum_{i=1}^{j} r_{ij} q_i = f_j - \sum_{i=1}^{2(j-1)} s_i - \sum_{i=1}^{j-1} \Delta r_{ij} q_i$$

where

$$\Delta r_{ij} = r_{ij}^{(1)} + r_{ij}^{(2)} - r_{ij}, \quad \text{so} \quad |\Delta r_{ij}| \le |r_{ij}^{(1)} + r_{ij}^{(2)}|u < 2(1+\delta)^2\|a_j\|u.$$

Using the above bounds for f_j, the s_i and Δr_{ij} gives $\|\Delta a_j\| < 10(1+\delta)^2 j\|a_j\|u$. If no reorthogonalization was done, then we have $s_i = 0$ for $i = j\!:\!2(j-1)$, and $\Delta r_{ij} = 0$, $\|f_j\| \le (1+\delta)\|a_j\|u$ for all j. Taking this into account yields $\|\Delta a_j\| < 4(1+\delta)j\|a_j\|u$. ∎

Remark 7.3.4. Suppose the perturbed QR factorization (7.9) was computed

using MGS. Then, taking $\delta = 1/10$ and assuming the conditions of Theorem 7.3.3 hold, yield that $\eta(n,k)$ in (7.10) is bounded by $\eta(n,k) \leq 5k$ if standard MGS is used, and $\eta(n,k) \leq 13k$ if MGS with one round of reorthogonalization is used. We point out that these bounds should not be interpreted as saying that standard MGS should be favored over MGS with reorthogonalization. In this context of shift-and-invert Arnoldi, the difference between the constants 5 and 13 is not significant, and, as we will see in the next section, retaining a well-conditioned basis (which is the effect of reorthogonalization) is of great importance to the shift-and-invert Arnoldi algorithm.

7.4 Errors in the shift-and-invert Arnoldi recurrence

Recall the perturbed Krylov recurrence

$$(A - \sigma I)^{-1}(V_k + F_k) = V_{j+1}\underline{H}_k, \tag{7.11}$$

where $F_k = [f_1\ f_2\ \cdots\ f_k]$ and f_j, for $j = 1\!:\!k$, is defined by $f_j = r_j - (A - \sigma I)g_j$. We discussed in sections 7.2 and 7.3 how to bound r_j and g_j, respectively. By using these bounds, we can now easily bound F_k. Assuming (7.6) and (7.10) yields

$$\|f_j\| \leq \|v_j\|\epsilon_1 + \|A - \sigma I\|\|w_j\|(\epsilon_2 + \eta(n,j)u). \tag{7.12}$$

Further, from (7.9) we see that

$$\|w_j\| = \|V_{j+1}h_{1:j+1,j} + g_j\| \leq \|V_{j+1}\|\|h_{1:j+1,j}\| + \eta(n,j)\|w_j\|u,$$

which in turn implies

$$\|w_j\| \leq \frac{\|V_{j+1}\|\|h_{1:j+1,j}\|}{1 - \eta(n,j)u},$$

assuming that $\eta(n,j)u < 1$. We get

$$\|f_j\| \leq \|v_j\|\epsilon_1 + \|A - \sigma I\|\|V_{j+1}\|\|h_{1:j+1,j}\|c_{jn}(\epsilon_2)$$

and further (assuming that $\eta(n,k)$ is monotonically increasing in k)

$$\|F_k\| \leq \sqrt{k}\|V_k\|\epsilon_1 + \sqrt{k}\|A - \sigma I\|\|V_{k+1}\|\|\underline{H}_k\|c_{kn}(\epsilon_2), \tag{7.13}$$

where

$$c_{kn}(\epsilon_2) := \frac{\epsilon_2 + \eta(n,k)u}{1 - \eta(n,k)u} \tag{7.14}$$

should be thought of as a tiny factor.

Similarly, if we assume the bound (7.8) instead of (7.6), we get

$$\|F_k\| \leq \sqrt{k}\|A - \sigma I\|\|V_{k+1}\|\|\underline{H}_k\|c_{kn}(2\epsilon_1 + \epsilon_2). \tag{7.15}$$

This is the same bound we get from (7.13) if we replace (ϵ_1, ϵ_2) by $(0, 2\epsilon_1 + \epsilon_2)$.

Having established (7.13) and (7.15), we are now ready to reshuffle equation (7.11) in order to derive backward error bounds for the shift-and-invert Krylov recurrence. We will derive perturbed recurrences of the form

$$V_k = (A + \Delta A - \sigma I)V_{k+1}\underline{H}_k. \tag{7.16}$$

If we look at this from a backward error perspective, (7.16) means that we have taken k steps, without errors, of a shift-and-invert Krylov algorithm applied to a perturbed matrix, and all linear systems that occurred in the process must have been consistent. However, in order to rewrite (7.16) as

$$(A + \Delta A - \sigma I)^{-1}V_k = V_{k+1}\underline{H}_k,$$

we need to ensure that $A + \Delta A - \sigma I$ is invertible. We need the following lemma to solve this technicality.

Lemma 7.4.1. *Let A and V be matrices of size $n \times n$ and $n \times k$ respectively, such that* $\operatorname{rank} AV = k$*. Then for any $\epsilon > 0$, there exists a matrix X with $\|X\| < \epsilon$ such that $A + X$ is nonsingular and $XV = 0$. Furthermore, if A is Hermitian, then we may take X to be Hermitian.*

Proof. Find a unitary matrix Q such that

$$Q^H V = \begin{bmatrix} 0 \\ V_2 \end{bmatrix} \tag{7.17}$$

for some $k \times k$ matrix V_2, and define $AQ = [A_1\ A_2]$ where A_2 is of size $n \times k$. From $\operatorname{rank} AV = k$, it follows A_2 has rank k. Define Y so its columns span the orthogonal complement to range of A_2, and set $Z = [Y - A_1\ 0]$. We have that $A + ZQ^H = [Y\ A_2]Q^H$ is nonsingular and $ZQ^HV = 0$. In particular, this means that the pencil $A + \lambda ZQ^H$ is regular. If λ is any value outside the spectrum of

the pencil such that $|\lambda| < \epsilon/\|Z\|$, then $X = \lambda Z Q^H$ satisfies the conditions of the theorem.

For the second part, suppose A is Hermitian and Q is such that (7.17) holds. Write

$$Q^H A Q = \begin{bmatrix} A_{11} & A_{12} \\ A_{12}^H & A_{22} \end{bmatrix} \quad \text{and} \quad W = \begin{bmatrix} \omega I - A_{11} & 0 \\ 0 & 0 \end{bmatrix}, \quad \omega > 0,$$

where A_{11} is of size $(n-k) \times (n-k)$. We have that QWQ^H is Hermitian, $QWQ^H V = 0$, and $Q(Q^H A Q + W)Q^H = A + QWQ^H$. Thus, for the same reason as above, it is enough to find one $\omega > 0$ such that $Q^H A Q + W$ is nonsingular. Let

$$A_{22} = U \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} U^H$$

be a spectral decomposition so D is of full rank and define $[B_1 \; B_2] = A_{12} U$ such that B_1 has as many columns as D. We have that $Q^H A Q + W$ is nonsingular if and only if

$$\begin{bmatrix} \omega I & B_1 & \omega B_2 \\ B_1^H & D & 0 \\ \omega B_2^H & 0 & 0 \end{bmatrix}$$

is nonsingular. Further, since $[A_{12}^T \; A_{22}^T]^T$ is of full rank, and

$$\begin{bmatrix} B_1 & B_2 \\ D & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & U^H \end{bmatrix} \begin{bmatrix} A_{12} \\ A_{22} \end{bmatrix} U,$$

it follows that B_2 is also of full rank. We have

$$\begin{bmatrix} \omega I & B_1 & \omega B_2 \\ B_1^H & D & 0 \\ \omega B_2^H & 0 & 0 \end{bmatrix} \begin{bmatrix} I & -\omega^{-1} B_1 & -\omega^{-1} B_2 \\ 0 & I & 0 \\ 0 & 0 & \omega^{-1} I \end{bmatrix}$$
$$= \begin{bmatrix} \omega I & 0 & 0 \\ B_1^H & D - \omega^{-1} B_1^H B_1 & -\omega^{-1} B_1^H B_2 \\ \omega B_2^H & -B_2^H B_1 & -B_2^H B_2 \end{bmatrix},$$

which is easily seen to be nonsingular for large enough values of ω. ∎

If we use the bound on F_k shown in (7.13), then we can deduce the following theorem.

Theorem 7.4.2. *Let $(A-\sigma I)^{-1}(V_k+F_k) = V_{k+1}\underline{H}_k$ be of full rank and assume F_k is bounded as in (7.13) and $\sqrt{k}\kappa(V_k)\epsilon_1 < 1$. Then there is a ΔA of rank at most k such that $V_k = (A+\Delta A-\sigma I)V_{k+1}\underline{H}_k$, and*

$$\|\Delta A\| \leq \sqrt{k}\|A-\sigma I\|\frac{\kappa(V_k)\epsilon_1 + \kappa(V_{k+1})\kappa(\underline{H}_k)c_{kn}(\epsilon_2)}{1-\sqrt{k}\kappa(V_k)\epsilon_1},$$

where $c_{kn}(\epsilon_2)$ is given by (7.14).

Proof. From $V_k+F_k = (A-\sigma I)V_{k+1}\underline{H}_k$ and $V_k = (A+\Delta A-\sigma I)V_{k+1}\underline{H}_k$ we see that any eligible ΔA has to satisfy $\Delta A V_{k+1}\underline{H}_k = -F_k$. We choose $\Delta A = -F_k(V_{k+1}\underline{H}_k)^\dagger$ (which is of rank at most k) implying $\|\Delta A\| \leq \|F_k\|/\sigma_{\min}(V_{k+1}\underline{H}_k)$. Substituting $\|F_k\|$ by the upper bound given in (7.13) yields

$$\begin{aligned}\|\Delta A\| &\leq \frac{\sqrt{k}\|V_k\|\epsilon_1 + \sqrt{k}\|A-\sigma I\|\|V_{k+1}\|\|\underline{H}_k\|c_{kn}(\epsilon_2)}{\sigma_{\min}(V_{k+1}\underline{H}_k)} \\ &\leq \frac{\sqrt{k}\|V_k\|\epsilon_1}{\sigma_{\min}(V_{k+1}\underline{H}_k)} + \sqrt{k}\|A-\sigma I\|\kappa(V_{k+1})\kappa(\underline{H}_k)c_{kn}(\epsilon_2).\end{aligned}$$

For the denominator we get

$$\begin{aligned}\sigma_{\min}(V_{k+1}\underline{H}_k) &\geq \sigma_{\min}\big((A+\Delta A-\sigma I)V_{k+1}\underline{H}_k\big)/\|A+\Delta A-\sigma I\| \\ &\geq \sigma_{\min}(V_k)/(\|A-\sigma I\| + \|\Delta A\|),\end{aligned}$$

where we used $\sigma_{\min}(XY) \leq \|X\|\sigma_{\min}(Y)$ which holds for any matrices X, Y. Thus

$$\|\Delta A\| \leq \frac{\sqrt{k}\|V_k\|(\|A-\sigma I\| + \|\Delta A\|)\epsilon_1}{\sigma_{\min}(V_k)} + \sqrt{k}\|A-\sigma I\|\kappa(V_{k+1})\kappa(\underline{H}_k)c_{kn}(\epsilon_2)$$

which can be reordered to the claimed bound. ∎

If the linear systems are solved up to a normwise backward error ϵ_{bw}, and (7.8) and (7.15) hold for $2\epsilon_1 + \epsilon_2 = 3\epsilon_{\text{bw}}$, then we get the following corollary.

Corollary 7.4.3. *Let $(A-\sigma I)^{-1}(V_k+F_k) = V_{k+1}\underline{H}_k$ be of full rank and assume F_k is bounded as in (7.15) with $2\epsilon_1 + \epsilon_2 = 3\epsilon_{\text{bw}}$. Then there is a ΔA of rank at most k such that $V_k = (A+\Delta A-\sigma I)V_{k+1}\underline{H}_k$, and*

$$\|\Delta A\| \leq \sqrt{k}\|A-\sigma I\|\kappa(V_{k+1})\kappa(\underline{H}_k)c_{kn}(3\epsilon_{\text{bw}}),$$

where $c_{kn}(\cdot)$ is given by (7.14).

A few remarks are in order.

Remark 7.4.4. If $A + \Delta A - \sigma I$ in Theorem 7.4.2 and Corollary 7.4.3 is singular, then we can invoke Lemma 7.4.1 with $V = V_{k+1}\underline{H}_k$ to obtain a backward error $\Delta\widehat{A}$, arbitrarily close to ΔA, such that $(A + \Delta\widehat{A} - \sigma I)^{-1}V_k = V_{k+1}\underline{H}_k$. The new backward error $\Delta\widehat{A}$ will in general have rank greater than k, but its numerical rank is still bounded by k. Here the definition of numerical rank can be arbitrarily strict, in the sense that we may define the numerical rank to be the number of singular values that are greater than $\epsilon > 0$, for an arbitrarily small ϵ.

Remark 7.4.5. If the orthonormalization is done properly, using, for instance, MGS with reorthogonalization, then $\kappa(V_{k+1}) \approx 1$. In this case we can ignore the factors $\kappa(V_{k+1})$ and $\kappa(V_k)$ when evaluating the bounds in Theorem 7.4.2 and Corollary 7.4.3. In particular this means that the bounds can be estimated cheaply as long as $\|A - \sigma I\|$ (or a good estimate of it) is known.

Remark 7.4.6. For the standard eigenvalue problem, shifts are used to find interior eigenvalues, so any sensible shift satisfies $|\sigma| \leq \|A\|$. Thus, we have $\|A - \sigma I\| \leq 2\|A\|$ in practice.

Remark 7.4.7. In view of [14], we note that our bounds do not contain the loss-of-orthonormality term $\|V_{k+1}^H V_{k+1} - I\|$. Instead we saw that the condition number of the computed basis V_{k+1} plays a role in the bounds of the backward error. We note, however, that a small value of $\|V_{k+1}^H V_{k+1} - I\|$ implies that V_{k+1} is well-conditioned:

$$\|V_{k+1}^H V_{k+1} - I\| < \epsilon < 1 \quad \Rightarrow \quad \kappa(V_{k+1}) < \sqrt{\frac{1+\epsilon}{1-\epsilon}}.$$

The next example shows how Theorem 7.4.2 can be used to derive a simple a posteriori backward error bound.

Example 7.4.8. Suppose a matrix A and a shift σ with $|\sigma| < \|A\|$ are given, and suppose we perform k steps of the shift-and-invert Arnoldi algorithm. To solve the linear systems we use an iterative method that employs (7.5) as stopping condition, that is, the linear systems are considered "solved" when the residuals are less than some tolerance ϵ_{tol} (we ignore the norm of the right hand side since it is approximately one). We use a rather crude tolerance so $\epsilon_{\text{tol}} \gg u$. For the orthogonalization we use MGS with one round of reorthogonalization so $c_{kn}(0) \lesssim 13ku$ (cf. Remark 7.3.4). If

$$\epsilon_{\text{tol}} \geq \kappa(\underline{H}_k)c_{kn}(0), \tag{7.18}$$

then Theorem 7.4.2, with $\epsilon_1 = \epsilon_{\text{tol}}$ and $\epsilon_2 = 0$, and the following remarks, yield that the computed quantities satisfy

$$(A + \Delta A - \sigma I)^{-1} V_k = V_{k+1} \underline{H}_k,$$

where

$$\|\Delta A\| \leq \frac{4\sqrt{k}\kappa(V_{k+1})\epsilon_{\text{tol}}}{1 - \sqrt{k}\kappa(V_k)\epsilon_{\text{tol}}} \|A\|. \tag{7.19}$$

Here we have used the fact that $\kappa(V_{k+1}) \geq \kappa(V_k)$. Since MGS with reorthogonalization was employed, we expect $\kappa(V_{k+1})$ to be close to one. Thus, (7.19) tells us that the relative backward error $\|\Delta A\|/\|A\|$ is of roughly the same size as the tolerance we used to solve the linear systems. So, in this setting the shift-and-invert Arnoldi algorithm is backward stable. ▲

We end this section with a numerical experiment. We consider the matrix

$$A = \begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & 1 & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & 1 & -2 \end{bmatrix},$$

of order $n = 1000$, and the shift $\sigma = -2$. It is well-known that the spectrum of A is a subset of the interval $(-4, 0)$, and the eigenvalues are given by $-2+2\cos(\pi k/(n+1))$, for $k = 1\colon n$. It follows that the shifted matrix $A - \sigma I$ is invertible and has norm $2\cos(\pi/(n+1)) \approx 2$.

We implemented the shift-and-invert Arnoldi algorithm in MATLAB R2013a. For orthonormalization we used MGS with one round of reorthogonalization. The matrix $A - \sigma I$ was stored in sparse format, and the linear systems was solved using MATLAB's "backslash" and `lu` routines. We took $k = 30$ steps with the starting vector $[1, 1, \ldots, 1]^T$, and in each iteration we computed the backward error shown in (7.3), where the residual was evaluated in extended precision (32 digits) and then rounded to double precision. We did this using the `vpa` function from the Symbolic Math Toolbox. We also computed the error $F_k = V_k - (A - \sigma I)V_{k+1}\underline{H}_k$ in extended precision and rounded the result to double precision. For each $j = 1\colon k$, we computed

$$\mathcal{B}(\|\Delta A^{(j)}\|) := \sqrt{j}\|A - \sigma I\|\kappa(\underline{H}_j)c_{jn}(3\epsilon_{\text{bw}}),$$

where ϵ_{bw} was set to be the largest backward error of the linear systems that was encountered in the algorithm, and $c_{jn}(3\epsilon_{\text{bw}}) := (3\epsilon_{\text{bw}} + 13ju)/(1 - 13ju)$ (cf.

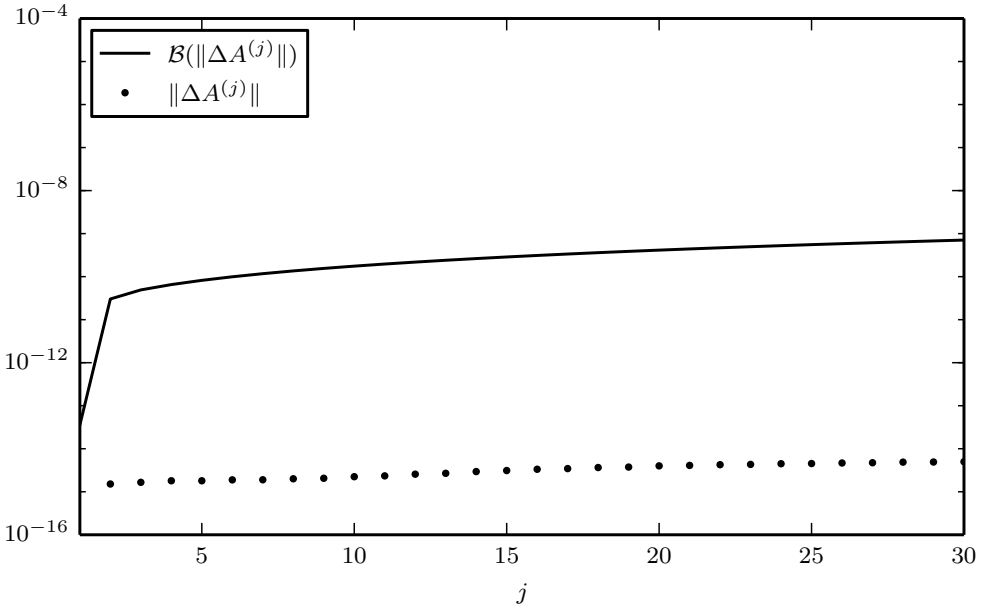

Figure 7.1: Computed backward errors and associated bound.

Remark 7.3.4). As is mentioned in Remark 7.4.5, the above quantity is a good estimate of the bound in Corollary 7.4.3. We also evaluated the expression for the backward error, $\Delta A^{(j)} = -F_k(V_{k+1}\underline{H}_k)^\dagger$, given in the proof of Theorem 7.4.2, and estimated its norm. We did this using the MATLAB routines `pinv` (for the Moore-Penrose pseudo-inverse) and `normest`. The quantities $\mathcal{B}(\|\Delta A^{(j)}\|)$ and $\|\Delta A^{(j)}\|$ are shown in Figure 7.1 for $j = 1\!:\!30$. Even though the (estimated) upper bound $\mathcal{B}(\|\Delta A^{(j)}\|)$ can be seen to be rather pessimistic, it does show that the backward error is less than $\sqrt{u}$. In other words, by evaluating $\mathcal{B}(\|\Delta A^{(k)}\|)$ (which is cheap), we can deduce that the computation is backward stable up to single precision.

7.5 Further topics

7.5.1 Implicit restarting

The bounds in Theorem 7.4.2 and Corollary 7.4.3 contain the factor $\kappa(\underline{H}_k)$, so if $\kappa(\underline{H}_k) \gg 1$ we cannot guarantee a small backward error. If we recall how Arnoldi locates eigenvalues [85, pp. 257–265], we have, unfortunately, reason to suspect that this is the case. Since Arnoldi does not target the largest eigenvalues, but *any* isolated eigenvalue cluster, $H_k := [I_k\ 0]\underline{H}_k$ is likely to have both large and small eigenvalues, which suggests that H_k may be ill-conditioned. We will now show that the situation can be much better than expected if we restrict our attention to the largest eigenvalues of H_k, that is, the ones corresponding to eigenvalues of A

closest to the shift σ. The idea is to do an implicit (thick) restart [74], and purge the small eigenvalues of H_k. In exact arithmetic, a thick restart of an Arnoldi recurrence

$$(A - \sigma I)^{-1} V_k = V_k H_k + h_{k+1,k} v_{k+1} e_k^T,$$

refers to a transformation to

$$(A - \sigma I)^{-1} U = UT + h_{k+1,k} v_{k+1} e_k^T Q,$$

where $H_k = QTQ^H$ is a Schur decomposition and $U = V_k Q$, and a truncation

$$(A - \sigma I)^{-1} U_\ell = U_\ell T_\ell + h_{k+1,k} v_{k+1} e_k^T Q_\ell,$$

where T_ℓ is the leading $\ell \times \ell$ submatrix of T, and U_ℓ and Q_ℓ denote the first ℓ columns of U and Q, respectively. The truncation is commonly referred to as *purging*. The idea behind purging is to filter out the Ritz values we are not interested in.

Now, since small eigenvalues of H_k correspond to eigenvalues of A far from the shift σ, it is reasonable to assume they are of less interest. Suppose

$$(A - \sigma I)^{-1}(V_k + F_k) = V_{k+1} \underline{H}_k$$

and consider a Schur form $H_k = QTQ^H$ such that t_{ii}, $i = \ell + 1 : k$, are the small eigenvalues to be purged. We have

$$(A - \sigma I)^{-1}(U + F_k Q) = [U \; v_{k+1}] \begin{bmatrix} T \\ h_{k+1,k} e_k^T Q \end{bmatrix},$$

where $U = V_k Q$. Throwing away the last $k - \ell$ columns yields

$$(A - \sigma I)^{-1}(U_\ell + F_k Q_\ell) = [U_\ell \; v_{k+1}] \begin{bmatrix} T_\ell \\ h_{k+1,k} e_k^T Q_\ell \end{bmatrix},$$

where $Q_\ell = Q(:, 1:\ell)$, $U_\ell = U(:, 1:\ell)$ and $T_\ell = T(1:\ell, 1:\ell)$. Defining $u_{\ell+1} = v_{k+1}$,

$$\underline{T}_\ell = \begin{bmatrix} T_\ell \\ h_{k+1,k} e_k^T Q_\ell \end{bmatrix},$$

and $E_\ell = F_k Q_\ell$, results in a compact recurrence

$$(A - \sigma I)^{-1}(U_\ell + E_\ell) = U_{\ell+1} \underline{T}_\ell, \tag{7.20}$$

where $\|E_\ell\| \leq \|F_k\|$. Note that our bound on E_ℓ depends on k and not ℓ. We can now repeat the proof of Theorem 7.4.2, and use the bounds $\|E_\ell\| \leq \|F_k\|$ and $\sigma_{\min}(U_{\ell+1}) \geq \sigma_{\min}(V_{k+1})$, and the recurrence (7.20) instead of the one assumed in the theorem. We get

$$U_\ell = (A + \Delta A - \sigma I)U_{\ell+1}\underline{T}_\ell,$$

where

$$\|\Delta A\| \leq \|A - \sigma I\| \frac{\sqrt{k}\kappa(V_k)\epsilon_1 + \sqrt{k}\kappa(V_{k+1})c_{kn}(\epsilon_2)\|\underline{H}_k\|/\sigma_{\min}(\underline{T}_\ell)}{1 - \sqrt{k}\kappa(V_k)\epsilon_1}. \tag{7.21}$$

Comparing this to the bound in Theorem 7.4.2 we see that $\kappa(\underline{H}_k)$ has been replaced by $\|\underline{H}_k\|/\sigma_{\min}(\underline{T}_\ell)$. Further, it holds that

$$\|\underline{H}_k\|/\sigma_{\min}(\underline{T}_\ell) \leq \|\underline{H}_k\|/\sigma_{\min}\left(\begin{bmatrix} T \\ h_{k+1,k}e_k^T Q \end{bmatrix}\right) = \kappa(\underline{H}_k).$$

It follows that if $\underline{H}_k$ is ill-conditioned due to the small eigenvalues we purged, then $\|\underline{H}_k\|/\sigma_{\min}(\underline{T}_\ell) \ll \kappa(\underline{H}_k)$ and (7.21) shows that the upper bound for the backward error corresponding to the part of the spectrum we care about is much smaller than the upper bound for the general backward error.

7.5.2 Hermitian backward errors

We now restrict the scope to the Hermitian matrix eigenvalue problem, that is, when $A = A^H$ and σ is real. Let us mention that we still consider the shift-and-invert Arnoldi algorithm, as it is shown in Algorithm 7.1, and *not* the shift-and-invert Lanczos algorithm with a three-term recurrence. In the Hermitian case, Algorithm 7.1 is also known as the shift-and-invert Lanczos algorithm with full orthogonalization, and it is used in, e.g., MATLAB's `eigs` command.

Is it, for a Hermitian A, possible to find a Hermitian backward error ΔA? We have seen in the proof of Theorem 7.4.2 that ΔA has to satisfy $\Delta A V_{k+1}\underline{\mathrm{H}}_k = -F_k$. Unfortunately the following Lemma rules out existence of such a Hermitian ΔA in general.

Lemma 7.5.1. *Let $X \in \mathbb{C}^{n\times k}$ and $F \in \mathbb{C}^{n\times k}$. Then there exists a Hermitian E with $EX = F$ if and only if $X^H F$ is Hermitian and $FX^\dagger X = F$. In that case, there is such an E with* $\mathrm{rank}(E) \leq 2k$ *and $\|E\|_* \leq 2\|F\|_*/\sigma_{\min}(X)$ where $\|\cdot\|_*$ denotes the spectral norm or the Frobenius norm.*

Proof. The proof is simple and, for $k = 1$, is contained in [58]. We give it for

completeness. Let E be any matrix such that $EX = F$. This implies $EXX^\dagger X = FX^\dagger X$ and (using $XX^\dagger X = X$) $EX = FX^\dagger X$, contradicting $EX = F$ if $F \neq FX^\dagger X$. Thus $F = FX^\dagger X$ is necessary for the existence of an E with $EX = F$. Now, if E is Hermitian, then so is $X^H EX = X^H F$. Hence, if $X^H F$ is not Hermitian, then there is no Hermitian E with $EX = F$.

On the other hand, if $X^H F$ be Hermitian and $F = FX^\dagger X$, then

$$E := FX^\dagger + (FX^\dagger)^H - X^{\dagger H} F^H X X^\dagger = FX^\dagger + (FX^\dagger)^H (I - XX^\dagger)$$

is also Hermitian. Furthermore, $\operatorname{rank}(E) \leq 2k$, $EX = F$, and (using that $I - XX^\dagger$ is an orthogonal projector)

$$\|E\|_* \leq 2\|FX^\dagger\|_* \leq 2\|F\|_* \|X^\dagger\|_2 = 2\|F\|_*/\sigma_{\min}(X). \qquad \blacksquare$$

The next result shows that one still gets a Hermitian backward error if one replaces the Hessenberg matrix $\underline{H}_k$ by some other $(k+1) \times k$ matrix $\underline{G}_k$. Before we state the theorem, we should clarify what we mean by "backward error" in this case. If we replace $\underline{H}_k$ by something else, we cannot say that the computed quantities (V_{k+1} and $\underline{H}_k$) satisfy an exact Krylov recurrence of a perturbed input matrix. We can, however, still say that the *computed subspace* is a Krylov subspace of a perturbed Hermitian input matrix. We refer to this Hermitian perturbation as the backward error.

Theorem 7.5.2. *Let A be Hermitian and $(A-\sigma I)^{-1}(V_k+F_k) = V_{k+1}\underline{H}_k$. Suppose it holds for $\underline{G}_k \in \mathbb{C}^{(k+1)\times k}$ that $V_k^H V_{k+1}\underline{G}_k$ is Hermitian and $V_{k+1}\underline{G}_k$ is of full rank. Then there is a Hermitian ΔA of rank at most $2k$ such that*

$$V_k = (A + \Delta A - \sigma I)V_{k+1}\underline{G}_k,$$

and

$$\|\Delta A\| \leq 2\frac{\|(A-\sigma I)\|\|V_{k+1}\|\|\underline{H}_k - \underline{G}_k\| + \|F_k\|}{\sigma_{\min}(V_{k+1}\underline{G}_k)}.$$

Proof. From $V_k = (A + \Delta A - \sigma I)V_{k+1}\underline{G}_k$ and

$$V_k + F_k = (A-\sigma I)V_{k+1}\underline{H}_k = (A-\sigma I)V_{k+1}\underline{G}_k + (A-\sigma I)V_{k+1}(\underline{H}_k - \underline{G}_k)$$

we see that any eligible ΔA has to satisfy

$$\Delta A V_{k+1}\underline{G}_k = (A-\sigma I)V_{k+1}(\underline{H}_k - \underline{G}_k) - F_k = V_k - (A-\sigma I)V_{k+1}\underline{G}_k.$$

Since it is assumed that $V_{k+1}\underline{G}_k$ is of full rank, Lemma 7.5.1 implies that such a Hermitian ΔA exists if

$$(V_{k+1}\underline{G}_k)^H(V_k - (A-\sigma I)V_{k+1}\underline{G}_k) = (V_{k+1}\underline{G}_k)^H V_k - (V_{k+1}\underline{G}_k)^H(A-\sigma I)V_{k+1}\underline{G}_k$$

is Hermitian. Since the first term on the right hand side is Hermitian by assumption, this is easily seen to be the case. Also by Lemma 7.5.1, ΔA is bounded by

$$\begin{aligned}\|\Delta A\| &\leq 2\|(A-\sigma I)V_{k+1}(\underline{H}_k - \underline{G}_k) - F_k\|/\sigma_{\min}(V_{k+1}\underline{G}_k)\\ &\leq 2(\|(A-\sigma I)\|_2\|V_{k+1}\|\|\underline{H}_k - \underline{G}_k\| + \|F_k\|)/\sigma_{\min}(V_{k+1}\underline{G}_k),\end{aligned}$$

and is of rank at most $2k$. ∎

Remark 7.5.3. If $A + \Delta A - \sigma I$ is singular, then we can use the second part of Lemma 7.4.1 to find a *Hermitian* backward error $\Delta\widetilde{A}$ arbitrarily close to ΔA such that $A + \Delta\widetilde{A} - \sigma I$ is invertible.

In order to obtain a small Hermitian backward error, we need to find a matrix $\underline{G}_k$ close to $\underline{\mathrm{H}}_k$ such that $V_k^H V_{k+1}\underline{G}_k$ is Hermitian. One possibility is

$$\underline{G}_k := R_{k+1}^{-1}\begin{bmatrix} T_k \\ h_{k+1,k}e_k^T \end{bmatrix} R_k, \tag{7.22}$$

where R_k, R_{k+1} are the upper triangular QR factors of V_k, V_{k+1}, respectively, and T_k is the tridiagonal part of the Hermitian part of H_k. Then $\underline{G}_k$ is Hessenberg and computing Ritz pairs is particularly easy: we need to find vectors z and scalars μ such that

$$V_k^H(A + \Delta A - \sigma I)^{-1}V_k z = \mu V_k^H V_k z.$$

Here we have used Remark 7.5.3 in order to ensure that $A + \Delta A - \sigma I$ is invertible. By using the Krylov relation $(A + \Delta A - \sigma I)^{-1}V_k = V_{k+1}\underline{G}_k$ we obtain

$$V_k^H V_{k+1}\underline{G}_k z = \mu V_k^H V_k z.$$

Inserting the QR factorizations $V_j = Q_j R_j$, $j = k, k+1$ and the formula for $\underline{G}_k$ shown in (7.22) yields

$$R_k^H[I\ 0]R_{k+1}R_{k+1}^{-1}\begin{bmatrix} T_k \\ h_{k+1,k}e_k^T \end{bmatrix} R_k z = \mu R_k^H R_k z,$$

which simplifies to $T_k\tilde{z} = \mu\tilde{z}$ where $\tilde{z} = R_k z$. So, the Ritz values are just the eigenvalues of T_k (which are real, since T_k is Hermitian). To obtain the Ritz vectors,

we would have to multiply $\tilde{z}$ with R_k^{-1}. However, since R_k is close to the identity matrix if the orthogonalization has been done properly (for instance, by using MGS with reorthogonalization) we can approximate $\tilde{z}$ by z. Thus, (approximations of) Ritz pairs for the choice (7.22) of $\underline{G}_k$ can be obtained without computing R_k, R_{k+1}. We also note that choosing the eigenpairs of T_k to construct Ritz pairs is what is done in practice.

7.5.3 Conditions for breakdown

We now discuss how to derive a sensible breakdown criterion based on our error analysis. We saw in Section 7.1.1 that the computed quantities V_{j+1} and $\underline{H}_j$ satisfy

$$(A - \sigma I)^{-1}(V_j + F_j) = V_{j+1}\underline{H}_j.$$

This recurrence can be rewritten as

$$(A - \sigma I)^{-1}(V_j + \widetilde{F}_j) = V_j H_j,$$

where $\widetilde{F}_j = F_j - (A - \sigma I)h_{j+1,j}v_{j+1}e_j^T$. Note that the first $j-1$ columns of $\widetilde{F}_j$ and F_j are identical. For the last column, we have

$$\widetilde{f}_j = r_j - (A - \sigma I)(g_j + h_{j+1,j}v_{j+1}),$$

where r_j is the residual from the linear system, and g_j the columnwise backward error from the orthonormalization. It is natural to declare breakdown when the error introduced by neglecting $h_{j+1,j}$ is of the same order as the errors that are present in the computation. This leads us to the following breakdown condition:

$$h_{j+1,j} < \|g_j\| + \|r_j\|/\|(A - \sigma I)v_{j+1}\|.$$

We can simplify this condition by replacing $\|g_j\|$ with its bound in (7.10). This yields

$$h_{j+1,j} < \eta(n,j)\|w_j\|u + \|r_j\|/\|(A - \sigma I)v_{j+1}\|. \tag{7.23}$$

We now discuss how to evaluate (7.23) in practice. For the residual term $\|r_j\|$ we consider two cases. The residual, or a good upper bound of the residual, may be given to us. This is the case if we, for instance, use an iterative linear system solver that guarantees a residual less than some tolerance. In this case, we can substitute $\|r_j\|$ in (7.23) by the given tolerance. If the residual, or any good bounds for it,

are not given, then we need to compute it. Let m be a constant such that the following forward error bound holds for an arbitrary vector x

$$\|\text{float}((A-\sigma I)x) - (A-\sigma I)x\| \leq mu\|A-\sigma I\|\|x\|.$$

If $A-\sigma I$ is given as a dense matrix, we have $m = n^{3/2}$ [36, p. 70]. For sparse matrices, m can be much smaller. The computed residual $\widehat{r}_j$ satisfies

$$\begin{aligned}\|\widehat{r}_j\| &\leq (1+u)\|\text{float}((A-\sigma I)w_j) - v_j\| \\ &\leq (1+u)(\|r_j\| + mu\|A-\sigma I\|\|w_j\|).\end{aligned}$$

By comparing to (7.4), we recognize $\|A-\sigma I\|\|w_j\|mu$ as a part of the norm of a residual associated with a computed solution with corresponding backward error mu. Thus, we can compute a satisfactory $\widehat{r}_j$ if we use an extended precision $\overline{u}$ such that $m\overline{u} < u$.

For the computation of vector $(A-\sigma I)v_{j+1}$, we have

$$\begin{aligned}\|\text{float}((A-\sigma I)v_j) - (A-\sigma I)v_j\| &\leq mu\|A-\sigma I\|\|v_j\| \\ &\leq mu\kappa(A-\sigma I)\|(A-\sigma I)v_j\|,\end{aligned}$$

and, using the reverse triangle inequality, that

$$\|(A-\sigma I)v_j\|\big(1 - mu\kappa(A-\sigma I)\big) \leq \|\text{float}((A-\sigma I)v_j)\|.$$

Thus the computed vector is accurate enough as long as $mu\kappa(A-\sigma I) \ll 1$. If $A-\sigma I$ is so ill-conditioned that this is not satisfied, then we can use an extended precision $\overline{u}$ such that $m\overline{u}\kappa(A-\sigma I) \ll 1$.

If (7.23) and (7.6) hold, then

$$\|\widetilde{f}_j\| \leq 2\big(\|v_j\|\epsilon_1 + \|A-\sigma I\|\|w_j\|(\epsilon_2 + \eta(n,j)u)\big).$$

By derivations similar to those leading to (7.13), we get

$$\|\widetilde{F}_j\| \leq 2\big(\sqrt{j}\|V_j\|\epsilon_1 + \sqrt{j}\|A-\sigma I\|\|V_{j+1}\|\|\underline{H}_j\|c_{jn}(\epsilon_2)\big). \tag{7.24}$$

From this we obtain the following "breakdown analogue" of Theorem 7.4.2.

Theorem 7.5.4. *Let $(A-\sigma I)^{-1}(V_j + \widetilde{F}_j) = V_j H_j$ be of full rank and assume $\widetilde{F}_j$ is bounded as in (7.24) and $\sqrt{j}\kappa(V_j)\epsilon_1 < 1$. Then there is a ΔA of rank at most j*

such that $V_j = (A + \Delta A - \sigma I)V_j H_j$ *and*

$$\|\Delta A\| \leq 2\sqrt{j}\|A - \sigma I\| \frac{\kappa(V_j)\epsilon_1 + \kappa(V_{j+1})\|\underline{H}_j\|/\sigma_{\min}(H_j)c_{jn}(\epsilon_2)}{1 - \sqrt{j}\kappa(V_j)\epsilon_1},$$

where $c_{jn}(\cdot)$ *is given by* (7.14).

The proof is omitted since it is essentially the same as the proof of Theorem 7.4.2. In a similar manner, we can get corresponding breakdown analogues to Corollary 7.4.3 and Theorem 7.5.2.

Chapter

8 Conclusion

In Chapter 3 we studied strongly damped quadratic matrix polynomials, or more precisely, matrix polynomials $M\lambda^2 + sD\lambda + K$, where all coefficient matrices are real and positive semi-definite, M and K positive definite, and the damping parameter s goes to infinity. We showed that such polynomials in many ways are similar to their undamped counterparts $M\lambda^2 + K$. In particular we extended some of Lancaster's early work [50], and furthermore, showed how the eigenvalues move as the damping gets stronger. We saw that strong damping leads to small negative eigenvalues close to zero, which are harmless in the sense that it does not cause any term in the response (3.22) to blow up, as long as the corresponding undamped problem does not have a tiny eigenvalue. This is interesting from a practical point of view. When no damping is present, the smallest eigenvalues are the most interesting ones. In this case, they are purely imaginary and correspond to the modes with the lowest frequencies. Engineers are concerned with these modes since external forces that commonly come into play (e.g., from car traffic or earthquakes) are likely to have low frequencies components $f_0 e^{i\omega t}$, $\omega \in \mathbb{R}$, themselves, and if an external force has a frequency close to that of a mode, we saw in Section 1.1 that resonance is likely to appear (see also [50, p. 125]). The fact that strong damping leads to small negative eigenvalues suggests that when damping is present, the region in which the "interesting" eigenvalues lie is much more complicated to describe than in the undamped case.

We also derived a new "symmetric" formula for the inverse of real symmetric matrix polynomials with invertible leading coefficient (Theorem 3.5.1), and used

this to write down the particular solutions for associated ODEs.

In Chapter 4 we first developed Algorithm 4.2 for solving definite generalized eigenvalue problems $Ax = \lambda Bx$, where both A and B are semidefinite. We assumed further that A and B were both real, but the algorithm is easily generalized to complex matrices. Algorithm 4.2 is to the author's knowledge the first one that computes all eigenvalues of such problems in a backward stable and symmetry preserving manner. Here symmetry preserving means that we can compute a nonnegative diagonal pencil $D_1 - \lambda D_2$ such that

$$X^H(D_1 - \lambda D_2)X = A + \Delta A - \lambda(B + \Delta B),$$

where ΔA and ΔB are symmetric and small with respect to A and B, respectively. This means that one backward error $\Delta A + \lambda \Delta B$ applies to all computed eigenvalues in homogeneous form. This should be compared to the QZ algorithm, which, applied to such problems, also computes all eigenvalues in a backward stable manner, but not with respect to symmetry. However, *if* a computed eigenvalue is real, then it can be shown that the symmetric backward error is also small. But contrary to Algorithm 4.2, different eigenvalues do in general have different backward errors, even when we consider them in homogeneous form [71, 35].

We used Algorithm 4.2 as a first step in the more involved eigensolver Algorithm 4.3, which computes the eigenvalues, and eigenvectors if desired, of certain QEPs where the damping matrix is of low rank. In particular, we used the fact that Algorithm 4.2 can diagonalize the undamped problem via congruence. Algorithm 4.3, which makes use of an Ehrlich-Aberth iteration, was shown to be both fast and accurate in numerical experiments. The algorithm is linearization free, which means that it does not have the same problem with strong damping as, for example `quadeig` [34]. This allowed us to confirm our theory in Chapter 3 by numerical examples, and we could compute all eigenpairs of a strongly damped vibrating beam so the worst backward error encountered was smaller than 10^{-14}. If also the eigenvectors are desired, our algorithm computes these in a final step using an inverse iteration that is specially designed for real symmetric matrix polynomials. To motivate why this iteration works, we used the Takagi factorization for complex symmetric matrices. Since our quadratic eigensolver computes all eigenvalues/eigenpairs, it is only suitable for small and moderately sized QEPs. For large scale quadratics, where only a few eigenvectors can be stored in memory, some "subspace based" algorithm needs to be applied to construct QEPs of a size that the new algorithm can handle. Possible future work includes the design

of such an algorithm. Further, for such an algorithm to be useful, we need to understand which eigenvalues to look for; as mentioned above, this is a nontrivial problem.

The degree deficiency $n \deg P(\lambda) - \deg \det P(\lambda)$, of a regular $n \times n$ matrix polynomial $P(\lambda)$, equals the algebraic multiplicity of the infinite eigenvalue of $P(\lambda)$ (cf. page 18). Similarly, $n \deg \operatorname{rev}(P) - \deg \det \operatorname{rev}(P)$ equals the algebraic multiplicity of the zero eigenvalue. Unfortunately, these expressions are impractical from a numerical point of view. For the special matrix polynomials considered in Chapter 4, we derived more convenient expressions for the algebraic multiplicity of zero and infinite eigenvalues, that could be computed numerically using the SVD. This result, which was summarized in Proposition 4.4.1, depends on the special structure of the coefficient matrices and cannot be applied to more general quadratics. To see this, take for example,

$$M\lambda^2 + D\lambda + K = \begin{bmatrix} \lambda^2 & \lambda^2 - \lambda \\ \lambda & \lambda + 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \lambda^2 + \begin{bmatrix} 0 & -1 \\ 1 & 1 \end{bmatrix} \lambda + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

The algebraic multiplicity of the eigenvalue at infinity is

$$4 - \deg \det(M\lambda^2 + D\lambda + K) = 2,$$

but

$$\dim \operatorname{null}(M) + \dim(\operatorname{null}(M) \cap \operatorname{null}(D)) = 1.$$

An interesting open problem is to derive "numerically computable" expressions, like the ones in Proposition 4.4.1, for more general matrix polynomials.

In Chapter 5 we solved inverse polynomial eigenproblems. In particular, we showed that the only additional constraint that is imposed on the eigenstructure of an $n \times n$ matrix polynomial if it has real instead of complex matrix coefficients, is that all elementary divisors (seen as a matrix polynomial over $\mathbb{C}$) of finite nonreal eigenvalues come in complex conjugate pairs. We conjectured that the same is true for regular Hermitian matrix polynomials, and supported the conjecture with proofs for some special cases. In particular, we saw that the conjecture was easily verified for pencils. This is of relevance to polynomial eigenproblems that are solved via linearizations. For instance, it follows immediately that any real regular matrix polynomial has a Hermitian linearization. Thus, in terms of the *eigenstructure* (defined on page 18) a Hermitian linearization is not richer in structure than an unstructured real linearization. Since a real linearization enforces the same constraints on the eigenstructure as a Hermitian linearization, one idea is to use

real (possibly nonsymmetric) linearizations of Hermitian matrix polynomials. If we use an algorithm that works in real arithmetic, then the symmetry (with respect to the real axis) in the eigenstructure will automatically be preserved even though we break all Hermitian structure. The hard part, which is left as an open problem, is to construct such linearizations in a cheap manner. If we can afford to double the problem size one approach is the following: let $A(\lambda) + iB(\lambda)$, with $A(\lambda), B(\lambda) \in \mathbb{R}^{n\times n}$, be a regular Hermitian matrix polynomial we want to find the eigenvalues of and note that $A(\lambda) + iB(\lambda) \sim A(\lambda) - iB(\lambda)$. We have that

$$\begin{bmatrix} & A(\lambda) - iB(\lambda) \\ A(\lambda) + iB(\lambda) & \end{bmatrix} \sim \begin{bmatrix} A(\lambda) & B(\lambda) \\ -B(\lambda) & A(\lambda) \end{bmatrix} =: P(\lambda),$$

so eigenstructure of $P(\lambda)$ is the same as that of $A(\lambda) + iB(\lambda)$ but each elementary divisor appears twice as many times. Furthermore, $P(\lambda)$ is real, so we may pick any real linearization, for example (2.7), and solve the associated GEP using an algorithm that works in real arithmetic.

In Chapter 6 we described a method for reducing monic matrix polynomials $P(\lambda)$ of degree ℓ to (block) triangular, (block) diagonal and Hessenberg form, by means of structure preserving similarity transformations applied to the left companion linearization. Our method made use of block Krylov matrices $[V\ AV\ \cdots\ A^{\ell-1}V]$, where A is the constant part of some monic linearization of $P(\lambda)$. This ill-conditioning is something we need to worry about, in particular as ℓ gets bigger. The main problem, however, is to identify applications of the simpler forms. The diagonal form has received some attention in the literature since it may be used to transform second order systems of ODEs to equivalent "decoupled" systems of second order [19, 20, 28, 63]. It is, however, unclear (at least to the author) what is gained in doing this. In the quadratic case, the diagonal form tells us what the eigenvalues are—neither more nor less. It was suggested in [63] to use diagonalization to decouple systems of ODEs and then solve decoupled scalar second order ODEs. However, since the standard way of solving scalar second order ODEs is via a linearization, we might as well linearize the system of ODEs and do the decoupling on the linearization level.

In Chapter 7 we showed that a floating point implementation of the shift-and-invert Arnoldi algorithm, where errors from all steps of the computation are taken into account, yields computed quantities that satisfy an exact shift-and-invert Krylov recurrence of a perturbed matrix. Here, the word "Krylov" is used instead of "Arnoldi" since the computed basis cannot be guaranteed to be perfectly orthogonal. We showed that the norm of the backward error depends on

the condition number of the computed Hessenberg matrix $\underline{H}_k$, and argued that even if this number is large, the restriction to the most important part of the recurrence (that is, what is left after purging the small eigenvalues of H_k) can have a small backward error. For Hermitian matrices A, we showed that there is an Hermitian backward error ΔA such that the computed basis V_{k+1} spans a Krylov subspace associated with $A + \Delta A$. However, as in the case of standard Arnoldi [44], the small $(k+1) \times k$ matrix associated with this subspace is generally not the computed Hessenberg matrix. Finally, we defined a new breakdown condition based on our error analysis. If this condition is met, we could derive a new set of backward error bounds, which show that an invariant subspace of a perturbed matrix has been found.

One problem that is interesting to look into is how the results in Chapter 7 can be generalized to pencils $A - \lambda B$. In this case the perturbed recurrence

$$(A - \sigma I)(V_k + F_k) = V_{k+1}\underline{H}_k,$$

studied in Chapter 7, has to be replaced by

$$(A - \sigma B)(\text{float}(BV_k) + F_k) = V_{k+1}\underline{H}_k.$$

Thus a new error, that is, the error from the computation of BV_k, has to be taken into account.

Appendix

A The principal angles and the gap

A.1 Two results on the principal angles

In this section we prove Theorem 3.2.1 and the equality in (3.5). We need the following two results; both are well known [41, Section 3.1].

Theorem A.1.1. *The singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p$ of a matrix A can be characterized recursively as follows:*

$$\begin{aligned} \sigma_i &= \max\{\, |x^H A y| \;:\; \|x\|_2 = \|y\|_2 = 1, \\ &\qquad x^H x_j = y^H y_j = 0,\ j = 1, 2, \ldots, i-1\} \\ &= |x_i^H A y_i|, \end{aligned}$$

where x_i and y_i are maximizing vectors (in fact singular vectors).

Proof. We have $|x^H A y| \leq \|x\|_2 \|A\|_2 \|y\|_2 = \sigma_1$ and $x_1^H A y_1 = \sigma_1$ where x_1 and y_1 can be any left and right first singular vectors, respectively. Hence the result is true for $i = 1$. Set $B = \sum_{k=i}^{p} \sigma_k x_k y_k^H$. For any vectors x and y in the ith set above, we have $|x^H A y| = |x^H B y| \leq \|x\|_2 \|B\|_2 \|y\|_2 = \sigma_i$. Since $x_i^H A y_i = \sigma_i$ for any ith left and right singular vectors x_i and y_i, the result follows. ∎

Theorem A.1.2. *The smallest singular value of a $p \times p$ matrix $A = U \Sigma V^H$, is given by*

$$\sigma_p = \min_{\|x\|_2 = 1} \max_{\|y\|_2 = 1} |x^H A y|.$$

Proof. Let f denote the right hand side above. We have

$$f \leq \max_{\|y\|_2=1} |x_p^H A y| = \sigma_p,$$

for any pth left singular vector x_p. For any x we can pick y such that $\|y\|_2 = 1$ and $V^H y = U^H x$. Hence

$$f \geq \min_{\|x\|_2=1} |(U^H x)^H \Sigma U^H x| = \min_{\|w\|_2=1} |w^H \Sigma w| = \sigma_p.$$

∎

Let $\langle \cdot, \cdot \rangle$ denote the A-inner product for some Hermitian positive definite matrix A, and let $\| \cdot \|$ be the induced norm. The proof of Theorem 3.2.1 we present is a straightforward generalization of the proof given in [13] for the standard inner product. We will make use of the following fact.

Fact A.1.3. *Let $f : X \to \mathbb{R}$ such that $f[X]$ is a closed interval of the real line. For any decreasing function $g : f[X] \to \mathbb{R}$ it holds that*

$$g\left(\max_{x \in X} f(x)\right) = \min_{x \in X} g(f(x)) \quad \textit{and} \quad g\left(\min_{x \in X} f(x)\right) = \max_{x \in X} g(f(x)).$$

Proof of Theorem 3.2.1. Cosine is a decreasing function on $[0, \pi/2]$. Hence, by using Fact A.1.3, we see that taking cosine on both sides of the definition of $\theta_i(\mathcal{U}, \mathcal{V})$ yields

$$\begin{aligned} \cos(\theta_i(\mathcal{U}, \mathcal{V})) &= \max\{\ |\langle u, v \rangle| \ :\ u \in \mathcal{U},\ v \in \mathcal{V}, \|u\| = \|v\| = 1, \\ &\qquad \langle u, u_j \rangle = \langle v, v_j \rangle = 0,\ j = 1, 2, \ldots, i-1\} \\ &= |\langle u_i, v_i \rangle|. \end{aligned}$$

By defining $u = Ux$, $v = Vy$ and $u_i = Ux_i$, $v_i = Vy_i$ we get

$$\begin{aligned} \cos(\theta_i(\mathcal{U}, \mathcal{V})) &= \max\{\ |x^H (U^H A V) y| \ :\ \|x\|_2 = \|y\|_2 = 1, \\ &\qquad x^H x_j = y^H y_j = 0,\ j = 1, 2, \ldots, i-1\} \\ &= |x_i^H (U^H A V) y_i|. \end{aligned}$$

The result now follows from Theorem A.1.1. ∎

We have the following characterization of the largest principal angle between subspaces of the same dimension.

Theorem A.1.4. *If $p = q$, then the largest principal angle is given by*

$$\theta_{\max}(\mathcal{U}, \mathcal{V}) = \max_{\substack{u \in \mathcal{U} \\ \|u\|=1}} \min_{\substack{v \in \mathcal{V} \\ \|v\|=1}} \measuredangle(u, v).$$

Proof. Using Theorem A.1.2 we get

$$\begin{aligned} \cos(\theta_{\max}(\mathcal{U}, \mathcal{V})) &= \sigma_p(U^H A V) = \min_{\|x\|_2=1} \max_{\|y\|_2=1} |x^H U^H A V y| \\ &= \min_{\substack{u \in \mathcal{U} \\ \|u\|=1}} \max_{\substack{v \in \mathcal{V} \\ \|v\|=1}} |u^H A v| = \min_{\substack{u \in \mathcal{U} \\ \|u\|=1}} \max_{\substack{v \in \mathcal{V} \\ \|v\|=1}} |\langle u, v \rangle|. \end{aligned}$$

Since arccos is a decreasing function on $[0, 1]$, Fact A.1.3 implies that

$$\begin{aligned} \theta_{\max}(\mathcal{U}, \mathcal{V}) &= \arccos\left(\min_{\substack{u \in \mathcal{U} \\ \|u\|=1}} \max_{\substack{v \in \mathcal{V} \\ \|v\|=1}} |\langle u, v \rangle| \right) = \max_{\substack{u \in \mathcal{U} \\ \|u\|=1}} \arccos\left(\max_{\substack{v \in \mathcal{V} \\ \|v\|=1}} |\langle u, v \rangle| \right) \\ &= \max_{\substack{u \in \mathcal{U} \\ \|u\|=1}} \min_{\substack{v \in \mathcal{V} \\ \|v\|=1}} \arccos(|\langle u, v \rangle|) = \max_{\substack{u \in \mathcal{U} \\ \|u\|=1}} \min_{\substack{v \in \mathcal{V} \\ \|v\|=1}} \measuredangle(u, v). \end{aligned}$$

∎

A.2 The gap

In this section we introduce the gap between subspaces of $\mathbb{C}^n$ and relate it to the largest principal angle. As in the previous section, all norms and angles are with respect to the A-inner product for some Hermitian positive definite matrix A. We have the following definitions [45, p. 7 and p. 197]:

$$\operatorname{dist}(u, \mathcal{V}) = \min_{v \in \mathcal{V}} \|u - v\|,$$

$$\delta(\mathcal{U}, \mathcal{V}) = \begin{cases} 0 & \text{if } \mathcal{U} = 0, \\ \max\limits_{\substack{u \in \mathcal{U} \\ \|u\|=1}} \operatorname{dist}(u, \mathcal{V}) & \text{otherwise,} \end{cases}$$

and

$$\operatorname{gap}(\mathcal{U}, \mathcal{V}) = \max\left(\delta(\mathcal{U}, \mathcal{V}), \delta(\mathcal{V}, \mathcal{U})\right).$$

Note that we in general have $\delta(\mathcal{U}, \mathcal{V}) \neq \delta(\mathcal{V}, \mathcal{U})$, so gap $\not\equiv \delta$. If $\dim \mathcal{U} = \dim \mathcal{V}$, however, we will see that it always holds that $\operatorname{gap}(\mathcal{U}, \mathcal{V}) = \delta(\mathcal{U}, \mathcal{V}) = \delta(\mathcal{V}, \mathcal{U})$. Since we assume that the norm is induced by an inner product, an equivalent

definition is given by

$$\operatorname{gap}(\mathcal{U}, \mathcal{V}) = \|\mathbf{P}_{\mathcal{U}} - \mathbf{P}_{\mathcal{V}}\|, \tag{A.1}$$

where $\mathbf{P}_{\mathcal{U}}$ and $\mathbf{P}_{\mathcal{V}}$ are orthogonal projections onto $\mathcal{U}$ and $\mathcal{V}$, respectively [31, Theorem 13.1.1]. From (A.1) we see that the gap is a metric on the set of subspaces of $\mathbb{C}^n$. Thus we may talk about continuous subspaces depending on a parameter. More precisely, we say that $\mathcal{U}(s)$ is continuous on $I \subset \mathbb{R}$, if for each $a \in I$, it holds that $\lim_{s \to a} \operatorname{gap}(\mathcal{U}(s), \mathcal{U}(a)) = 0$.

Variants of the next two lemmas can be found in [73, p. 249–250].

Lemma A.2.1. *Let the columns of V and $V_\perp$ be A-orthonormal bases of $\mathcal{V}$ and $\mathcal{V}_\perp$ respectively. If $\|u\| = 1$, then $\sin \measuredangle(u, \mathcal{V}) = \|V_\perp^H A u\|_2$.*

Proof. Define

$$\begin{bmatrix} V^H \\ V_\perp^H \end{bmatrix} Au = \begin{bmatrix} c \\ s \end{bmatrix}.$$

Since VV^HA and $V_\perp V_\perp^H A$ are A-orthogonal projectors onto $\mathcal{V}$ and $\mathcal{V}_\perp$, respectively, we have

$$c^H c + s^H s = u^H A(VV^H A + V_\perp V_\perp^H A)u = u^H A u = 1.$$

Now, $c^H c$ and $s^H s$ are scalars and equals the squares of 2-norms of $V^H Av$ and $V_\perp^H Av$, respectively. By Theorem 3.2.1, $\cos^2\theta = c^H c$, where θ is the principal angle between $\operatorname{span}\{u\}$ and $\mathcal{V}$. It follows that $\sin\theta = (s^H s)^{1/2}$. ∎

Using this result we can prove the next lemma.

Lemma A.2.2. *If $\|u\| = 1$, then $\sin \measuredangle(u, \mathcal{V}) = \min_{v \in \mathcal{V}} \|u - v\| = \operatorname{dist}(u, \mathcal{V})$.*

Proof. Let V and $V_\perp$ be as in Lemma A.2.1. Write $V^H Au = \widetilde{u}$ and $V_\perp^H Au = \widetilde{u}_\perp$, and note that $V^H AVx = x$ and $V_\perp^H AVx = 0$, for any x. We have

$$\begin{bmatrix} V^H \\ V_\perp^H \end{bmatrix} A(u - Vx) = \begin{bmatrix} \widetilde{u} - x \\ \widetilde{u}_\perp \end{bmatrix}.$$

Further,

$$\|u - Vx\| = \left\| \begin{bmatrix} V^H \\ V_\perp^H \end{bmatrix} A(u - Vx) \right\|_2 = \left\| \begin{bmatrix} \widetilde{u} - x \\ \widetilde{u}_\perp \end{bmatrix} \right\|_2$$

which is minimized for $x = \widetilde{u}$, with minimum $\|\widetilde{u}_\perp\|_2 = \|V_\perp^H Au\|_2$. By Lemma A.2.1, $\|V_\perp^H Au\|_2 = \sin \measuredangle(u, \mathcal{V})$. ∎

Now, if $\dim\mathcal{U} = \dim\mathcal{V} > 0$, then we have

$$\begin{aligned}\max_{\substack{u\in\mathcal{U}\\ \|u\|=1}} \operatorname{dist}(u,\mathcal{V}) &= \max_{\substack{u\in\mathcal{U}\\ \|u\|=1}} \sin\measuredangle(u,\mathcal{V}) \\ &= \sin\max_{\substack{u\in\mathcal{U}\\ \|u\|=1}} \measuredangle(u,\mathcal{V}) \\ &= \sin\max_{\substack{u\in\mathcal{U}\\ \|u\|=1}} \min_{\substack{v\in\mathcal{V}\\ \|v\|=1}} \measuredangle(u,v) \\ &= \sin\theta_{\max}(\mathcal{U},\mathcal{V}),\end{aligned}$$

where the first equality follows from Lemma A.2.2; the second from the fact that sine is an increasing function on $[0,\pi/2]$; the third from the definition of the angle between a vector and a subspace; and the fourth from Corollary A.1.4. Since $\theta_{\max}(\mathcal{U},\mathcal{V}) = \theta_{\max}(\mathcal{V},\mathcal{U})$, we have proved the following theorem.

Theorem A.2.3. *If* $\dim\mathcal{U} = \dim\mathcal{V}$, *then* $\operatorname{gap}(\mathcal{U},\mathcal{V}) = \delta(\mathcal{U},\mathcal{V}) = \sin\theta_{\max}(\mathcal{U},\mathcal{V})$.

Now, consider two gap functions $\operatorname{gap}(\cdot,\cdot)$ and $\operatorname{gap}'(\cdot,\cdot)$ associated with the norms $\|\cdot\|$ and $\|\cdot\|'$, respectively. (These norms need not be induced by inner products.) Since all norms on $\mathbb{C}^n$ are equivalent, there exist α and β such that

$$\alpha\|u\| \le \|u\|' \le \beta\|u\|$$

for any $u \in \mathbb{C}^n$. Using this and the definition of the gap, one can show that

$$\frac{\alpha}{\beta}\operatorname{gap}(\mathcal{U},\mathcal{V}) \le \operatorname{gap}'(\mathcal{U},\mathcal{V}) \le \frac{\beta}{\alpha}\operatorname{gap}(\mathcal{U},\mathcal{V}).$$

for any two subspaces $\mathcal{U},\mathcal{V} \subseteq \mathbb{C}^n$ [75, Theorem 4.4]. See also [31, Theorem 13.8.3]. Thus, if $\mathcal{U}(s)$ and $\mathcal{V}(s)$ are two subspaces of the same dimension k, depending on a real parameter s, it follows that a condition like

$$\lim_{s\to\infty} \theta_{\max}(\mathcal{U}(s),\mathcal{V}(s)) = 0$$

is independent of which positive definite inner product $\theta_{\max}(\cdot,\cdot)$ refers to.

Appendix

B

Roundoff error in complex arithmetic

Let $\mathbb{F}$ denote the set of finite floating point numbers with base β and precision t. Consider a sum $s = x_1 + x_2 + \cdots + x_n$, where the $x_k \in \mathbb{F}$, $k = 1\colon n$, and let $\widehat{s}$ denote the computed sum. Assuming $nu < 1$, where $u = \frac{1}{2}\beta^{1-t}$ is the unit roundoff, the forward error associated with this computation is commonly bounded as

$$|\widehat{s} - s| \leq \gamma_{n-1} \sum_{k=1}^{n} |x_k|, \quad \text{where} \quad \gamma_k := \frac{ku}{1-ku}.$$

Recently, Jeannerod and Rump [43] simplified this bound and showed that

$$|\widehat{s} - s| \leq (n-1)u \sum_{k=1}^{n} |x_k|, \tag{B.1}$$

with no constraints on n. Using this result, they then deduced the following forward error bound for the computed inner products of $x, y \in \mathbb{F}^n$:

$$|x^T y - \mathrm{float}(x^T y)| \leq nu|x|^T|y|. \tag{B.2}$$

When using complex numbers, one can often use the rounding error analysis from the analogous real computation, if one simply redefines the unit roundoff u to be a slightly larger value. See e.g., [36, Section 3.6]. We now prove that this can be done for (B.1) and (B.2). For (B.1), the following theorem shows that we can keep the same u when summing up complex numbers.

Theorem B.0.1. *If neither underflow nor overflow occurs, then* (B.1) *also holds for complex floating point numbers, that is, when* $x_k \in \mathbb{F} + i\mathbb{F}$ *for* $k = 1\colon n$.

Proof. Write $x_k = a_k + ib_k$. By (B.1), we have

$$\begin{aligned}
|x - \widehat{x}|^2 &\le \Big((n-1)u\sum_{k=1}^{n}|a_k|\Big)^2 + \Big((n-1)u\sum_{k=1}^{n}|b_k|\Big)^2 \\
&= ((n-1)u)^2\Big|\sum_{k=1}^{n}|a_k| + i\sum_{k=1}^{n}|b_k|\Big|^2 \\
&\le ((n-1)u)^2\Big(\sum_{k=1}^{n}\big||a_k| + i|b_k|\big|\Big)^2 \\
&= ((n-1)u)^2\Big(\sum_{k=1}^{n}|x_k|\Big)^2.
\end{aligned}$$

∎

For complex inner product we need to consider complex multiplication. Recently, Brent, Percival and Zimmermann [15] showed that the forward error for multiplying two complex number z and w can be bounded as

$$|zw - \mathrm{float}(zw)| \le \sqrt{5}|z||w|u. \tag{B.3}$$

We use this to prove the next theorem.

Theorem B.0.2. *If neither underflow nor overflow occurs, then* (B.2) *also holds for complex floating point vectors* $x, y \in (\mathbb{F} + i\mathbb{F})^n$ *if we replace* u *by* $\sqrt{5}u$.

Proof. Assume $n > 1$, otherwise we are done. If we define $f_k = \mathrm{float}(\overline{x}_k y_k) - \overline{x}_k y_k$, then (B.3) and Theorem B.0.1 yield

$$\begin{aligned}
|x^H y - \mathrm{float}(x^H y)| &= \Big|x^H y - \mathrm{float}\Big(\sum_{k=1}^{n}(\overline{x}_k y_k + f_k)\Big)\Big| \\
&\le \Big|x^H y - \sum_{k=1}^{n}(\overline{x}_k y_k + f_k)\Big| + nu\sum_{k=1}^{n}\big|\overline{x}_k y_k + f_k\big| \\
&\le \sum_{k=1}^{n}|f_k| + nu\sum_{k=1}^{n}(|x_k||y_k| + |f_k|) \\
&\le (1 + \sqrt{5}/n + u\sqrt{5})nu\sum_{k=1}^{n}|x_k||y_k| \\
&\le \sqrt{5}nu\sum_{k=1}^{n}|x_k||y_k|.
\end{aligned}$$

∎

Bibliography

[1] Structural applications of fluid viscous dampers. October 2012. Retrieved from `http://taylordevices.com/dampers-seismic-protection.html` Accessed: 2014-05-13.

[2] N. N. Abdelmalek. Round off error analysis for Gram-Schmidt method and solution of linear least squares problems. *BIT*, 11(4):345–368, 1971.

[3] O. Aberth. Iteration methods for finding all zeros of a polynomial simultaneously. *Math. Comp.*, 27(112):339–344, 1973.

[4] J. Ackermann. Der Entwurf linearer Regelungssysteme im Zustandsraum. *Regelungstechnik und Prozessdatenverarbeitung*, 7:297–300, 1972.

[5] M. Arioli, I. Duff, and D. Ruiz. Stopping criteria for iterative solvers. *SIAM J. Matrix Anal. Appl.*, 13(1):138–144, 1992.

[6] M. Arioli and C. Fassino. Roundoff error analysis of algorithms based on Krylov subspace methods. *BIT*, 36(2):189–206, 1996.

[7] W. E. Arnoldi. The principle of minimized iterations in the solution of the matrix eigenvalue problem. *Quart. Appl. Math.*, 9:17–29, 1951.

[8] Z. Bai and J. W. Demmel. On swapping diagonal blocks in real Schur form. *Linear Algebra Appl.*, 186:73–95, 1993.

[9] Z. Bai and Y. Su. SOAR: A second-order Arnoldi method for the solution of the quadratic eigenvalue problem. *SIAM J. Matrix Anal. Appl.*, 26(3): 640–659, 2005.

[10] T. Betcke, N. J. Higham, V. Mehrmann, C. Schröder, and F. Tisseur. NLEVP: A collection of nonlinear eigenvalue problems. *ACM Trans. Math. Software*, 39(2):7:1–7:28, 2013.

[11] D. A. Bini and V. Noferini. Solving polynomial eigenvalue problems by means of the Ehrlich-Aberth method. *Linear Algebra Appl.*, 439(4):1130–1149, 2013.

[12] Å. Björck. Solving linear least squares problems by Gram-Schmidt orthogonalization. *BIT*, 7(1):1–21, 1967.

[13] Å. Björck and G. H. Golub. Numerical methods for computing angles between linear subspaces. *Math. Comp.*, 27(123):579–594, 1973.

[14] T. Braconnier, P. Langlois, and J. Rioual. The influence of orthogonality on the Arnoldi method. *Linear Algebra Appl.*, 309(1–3):307–323, 2000.

[15] R. Brent, C. Percival, and P. Zimmermann. Error bounds on complex floating-point multiplication. *Math. Comp.*, 76(259):1469–1481, 2007.

[16] R. Byers, V. Mehrmann, and H. Xu. Trimmed linearizations for structured matrix polynomials. *Linear Algebra Appl.*, 429(10):2373–2400, 2008.

[17] D. H. Carlson. On real eigenvalues of complex matrices. *Pacific J. Math.*, 15 (4):1119–1129, 1965.

[18] G. Chrystal. A fundamental theorem regarding the equivalence of systems of ordinary linear differential equations, and its application to the determination of the order and the systematic solution of a determinate system of such equations. *Trans. Roy. Soc. Edin.*, 38(1):163–178, 1897. Issued separately 1895.

[19] M. T. Chu and N. D. Buono. Total decoupling of general quadratic pencils, part I: Theory. *Journal of Sound and Vibration*, 309(1–2):96–111, 2008.

[20] M. T. Chu and N. D. Buono. Total decoupling of general quadratic pencils, part II: Structure preserving isospectral flows. *Journal of Sound and Vibration*, 309(1–2):112–128, 2008.

[21] J. Drkošová, A. Greenbaum, M. Rozložník, and Z. Strakoš. Numerical stability of GMRES. *BIT*, 35(3):309–330, 1995.

[22] I. S. Duff, A. M. Erisman, and J. K. Reid. *Direct Methods for Sparse Matrices*. Oxford University Press, New York, 1986. ISBN 0-198-53408-6.

[23] L. W. Ehrlich. A modified Newton method for polynomials. *Comm. ACM*, 10(2):107–108, 1967.

[24] H.-Y. Fan, W.-W. Lin, and P. Van Dooren. Normwise scaling of second order polynomial matrices. *SIAM J. Matrix Anal. Appl.*, 26(1):252–256, 2004.

[25] J. B. Fraleigh. *A First Course in Abstract Algebra*. Addison Wesley, Boston, MA, USA, 7th edition, 2002. ISBN 0-321-15608-0.

[26] F. R. Gantmacher. *The Theory of Matrices*, volume one. Chelsea, New York, 1959. ISBN 0-8284-0131-4.

[27] F. R. Gantmacher. *The Theory of Matrices*, volume two. Chelsea, New York, 1959. ISBN 0-8284-0133-0.

[28] S. Garvey, M. Friswell, and U. Prells. Co-ordinate transformations for second order systems. part I: Ggeneral transformations. *Journal of Sound and Vibration*, 258(5):885–909, 2002.

[29] L. Giraud, S. Gratton, and J. Langou. Convergence in backward error of relaxed GMRES. *SIAM J. Sci. Comput.*, 29(2):710–728, 2007.

[30] I. Gohberg, P. Lancaster, and L. Rodman. Spectral analysis of selfadjoint matrix polynomials. *Ann. Math.*, 112(1):33–71, 1980.

[31] I. Gohberg, P. Lancaster, and L. Rodman. *Invariant Subspaces of Matrices with Applications*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2006. ISBN 0-89871-608-X.

[32] I. Gohberg, P. Lancaster, and L. Rodman. *Matrix Polynomials*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2009. ISBN 0-898716-81-8. Unabridged republication of book first published by Academic Press in 1982.

[33] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, USA, 4th edition, 2013. ISBN 978-1421407944.

[34] S. Hammarling, C. J. Munro, and F. Tisseur. An algorithm for the complete solution of quadratic eigenvalue problems. *ACM Trans. Math. Software*, 38 (3):18:1–18:19, 2013.

[35] D. J. Higham and N. J. Higham. Structured backward error and condition of generalized eigenvalue problems. *SIAM J. Matrix Anal. Appl.*, 20(2):493–512, 1998.

[36] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2nd edition, 2002. ISBN 0-89871-521-0.

[37] N. J. Higham, D. S. Mackey, F. Tisseur, and S. D. Garvey. Scaling, sensitivity and stability in the numerical solution of quadratic eigenvalue problems. *Internat. J. Numer. Methods Eng.*, 73(3):344–360, 2008.

[38] N. J. Higham and F. Tisseur. A block algorithm for matrix 1-norm estimation, with an application to 1-norm pseudospectra. *SIAM J. Matrix Anal. Appl.*, 21(4):1185–1201, 2000.

[39] M. E. Hochstenbach. Probabilistic upper bounds for the matrix two-norm. *J. Sci. Comput.*, 57(3):464–476, 2013.

[40] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, UK, 1985. ISBN 0-521-30586-1.

[41] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991. ISBN 0-521-46713-6.

[42] *IEEE Standard for Floating-Point Arithmetic, IEEE Std 754-2008 (revision of IEEE Std 754-1985)*. IEEE Computer Society, New York, 2008. ISBN 978-0-7381-5752-8.

[43] C.-P. Jeannerod and S. M. Rump. Improved error bounds for inner products in floating-point arithmetic. *SIAM J. Matrix Anal. Appl.*, 34(2):338–344, 2013.

[44] U. Kandler and C. Schröder. Backward error analysis of an inexact Arnoldi method using a certain Gram-Schmidt variant. Technical report, 2013.

[45] T. Kato. *Pertubation Theorey for Linear Operators*. Springer, Berlin, 1995. ISBN 3-540-58661X. Reprint of the 1980 Edition.

[46] K. Kim. A review of mass matrices for eigenproblems. *Comput. Struct.*, 46 (6):1041–1048, 1993.

[47] K. Knopp. *Theory of Functions*. Dover, New York, 1947. English translation, Part II.

[48] P. Kunkel and V. Mehrmann. *Differential-Algebraic Equations: Analysis and Numerical Solutions*. European Mathematical Society, Zürich, Switzerland, 2006. ISBN 3-03719-017-5.

[49] P. Lancaster. On eigenvalues of matrices dependent on a parameter. *Numer. Math.*, 6:377–387, 1964.

[50] P. Lancaster. *Lambda-matrices and Vibrating Systems*. Pergamon Press, Oxford, UK, 1966. ISBN 0-486-42546-0. Reprinted by Dover, New York, USA, 2002.

[51] P. Lancaster and M. Tismenetsky. *The Theory of Matrices*. Academic Press, London, 2nd edition, 1985. ISBN 0-12-435560-9.

[52] P. Lancaster and I. Zaballa. Diagonalizable quadratic eigenvalue problems. *Mech. Syst. Signal Pr.*, 23(4):1134–1144, 2009.

[53] N. B. Langer, H. and K. Veselić. Perturbation of the eigenvalues of quadratic matrix polynomials. *SIAM J. Matrix Anal. Appl.*, 13(2):474–489, 1992.

[54] R. B. Lehoucq and K. Meerbergen. Using generalized Cayley transformations within an inexact rational Krylov sequence method. *SIAM J. Matrix Anal. Appl.*, 20(1):131–148, 1998.

[55] D. Lu, X. Huang, Z. Bai, and Y. Su. A Padé approximate linearization for solving the quadratic eigenvalue problem with low-rank damping. Technical report, UC Davis: College of Engineering, June 2014.

[56] D. S. Mackey, N. Mackey, C. Mehl, and V. Mehrmann. Jordan structures of alternating matrix polynomials. *Linear Algebra Appl.*, 432(4):867–891, 2010.

[57] D. S. Mackey, N. Mackey, C. Mehl, and V. Mehrmann. Möbius transformations of matrix polynomials. MIMS EPrint 2014.2, Manchester Institute for Mathematical Sciences, The University of Manchester, UK, January 2014.

[58] D. S. Mackey, N. Mackey, and F. Tisseur. Structured mapping problems for matrices associated with scalar products. part I: Lie and Jordan algebras. *SIAM J. Matrix Anal. Appl.*, 29(4):1389–1410, 2007.

[59] D. S. Mackey and F. Tisseur. The Hermitian quadratic realizability problem. Technical report. In preparation.

[60] E. Marques de Sá. Imbedding conditions for λ-matrices. *Linear Algebra Appl.*, 24:33–50, 1979.

[61] K. Meerbergen and R. Morgan. Inexact methods (section 11.2). In 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.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2000.

[62] D. Meyer and K. Veselić. On some new inclusion theorems for the eigenvalues of partitioned matrices. *Numer. Math.*, 34(4):431–437, 1980.

[63] M. Morzfeld, F. Ma, and B. N. Parlett. The transformation of second-order linear systems into independent equations. *SIAM J. Math. Anal.*, 71(4): 1026–1043, 2011.

[64] A. Neumaier. Residual inverse iteration for the nonlinear eigenvalue problem. *SIAM J. Numer. Anal.*, 22(5):914–923, 1985.

[65] G. Peters and J. H. Wilkinson. Inverse iteration, ill-conditioned equations and Newton's method. *SIAM Rev.*, 21(3):339–360, 1979.

[66] J. L. Rigal and J. Gaches. On the compatibility of a given solution with the data of a linear system. *J. Assoc. Comput. Mach.*, 14(3):543–548, 1967.

[67] J. Rotman. *Advanced Modern Algebra.* American Mathematical Society, Providence, RI, USA, 2010. ISBN 978-0-8218-4741-1.

[68] C. Schröder and L. Taslaman. Backward error analysis of the shift-and-invert Arnoldi algorithm. MIMS EPrint 2014.53, Manchester Institute for Mathematical Sciences, The University of Manchester, UK, October 2014.

[69] A. A. Shabana. *Vibration of Discrete and Continuous Systems, Volume 2.* Springer, New York, 2nd edition, 1997. ISBN 978-0-387-94744-0.

[70] V. Simoncini and D. B. Szyld. Theory of inexact Krylov subspace methods and applications to scientific computing. *SIAM J. Sci. Comput.*, 25(2): 454–477, 2003.

[71] A. Smoktunowicz. The strong stability of algorithms for solving the symmetric eigenproblem. *J. Korean Soc. Ind. Appl. Math.*, 7(1):25–31, 2003.

[72] A. Smoktunowicz, J. L. Barlow, and J. Langou. A note on the error analysis of classical Gram-Schmidt. *Numer. Math.*, 105(2):299–313, 2006.

[73] G. Stewart. *Matrix Algorithms, Volume II: Eigensystems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2001. ISBN 978-0-898714-14-2.

[74] G. Stewart. A Krylov–Schur algorithm for large eigenproblems. *SIAM J. Matrix Anal. Appl.*, 23(3):601–614, 2002.

[75] G. W. Stewart and J. Sun. *Matrix Perturbation Theory*. Academic Press, Boston, MA, USA, 1990. ISBN 0-12-670230-6.

[76] D. B. Szyld. The many proofs of an identity on the norm of oblique projections. *Numer. Algorithms*, 42(3-4):309–323, 2006.

[77] L. Taslaman. An algorithm for quadratic eigenproblems with low rank damping. MIMS EPrint 2014.21, Manchester Institute for Mathematical Sciences, The University of Manchester, UK.

[78] L. Taslaman. Strongly damped quadratic matrix polynomials. MIMS EPrint 2014.10, Manchester Institute for Mathematical Sciences, The University of Manchester, UK.

[79] L. Taslaman, F. Tisseur, and I. Zaballa. Triangularizing matrix polynomials. *Linear Algebra Appl.*, 439(7):1679–1699, 2013.

[80] F. Tisseur. Backward stability of the QR algorithm. Technical report 239, Equipe d'Analyse Numerique, Université Jean Monnet de Saint-Etienne, Cedex 02, France, October 1996.

[81] F. Tisseur. Backward error and condition of polynomial eigenvalue problems. *Linear Algebra Appl.*, 309:339–361, 2000.

[82] F. Tisseur. Newton's method in floating point arithmetic and iterative refinement of generalized eigenvalue problems. *SIAM J. Matrix Anal. Appl.*, 22(4):1038–1057, 2001.

[83] F. Tisseur and K. Meerbergen. The quadratic eigenvalue problem. *SIAM Rev.*, 43(2):235–286, 2001.

[84] F. Tisseur and I. Zaballa. Triangularizing quadratic matrix polynomials. *SIAM J. Matrix Anal. Appl.*, 34(2):312–337, 2013.

[85] L. N. Trefethen and D. Bau. *Numerical Linear Algebra*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1997. ISBN 0-89871-361-7.

[86] A. Van der Sluis. Condition numbers and equilibrium of matrices. *Numer. Math.*, 14(1):14–23, 1969.

[87] P. Van Dooren and P. Dewilde. The eigenstructure of an arbitrary polynomial matrix: computational aspects. *Linear Algebra Appl.*, 50:545–579, 1983.

[88] K. Veselić. *Damped Oscillations of Linear Systems*. Springer, Berlin, 2011. ISBN 978-3642213342.

[89] S. Wang and S. Zhao. An algorithm for $Ax = \lambda Bx$ with symmetric and positive-definite A and B. *SIAM J. Matrix Anal. Appl.*, 12(4):654–660, 1991.

[90] D. S. Watkins. *The Matrix Eigenvalue Problem: GR and Krylov Subspace Methods*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2007. ISBN 987-0-898716-41-2.

[91] I. Zaballa and F. Tisseur. Finite and infinite elementary divisors of matrix polynomials: a global approach. MIMS EPrint 2012.78, Manchester Institute for Mathematical Sciences, The University of Manchester, UK, August 2012.

[92] L. Zeng and Y. Su. A backward stable algorithm for quadratic eigenvalue problems. *SIAM J. Matrix Anal. Appl.*, 35(2):499–516, 2014.

[93] J. C. Zúñiga Anaya. Diagonalization of quadratic matrix polynomials. *Syst. Control Lett.*, 59(2):105–113, 2010.

Index